RetiFluidNet: A Self-Adaptive and Multi-Attention Deep Convolutional Network for Retinal OCT Fluid Segmentation

Reza Rasti, Member, IEEE, Armin Biglari, Mohammad Rezapourian, Ziyun Yang, and Sina Farsiu, Fellow, IEEE

Abstract—Optical coherence tomography (OCT) helps ophthalmologists assess macular edema, accumulation of fluids, and lesions at microscopic resolution. Quantification of retinal fluids is necessary for OCT-guided treatment management, which relies on a precise image segmentation step. As manual analysis of retinal fluids is a time-consuming, subjective, and error-prone task, there is increasing demand for fast and robust automatic solutions. In this study, a new convolutional neural architecture named RetiFluidNet is proposed for multi-class retinal fluid segmentation. The model benefits from hierarchical representation learning of textural, contextual, and edge features using a new self-adaptive dual-attention (SDA) module, multiple self-adaptive attention-based skip connections (SASC), and a novel multi-scale deep self-supervision learning (DSL) scheme. The attention mechanism in the proposed SDA module enables the model to automatically extract deformation-aware representations at different levels, and the introduced SASC paths further consider spatial-channel interdependencies for concatenation of counterpart encoder and decoder units, which improve representational capability. RetiFluidNet is also optimized using a joint loss function comprising a weighted version of dice overlap and edge-preserved connectivity-based losses, where several hierarchical stages of multi-scale local losses are integrated into the optimization process. The model is validated based on three publicly available datasets: RETOUCH, OPTIMA, and DUKE, with comparisons against several baselines. Experimental results on the datasets prove the effectiveness of the proposed model in retinal OCT fluid segmentation and reveal that the suggested method is more effective than existing state-of-the-art fluid segmentation algorithms in adapting to retinal OCT scans recorded by various image scanning instruments.

Index Terms—Medical image segmentation, convolutional neural network, retinal disease, fluid segmentation.

I. INTRODUCTION

The retina in human eyes receives focused light and transforms it into neural signals. Located in the central region of the retina, the macula is responsible for detecting color, light intensity, and visual details. Optical coherence tomography (OCT) imaging captures 3-D images at micrometer resolution and provides cross-sectional scans for the retinal structure and its biological tissues [1]. In clinical ophthalmology, the OCT imaging technique has become the preferred scanning method for the macula because of its high resolution and visualization ability for the anatomical constituent layers of the photoreceptors [2], [3].

Diabetic macular edema (DME) is a major cause of blindness and visual impairments in patients with diabetic retinopathy (DR). Quality degradation of vision due to DME is related to damaged blood vessels and fluid accumulation between retinal layers. Thus, to properly manage the treatment of DME and prevent its progression to blindness, DME patients’ retinal layer morphology and fluid accumulation must be thoroughly monitored [4], [5]. For the accurate assessment of the macula and fluid regions, precise segmentation of retinal OCT images is required[6]. Manual segmentation, although sufficient, is a time-consuming, subjective, and error-prone task; hence, fast and robust automatic alternatives are desired.

To analyze macular fluids automatically and accurately, different image processing and machine learning algorithms have been developed [7], [8], [9]. Of the advanced solutions in the field, deep learning provides the advantage of learning discriminative representations to obtain acceptable segmentation performance [10], [11], [12]. A review of the related works is presented in Section II.

Despite these recent advances, accurate retinal fluid segmentation in macular OCT imaging remains a challenging task due to variations in the structure of the retina, disease-related variations in fluid shape and size, blurred boundaries of the suspected regions, different levels of disruptive speckle noise,
and low contrast of the OCT image data. These difficulties give rise to imperfect segmentation, especially for near-edge regions, and low spatial coherence. Moreover, the lack of sufficient fluid patterns with pixel-level annotation complicates the fluid segmentation task. In the present study, to address these challenges and improve performance, we propose a novel self-adaptive and multi-level attention convolutional U-shaped network called RetiFluidNet. Our main technological contributions in this work are summarized as follows:

1. A novel self-adaptive, dual-attention (SDA) module is proposed and designed to jointly capture the relevant contextual information in the spatial and channel domains. The SDA module imposes pixel-group attention on input tensors by introducing an attention mechanism that accounts for spatial-channel interdependencies efficiently and adaptively. With only two trainable free parameters, the module can weigh and highlight spatial and channel information adaptively in an automatic and end-to-end manner to improve the attention process performance.

2. A novel self-adaptive, attention-based skip connection (SASC) is proposed and inserted between the encoder and decoder counterpart blocks using the SDA module. The SASC paths improve the ability of the overall model to learn global and multi-semantic contextual representative features.

3. By combining SDA and SASC modules, a novel convolutional architecture, RetiFluidNet, is designed based on the U-Net structure and used for fluid segmentation in macular OCT images.

4. We further proposed and evaluated a deep self-supervision learning (DSL) scheme for the model’s optimization in our multi-class segmentation problem. A new joint loss function comprising a multi-scale weighted version of dice overlap losses and edge-preserved connectivity-based losses is introduced. Several hierarchical stages of local losses are computed according to the consecutive decoder blocks’ outputs and merged into the joint loss function. For this, multi-scale local losses of dice loss components (DLC) and connectivity-based loss components (CLC) are used for texture- and edge-specific feature enhancement, respectively. More specifically, in CLC, local connectivity masks are used together with annotated masks as labels for effective modeling of inter-pixel relationships and edge feature enhancement for each class of fluids. Finally, a linear combination of DLC and CLC terms is integrated into the optimization process to compute the overall objective function and obtain enhanced pixel-wise segmentation performance.

5. We also conducted extensive experiments and ablation studies using three publicly available retinal fluid segmentation datasets (RETOUCH [13], OPTIMA [14], and DUKE [15]) to evaluate the effectiveness of the proposed model. Compared with other baselines and state-of-the-art deep learning-based fluid segmentation methods, RetiFluidNet performance was significantly better with comparative and acceptable execution time.

The remainder of this paper is organized as follows: Section II describes the latest research studies on the segmentation of fluid regions in retinal OCT images. Section III introduces the materials and formulates our proposed method in detail. Section IV shows the experimental study and results. Section V discusses our findings in this study and draws potential future research directions. Finally, in section VI, the study’s conclusion is presented.

