Context attention-and-fusion network for multiclass retinal fluid segmentation in OCT images

Ye, Yanqing, Chen, Xinjian, Shi, Fei, Xiang, Dehui, Pan, Lingjiao, et al.

Yanqing Ye, Xinjian Chen, Fei Shi, Dehui Xiang, Lingjiao Pan, Weifang Zhu, "Context attention-and-fusion network for multiclass retinal fluid segmentation in OCT images," Proc. SPIE 11596, Medical Imaging 2021: Image Processing, 1159622 (15 February 2021); doi: 10.1117/12.2580527

Event: SPIE Medical Imaging, 2021, Online Only

Context attention-and-fusion network for multiclass retinal fluid segmentation in OCT images

Yanqing Ye[1], Xinjian Chen[1,2], Fei Shi[1], Dehui Xiang[1], Lingjiao Pan[3], Weifang Zhu[1,*]

[1]School of Electronics and Information Engineering, Soochow University, Suzhou, 215006, China
[2]State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, 215123, China
[3] School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, Jiangsu Province, 213000, China

ABSTRACT

Optical coherence tomography (OCT) is an imaging modality that is extensively used for ophthalmic diagnosis and treatment. OCT can help reveal disease-related alterations below the surface of the retina, such as retinal fluid which can cause vision impairment. In this paper, we propose a novel context attention-and-fusion network (named as CAF-Net) for multiclass retinal fluid segmentation, including intraretinal fluid (IRF), subretinal fluid (SRF) and pigment epithelial detachment (PED). To deal with the seriously uneven sizes and irregular distributions of different types of fluid, our CAF-Net proposes the context shrinkage encode (CSE) module and context pyramid guide (CPG) module to extract and fuse global context information. The CSE module embedded in the encoder path can ignore redundant information and focus on useful information by a shrinkage function. Besides, the CPG module is inserted between the encoder and decoder, which can dynamically fuse multi-scale information in high-level features. The proposed CAF-Net was evaluated on a public dataset from RETOUCH Challenge in MICCAI2017, which consists of 70 OCT volumes with three types of retinal fluid from three different types of devices. The average of Dice similarity coefficient (DSC) and Intersection over Union (IoU) are 74.64% and 62.08%, respectively.

KEYWORDS: Retinal fluid segmentation, Optical coherence tomography, Convolutional neural network, Context shrinkage encode module, Context pyramid guide module

1. INTRODUCTION

Retinal fluid refers to the accumulation of the leaked fluid within the intercellular space of the retina due to the disruptions in blood-retinal barrier. It occurs secondary to many retinal diseases such as diabetic retinopathy (DR), age-related macular degeneration (AMD) and retinal vain occlusion (RVO) [1]. These three retinal diseases are currently the most common causes of vision loss in the world, affecting hundreds of millions of people. Optical coherence tomography (OCT) is widely used in clinical diagnosis of ophthalmic diseases due to a series of advantages such as high resolution, non-invasive, and fast speed [2]. Therefore, as typical imaging biomarkers, accurate segmentation and quantification of fluid in OCT images is vital for the assessment of retinal diseases.

Three types of retinal fluid including intraretinal fluid (IRF), subretinal fluid (SRF) and pigment epithelial detachment (PED) are clinically distinguishable in optical coherence tomography (OCT) images [3], as shown in Fig.1. IRF appears as separated hyporeflective cystoid pockets located in the inner and outer nuclear layers that increase the overall retinal thickness. SRF corresponds to the accumulation of exudate between the neurosensory retina and retinal pigment epithelium (RPE), leading to retinal detachment. PED is the separation of RPE from the Bruch's membrane (BM), which can be subdivided into serous, fibrovascular and drusenoid types. In fact, IRF, SRF and PED may occur simultaneously in clinical cases.

[*]Corresponding author: Weifang Zhu, E-mail: wfzhu@suda.edu.cn

Medical Imaging 2021: Image Processing, edited by Ivana Išgum,
Bennett A. Landman, Proc. of SPIE Vol. 11596, 1159622 · © 2021
SPIE · CCC code: 1605-7422/21/$21 · doi: 10.1117/12.2580527

