Medical Image Analysis 54 (2019) 100–110

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network

Donghuan Lu[a], Morgan Heisler[a], Sieun Lee[a], Gavin Weiguang Ding[a], Eduardo Navajas[b], Marinko V. Sarunic[a], Mirza Faisal Beg[a,*]

[a] Simon Fraser University, School of Engineering Science, Burnaby V5A 1S6, Canada
[b] University of British Columbia, Department of Ophthalmology and Visual Sciences, Vancouver V6T 1Z4, Canada

ARTICLE INFO

Article history:
Received 11 January 2018
Revised 15 February 2019
Accepted 15 February 2019
Available online 22 February 2019

Keywords:
Retinal fluid
Optical coherence tomography
Fully convolutional network
Multiclass segmentation and detection

ABSTRACT

As a non-invasive imaging modality, optical coherence tomography (OCT) can provide micrometer-resolution 3D images of retinal structures. These images can help reveal disease-related alterations below the surface of the retina, such as the presence of edema, or accumulation of fluid which can distort vision, and are an indication of disruptions in the vasculature of the retina. In this paper, a new framework is proposed for multiclass fluid segmentation and detection in the retinal OCT images. Based on the intensity of OCT images and retinal layer segmentations provided by a graph-cut algorithm, a fully convolutional neural network was trained to recognize and label the fluid pixels. Random forest classification was performed on the segmented fluid regions to detect and reject the falsely labeled fluid regions. The proposed framework won the first place in the MICCAI RETOUCH challenge in 2017 on both the segmentation performance (mean Dice: 0.7667) and the detection performance (mean AUC: 1.00) tasks.

1. Introduction

Optical coherence tomography (OCT) is an established imaging modality in ophthalmology providing micrometer-resolution 3D images of sub-surface biological tissue. It can be used to monitor disease progression, such as in diabetic macular edema (DME) or age-related macular degeneration (AMD) (Hee et al., 1995). High quality visualizations of the retinal structures provided by OCT images can improve the understanding of the onset and development of these retinal diseases, which are major causes of visual impairment (Joussen et al., 2010). Because the high-resolution imaging volumes provided by OCT are large and information-rich for ophthalmologists to screen quickly while assessing a large number of patients, automated tools to evaluate the health of the retina based on these non-invasive images are urgently needed. Macular edema is the accumulation of fluid in the center of the retina due to the disrupted blood-retina barrier. It is commonly caused by diseases that involve damage and leakage of retinal blood vessels, such as DME and AMD. Accurate segmentation of the regions of macular edema (henceforth denoted as "fluid") is of interest to ophthalmologists. Although 3D OCT images can visualize these regions of fluid, accurate, automated segmentation and quantification of the fluid regions can provide clinicians precise characterization of the disease state for tracking progression and response to treatment.

In recent years, several efforts have been made to automate segmentation of the fluid regions in retinal OCT images, such as using intensity thresholding and size-based criteria (Wilkins et al., 2012), k-NN classification (Pilch and Stieger, 2013), random forest classification (Lang and Carass, 2015), layer-dependent stratified sampling (Xu et al., 2015), kernel regression (Chiu et al., 2015), fuzzy level-set methods (Wang et al., 2016), AdaBoost classification (Sun et al., 2016), continuous max flow optimization (Wu et al., 2017), kernel graph cut (Rashno et al., 2017) and convolutional neural network (Lee et al., 2017c). There are also some studies which tried to jointly segment retinal layer and fluid (Novosel et al., 2017; Montuoro et al., 2017; Roy et al., 2017). However, these methods were all built for binary segmentation tasks, and limited to identifying a single type of retinal fluid, when there are often different types of fluid present in a patient's retina. In comparison to a binary fluid segmentation method, a multiclass segmentation approach has better potential to segment fluid regions with different properties.

There have been few studies on multiclass fluid segmentation (Schlegl et al., 2015, 2017) and these have been limited to

* Corresponding author.
E-mail address: mfbeg@sfu.ca (M.F. Beg).

https://doi.org/10.1016/j.media.2019.02.011
1361-8415/

intraretinal fluid (IRF) and subretinal fluid (SRF) segmentation.In OCT images the fluid regions appear as contiguous spaces that have different reflectivity than the surrounding tissue. IRF is a hyporeflective space predominantly located in the inner and outer nuclear layers that increases the overall retinal thickness. SRF is a hyporeflective space between the neurosensory retina and retinal pigment epithelium (RPE), resulting in retinal detachment. Pigment epithelial detachment (PED) is separation of RPE from the Bruch's membrane (BM). It can be subdivided into serous (hyporeflective, dome-shaped), fibrovascular (heterogenous reflectivity, irregularly shaped), drusenoid (medium to high homogenous reflectivity, smooth contour) and hemorrhagic types. Hemorrhagic PEDs appear as dome-shaped hyporeflective elevations similar to serous PEDs, but the structure below the PED, such as choroid, is less well visualized with lower intensity due to the presence of hemorrhage. Most PEDs are of a mixed configuration (Told et al., 2017; Lee et al., 2012; Arevalo et al., 2013). The IRF and SRF fluid regions usually show obvious lower intensity compared to nearby retinal tissue, while PED can have similar or even higher intensity, as shown in Fig. 2. To the best of our knowledge, there is no previous publication on achieving simultaneous three class retinal fluid (IRF, SRF and PED) segmentation or detection. In addition, previous work used private datasets, making it difficult to compare the performance across different methods.

Deep learning is a subfield of machine learning in which artificial neural networks with multiple layers are used for hierarchical feature learning. As increasingly large databases and computational power are available, deep learning has come to excel in classification tasks involving images and speech (Hinton et al., 2006). In ophthalmology, interest has been building to use deep learning to take advantage of the rich resource of fundus photography and OCT images that are routinely taken in clinics (Lee et al., 2017a), in particular for computer-aided diagnosis. Google has recently created a dataset of 128,000 fundus photographs, and several works have been published on deep learning-based automated detection of DME using fundus photography (Abràmoff et al., 2016; Gulshan et al., 2016; Gargeya and Leng, 2017; Quellec et al., 2017). Lee et al. linked electronic medical records to approximately 100,000 OCT scans from 9182 subjects to automatically classify normal and age-related macular degeneration cases (Lee et al., 2017b), and deep learning-based automated detection of glaucoma has been performed on wide-field OCT scans (Muhammad et al., 2017) and preperimetric glaucoma visual field (Asaoka et al., 2016). Deep learning has been also used for segmentation and classification of retinal vasculature in fundus photography (Leopolda et al., 2017; Welikala et al., 2017) and OCT (Prentašić et al., 2016) and retinal layers (Fang et al., 2017).

Fully convolutional neural networks (FCN) have demonstrated excellent performance for image segmentation tasks (Long et al., 2015). FCN and its advanced version, the U-net (Ronneberger et al., 2015), have proven to outperform other methods in the application of segmenting small datasets of medical images, such as neural structures (Ronneberger et al., 2015), the kidney (Çiçek et al., 2016) and liver tumors (Christ et al., 2017). Their applications on OCT retinal image have also shown superior performance, such as blood vessel (Prentašić et al., 2016; Fu et al., 2016) segmentation, layer segmentation (Fang et al., 2017). Although there has been a few efforts to apply it for retinal fluid segmentation (Lee et al., 2017c; Roy et al., 2017; Venhuizen et al., 2018), all the studies were limited to binary segmentation of IRF or SRF. Therefore, it is worth exploring developing an application of FCN towards the task of OCT multiclass fluid segmentation.

In this paper, we present a novel FCN-based framework for the segmentation and detection of retinal fluid in OCT images. The network used for segmentation followed the U-net structure detailed in Section 2.2, but took additional spatial information as an input. Segmentation results were further improved by random forest classifiers trained on the candidate fluid regions. The probability of random forest classifiers was also used to determine the presence of fluid in each volume. The proposed framework won the first prize in the Retinal OCT Fluid Challenge (RETOUCH) in the 2017 Medical Image Computing and Computer Assisted Interventions Conference (MICCAI) on both segmentation and detection. In addition, we evaluated the network trained on RETOUCH data for performance on an unseen, independent, publicly available dataset (Kermany et al., 2018a,b) with OCT images from 530 subjects and obtained results comparable to that found in testing on the RETOUCH dataset. This suggests the generalizability of the proposed framework to unseen data and hence its potential for usage in clinical application.