II. RELATED WORK

Segmentation of disease-affected regions in retinal OCT imaging is an active and demanded area of research. Many machine learning methods have been proposed to segment retinal fluids and layers [2], [16], [17], and classify retinal diseases [18], [19], [20], [21]. Recently, several deep-learning techniques have been applied to segment retinal lesions and layers. U-Net [22], FCN [23], Seg-Net [24], DeepLabv3+ [25], and U-Net++ [26] are among the most common and interesting deep learning architectures developed for image segmentation and used in medical applications. Table I summarizes recent studies that have applied deep learning models to retinal fluid segmentation.

Fluid segmentation algorithms are often affected by a high rate of false detections in the background regions [27], [28]. To overcome this issue, Chen et al. [29] and Fu et al. [30] have suggested the use of larger convolutional kernel sizes in convolutional neural network (CNN)-based methods to benefit from larger receptive fields, while [31], [32] studies have proposed using weighted versions of loss functions in the optimization process. Since enlarged receptive fields and weighted loss functions cannot be dynamically adjusted to fit fluid targets of different sizes, a variety of attention mechanisms have been proposed to address this challenge and develop scale-aware models [27], [33], [34].

In the study of human vision, the attention mechanism is often referred to as the process by which humans selectively focus on some parts of visual information while ignoring others, which results in enhanced visual information processing. To improve the performance of deep learning-based segmentation models, several attention mechanisms have been suggested and used in recent years [28], [34]. W. Liu et al. [27] proposed an attention-based method for fluid segmentation, and they integrated multi-scale inputs, multi-scale side-outputs, and attention methods into U-Net++ for better performance. In a study by X. Liu et al. [34] an attention-based U-Net model was developed for intraretinal cystoid fluid (IRC) segmentation in DME images with an attention structure that automatically considers suspected regions. Feng et al. [35] proposed the CPNFNet model which explored global and multi-scale contextual information mainly through a soft spatial attention mechanism that learned features across different receptive fields using dilated convolutions with shared weights. They also tested the model’s performance for automated segmentation of retinal edema lesions on the AI-chALLENGER 2018 dataset with acceptable performance. Recently, Wang et al. [36] developed the MsTGANet model, a multi-scale transformer global attention network for drusen segmentation in retinal OCT images that combines non-local and multi-semantic information.

In this study, we present a multi-attention and self-adaptive neural structure that adaptively tunes the attention processes based on an end-to-end training scheme to reduce the number
TABLE I
SUMMARY OF THE RECENT DEEP LEARNING METHODS APPLIED TO RETINAL FLUID SEGMENTATION

| Reference       | Data                                      | Pre-Processing                  | Model   | Validation Method | Evaluation Metric | Results* (%) |
|-----------------|-------------------------------------------|----------------------------------|---------|------------------|-------------------|--------------|
| Roy et al. [2]  | DUKE: Spectralis                          | 10 IRC                           |         | ReLayNet, 2D     | F-CV              | 81           |
| Alsah et al. [3]| RETOUCH, OPTIMA, DUKE: Cirrus, Nidek, Spectralis, Topcon | 112+30 IRC, 10 All              |         | Deeplabv3+ext. patch, 2.5D | F-CV H.E O        | RETOUCH-80.82; OPTIMA-72.62, 70.78; DUKE-84.80-86.80 |
| Alsah et al. [5]| RETOUCH, OPTIMA: Cirrus, Nidek, Spectralis, Topcon | 112+30 IRC, 10 All              |         | Pa.Deeplabv3+, 2D | F-CV              | RETOUCH-84.79; OPTIMA-85.63 |
| Chiu et al. [15]| DUKE: Spectralis                          | 10 IRC                           |         | Kernel Reg., 2D   | F-CV              | 53           |
| W.Liu et al. [27]| RETOUCH: Cirrus, Spectralis, Topcon       | 112 IRC, 10 All, PED            |         | MDAN U-Net, 2D    | H.O              | 67, 69, 65 |
| X.Liu et al. [34]| OPTIMA: Cirrus, Nidek, Spectralis, Topcon | 30 IRC                           |         | Att-U-Net, 2D     | F-CV              | 83, 80, 84, 84 |
| Feng et al. [35]| AI Challenger 2018                        | 83 REA                           |         | CPPNet, 2.5D      | F-CV              | 80           |
| Lee et al. [37] | Local Data: Spectralis                    | 30 IRC                           |         | U-Net, 2D         | H.O              | 75           |
| Lu et al. [38]  | RETOUCH: Cirrus, Spectralis, Topcon       | 112 IRC, 10 All, PED            |         | U-Net, 2D         | L.O.O CV          | 71, 77, 64 |
| Venhuize et al. [39]| OPTIMA: Cirrus, Nidek, Spectralis, Topcon | 30 IRC                           |         | CNN, 2D           | H.O              | 65, 66, 83, 81 |
| Girish et al. [40]| OPTIMA: Cirrus, Nidek, Spectralis, Topcon | 30 IRC                           |         | U-Net, 2D         | H.O              | 63, 75, 74, 77 |
| Gopinath et al. [41]| OPTIMA: Cirrus, Nidek, Spectralis, Topcon | 30 IRC                           |         | U-Net, 3D         | H.O              | 62, 73, 76, 76, DUKE-59 |
| Reis et al. [42] | RETOUCH: Cirrus, Spectralis, Topcon       | 112 IRC, 10 All, PED            |         | Semi-Weakly Supervised U-Net, 2D | F-CV | IoU | 51, 42 |
| Ren et al. [43]  | RETOUCH: Cirrus, Spectralis, Topcon       | 112 IRC, 10 All, PED            |         | CGAN, 2D          | H.O              | 62, 62, 52, 51, 48, 43 |
| Mahapatra et al. [44]| RETOUCH (Pathological Images): Cirrus, Spectralis, Topcon | 112 IRC, 10 All, PED            |         | GeoGAN, 2D        | H.O              | DSC; HD Average=90; 7.9 |
| Schlegl et al. [45]| Local Data: Cirrus, Spectralis            | 1200 IRC, 10 All, PED           |         | U-Net, 2D         | 4-F-CV            | Pr, Re     | 80, 86, 83, 75 |
| Yadav et al. [46] | Local Data: Cirrus, Spectralis, Topcon    | 112 IRC, 10 All, PED            |         | 8-F-CV            | DSC; AUC          | 72, 68, 66, 84, 82, 87 |
| Hassan et al. [47]| RETOUCH: Cirrus, Spectralis, Topcon       | 112 IRC, 10 All, PED            |         | RFS-Net, 2D       | H.O              | DSC; IoU   | 75, 79, 80, 61, 65, 67 |