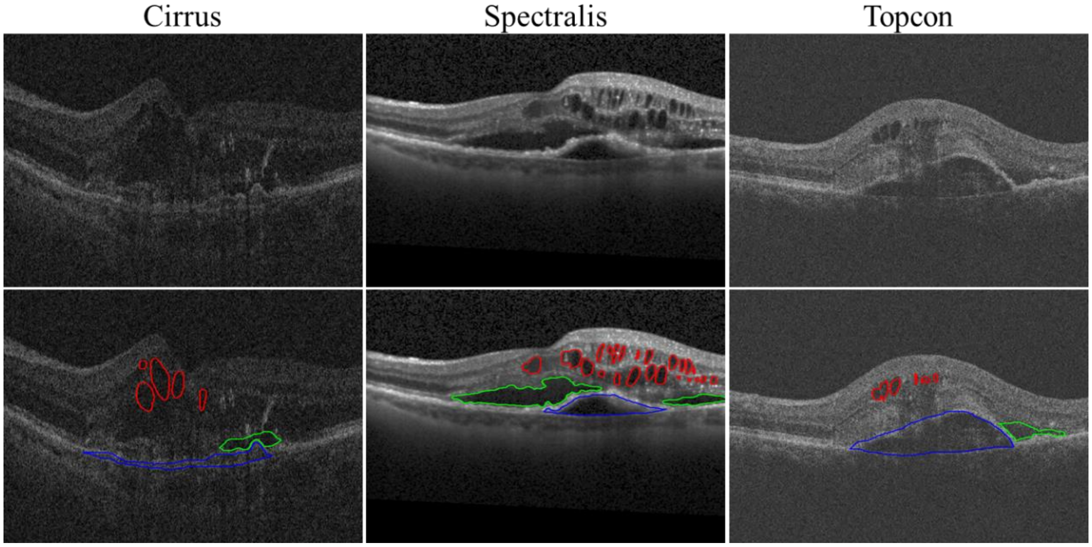


Fig. 1. Retina imaged with OCT scanners from three different vendors: Cirrus, Spectralis and Topcon. The images on the first row represent the original OCT B-scans, while the regions surrounded by red, green and blue curves on the second row represent IRF, SRF and PED respectively.

In recent years, there are some studies focused on the automatic segmentation of retinal fluid in OCT images. Lee *et al*. developed the AMD intelligent assisted diagnostic system, which can segment multiple lesions, including IRF, SRF, PED, and subretinal hyperreflective lesion area [4]. Lu *et al*. applied Graph-Cut to perform layer segmentation as pre-processing and utilized convolutional neural network (CNN) to segment retinal fluid [5]. Rashno *et al*. presented a fully-automated method based on graph shortest path layer segmentation and fully convolutional networks (FCNs) for fluid segmentation [6]. Furthermore, ReLayNet proposed by Roy *et al*. can segment the retinal layer and fluids simultaneously in abnormal and normal OCT images [7]. Compared to the binary segmentation tasks, multiclass segmentation will be more difficult because different diseases may interfere with each other. Multiclass retinal fluid segmentation, which means the joint segmentation of IRF, SRF and PED, is a great challenge because of the high variability in the appearance of these three lesions in OCT images. In addition, the acquired image quality and scan density vary widely between different types of OCT devices that will reduce the versatility of the algorithm.

To deal with these problems, we propose an end-to-end context attention-and-fusion network named as CAF-Net for the automatic segmentation of multiclass retinal fluid in OCT images. The context shrinkage encode (CSE) module and context pyramid guide (CPG) module are proposed and integrated into a U-shape convolutional neural network. Experiments show that these two modules are able to extract and fuse global context information, thereby effectively improving the segmentation results of the network.

2. METHODS

We introduce the proposed method in the following four parts: the structure of the proposed CAF-Net, the context shrinkage encode (CSE) module, the context pyramid guide (CPG) module and the loss function.

2.1 Overall structure of the proposed CAF-Net

U-Net [8] has pioneered the use of U-shape architecture for biomedical image segmentation, which comprises encoder and decoder paths and provides dense output for each pixel. Without changing the depth of U-Net, we reduce the number of feature maps by half at each stage to get a baseline network with compact size, and improve it with context shrinkage encode (CSE) module and context pyramid guide (CPG) module for multiclass retinal fluid segmentation. Fig.2 shows the overall structure of the proposed CAF-Net. In order to discard redundant information and focus on useful information, the CSE module is designed and embedded in the encoder path of the network. Moreover, the CPG module is inserted between the encoder and decoder to dynamically guide the fusion of multiscale information in high-level features.