2. Methods

Our framework for the segmentation and classification of retinal fluid consisted of three steps as displayed in Fig. 1: 1) Layer segmentation: pre-process the image and segment the retina by delineating the internal limiting membrane (ILM) and the Bruch's membrane (BM); 2) Multiclass fluid segmentation: segment all pixels of each B-scan into background (non-fluid region), IRF, SRF and PED using fully convolutional neural network (FCN); 3) Classification: extract features from potential fluid regions and train a classifier to reject false fluid regions. Instead of first classifying B-scans by the presence of fluid and then segmenting the fluids in the selected scans, we performed fluid segmentation in all scans and labelled the scans according to the resulting segmentation.

2.1. Materials

Two datasets were used in this study, the RETOUCH dataset (Bogunovic et al., 2017) and another independent dataset provided by Kermany et al., which is referred to here as the Kermany dataset (Kermany et al., 2018a,b).

The RETOUCH dataset was released in two steps. First, 3 training datasets were released with manual segmentation to develop the method. Each dataset corresponded to a different OCT device. A total of 70 volumes, with 24 volumes acquired with each of the two OCT imaging devices: Cirrus (Zeiss) and Spectralis (Heidelberg), and 22 volumes acquired with T-1000 and T-2000 (Topcon) were provided. These images are collectively referred to by the manufacturer name in the rest of paper. For each volume from these three devices, the numbers of B-scans were 128, 49 and 128, respectively.

Three different types of fluid, namely the IRF, SRF and PED, were manually labeled and provided as ground truth. Although not all B-scans contained fluid, there was at least one type of fluid in each volume. Images with and without fluid are shown in Fig. 2 for each commercial device.

Then 3 testing datasets with 14 volumes acquired with each of the three devices were released for the validation of proposed method. The results of both the segmentation and detection tasks were evaluated by the RETOUCH organizers upon submission by the research teams, and hence, the ground truth of the RETOUCH test data remains unknown to public.

To test the generalizability of the networks trained on RETOUCH data to segment an unknown test dataset, the publicly available Kermany dataset, which was used only for further evaluation (and not used for training). We included all the 530 OCT images from DME, DRUSEN and NORMAL groups from their test set in this experiment. The images were acquired with Spectralis (Heidelberg) imaging system, and manual segmentation for fluid was performed by three trained raters (TY, NP and MB). The segmentations were

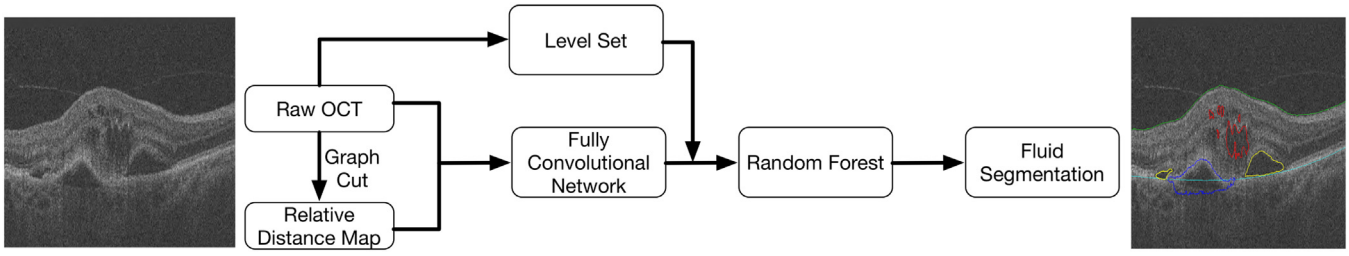
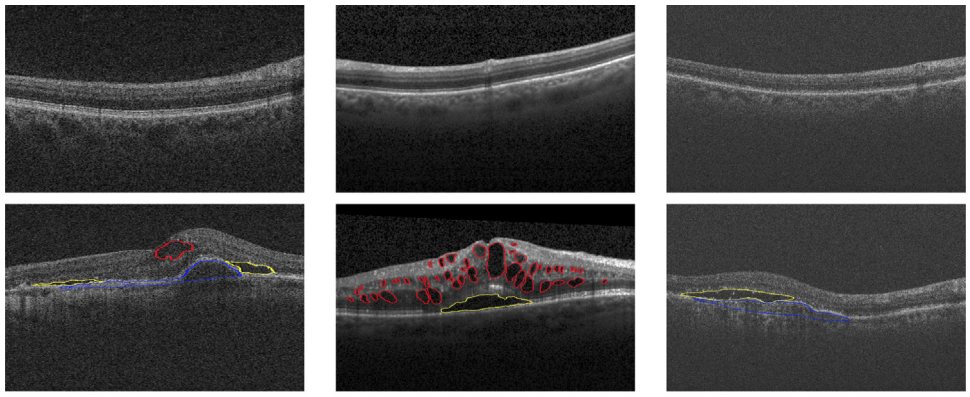


Fig. 1. Flowchart of multiclass fluid segmentation. A novel framework is proposed comprising a fully convolutional network (FCN) core and incorporating prior information via fluid features, a relative distance map, as well as post-processing by level sets and random forest based classifiers to deliver the final three-class (IRF, SRF, PEF) fluid segmentation.



(a) Cirrus (b) Spectralis (c) Topcon

Fig. 2. OCT images with manual segmentations. The images on the first row have no macular edema ("fluid"), while the red, blue and yellow segmentations on the second row represent the IRF, PED and SRF respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

reviewed by opthalmology clinician scientist VV and EN for accuracy. The raw images are available from https://data.mendeley.com/datasets/rscbjbr9sj/2 and our manual segmentations, in the same reference frame as the original images, is available for download at this link: https://vault.sfu.ca/index.php/s/0JwlgZ4IwRhmBkI?path=%2FFINAL_ATTEMPT%2FSegmentation. In the Kermany dataset, the number of B-scans for each volume ranged from 1 to 13 with most volumes only containing 1 or 2 B-scans. There are 168 volumes in the DME group, 142 volumes in the DRUSEN group and 220 volumes in the NORMAL group. The deep network trained on MICCAI RETOUCH Spectralis images was applied to a total of 750 B-scans taken from the 530 volumes of the Kermany data set for assessing the generalizability of performance for an unseen dataset.

2.2. Layer segmentation

Axial motion between B-scans was corrected using cross-correlation. Bounded variation (BV) 3D smoothing was applied to the motion-corrected intensity B-scans in order to reduce the effect of speckle while preserving and enhancing the boundaries between retinal layers. Next, the ILM and BM were automatically segmented using a 3D graph-cut based algorithm as smooth 3D surfaces (Li et al., 2006; Lee et al., 2013). For each volume, a graph was generated with nodes at each voxel and arcs based on the surface smoothness and between-surface distance, with the cost function of axial intensity gradient. In cases where the BM was not visible because of a large PED, a smooth curve was fitted to approximate BM.

2.3. Fully convolutional neural network

The network architecture is illustrated in Fig. 3. It shares a similar structure as the standard U-net (Ronneberger et al., 2015) except the input image contains a second channel in addition to the raw image. Relative distance maps were concatenated to the intensity of each B-scan as the second channel based on the assumption that the location of the fluid within the retina was an important property to determine the type of fluid. Since a network that classifies each pixel solely by the intensities of its neighbours cannot capture this information, additional inputs were necessary to better classify the different fluid types. For a pixel (x, y) in the relative distance map, its intensity in the relative distance map is defined as:

$$I(x,y) = \frac{y - Y_1(x)}{Y_1(x) - Y_2(x)}, \tag{1}$$

where $Y_1(x)$ and $Y_2(x)$ represent the y-coordinate of ILM and BM, respectively.