IRF=Intraretinal Fluid, SRF=Subretinal Fluid, PED=Pigment Epithelial Detachment, IRC=Intraretinal Cystoid Fluid, REA=Retinal Edema Area, DSC=Dice Coefficient Score, IoU=Intersection over Union, CV=Cross-Validation, H.O=Hold Out, L.O.O=Leave-One-Out, Pr=Precision, Re=Recall, HD=Hausdorff Distance, GAN= Generative Adversarial Network, Att=Attention. *The results are reported based on OCT scanners analyses, respectively. In [44], the results are only available for whole scanners.

of false detections. Our model takes into account multi-scale textural, contextual, and edge information for better segmentation of noisy and blurry retinal fluid regions in OCT images. The suggested CNN architecture can efficiently manage large variations in location, size, and shape of diverse pathological fluid lesions. The efficiency of our proposed model is supported by extensive empirical results in Section IV.

III. MATERIALS AND METHODS

In this section, we first describe the datasets used in this study for model evaluation and generalization, and then we present our data analysis pipeline and introduce in detail the pre-processing stage, model structure, and core attention components. Finally, we formulate the joint loss function and learning scheme integrated for optimizing the proposed model.

A. Datasets

We used three publicly available datasets: the RETOUCH challenge dataset [13], OPTIMA challenge dataset [14], and DUKE dataset [15]. The RETOUCH dataset was collected to distinguish intraretinal fluid (IRF), subretinal fluid (SRF), and pigment epithelial detatchment (PED) regions. This dataset contains three OCT scanners: Zeiss Circus, Heidelberg Spectralis, and Topcon. The OPTIMA dataset includes IRC regions for segmentation tasks. In addition to the previous three scanners, this dataset also has a Nidek scanner. The DUKE dataset was collected to segment DME-relevant fluids and only used the Spectralis scanner. Fig.1 provides sample illustrations of OCT B-scans for the RETOUCH, DUKE, and OPTIMA datasets. As the ground truth (GT) for the test set is not made available by the challenge organizers, the RETOUCH test data was not used in our study. Therefore, we considered the training set with cross-validation for evaluating our method and the baselines. Moreover, the models were further externally evaluated over the OPTIMA and the DUKE datasets. The training and testing sub-sets of the OPTIMA and DUKE datasets were both supplied with labels. The details of the datasets are summarized in Table II.
In the present study, the RETOUCH challenge dataset was used to train and validate the proposed RetiFluidNet and baseline models. The OPTIMA and DUKE datasets were also used for comparison with the recent methods developed for the retinal fluid segmentation problem.

B. Pre-Processing

All B-scan images in various volumes were initially resized to 256 × 256 pixels to obtain a unique field of view. Furthermore, a min-max normalization step was performed to map the intensity values of each B-scan between 0 and 1.

C. RetiFluidNet

Fig. 2 demonstrates the proposed RetiFluidNet architecture, which adopts an encoder-decoder U-shaped topology as the basic framework. RetiFluidNet consists of five major components, including (1) encoder blocks, (2) SDA module, (3) SASC path, (4) residual multi-kernel pooling (RMP) module, and (5) decoder blocks.

1) Encoder Blocks: As shown in Fig. 2 and like the U-Net structure, the encoder part of the proposed RetiFluidNet contains five blocks. Each block consists of two consecutive blocks of Conv2D-BN-ReLU layers with 3 × 3 kernels followed by a 2 × 2 Max-pooling operation. The Conv2D-BN-RelU layers are used to extract the representative features in different stages, and the Max-pooling operation is adopted to abstract data and reduce the complexity of the model.

2) SDA Module: In the present work, we introduce a soft attention module inspired by the attention mechanisms in [48] to guide the network where to look for the fluid regions in the retinal layers. The proposed self-adaptive and dual attention module, SDA, enables the neural model to automatically extract deformation-aware representations at multiple levels. Several preceding studies have developed self-attention modules that demonstrate the positive impact of spatial- and/or channel attentions in improving various image analysis tasks [48], [49], [50], [51]. Different from the earlier works, we extend the self-attention mechanism to the task of fluid segmentation, where the two types of attention are embedded in an integrated module to capture rich contextual relationships with intra-class compactness. The module aggregates spatial- and channel attention processes selectively and adaptively with only two trainable parameters for enhanced feature representations. The efficiency of the proposed SDA module is supported by experimental results in Section IV. The SDA module performs two attention mechanisms on the given tensor: a channel-wise attention process and a pixel-wise attention process. Fig. 3 demonstrates the structure of the attention block. Using perturbed and augmented spatial attention, the proposed SDA module incorporates spatial and channel interactions for more efficient feature representation. The perturbation mechanism is embedded into the module where $K_{\text{pixel}}$ is computed in the processing pipeline. It extracts and attends spatial features from relevant parts of input feature maps (FMs). The representation process is then directed to generate relevant discriminative FMs by the perturbed region amplification. Meanwhile, the module calculates the $K_{\text{ch}}$ to attend various channel dimensions of the input FMs and efficiently encodes FM inter-channel interactions. By adopting higher values of weights over the signal flows, the model adaptively integrates the mechanisms to further converge on the relevant regions of the input tensor for better attention performance. By modeling rich contextual dependencies over local features, it significantly improves segmentation performance. The RetiFluidNet takes advantage of several SDA processes at different levels, gaining coarse and smooth attentive features from two intrinsic attention processes embedded into the SDA module design. The SDA module is used between the encoder and the decoder counterparts through SASC paths and is further embedded into the top of the final encoder block to enhance latent-space representation ability.