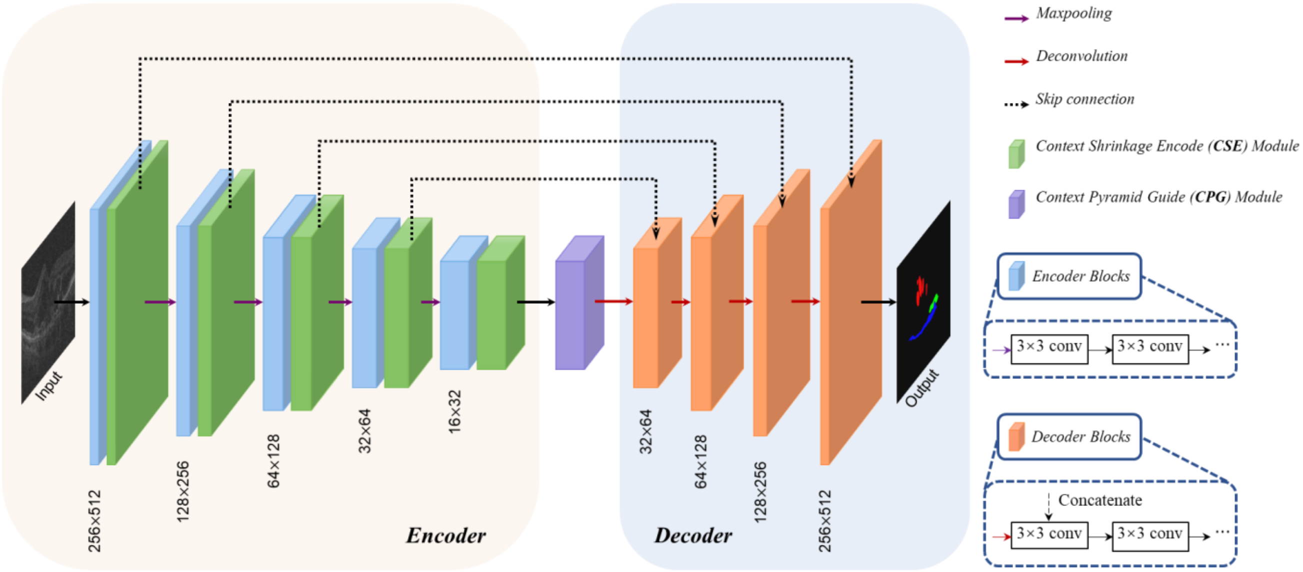


Fig. 2. An overview of the context attention-and-fusion network (CAF-Net).

2.2 Context shrinkage encode module

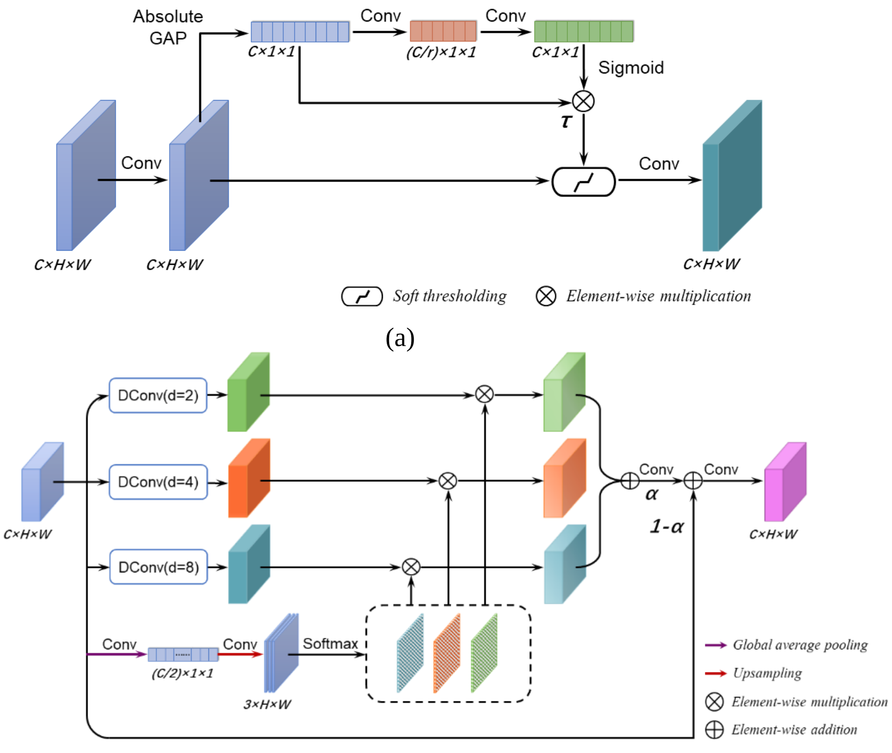
Soft thresholding [9] is a function that shrinks the inputs toward zero and has been used as a key step in many signal denoising methods, which can be expressed by

$$y = \begin{cases} x - \tau & x > \tau \\ 0 & -\tau \le x \le \tau \\ x + \tau & x < -\tau \end{cases} \tag{1}$$