There were two paths in the network architecture: the contracting path (left side) and the expansive path (right side). Each path consisted of 4 blocks. In each block there were two convolutional layers with kernel size 3×3 and a rectified linear unit (ReLU) after each convolution operation. A 2×2 max pooling layer with stride 2 was then added to the contracting path and a 2×2 up-convolution layer was added to the expansive path. Shortcut connections were added to the layers with the same resolution from the contracting path to the expansive path to provide high-resolution features. After the expansive path, a 1×1 convolutional layer was used to map the features to a 4 channel probability map corresponding to background, IRF, SRF and PED. For each pixel, the channel with the highest probability was chosen as the segmentation result.

The network was trained end-to-end with a pixel-wise softmax function $p_j(\vec{z}) = e^{z_j} / \sum_{k=1}^{4} e^{z_k}$ combined with a cross entropy loss

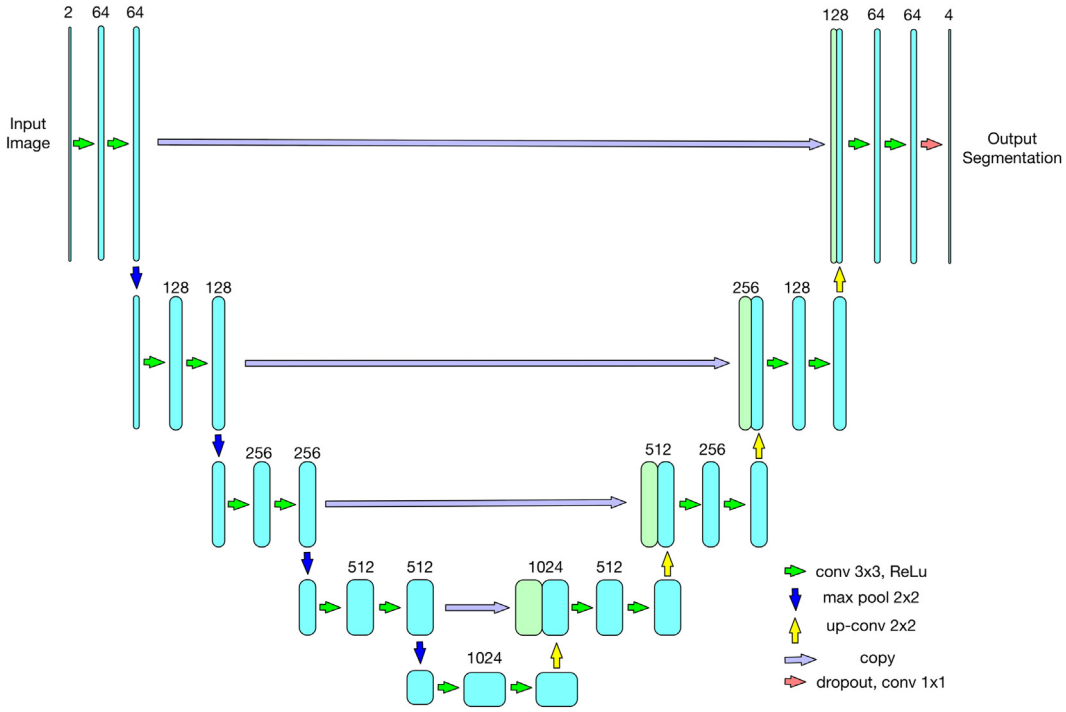


Fig. 3. Fully convolutional neural network. Each number above the cyan box represents the number of channels of the feature map.

function:

$$H(i) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4}[\mathbb{1}\{y^i = j\}log(h(x^i)_j], \quad (2)$$

where z_j denotes the probability in channel j, N is the number of input samples, j represents the class of samples, x^i, y^i are the feature vector and label of the i_{th} sample and h represents the network function.

To avoid overfitting, a dropout layer (Srivastava et al., 2014) was inserted before the 1×1 convolutional layer. During the training stage, only half of the units were randomly retained to feed features to the next layer, while in the testing stage, all the units were harnessed to generate the segmentation. By avoiding training all units on every sample, the chances of overfitting are reduced by preventing co-adaption on the training data. Because there were far more background pixels than fluid pixels, the training set was highly imbalanced. Training with Eq. (2) would result in a network tending to predict most pixels as background. Therefore, the network was trained only with true positive, false positive and false negative pixels, which means if a background pixel is correctly classified, it would not be used for the calculation of loss so that the absolute value of average cross entropy would be higher to guarantee the network being fully trained for the detection of fluid pixels. Because of the limitation of GPU memory, batch size was set as 3. The weight parameters for each layer were initialized with a uniform distribution while all bias started with 0 (Glorot and Bengio, 2010). Stochastic gradient descent (SGD) was used for optimization with a fixed learning rate of 10^{-4} and network training was stopped if the training accuracy ceased to increase during 10 epochs.

One important advantage of FCN is that without any fully connected layer, the network can be applied to images of arbitrary size. In this study, although the sizes of images from 3 devices were not the same, the network trained on one dataset was used to initialize the networks for the other two datasets to accelerate the training process. Due to the limited number of training samples, data augmentation is essential to prevent overfitting and increase the robustness and invariance properties of the network. Three processes - flip, rotation and zooming - were applied to the training samples. The rotation degree was from -25° to 25° and the maximum zooming ratio was 0.5.

Because FCN is a pixel-wise classification based segmentation approach, it may fail to segment the whole region of large fluid due to lack of gradient information and similarity between fluid and background, as shown in the left panel of Fig. 4. The retina was severely deformed due to a large PED and there was not enough similar volumes for training. The network was able to detect the PED pixels in the boundary area but failed to classify the central part of the fluid region. Therefore, a post-processing step was added by combining the FCN segmentation with a level-set method (Goldstein et al., 2010). Because the level-set method is able to detect the regions with low intensity, it can distinguish the whole fluid region from background as shown in the right panel of Fig. 4. For implementation, the level-set segmentation was applied to each B-scan to detect low intensity region which was considered as potential fluid region. Because it was a binary segmentation approach and may over segment, the detected fluid region was kept only when it had any overlap with PED or SRF segmented by FCN, and if so, the whole region was labelled as PED or SRF. This is shown in Fig. 4 where the PED in the left panel was expanded to the region in the right panel. This extra step was not necessary for IRF because the fluid regions were sufficiently small.

2.4. Random forest classification

Because the network tended to over-segment, random forest classifiers (Breiman, 2001) were trained to rule out false positive regions and determine the presence of fluid for each volume (Lu et al., 2017). In every B-scan of the given volume, potential fluid pixels with 8-connectivity were defined as a candidate region, and the regions with less than 3 pixels were removed. For each candidate region, a bounding rectangle with the edges 1.2 times longer than that of the tight bounding box was extracted, and from this a 16-dimension feature vector was further extracted.

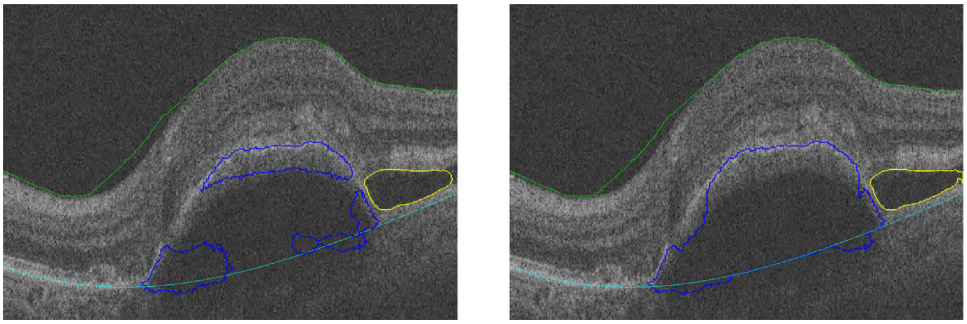


Fig. 4. Segmentation result before post-processing and after post-processing. The left panel shows the result of FCN, while the right panel displays the segmentation result combined with level-set method. The green and cyan lines are ILM and BM segmented with graph-cut based algorithm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The features consisted of the major and minor axis lengths, ratio of the major and minor axis lengths, perimeter, area, ratio of the perimeter and area, eccentricity, orientation, variance of the fluid height at each column, average intensity inside the fluid, average intensity outside of the fluid, intensity difference of inside and outside, intensity variance inside, intensity kurtosis inside, intensity skewness inside, and the relative distance of the center pixel as defined in Eq. (1). The label of the candidate region was defined by $r = \frac{area(S_1 \cap S_2)}{min(area(S_1), area(S_2))}$, where S_1 is the segmented region and S_2 is the manual segmentation. The candidate region was labeled as truth when $r > 0.7$, or vice versa.