Let the input tensor for the SDA module be $X \in \mathbb{R}^{H \times W \times C}$, where $W$, $H$, and $C$ stand for width, height, and channel dimensions for the given tensor. The SDA generates...
two highlighted tensor maps: $X_{att,\text{pixel}} \in \mathbb{R}^{H \times W \times C}$ and $X_{att,\text{channel}} \in \mathbb{R}^{H \times W \times C}$, for pixel-wise attention and channel-wise attention, respectively. They are added to the input tensor using the following expression:

$$X_{att} = X + \frac{1}{2} (\alpha X_{att,\text{pixel}} + \beta X_{att,\text{channel}})$$

(1)

Here, $X_{att} \in \mathbb{R}^{H \times W \times C}$ is the output tensor with multi-dimension attention, which represents the relevant contextual information in the spatial and channel domains. In practice, the SDA adds a highlighted version of the input, which is a weighted average of $X_{att,\text{pixel}}$ and $X_{att,\text{channel}}$. In this equation, $\alpha, \beta \geq 0$ are two trainable parameters that adaptively weigh the contribution of the pixel-wise and channel-wise attention mechanisms for driving the models’ focus on more relevant representative parts of the input data. To generate $X_{att,\text{pixel}}$ and $X_{att,\text{channel}}$, the input tensor is fed to a Max-pooling layer by a pooling factor of $p$ and the resultant down-sampled tensor, i.e., $X_d \in \mathbb{R}^{H_p \times W_p \times C}$, is followed by a reshape layer that transforms the 3-D tensor into a 2-D matrix, $X_{dr} \in \mathbb{R}^{(H_p \times W_p) \times C}$, as shown here:

$$X_d = \text{Maxpooling} (X, p),$$

(2)

$$X_{dr} = \text{Reshape} (X_d).$$

(3)

The pixel-wise attention coefficient $K_{pixel} \in \mathbb{R}^{(H_p \times W_p) \times (H_p \times W_p)}$ is calculated by:

$$K_{pixel} = \text{Softmax} \left( \frac{1}{\sqrt{H_p \times W_p}} X_{dr} \otimes X_{dr}^T \right),$$

(4)

In this equation, $T$ denotes transpose operation, the operator of $\otimes$ indicates matrix-multiplication, the normalization factor of $\frac{1}{\sqrt{H_p \times W_p}}$ limits the magnitude of the product output, and $\text{Softmax} \left( x_j \right) = \frac{\exp(x_j)}{\sum_{i=1}^{n} \exp(x_i)}$, $j = 1, 2, \ldots, n$ where $n$ is the number of matrix elements. The normalization factor, here, is used so that the Softmax function’s input does not push it into regions where it has extremely small gradients in the optimization process. The matrix multiplication in Eq.4, amplifies the spatial similarity of $X_{dr}$ and $X_{dr}^T$ whereas the Softmax function yields context-aware attention coefficients by assigning greater probabilities to higher entropy regions.

The generated pixel-attention coefficient $K_{pixel}$ acts as a weighing function on $X_{dr}$ using Eq.5 and yields:

$$X_{dr,\text{att,\text{pixel}}} = K_{pixel} \otimes X_{dr}; X_{dr,\text{att,\text{pixel}}} \in \mathbb{R}^{(H_p \times W_p) \times C}$$

(5)

Subsequently, the resultant weighted matrix is reshaped back and resized to obtain $X_{att,\text{pixel}} \in \mathbb{R}^{H \times W \times C}$ feature tensor, given by:

$$X_{att,\text{pixel}} = \text{Unpooling} \left( \text{Reshape} \left( X_{dr,\text{att,\text{pixel}}}, p \right) \right).$$

(6)

The second path in the SDA module (see Fig.3) exploits interdependencies between high-level FMs or channels of the input tensor. By explicitly modeling the correlations between high-level FMs, it emphasizes interdependent channels, leading to improved feature representation ability and better attention. Like Eq.4, channel-attention coefficients $K_{ch}$ can be mathematically written as:

$$K_{ch} = \text{Softmax} \left( \frac{1}{C} X_{dr}^T \otimes X_{dr} \right); K_{ch} \in \mathbb{R}^{C \times C}$$

(7)

Before applying the Softmax function over the resultant matrix production, the scores are normalized using the number of FMs, i.e., $C$ value. Please note the order of multiplicand and multiplier in the matrix multiplication here. Then, $X_{dr,\text{att,\text{ch}}}$ is calculated by performing multiplication between $K_{ch}$ and $X_{dr}$ matrices using Eq.8.

$$X_{dr,\text{att,\text{ch}}} = K_{ch} \otimes X_{dr}; X_{dr,\text{att,\text{ch}}} \in \mathbb{R}^{(H_p \times W_p) \times C}$$

(8)

Therefore, as shown in Fig.3, the resultant weighted matrix is reshaped back and resized to obtain $X_{att,\text{channel}} \in \mathbb{R}^{H \times W \times C}$ feature tensor, which is given by:

$$X_{att,\text{channel}} = \text{Unpooling} \left( \text{Reshape} \left( X_{dr,\text{att,\text{ch}}}, p \right) \right).$$

(9)

It should be noted that the prior Max-pooling layer ($p > 1$) in the SDA structure helps the overall attention process to be much faster and more space-efficient in practice than when $p = 1$.

3) Self-Adaptive Attention-Based Skip Connection: The SASC paths are adopted to replace the regular skip-connection (SC) between encoder and decoder counterpart blocks. Their process is based on the SDA module. The SASCs concatenate the decoder blocks’ output feature maps with their SDA-processed versions, and the resultant tensors are passed to the decoder blocks for further processing. The SASCs provide a mechanism to adaptively highlight and transfer global information from encoder to decoder parts for better data reconstruction. Accordingly, lower $\alpha$ and larger $\beta$ values for the SADs’ weighting parameters are adaptively assigned at each scale to capture richer contextual dependencies and model semantic interdependencies in the channel dimension. This guides the model to fuse multi-semantic global contextual features and improves the ability of the model to pay attention to global salient features and suppress the interference of irrelevant local features adaptively.