Where x is the input feature, y is the output feature, and τ is the threshold, i.e., a positive parameter. In fact, the noise can be broadly understood as information that is irrelevant to the task at present. As we all know, the encoder path of network can transform useful information to very positive or negative features, and redundant information to near-zero features. In hence, inspired by soft thresholding, we design a context shrinkage encode (CSE) module to convert redundant information to zero while retaining useful information. As shown in Fig.3(a), we first use 3×3 convolution to transform the input feature maps to extract semantic information. Then the absolute value of the transformed feature maps go through global average pooling, two 1×1 convolution and sigmoid excitation to obtain a scaling parameter between 0 and 1. After that, the scaling parameter is multiplied by the average of the absolute value of transformed feature maps to get the channel-wise threshold τ. Features can be distinguished by the threshold vector which is adaptively determined based on global context information. The CSE module can effectively enhance the ability of network to extract key information from noise-containing feature maps.

2.3 Context pyramid guide module

Spatial pyramid structure [10] can effectively utilize multi-scale information to improve the performance of semantic segmentation tasks. Inspired by this, we propose a context pyramid guide (CPG) module which is shown in Fig.3(b). We use three parallel dilated convolutions with different dilation rates of 2, 4 and 8 to obtain three groups of feature maps with multi-scale information. At the same time, the input feature maps go through global average pooling and 1×1 convolution to extract global context information and reduce the number of feature maps by half. Then we use upsampling, 3×3 convolution and softmax operation to obtain three feature maps with spatial weights, which are multiplied to the previously obtained three groups of feature maps, respectively. The element-wise addition operation followed by 3×3 convolution is used to fuse them. Finally, a residual connection with learnable parameter α is employed to obtain the output of the CPG module. The function of CPG module is to utilize spatial context information to guide the network to dynamically select appropriate scale features and fuse them by self-learning.



(a)

(b)

Fig. 3. Components of the proposed two modules. (a) Context shrinkage encode (CSE) module. (b) Context pyramid guide (CPG) module.

2.4 Loss function

We employ a hybrid loss consisting of cross-entropy loss L_{CE} and multi-class exponential logarithmic Dice loss L_{ELDice}. L_{ELDice} can address the pixel imbalance problem, while L_{CE} is employed to alleviate the problem that exponential logarithmic Dice loss is too sensitive. The hybrid loss L makes the segmentation of small targets more accurate and guarantees the stability of the training process to some extent. The loss functions are defined as follows:

$$L_{CE} = -\frac{1}{N}\sum_{i}^{N}\sum_{c=0}^{C-1} y_{i,c}\log(p_{i,c}) \tag{2}$$

$$L_{ELDice} = \frac{1}{C}\sum_{c=0}^{C-1}\left(-\ln\left(\frac{2\sum_i y_{i,c}\,p_{i,c}+\varepsilon}{\sum_i(y_{i,c}+p_{i,c})+\varepsilon}\right)^{\gamma}\right) \quad (\gamma = 0.5) \tag{3}$$

$$L = L_{CE} + L_{ELDice} \tag{4}$$

where $y_{i,c}$ denotes the ground truth for pixel i being class c, $p_{i,c}$ denote the predicted probabilities for pixel i being class c, N is the is the total size of feature map, C is the total number of class, and ε is the small smoothing factor.

3. RESULTS

3.1 Dataset

The dataset [11] used in this paper was acquired from a public competition: RETOUCH Challenge in MICCAI2017. Since the label of the test set of the competition is not available, a total of 70 OCT volumes (labeled as IRF, SRF, PED and normal), with 24, 24 and 22 volumes acquired from three different types of devices, including Cirrus, Spectralis and Topcon, are used in our experiments. For each volume from these three devices, the numbers of B-scans are 128, 49, 128, respectively. Although not all B-scans contain fluid, there is at least one type of fluid in each volume.