Three random forest classifiers were trained separately for each fluid type. To avoid the effect of imbalanced training samples, the weights of the positive samples and negative samples were set as $N(negative)/(N(negative)+N(positive))$ and $N(positive)/(N(negative)+N(positive))$ respectively, where $N(\cdot)$ represents the number of samples. For each random forest, 100 decision trees were trained with 4 randomly selected features for each tree. The output probability of the random forest classifier was compared with a threshold to determine the label of each sample. A threshold of 0 to 1 with an interval of 0.01 was tested by a 5-fold cross validation on the training set and the one which resulted in the maximum $Fmeasure = \frac{2 \times precision \times recall}{precision + recall}$ was selected.

It is worth noting that the previous step of FCN yielded pixel-wise segmentation, and this was followed by the random forest classification on each candidate fluid region segmented by FCN. Therefore, if a region was classified as negative, all pixels in that region would be labeled as background; otherwise, they would remain the same as the FCN output.

2.5. Volume-wise fluid presence

To determine the presence of fluid in each volume, first the probability map of each B-scan was created based on the random forest classification result. Then the probability of each B-scan containing fluid was defined as the highest probability of all pixels within the probability map. Based on the assumption that fluid exists within multiple B-scans, the mean of the 10 highest probabilities over all B-scans in a volume was calculated and used as the probability of fluid presence for this volume. For example, to determine the presence of IRF in volume k, probability maps $P_1, P_2, \ldots, P_N$ was created for B-scans $I_1, I_2, \ldots, I_N$, respectively. In $P_i (i \in [1, \ldots, N])$, the intensities of the pixels segmented as background, SRF, or PED were set to 0, and the intensities of the remaining pixels were set to the probabilities from the random forest classification. Since the random forest classification was performed on each candidate region, pixels close to each other have the same intensity. Then the probability pb_n of B-scan I_n containing IRF was defined as the highest intensity of P_n. The average of 10 highest values across $pb_1, pb_2, \ldots, pb_n$ was defined as the probability of volume k containing IRF.

Because most volumes in the Kermany dataset contain less than 10 B-scans, the probability of fluid presence for each volume in this dataset was defined as the average probability of all B-scans in the volume.

3. Experiments and results

The deep neural network was built with an open source deep learning toolbox, Tensorflow (Abadi et al., 2015). Because of the difference between images acquired with different devices, as shown in Fig. 2, the experiments were performed on different device datasets independently. Three experiments were conducted to evaluate the performance of our proposed approach. For the RETOUCH dataset acquired with each device, a leave-one-out cross validation was first performed on the training data and the results were evaluated based on the ground truth provided by RETOUCH organizers. Then a classifier was trained on all training data and used to generate the segmentation and detection result for RETOUCH testing data. Those results were uploaded to the RETOUCH website and evaluated by the organizers. For the Kermany dataset, because the images were acquired with Spectralis imaging system, the classifier trained on the RETOUCH Spectralis dataset was applied for the fluid segmentation and detection. The segmentation performance was evaluated by the Dice index:

$$Dice = \frac{2area(S_1 \cap S_2)}{area(S_1) + area(S_2)} \tag{3}$$

The Dice index was computed per volume. For each type of fluid, the segmentation was measured separately, and the volumes that did not contain the particular type of fluid were ignored in the measurement.

For the detection of fluid in each volume, the receiver operating characteristic (ROC) curve was created by comparing the probabilities to the manual label and the area under the curve (AUC) was calculated. The probability of each volume was defined as the mean probability of 10 B-scans that mostly likely contained that type of fluid.

3.1. Leave-one-out cross validation

For the leave-one-out cross validation experiment, a single volume was used for testing at every iteration and the rest were used to train the neural network and the random forest classifiers. The experiments were repeated 24 times for both the Cirrus and Spectralis dataset, and 22 times for the Topcon dataset so that each volume would be used for testing exactly once.

The Dice index is displayed in Fig. 5. The ROC curves of fluid detection are displayed in Fig. 6. For each type of fluid in each

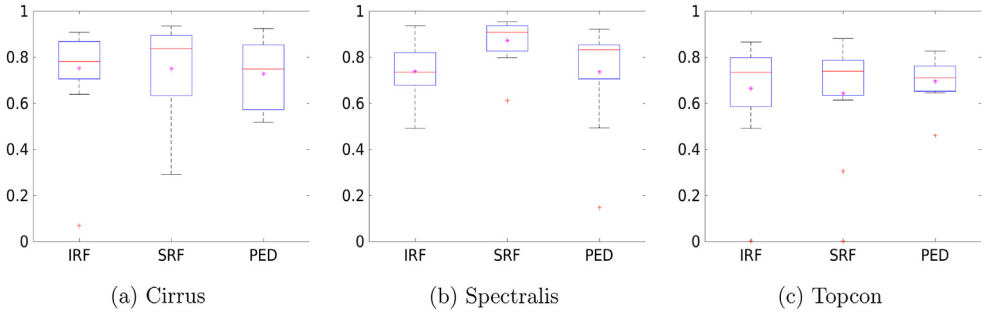


Fig. 5. Boxplot of the Dice index for fluid segmentation for each of the three commercial devices. The stars in each box are the mean Dice index, the line represents the median Dice index and the red '+' indicates outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

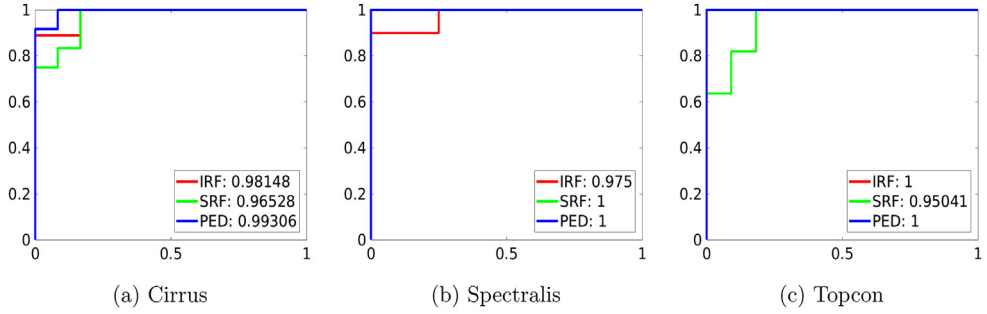


Fig. 6. ROC curves for identification of fluid-containing volumes between the automated and manual segmentations. The numbers in legend are the AUC of IRF, SRF and PED.

Table 1
Dice index and AUC after different steps. FCN-1channel represents the segmentation and detection result of the deep neural network with only intensity B-scan, while FCN-2channel represents the results with the relative distance map used as the second channel of input. The level-set result was combined with the FCN-2channel segmentation, and random forest classification was performed after the level-set process. To calculate AUC for the steps before random forest classification, the probability map created by FCN was directly used for each B-scan.

Measurement		Cirrus			Spectralis			Topcon		
		IRF	SRF	PED	IRF	SRF	PED	IRF	SRF	PED
Dice	FCN-1channel	0.587	0.563	0.5214	0.4837	0.7284	0.5872	0.5763	0.4834	0.4965
	FCN-2channel	0.7294	0.7183	0.6912	0.7418	0.8332	0.73	0.65	0.601	0.6634
	Level-set	0.7294	0.7221	0.7167	0.7418	0.8593	0.7324	0.65	0.6038	0.7069
	Random Forest	0.7525	0.7499	0.7286	0.7396	0.8731	0.7367	0.6651	0.643	0.6966
ACU	FCN-1channel	0.731	0.7282	0.7224	0.7283	0.86	0.8153	0.78	0.7357	0.7264
	FCN-2channel	0.912	0.8333	0.9167	0.8375	1	1	0.9	0.8967	0.8846
	Level-set	0.912	0.8333	0.9167	0.8375	1	1	0.9	0.8967	0.8846
	Random Forest	0.9815	0.9653	0.9931	0.975	1	1	1	0.9504	1

dataset, the AUC was always above 0.95, indicating an excellent detection accuracy.