4) Residual Multi-Kernel Pooling: Similar to [52], we propose to use the RMP block in the encoder part of the model which handles the large variety of fluid sizes in retinal OCT image representations. RMP involves multiple effective field-of-views for data processing and encodes global context information with different-sized receptive fields. In our model, the SAD and RMP modules are inserted into the top of the final encoder block to enhance latent-space representation potential.
5) Decoder Part: There are five decoder blocks in the decoder part, as illustrated in Fig.2. The key components of each decoder block include a Conv2DTranspose layer for up-sampling, a feature concatenation operation from the associated SASC path, and three consecutive blocks of Conv2D-BN-ReLU layers. The decoder part is primarily used to restore spatial information provided by SDA-r-RMP with strong multi-scale semantic features and to gradually fuse multi-semantic global contextual information from SASC paths.

Deep Supervision Scheme and Loss Function: In retinal OCT data analysis, the complicated pathological manifestations of retinal fluids with variations in size, shape, and location, and the presence of interferences such as other lesions, speckle noise, and also blurry boundaries for the suspected areas, pose great challenges to precisely segment fluid regions. Therefore, learning multi-scale non-local features along with key local features is essential for segmentation performance improvement. For this purpose, in addition to the previous model development considerations, we also proposed a DSL scheme for RetiFluidNet model training where misrepresentations on different scales are incorporated into the computation of the overall cost function. In this scheme, as shown in Fig.2, the FMs with different scale information from the middle decoder blocks are employed to generate auxiliary output tensors, which include standard convolutional outputs and Bicon output predictions at different scales. The tensors are then used with a scale-dependent loss weighting method to compute the DLC and CLC terms, while the weighting method regulates the impact of local losses for texture- and edge-specific penalties, respectively.

6) (6-A) DLC Term Computation: We used an up-sampling layer by a factor of P followed by a Conv2D layer with C filters, 1 × 1 kernel size, and a Softmax activation function after each decoder block. Here, P and C represent the scale-dependent up-sampling factor and the number of fluid classes, respectively. For each block, the size of the local FMs is compared to the size of the original ground truth mask, and P is chosen accordingly. Then, we integrated the decoder blocks’ outputs with the original output dice loss into the DLC component in the proposed compound objective function for model training. First, we define the dice loss function:

\[
L_{d} (y_{pred}, y_{true}) = 1 - \frac{2 \sum_{i,j} H_{i,j} W (y_{pred} \times y_{true})}{\sum_{i,j} H_{i,j} W (y_{pred}) + \sum_{i,j} H_{i,j} W (y_{true})},
\]

where H and W are the height and width of the input image, respectively. Thus, the DLC term, which is a weighted-sum version of the multi-scale dice loss components, is given by:

\[
DLC = L_{d} (y_{pred}, y_{true}) + \sum_{l=1}^{L} \frac{1}{2^{l-1}} L_{d} (y_{pred,l}, y_{true}),
\]

Here, \(l\) indicates the index of auxiliary output, \(y_{pred,l}\) is the \(l^{th}\) auxiliary predicted output and \(y_{true}\) is the annotated ground truth. Fig.2 illustrates more details about the local loss definitions. As seen from the figure, our proposed model mainly consists of \(L = 4\) middle encoder blocks for auxiliary output generation, and \(l_0\) indicates the original output scale.

7) (6-B) CLC Term Computation: To further incorporate edge information into our local loss computations, we employed a modified and multi-class version of the Bicon loss (BL) function proposed in [53]. Therefore, the CLC term is defined as a weighted-sum version of the multi-scale BL local loss components and can be expressed by:

\[
CLC = BL (y_{pred}, y_{true}) + \sum_{l=1}^{L} \frac{1}{2^{l-1}} BL_{scale_l} (y_{pred,l}, y_{true}),
\]

Here:

\[
BL = \frac{1}{2} L_{decouple} + L_{con.map} + L_{con.dice}.
\]

The first term in Eq.13, \(L_{decouple}\), is the edge-decoupled loss and is computed using the categorical cross-entropy (CCE) loss between the edge-decoupled map \(S_{decouple}\) and the ground truth segmentation mask \(y_{true}\) in the one-hot coding format, given by: (Eq.14), as shown at the bottom of the next page.

In this equation, \(L_{BCE}\) is the binary cross-entropy loss function, \(BCM\) indicates the bilateral connectivity map [53], and \(P_{edge,c}\) refers to the set of ground truth edge pixels that are obtained from the connectivity mask [56] for class \(c\).

In Eq.13, we use the following definition of connectivity-map loss \(L_{con.map}\):

\[
L_{con.map} = L_{CCE} (CM, y_{true,con}) + \sum_{c=1}^{C} L_{BCE} (CM_{c}, y_{true,con,c}),
\]

where \(CM\) and \(y_{true,con}\) are the connectivity map and connectivity mask, respectively [53]. \(y_{true,con}\) is generated from the original segmentation mask \(y_{true}\) according to [53].

The third term in Eq.13, \(L_{con.dice}\), is the connectivity-based dice loss, defined by Eq.10 when \(y_{pred}\) is replaced by \(S_{global}\). Here, \(S_{global}\) is constructed by the maximum connection probability across main output channels [53].

Finally, we define the overall objective function of our network as:

\[
Loss_{joint} = DLC + \lambda \cdot CLC
\]

IV. EXPERIMENTAL STUDY AND RESULTS
A. Experimental Setup

In this section, we first introduce our strategy for data augmentation, and then implementation details are presented.

1) Data Augmentation Strategy: The image data was augmented with random translations, rotations, contrast adjustments, and mirroring, resulting in eight augmented images for each image in the training datasets.
2) Implementation Details: Our implementation is based on TensorFlow v2.4.0, and Python 3.7. All tests were run on a PC with NVidia GeForce Titan-V GPU (12GB), and 128 GB of RAM. For model training, the hyperparameters, including the maximum number of epochs (30), learning rate (2e-4), learning rate decay schedule (learning rate was multiplied by 0.8 every 5 epochs), batch size (4), and optimization method (the RMSprop optimizer) were the same for all of the baselines. In addition, the seed point for random data partitioning in the cross-validation method was preserved the same for all the networks to ensure the evaluation results have no bias due to the data splitting step. Throughout the experiments, the lambda hyperparameter was set to 0.05 and was optimized using the grid search method with a range of [0.01, 0.05, 0.1, 0.5, 1] over the RETOUCH training subsets. The code scripts will be available at: https://github.com/aiddialab/RetiFluidNet.