3.2 Parameter settings

The implementation of our network is based on the Pytorch framework. The dataset is randomly divided into 3 folds according to subjects and 3-fold cross-validation strategy is adopted to evaluate the performance of the proposed method. In order to improve the efficiency of training, the size of OCT B-scans is cropped to 256×512. In the training process, stochastic gradient descent (SGD) algorithm with poly learning rate policy is used to optimize the weight of the network. The base learning rate and momentum are set to 0.01 and 0.9, respectively. The batch size is set to 8 and the proposed model is trained for 50 epochs in the experiment. To increase the generalization of the model, some data augmentation strategies, including random horizontal flip, random rotation and Gaussian noise addition are applied during training process.

3.3 Evaluation metrics

To show the performance of the proposed method, it has been objectively evaluated with three metrics, including Dice Similarity Coefficient (DSC) [12], intersection-over-union (IoU) and Accuracy (Acc). The metrics are computed from TP, FP, TN and FN, which represent true positive, false positive, true negative and false negative predictions, respectively.

$$DSC = 2TP/(2TP + FP + FN) \tag{5}$$

$$IoU = TP/(TP + FP + FN) \tag{6}$$

$$Acc = (TP + TN)/(TP + FP + TN + FN) \tag{7}$$

3.4 Experimental results

To prove the advantages of our proposed method, we perform comparison experiments with other networks which are widely used in the semantic segmentation field, including FCN [13], PSPNet [14], CE-Net [15] and U-Net. In order to verify the effectiveness of CSE and CPG module, we use U-Net as baseline and perform ablation experiments.

As can be seen from Table 1, in terms of Dice and IoU, U-Net achieves better segmentation results than FCN, PSPNet and CE-Net. After the CSE and CPG modules are inserted into U-Net respectively, the average DSC are improved by 1.54% and 0.86% respectively, which implies the necessary and effectiveness of the proposed CSE and CPG module. Finally, the proposed CAF-Net performs best in all evaluation metric with average DSC of 74.64%, IoU of 62.08% and Acc of 99.26%, which shows the combination of CSE and CPG module contribute well to the final performance.

Table 1 The performance of retinal fluid segmentation with different methods

	DSC (%)				IoU (%)				Acc (%)
Methods	IRF	SRF	PED	Ave	IRF	SRF	PED	Ave	Glob
FCN	65.24	72.95	64.19	67.46	49.87	61.41	50.42	53.90	98.98
PSPNet	66.28	74.09	66.40	68.92	51.14	62.68	53.25	55.69	99.08
CE-Net	69.02	75.12	68.34	70.83	54.70	64.07	55.98	58.25	99.16
U-Net	71.09	75.26	68.50	71.62	56.74	64.41	55.63	58.93	99.13
U-Net + CSE	72.76	78.12	68.60	73.16	58.72	66.92	56.11	60.58	99.22
U-Net + CPG	72.32	75.84	69.29	72.48	58.08	64.61	56.64	59.78	99.19
CAF-Net	**73.17**	**79.70**	**71.06**	**74.64**	**59.21**	**68.12**	**58.92**	**62.08**	**99.26**

Fig.4 shows some retinal fluid segmentation results of different methods. Due to multiclass segmentation training at the same time, the results of joint segmentation are affected by the interference of different class fluids. It can be seen that the results of the proposed CAF-Net are more accurate in terms of boundaries and details with different sizes, shapes and intensities of fluids. Compared with other methods, our method can predict more true positives and less false positives, which indicates that exploit rich global context information is conducive to image segmentation.