The result after different steps is shown in Table 1. The FCN-1channel and FCN-2channel process are completely independent, while the other two processes are based on the output of their previous steps, i.e., level-set segmentation was combined with FCN-2channel result and random forest was performed after FCN-2channel and level-set segmentation. For the calculation of AUC, strategy described in Section 2.5 was used to determine the probability of fluid presence in each volume. But for the steps before random forest classification, the probability map created by FCN was directly used for each B-scan.

3.2. Experiment on the RETOUCH testing sets

In this experiment, a classifier with all training images was trained for each dataset and the results on the testing set provided were submitted to, and were evaluated by, the RETOUCH organizers. The results were compared with 7 other methods proposed by other participants of the RETOUCH challenge. All proposed segmentation methods were deep learning based approach with different network architectures. The description of these methods can be found on https://retouch.grand-challenge.org/workshop/. As in the previous experiment, segmentation performance was measured by Dice index and detection accuracy by AUC. The learning curves of FCN for different datasets are shown as in Fig. 7.

The Dice indices for the segmentation of IRF, SRF and PED of each test set is displayed as in Fig. 8, in which the result of each group is represented by a different color. The first 9 points indicate individual tasks corresponding to different datasets and different types of fluid, and the last point indicates the mean Dice index across all tasks.

For the presence of fluid in each volume, the AUC of the 8 methods are displayed in Table 2, in which each type of fluid was measured separately across all 3 datasets. For the detection of every type of fluid, our method has a perfect AUC score (1.00) for

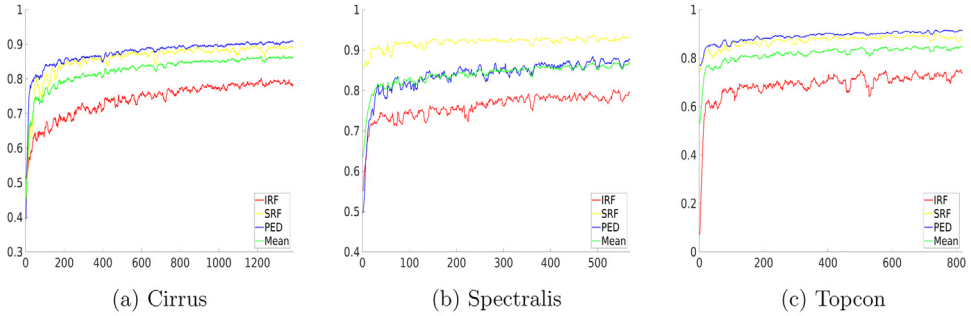


Fig. 7. Learning curve for different datasets with respect to iteration. The y axis represents the Dice index, while x axis represents the number of iteration. The green curve indicates the average Dice of 3 fluids. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

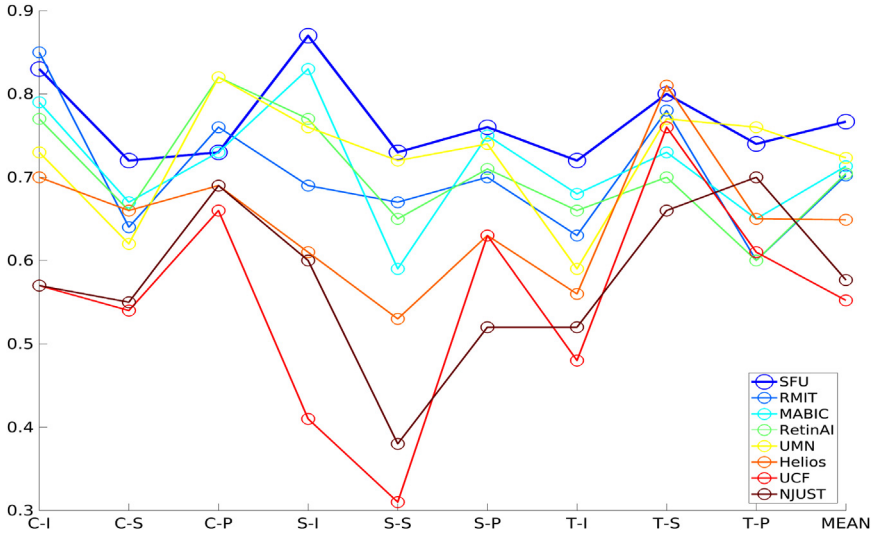


Fig. 8. Dice index of testing result. In X axis, 'C', 'S' and 'T' represents Cirrus, Spectralis and Topcon, while 'I', 'S' and 'P' represents IRF, SRF, PED. 'Mean' denotes the average index value for all tasks.

Table 2
AUC of different methods for the detection of IRF, SRF and PED. The result was evaluated across all 3 test sets. The last line represents the mean AUC of IRF, SRF and PED.

Fluid	SFU	RMIT	MABIC	RetinAI	UMN	Helios	UCF	NJUST
IRF	**1.00**	0.71	0.86	0.99	0.91	0.93	0.94	0.70
SRF	**1.00**	0.92	1.00	0.78	0.92	1.00	0.92	0.83
PED	**1.00**	1.00	0.97	0.82	0.95	0.97	1.00	0.98
Mean	**1.00**	0.88	0.94	0.86	0.93	0.97	0.95	0.84

the RETOUCH testing sets. This indicates that with the probability for each volume given by our method, there is a threshold which can separate the positive samples and negative samples with 100% accuracy as shown in Fig. 9. The probabilities of each volume were sorted from the lowest to highest. The first red line indicates the highest probability of negative samples, while the second one represents the lowest probability of positive samples.

3.3. Experiment on the unseen Kermany dataset

In this experiment, the classifier trained on RETOUCH Spectralis dataset was used for fluid segmentation and detection of the unseen Kermany dataset. As discussed in Section 2.5, because most volumes in this dataset only contain 1 or 2 B-scans, the probability of fluid presence in each volume was defined as the average probability of all the B-scans in the volume. As shown in Fig. 10, both segmentation and detection performance were comparable to that of the previous two experiments with the RETOUCH datasets.

4. Discussion

4.1. Leave-one-out cross validation

Deep neural networks have proven to be a powerful tool for pattern recognition. However, acquiring millions of labelled data samples to train the network can be difficult for medical image processing tasks. Therefore, incorporating a-priori, domain-specific knowledge with deep learning tools is important to achieve good performance specifically for medical imaging based tasks which are data-limited. Based on the cross validation results displayed in Table 1, adding the relative distance map as the second channel of input has improved the performance of FCN significantly (approximately 0.15 for average Dice and AUC). The level-set process improved the segmentation of SRF and PED (about 0.02 for average Dice), because it was only used to expand the FCN segmentation of the volumes which contained very large fluid regions. Random forest classification also did not improve the segmentation relatively as much, and even reduced the Dice index of IRF for Sepctralis and

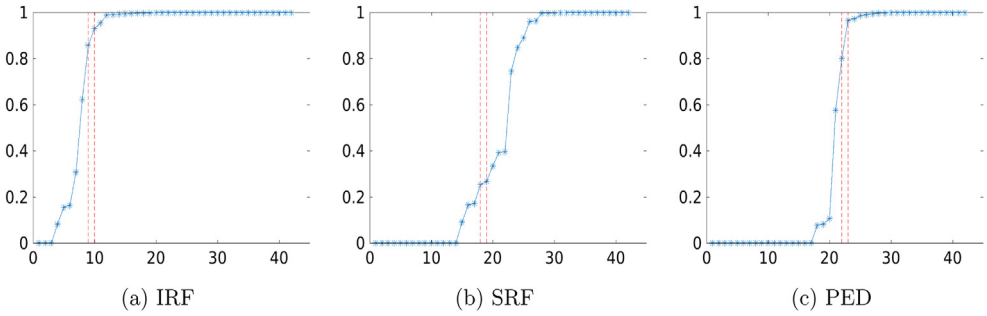


Fig. 9. Probabilities of each volume sorted from lowest to highest. The x-axis represents different volume, and y-axis is the probability of fluid presence. The red lines are used to divide negative and positive samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