B. Performance Measures

The following evaluation measures were used to calculate the segmentation performance of different methods in this study. Dice similarity coefficient (DSC) and balanced accuracy (ACC) were used for evaluating the similarity between OCT images and ground truths and pixel-wise classification performance according to the following equations:

\[
DSC = \frac{2|y_{pred} \cap y_{true}|}{|y_{pred}| + |y_{true}|},
\]

\[
Acc = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right),
\]

For pixel classification, TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. The segmentation result and ground truth are represented by \( y_{pred} \) and \( y_{true} \), respectively.

C. Validation Method

To evaluate and generalize the performance of the deep learning models, we used the unbiased k-fold cross-validation (CV) method. Folding was performed at the subject level to ensure that the B-scans from the same OCT were not used in both the training and testing sets throughout each iteration. The model’s final performance was obtained by averaging the evaluation metrics on testing sets over iterations.

D. Study Design

The following studies were conducted in this work to demonstrate the competence of the proposed model and to obtain a benchmark for comparing the performance of RetiFluidNet in the retinal fluid segmentation task.

1) Baseline Study: We compared the segmentation performance of the proposed model to those of the common image segmentation CNN models, including U-Net [22], FCN [23], Seg-Net [24], Deeplabv3+ [25], APC-Net [54], DA-Net [48], DDR-Net [55], and also the recently developed medical image segmentation models including U-Net++ [26], Attention U-Net [33], CE-Net [52], CPP-Net [35], nnU-Net [56], and MsTGANet [36]. All the networks were trained using the same configuration to ensure fairness in the evaluation step. The cross-validated results on the RETOUCH challenge dataset are reported in Table III.

2) State-of-the-Art Analysis: The proposed method was also compared to several recently developed fluid region segmentation algorithms, including patch-based U-Net segmentation method [3]; multi-scale and dual attention enhanced nested U-Net (MDAN-UNet) segmentation model [27]; multi-label deep supervision segmentation method [42]; attention-based U-Net segmentation approach [34]; generative adversarial network (GAN)-based segmentation-renormalized image translation framework [43], and retinal fluids segmentation network (RFS-Net) proposed in [47]. For fair comparisons, the target datasets, pre/post-processing pipelines, and evaluation methodologies of the original studies were meticulously followed. The comparative results are summarized in Table IV.

3) Ablation Study: In the proposed method, the SDA attention module, the SASC paths, and the multi-scale DSL scheme are embedded into the basis U-Net architecture. To further understand the influence of each component in our model, we conducted a comprehensive ablation study, and the experimental results are reported in Table V. The customized U-Net (cU-Net) model used in RetiFluidNet was adopted as our baseline model to evaluate the effectiveness of the SDA, SASC, and DSL scheme. The cU-Net is the backbone of the proposed RetiFluidNet model in Fig.2, whereas the SDA modules and DSL scheme are eliminated from the RetiFluidNet structure, and the SASC paths are replaced by simple SCs similar to the standard U-Net [33]. In the first step, we explored the contribution of the attention module to the cU-Net performance by replacing the proposed SDA (in Fig.2) with the competitive squeeze and excitation (SqEx) [49], squeeze-and-attention (SqAt) [50], adaptive context module (ACM) [54], or dual-attention (DA) [48] mechanisms. The attention-based SCs were also modified accordingly. For this experiment, the loss function in Eq. 10 was considered for the optimization process. Next, the DSL scheme contribution was analyzed on the best contributing model. For this purpose, different combinations of local DLC, CLC, or joint losses were evaluated. The ablation experiments were carried out on the RETOUCH dataset using the 3-fold CV method.

4) Statistical Analysis: The statistical significance of the performance improvement for the proposed RetiFluidNet versus the best alternative models was evaluated using the Wilcoxon
TABLE III
BASELINE PERFORMANCE EVALUATION OVER RETOUCH DATASET USING DSC & ACC METRICS AND THE 3-FOLD CROSS-VALIDATION METHOD

| Scanner | Metrics |
|---------|---------|
| DSC | SRF | PFD |
| U-Net | FCN | Seg-Net | DeepLabv3+ | AHC-Net | DA-Net | DOR-Net | U-Net++ | Attention-Net | CS-Net | CRF-Net | uCLINet | RetiFluidNet |
| 52.0±19.9 | 69.8±5.7 | 68.6±7.1 | 63.3±9.9 | 60.6±7.5 | 67.6±6.7 | 76.4±6.5 | 71.3±5.4 | 76.5±5.1 | 73.3±7.4 | 70.9±3.1 | 78.7±2.5 | 80.0±2.1 | 84.5±2.5 |
| 75.5±20.7 | 77.3±8.8 | 76.6±8.4 | 88.6±6.6 | 77.4±8.8 | 76.4±7.8 | 79.5±7.6 | 81.6±5.6 | 71.1±6.3 | 80.9±4.2 | 85.6±2.2 | 84.8±0.5 | 83.5±3.2 | 90.7±2.2 |
| 77.3±27.7 | 76.8±6.5 | 78.5±4.5 | 84.5±5.0 | 79.9±4.5 | 79.2±4.1 | 82.9±10.0 | 79.2±4.7 | 74.3±13.1 | 76.1±4.0 | 83.9±5.8 | 85.1±4.5 | 86.4±4.9 | 91.5±1.7 |

TABLE IV
TABLE OF THE-ART PERFORMANCE EVALUATION OVER STUDY DATASETS

| Reference | Method Configuration | Validation Method | Dataset | Metric | Ave-Result(s) (%) |
|-----------|----------------------|------------------|---------|--------|-------------------|
| Aslani et al. [2020] | 3D CNN | 3-Fold CV | RETOUCH (OCTA) | DSC | 84.3±2.3 |
| Wu et al. [2020] | CNN-based Layer Segmentation | 3-Fold CV | DSC | DSC | 86.1±3.5 |
| Xue et al. [2021] | 3D CNN | 3-Fold CV | OCTA | DSC | 85.1±4.5 |
| Ru et al. [2021] | Attention-Based Segmentation | 3-Fold CV | OCTA | DSC | 84.9±2.8 |
| Haas et al. [2021] | NLM Filtering, Total Retic Segmentation | 3-Fold CV | OCTA | DSC | 84.9±2.8 |