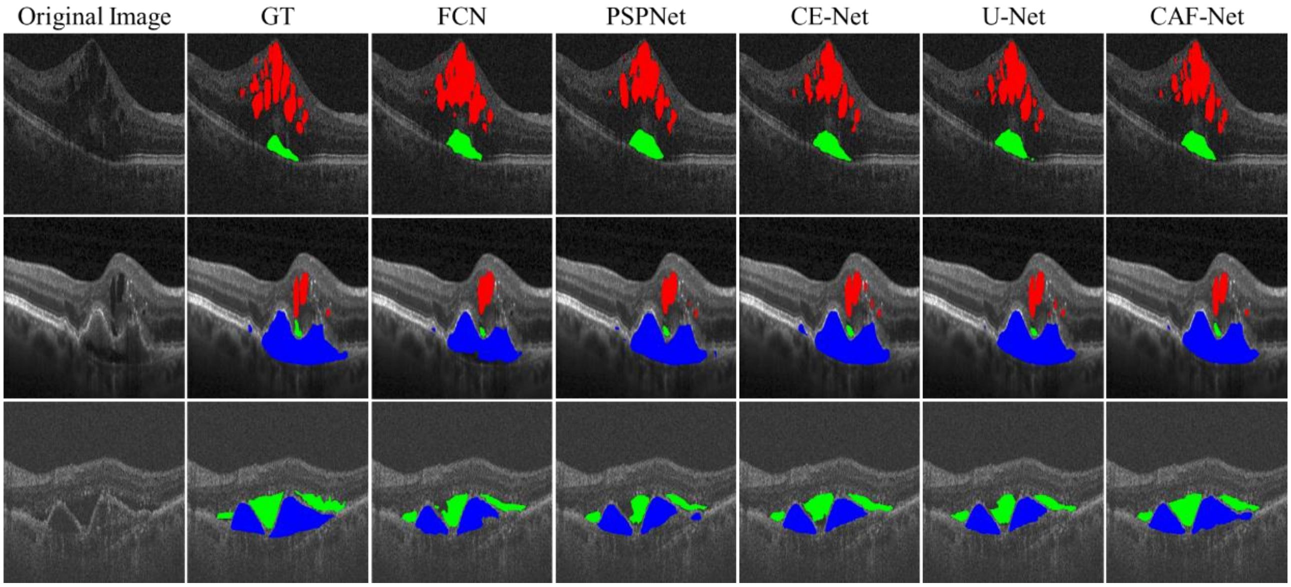


Fig. 4. Examples of retinal fluid segmentation results with different methods, where the red, green and blue regions denote the IRF, SRF and PED, respectively. From left to right: original image (cropped for illustration), ground truth (GT), FCN, PSPNet, CE-Net, U-Net and our CAF-Net. From up to down: Cirrus, Spectralis, and Topcon.

4. CONCLUSIONS

In this paper, we propose a novel context attention-and-fusion network for multiclass retinal fluid segmentation in OCT images. Two modules, including CSE module and CPG module, are designed to exploit rich global context information to dynamically guide the feature extraction and fusion. The segmentation results of the proposed method are in good agreement with the ground truth, which indicates that the proposed method can provide quantitative information for the analysis of retinal fluid in clinical practice.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700, in part by the National Nature Science Foundation of China under Grant 61622114 and 62001196, and in part by the National Basic Research Program of China under Grant 2014CB748600.

6. REFERENCE

[1] A.M. Joussen, T.W. Gardner, B. Kirchhof, and S.J. Ryan, "Retinal vascular disease," *Heidelberg: Springer*, 2007.

[2] D. Huang, E.A. Swanson, C.P. Lin *et al*, "Optical coherence tomography," *Science*, vol. 254(5035), pp. 1178-1181, 1991.

[3] M. Marmor, "Mechanisms of fluid accumulation in retinal edema," *Macular Edema, Springer, Dordrecht*, pp. 35-45, 2000.

[4] H. Lee, K.E. Kang, H. Chung, and H.C. Kim, "Automated segmentation of lesions including subretinal hyperreflective material in neovascular age-related macular degeneration," *American journal of ophthalmology*, vol. 191, pp. 64-75, 2018.

[5] D. Lu, M. Heisler, S. Lee *et al*, "Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network," *Medical image analysis*, vol. 54, pp.100-110, 2019.

[6] A. Rashno, D.D. Koozekanani, and K.K. Parhi, "OCT Fluid Segmentation using Graph Shortest Path and Convolutional Neural Network," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2018, pp. 3426-3429.

[7] A.G. Roy, S. Conjeti, S.P.K. Karri *et al*, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical Optics Express*, vol. 8, pp. 3627-3642, 2017.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241: Springer.

[9] D.L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41(3), pp. 613-627, 1995.

[10] L.C. Chen, G. Papandreou, I. Kokkinos *et al*, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40(4), pp. 834-848, 2017.

[11] H. Bogunovic, F. Venhuizen, and S. Klimscha, "RETOUCH - The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge," *IEEE Transactions on Medical Imaging*, vol. 38(8), pp. 1858-1874, 2019.

[12] S. Ye, J. Ye, "Dice Similarity Measure between Single Valued Neutrosophic Multisets and Its Application in Medical Diagnosis," *Neutrosophic Sets and Systems*, vol. 6, pp. 49-54, 2014.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881-2890.

[15] Z. Gu, J. Cheng and H. Fu, "CE-Net: Context encoder network for 2d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 38(10), pp. 2281-2292, 2019.