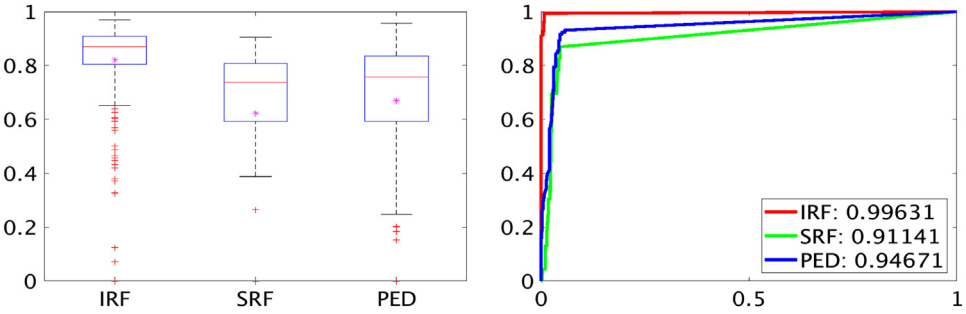


Fig. 10. Dice boxplot and ROC curves for fluid segmentation and detection of the Kermany dataset with classifier trained on RETOUCH Spectralis dataset. The left panel is a boxplot of the Dice index. The stars in each box are the mean Dice index, the line represents the median Dice index and the red '+' indicates outliers. The right panel is the ROC curves for the identification of fluid presence with respect to each volume. The numbers in legend are the AUC of IRF, SRF and PED. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

PED for Topcon slightly. This may be due to some true fluid regions being removed. However, by ruling out the false fluid regions, it increased the detection accuracy significantly (about 0.08 for average AUC with some results already at maximum value).

4.2. Experiment on the unseen RETOUCH testing sets

Each of the previous studies on fluid segmentation used their own datasets and the results were evaluated with different measures, making it difficult to compare the performance across methods. In this study, we used the datasets provided by the RETOUCH group as part of an international competition, competing with other groups using the same training data (including the same manual segmentation) and the same testing data whose ground truth remained unknown to all participants. This aspect of the RETOUCH challenge enables a direct comparison of different methods. Based on the evaluation of our network output provided by RETOUCH organizers, our proposed novel method achieved the best performance on both segmentation and detection tasks among 8 teams in the 2017 MICCAI RETOUCH challenge. As shown in Fig. 8 and Table 2, in 6 out of 9 individual segmentation tasks our method out-performed all other methods with the highest overall Dice index of 0.7667. The AUC for fluid detection was 1 for all three types of fluid. This provides supportive evidence for the robustness and stability of our method regarding different types of fluid and different devices used for acquiring the data.

It is worth mentioning that most participating teams used deep learning-based methods for the segmentation of all types of fluid with only two teams using non-deep learning based approaches to segment PED. Two strategies, which were both allowed by the RETOUCH challenge, have been used in the network training by different teams: training separate networks for datasets from different imaging systems, and training a single network with all datasets. Because machine learning based approaches, especially deep neural networks, are sensitive to training data, the performance of trained networks on an unknown dataset depends heavily on its similarity with the data used for training. A domain transfer approach to convert images into similar space is in development such that manually labelled samples may not be necessary to develop deep networks for classification of data acquired with other OCT devices.

4.3. Experiment on the unseen Kermany dataset

For the evaluation of generalizability and potential clinical application of the proposed framework, an additional experiment was performed on an unseen, independent Kermany dataset which was analyzed by using the network trained exclusively on the RETOUCH Spectralis dataset. As displayed in the left panel of Fig. 10, the Dice index for fluid segmentation is comparable to that of the RETOUCH dataset. The appearance of more outliers compared with the boxplots in Fig. 5 is likely due to increased volume numbers (530 vs. 24). There was a slightly decrease in the overall AUC compared with the previous experiments as shown in the right panel of Fig. 10. This may be because most volumes in the Kermany dataset only contained 1 or 2 B-scans, and a segmentation mistake in a single B-scan could lead to the misclassification of the entire volume. In the RETOUCH dataset, the fluid presence in each volume was determined by 10 B-scans with highest probabilities, which considerably reduced the effect of errors in a single B-scan. The Dice and AUC of SRF was slightly worse than IRF and PED due to "shadow regions" which were mistakenly segmented as SRF, as shown in Fig. 11. Because there was no such region in the MICCAI RETOUCH dataset segmentation protocol, the MICCAI RETOUCH Spectralis classifier was not trained to recognize it. The overall excellent performance on this unseen, independent dataset with 530 volumes suggests that our proposed framework has potential for

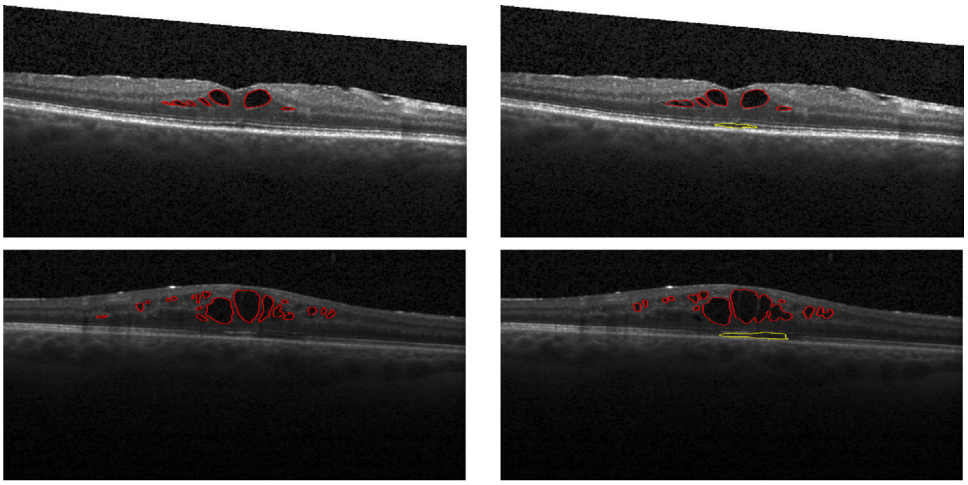


Fig. 11. Illustration of shadow regions mistakenly segmented as SRF. Images on the left side are the ground truth, while the images on right side show the segmentation result.

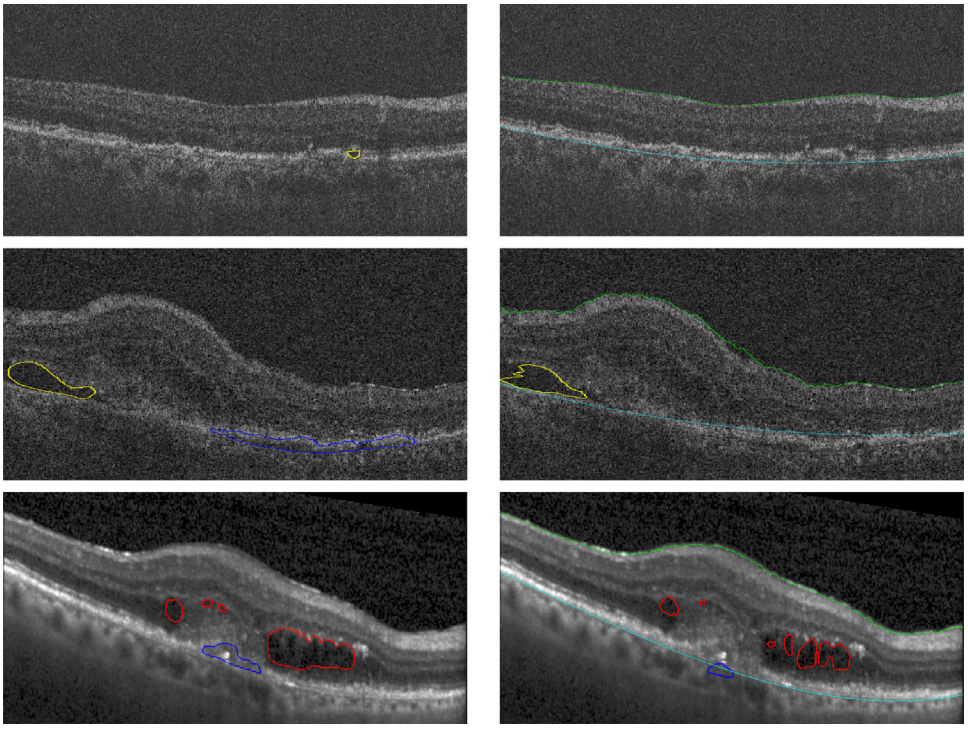


Fig. 12. Illustration of failed segmentation because of reasons 1 to 3. Images on the left side are the ground truth, while the images on right side show the segmentation result. IRF, SRF, PED, ILM and BM are represented by red, yellow, blue, green and cyan lines, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

being incorporated in a clinical pipeline for the automated segmentation and detection of macular edema in OCT images.