TABLE V
ABLATION STUDY RESULTS OVER RETOUCH DATASET USING DSC METRIC AND THE 3-FOLD CROSS-VALIDATION METHOD

| Method | Results(s) (%) | Ave. (%) |
|--------|----------------|----------|
| IRF | 70.3±13.7 | 73.1±13.2 |
| SRF | 70.3±13.7 | 73.1±13.2 |
| PFD | 70.3±13.7 | 73.1±13.2 |

V. DISCUSSION

A. Quantitative Evaluation

The mean and standard deviation values of the DSC and balanced ACC measures were used to quantitatively evaluate the retinal fluid segmentation performance on the RETOUCH dataset in the baseline study. As shown in Table III, our proposed model outperformed the comparative baseline methods, with average DSC and ACC values of 86.6% and 94.9%, respectively, over different fluid types and scanners. When compared to MsTGANet, the best performer of all baseline methods, the average DSC and ACC values for RetiFluidNet were 3.1% and 0.7% higher, respectively. Despite the fact that

signed-rank test. We determined the p-values for comparisons of the DSC metric, in which ten repetitions of the 3-fold CV method were employed for statistical significance assessment. Table VI shows the results of the statistical analysis.

V. DISCUSSION

The mean and standard deviation values of the DSC and balanced ACC measures were used to quantitatively evaluate the retinal fluid segmentation performance on the RETOUCH dataset in the baseline study. As shown in Table III, our proposed model outperformed the comparative baseline methods, with average DSC and ACC values of 86.6% and 94.9%, respectively, over different fluid types and scanners. When compared to MsTGANet, the best performer of all baseline methods, the average DSC and ACC values for RetiFluidNet were 3.1% and 0.7% higher, respectively. Despite the fact that
RetiFluidNet; the results are shown in Table V (Attention Module Analysis section). When cU-Net+SDA+SA.SC+DSC is compared to cU-Net+DSC, the average DSC improves from 73.5% to 85.3%, as the proposed SDA module and SASC paths can adaptively guide the model to learn multiscale local/non-local features with long dependency information. Additionally, we conducted experiments to compare the performance of the SqEx [49], SqAt [50], ACM [54], and DA [48] attention modules. When compared to cU-Net+SqEx+SqAt.SC, cU-Net+SqAt+SqAt.SC, cU-Net+ACM+ACM.SC, and cU-Net+DA+DA.SC, our proposed cU-Net+SDA+SA.SC achieves superior performance in DSC indices (with 9.1%, 7.9%, 2.6%, and 3.1% increases, respectively) while maintaining the same time efficiency, demonstrating the effectiveness of the proposed SDA module with adaptive fusion. Furthermore, Table V (DSL Scheme Analysis section) demonstrates that incorporating the DSL scheme into the cU-Net improved segmentation performance, with cU-Net+Multi-scale-DLC outperforming cU-Net in terms of the DSC indicator. When the CLC term was further integrated into the objective function, the DSC of cU-Net+Multi-scale-JointLoss was improved from 73.5% to 80.5%. These results further indicate that the proposed DSL scheme considerably improves the performance of the model. Additionally, as shown in Table V, the segmentation performance of the RetiFluidNet is significantly better than that of cU-Net+Multi-scale-JointLoss: the average DSC for our model improved from 75.3%, 84.9%, and 81.2% to 91.1%, 93.6%, and 92.8% respectively, for IRF, SRF, and PED segmentation. Compared with RetiFluidNet without CLC, the DSC of RetiFluidNet improved considerably by 6.1%, 0.9%, and 5.5% for IRF, SRF, and PED segmentation, respectively. RetiFluidNet also outperformed the method that only used CLC without DLC, with the DSC of RetiFluidNet increasing on average by 3% (p-value<0.05) for fluid segmentation. These results demonstrate that the proposed RetiFluidNet with multi-scale joint-loss can significantly improve the fluid segmentation performance by further leveraging the edge-connectivity-based information and the key textural and contextual information.

### B. Qualitative Assessment

Fig.4 shows sample segmentation results for the proposed RetiFluidNet on the RETOUCH test datasets and comparisons with six high-performance baseline networks and the RetiFluidNet without CLC. The IRF, SRF, and PED fluid segmentations are depicted in yellow, red, and green, respectively. As shown in Fig.4, the proposed model obtains higher segmentation performance, especially when segmenting small-sized retinal fluids. The competitive U-Net++ model (top and middle rows) suffers from false positives caused by the interference of non-correlated local information and speckle noise. Moreover, the proposed RetiFluidNet achieves better segmentation performance than nnU-Net, CE-Net, MsTGANet, and RetiFluidNet without the CLC term, in which speckle noise interferences and blurry edges negatively affect the results. Taken together, these results demonstrate the effectiveness and robustness of our proposed method.

### C. Statistical Comparison

The statistical significance of the performance improvement for the proposed RetiFluidNet versus the best alternative models is shown in Table IV. All gains in DSC for RetiFluidNet (compared to other methods) are statistically significant with p-values less than 0.05, except for the methods described by Alsaih et al. [3], and Ren et al. [43] for IRF segmentation. These findings support the effectiveness of the proposed attention module that was embedded in the RetiFluidNet. The table also reports the p-values of the proposed DSL scheme in comparison to the other competitive alternatives in the ablation study. All DSC improvements for RetiFluidNet are statistically significant, with p-values less than 0.05, with the exception of RetiFluidNet without CLC for IRF segmentation and RetiFluidNet without DLC for PED analysis. These results demonstrate that the suggested DSL scheme would significantly improve segmentation performance by leveraging edge-connectivity information for multi-scale representing fluid regions. Comparison to other segmentation models showed that the proposed RetiFluidNet benefits from hierarchical representation learning of textural, contextual, and edge features using the SDA module, multiple SASC connections, and the multi-scaled DSL scheme. The attention mechanism of the proposed SDA module enables the model to automatically extract deformation-aware representations at multiple levels, and the introduced SASC paths take spatial-channel interdependencies into account when concatenating counterpart encoder and decoder units, thereby improving local and global representational capability. The SDA’s contribution is to highlight the spatial or channel