4.4. Analysis of errors

There were some volumes for which our method showed lower segmentation accuracy despite its overall excellent performance. This is likely due to following reasons. 1) Very small area of SRF: when a volume only contained SRF of a much smaller size than most SRF regions in the training data, such as the first row in Fig. 12, the proposed method failed to detect the tiny SRF region because of its similarity with shadow regions in normal tissue. 2) Incorrect layer segmentation: the layer segmentation was not always perfect due to severe deformation in some retinas, such as in the second row in Fig. 12. The network failed to detect the PED region because the BM boundary was incorrectly segmented. Retinal layer segmentation using deep neural networks may be explored to improve the accuracy of the layer segmentation. 3) Varied subtypes of PED: some PED regions are heterogeneous in reflectivity, containing hyperreflective foci as shown in the third row in Fig. 12. With manual labeling of PED further by subtypes, the proposed algorithm could have a better performance on their segmentation. 4) Other abnormalities and artifacts: in Fig. 13, the top row shows a inner limiting membrane detachment segmented as IRF. The middle row shows heavy shadowing caused by large IRF regions segmented as SRF. The third row shows structural disturbance in the RPE band segmented as PED. 5) Limitations of the ground truth data: Some of the errors may be attributed to the inconsistencies in RETOUCH data, which was manually segmented and labeled by multiple experts. The area segmented as fluid in the first row of

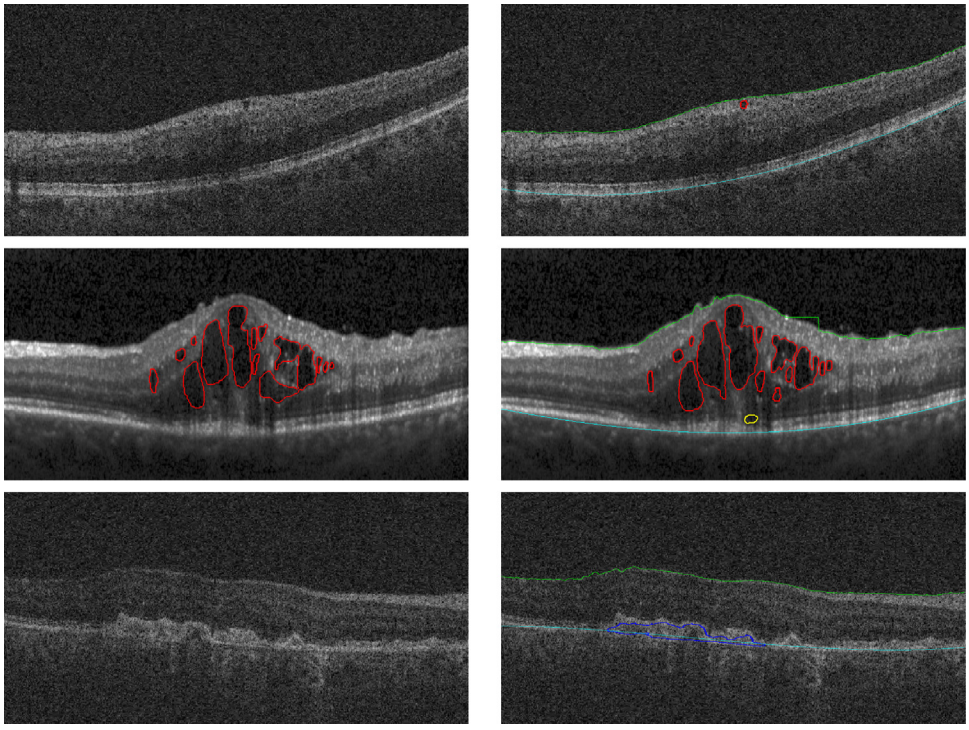


Fig. 13. Illustration of failed segmentation because of reason 4. From top to bottom, there are abnormal regions labelled as IRF, SRF and PED, respectively.

Fig. 12 may not be considered as SRF by some experts, and the hyporeflective PED in the third row of same figure may be explained by sub-RPE exudate or migrating RPE.

5. Conclusion

In this paper, we have described a novel framework to automatically segment multiple types of fluid regions in 3D retinal OCT images and identify volumes containing specific fluid types. Using ILM and BM boundaries segmented by a graph-cut technique, a relative distance map was generated and concatenated with each B-scan as an input to train a fully convolutional deep neural network. Pixels classified as potential fluid by the network were then grouped into different regions based on their 8-connectivity. For each type of fluid, a random forest classifier was trained on those candidate regions to rule out false positive samples and determine the fluid presence in each volume. The proposed method showed good performance in the leave-one-out classification experiments with 3 different training sets, and had the best performance on the testing set especially for the detection of fluid presence in each volume. The excellent performance on the unseen, independent Kermany dataset with 530 volumes suggests the potential for generalizability of the proposed framework for unseen data in clinical applications. Due to the limited number of training samples in the given datasets, the segmentation results are comparatively not as high as the detection results. As more data is accumulated in the future, it is likely that further improvements in accuracy for the task of fluid segmentation in retina OCT images may be achieved.

Conflicts of interest

None.

Acknowledgments

The authors would like to acknowledge Timothy Yu, Natalia Page (under graduate students of Simon Fraser University), Mahadev Bhalla (MD student of The University of British Columbia) and Dr. Vinicus Vanzan (a senior ophthalmology resident of The University of British Columbia) for the manual segmentation of the Kermany dataset. This research received funding support from the Natural Sciences and Engineering Research Council of Canada (NSERC), Canadian Institutes for Health Research (CIHR), the Brain Canada Foundation, Alzheimer Society of Canada, the Pacific Alzheimer Research Foundation, Genome British Columbia, and the Michael Smith Foundation for Health Research (MSFHR).

References

Abadi, M., Agarwal, A., Barham, P., et al., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Abràmoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M., 2016. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investig. Opthalmol. Vis. Sci. 57 (13), 5200. doi:10.1167/iovs.16-19964.

Arevalo, J.F., Lasave, A.F., Arias, J.D., Serrano, M.A., Arevalo, F.A., 2013. Clinical applications of optical coherence tomography in the posterior pole: the 2011 josã©-manuel espino lecture — part ii. Clin. Ophthalmol. 7, 2181–2206.

Asaoka, R., Murata, H., Iwase, A., Araie, M., 2016. Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. Ophthalmology 123 (9), 1974–1980.

Bogunovic, H., Venhuizen, F., Klimscha, S., Sánchez, C., Yanwu, X., Sebastian, M. W., Bianca, S. G., Klaver, C., Ursula, S.-E., 2017. MICCAI Retinal OCT Fluid Challenge.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Chiu, S.J., Allingham, M.J., Mettu, P.S., Cousins, S.W., Izatt, J.A., Farsiu, S., 2015. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomed. Opt. Express 6 (4), 1172–1194.

Christ, P. F., Ettlinger, F., Grün, F., Elshaera, M. E. A., Lipkova, J., Schlecht, S., Ahmaddy, F., Tatavarty, S., Bickel, M., Bilic, P., et al., 2017. Automatic Liver and Tumor Segmentation of CT and MRI Volumes using Cascaded Fully Convolutional Neural Networks. arXiv:1702.05970.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 424–432.