| Compared Methods | P-value | IRF | SRF | PED | IRC*
|---|---|---|---|---|---|
| Deeplabv3+ [25] | 0.026 | <0.001 | <0.001 | - |
| U-Net++ [26], DA-Net [48], CE-Net [52], APC-Net [34], DDR-Net [55] | <0.001 | <0.001 | <0.001 | - |
| CPF-Net [35] | <0.001 | <0.001 | 0.015 | - |
| MsTGANet [36] | 0.033 | 0.027 | 0.041 | - |
| nnU-Net [56] | <0.001 | 0.019 | 0.003 | - |
| Alsaïh et al. [3] | 0.088 | <0.001 | 0.002 | <0.001 |
| X.Liu et al. [34] | - | - | - | 0.038 |
| Ren et al. [43] | <0.001 | <0.001 | - | - |
| Hassan et al. [47] | | <0.001 | <0.001 | - |
| RetiFluidNet without CLC | 0.051 | 0.021 | 0.003 | - |
| RetiFluidNet without DLC | 0.033 | 0.040 | 0.009 | - |

*The average DSC metric for the mixed OPTIMA and DUKE datasets was used to estimate the p-values.
information adaptively for better representation. The $\alpha$ and $\beta$ values are optimized at each processing block/scale to adaptively capture cross-pixel and cross-channel dependencies in order to effectively highlight local and global structures for more precise segmentation performance.

Additionally, RetiFluidNet was optimized using a joint loss function that incorporated a weighted version of dice overlap and edge-preserved connectivity-based losses as well as multiple hierarchical levels of multi-scaled local losses, all of which resulted in improved segmentation.

**D. Combine With Pre-Trained CNN Backbones**

RetiFluidNet can be combined with any convolutional segmentation backbone. As an alternative backbone, we also employed ResNet50 for the encoder block of the model, which is frequently used in natural image semantic segmentation. The experimental results showed that the segmentation performance of res50-RetiFluidNet in the best training mode (i.e., fine-tuning) is not significantly superior to that of RetiFluidNet on the RETOUCH dataset. RetiFluidNet and res50-RetiFluidNet have shown comparable average dice performance of 92.5%±3.2 and 92.2%±2.1, respectively. The fact that the RETOUCH dataset was acquired from three different imaging devices is a challenging factor. It is difficult for the ResNet50 encoder block to manage fluid images from different data sources with varying shapes, while the OCT fluid segmentation datasets are quite small compared to other natural image segmentation datasets. Therefore, utilizing a pre-trained CNN model with a large number of parameters as the encoder on a limited-size dataset may result in performance degradation due to overfitting. Therefore, to apply our methodology to a real-world dataset, one can therefore select a suitable backbone network to adapt to various datasets. In future work, we plan to develop a 3-D version of our model for volumetric fluid region segmentation and accelerate the multi-class training phase of our proposed segmentation network.

**VI. CONCLUSION**

In this study, we proposed a deep learning framework called the RetiFluidNet model for segmenting different types of fluid in macular OCT images. We trained and evaluated this model on publicly available datasets acquired on different OCT scanners. RetiFluidNet is well-suited for clinical applications where quantification of retinal fluids is required by providing a fast and robust automated technique for precise fluid segmentation. Experimental results showed that RetiFluidNet requires minimal image pre-processing steps, exhibits high performance in terms of Dice and ACC metrics, and is a reliable method for segmenting retinal OCT images even in the presence of disruptive speckle noise and severe pathologies.

**REFERENCES**

[1] M. E. J. van Velthoven, D. J. Faber, F. D. Verbraak, T. G. van Leeuwen, and M. D. de Smet, “Recent developments in optical coherence tomography for imaging the retina,” *Prog. Retinal Eye Res.*, vol. 26, no. 1, pp. 57–77, 2007.

[2] 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.

[3] K. Alsaih, M. Z. Yusoff, I. Faye, T. B. Tang, and F. Meriaudeau, “Retinal fluid segmentation using ensembled 2-dimensionally and 2.5-dimensionally deep learning networks,” *IEEE Access*, vol. 8, pp. 152452–152464, 2020.

[4] R. Rasti et al., “Deep learning-based single-shot prediction of differential effects of anti-VEGF treatment in patients with diabetic macular edema,” *Biomed. Opt. Exp.*, vol. 11, no. 2, pp. 1139–1152, Feb. 2020.

[5] K. Alsaih, M. Z. Yusoff, T. B. Tang, I. Faye, and F. Meriaudeau, “Deep learning architectures analysis for age-related macular degeneration segmentation on optical coherence tomography scans,” *Comput. Methods Programs Biomed.*, vol. 195, Oct. 2020, Art. no. 105566.

[6] D. C. DeBuc and G. M. Somfai, “Early detection of retinal thickness changes in diabetes using optical coherence tomography,” *Med. Sci. Monitor*, vol. 16, no. 3, pp. MT15–MT21, Mar. 2010.

[7] V. G. Vidotti et al., “Sensitivity and specificity of machine learning classifiers and spectral domain OCT for the diagnosis of glaucoma,” *Eur. J. Ophthalmol.*, vol. 23, no. 1, pp. 61–69, Jan. 2013.

[8] M. Zhang et al., “Advanced image processing for optical coherence tomographic angiography of macular diseases,” *Biomed. Opt. Exp.*, vol. 6, no. 12, pp. 4661–4671, 2015.

[9] Y. Wang, Y. Zhang, Z. Yao, R. Zhao, and F. Zhou, “Machine learning based detection of age-related macular degeneration (AMD) and diabetic macular edema (DME) from optical coherence tomography (OCT) images,” *Biomed. Opt. Exp.*, vol. 7, no. 12, pp. 4928–4940, 2016.

[10] M. Awais, H. Müller, T. B. Tang, and F. Meriaudeau, “Classification of SD-OCT images using a deep learning approach,” in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Sep. 2017, pp. 489–492.

[11] O. Perdomo, S. Otalora, F. A. Gonzalez, F. Meriaudeau, and H. M’uller, “OCT-NET: A convolutional network for automatic classification of normal and diabetic macular edema using SD-OCT volumes,” in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1423–1426.

[12] M. Pekala, N. Joshi, T. Y. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, “Deep learning based retinal OCT segmentation,” *Comput. Biol. Med.*, vol. 114, Nov. 2019, Art. no. 103445.