Fang, L., Cunefare, D., Wang, C., Guymer, R.H., Li, S., Farsiu, S., 2017. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. Biomed. Opt. Express 8 (5), 2732–2744.

Fu, H., Xu, Y., Wong, D.W.K., Liu, J., 2016. Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. In: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on. IEEE, pp. 698–701.

Gargeya, R., Leng, T., 2017. Automated identification of diabetic retinopathy using deep learning. Ophthalmology 124 (7), 962–969. doi:10.1016/j.ophtha.2017.02.008.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256.

Goldstein, T., Bresson, X., Osher, S., 2010. Geometric applications of the split bregman method: segmentation and surface reconstruction. J. Sci. Comput. 45 (1–3), 272–293.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316 (22), 2402. doi:10.1001/jama.2016.17216.

Hee, M., Izatt, J., Swanson, E., Huang, D., Schuman, J., Lin, C., Puliafito, C., Fujimoto, J., 1995. Optical coherence tomography of the human retina. Arch. Ophthalmol. 113 (3), 325–332.

Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.

Joussen, A., Gardner, T., Kirchhof, B., Ryan, S., 2010. Retinal Vascular Disease. Springer Berlin Heidelberg.

Kermany, D., Zhang, K., Goldbaum, M., 2018a. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018b. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.

Lang, A., Carass, A.e.a., 2015. Automatic segmentation of microcystic macular edema in OCT. Biomed. Opt. Express 6 (1), 155–169.

Lee, A., Taylor, P., Kalpathy-Cramer, J., Tufail, A., 2017a. Machine learning has arrived!. Ophthalmology 124 (12), 1726–1728. doi:10.1016/J.OPHTHA.2017.08.046.

Lee, C.S., Baughman, D.M., Lee, A.Y., 2017b. Deep learning is effective for classifying normal versus age-Related macular degeneration optical coherence tomography images. Ophthalmol. Retina 1 (4), 322–327.

Lee, C.S., Tyring, A.J., Deruyter, N.P., Wu, Y., Rokem, A., Lee, A.Y., 2017c. Deep-Learning based, automated segmentation of macular edema in optical coherence tomography. bioRxiv 135640.

Lee, S., Fallah, N., Forooghian, F., Ko, A., Pakzad-Vaezi, K., Merkur, A.B., Kirker, A.W., Albiani, D.A., Young, M., Sarunic, M.V., et al., 2013. Comparative analysis of repeatability of manual and automated choroidal thickness measurements in nonneovascular age-related macular degeneration. Investig. Ophthalmol. Vis. Sci. 54 (4), 2864–2871.

Lee, S.Y., Stetson, P.F., Ruiz-Garcia, H., Heussen, F.M., Sadda, S.R., 2012. Automated characterization of pigment epithelial detachment by optical coherence tomography. Investig. Ophthalmol. Vis. Sci. 53 (1), 164–170.

Leopolda, H., Orchardb, J., Zeleka, J., Lakshminarayanana, V., 2017. Use of gabor filters and deep networks in the segmentation of retinal vessel morphology. In: Proc. of SPIE Vol, 10068, pp. 100680R–1.

Li, K., Wu, X., Chen, D.Z., Sonka, M., 2006. Optimal surface segmentation in volumetric images-a graph-theoretic approach. IEEE Trans. Pattern Anal. Mach. Intell. 28 (1), 119–134.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lu, D., Ding, W., Merkur, A., Sarunic, M.V., Beg, M.F., 2017. Multiple instance learning for age-related macular degeneration diagnosis in optical coherence tomography images. In: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on. IEEE, pp. 139–142.

Montuoro, A., Waldstein, S.M., Gerendas, B.S., Schmidt-Erfurth, U., Bogunović, H., 2017. Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context. Biomed. Opt. Express 8 (3), 1874–1888.

Muhammad, H., Fuchs, T.J., De, N.C., De, C.M., Blumberg, D.M., Liebmann, J.M., Ritch, R., Hood, D.C., 2017. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. J. Glaucoma.

Novosel, J., Vermeer, K.A., de Jong, J.H., Wang, Z., van Vliet, L.J., 2017. Joint segmentation of retinal layers and focal lesions in 3-D OCT data of topologically disrupted retinas. IEEE Trans. Med. Imaging 36 (6), 1276–1286.

Pilch, M., Stieger, K.e.a., 2013. Automated segmentation of pathological cavities in optical coherence tomography scans. Investig. Ophthalmol. Vis. Sci. 54 (6), 4385–4393.

Prentašić, P., Heisler, M., Mammo, Z., Lee, S., Merkur, A., Navajas, E., Beg, M.F., Šarunić, M., Lončarić, S., 2016. Segmentation of the foveal microvasculature using deep learning networks. J. Biomed. Opt. 21 (7). 075008–075008.

Quellec, G., Charrière, K., Boudi, Y., Cochener, B., Lamard, M., 2017. Deep image mining for diabetic retinopathy screening. Med. Image Anal. 39, 178–193.

Rashno, A., Nazari, B., Koozekanani, D.D., Drayna, P.M., Sadri, S., Rabbani, H., Parhi, K.K., 2017. Fully-automated segmentation of fluid regions in exudative age-related macular degeneration subjects: kernel graph cut in neutrosophic domain. PLoS ONE 12 (10), e0186949.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roy, A.G., Conjeti, S., Karri, S.P.K., Sheet, D., Katouzian, A., Wachinger, C., Navab, N., 2017. Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. Biomed. Opt. Express 8 (8), 3627–3642.

Schlegl, T., Waldstein, S.M., Bogunovic, H., Endstraßer, F., Sadeghipour, A., Philip, A.-M., Podkowinski, D., Gerendas, B.S., Langs, G., Schmidt-Erfurth, U., 2017. Fully automated detection and quantification of macular fluid in oct using deep learning. Ophthalmology 125 (4), 549-558.

Schlegl, T., Waldstein, S.M., Vogl, W.-D., Schmidt-Erfurth, U., Langs, G., 2015. Predicting semantic descriptions from medical images with convolutional neural networks. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 437–448.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Sun, Z., Chen, H., Shi, F., Wang, L., Zhu, W., Xiang, D., Yan, C., Li, L., Chen, X., 2016. An automated framework for 3D serous pigment epithelium detachment segmentation in SD-OCT images. Sci. Rep. 6, 21739.

Told, R., Waldstein, S.M., Schmidt-Erfurth, U., 2017. Neovascular Age-Related Macular Degeneration. Springer, New Delhi, pp. 183–203. doi:10.1007/978-81-322-3610-8_15.

Venhuizen, F.G., van Ginneken, B., Liefers, B., van Asten, F., Schreur, V., Fauser, S., Hoyng, C., Theelen, T., Sánchez, C.I., 2018. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. Biomed. Opt. Express 9 (4), 1545–1569.

Wang, J., Zhang, M., Pechauer, A.D., Liu, L., Hwang, T.S., Wilson, D.J., Li, D., Jia, Y., 2016. Automated volumetric segmentation of retinal fluid on optical coherence tomography. Biomed. Opt. Express 7 (4), 1577–1589.

Welikala, R., Foster, P., Whincup, P., Rudnicka, A., Owen, C., Strachan, D., Barman, S., et al., 2017. Automated arteriole and venule classification using deep learning for retinal images from the UK biobank cohort. Comput. Biol. Med. 90, 23–32.

Wilkins, G.R., Houghton, O.M., Oldenburg, A.L., 2012. Automated segmentation of intraretinal cystoid fluid in optical coherence tomography. IEEE Trans. Biomed. Eng. 59 (4), 1109–1114.

Wu, M., Fan, W., Chen, Q., Du, Z., Li, X., Yuan, S., Park, H., 2017. Three-dimensional continuous max flow optimization-based serous retinal detachment segmentation in SD-OCT for central serous chorioretinopathy. Biomed. Opt. Express 8 (9), 4257–4274.

Xu, X., Lee, K., Zhang, L., Sonka, M., Abràmoff, M.D., 2015. Stratified sampling voxel classification for segmentation of intraretinal and subretinal fluid in longitudinal clinical oct data. IEEE Trans. Med. Imaging 34 (7), 1616–1623.