Benchmarking Automated Detection Of The Retinal External Limiting Membrane In A 3D Spectral Domain Optical Coherence Tomography Image Dataset of Full Thickness Macular Holes

Vivek Kumar Singh\textsuperscript{a}, Burak Kucukgoz\textsuperscript{a}, Declan C Murphy\textsuperscript{b}, Xiaofan Xiong\textsuperscript{a}, David H Steel\textsuperscript{b} and Boguslaw Obara\textsuperscript{a,b,*}

\textsuperscript{a}School of Computing, Newcastle University, Newcastle upon Tyne, UK, \textsuperscript{b}Bioscience Institute, Newcastle University, Newcastle upon Tyne, UK.

ARTICLE INFO

Keywords: Image Analysis Machine Learning Annotation External Limiting Membrane Optical Coherence Tomography

ABSTRACT

In this article, we present a new benchmark for the segmentation of the retinal external limiting membrane (ELM) using an image dataset of spectral domain optical coherence tomography (OCT) scans in a patient population with idiopathic full-thickness macular holes. Specifically, the dataset used contains OCT images from one eye of 107 patients with an idiopathic full-thickness macular hole. In total, the dataset contains 5243 individual 2-dimensional (2-D) OCT image slices, with each patient contributing 49 individual spectral-domain OCT tagged image slices. We display precise image-wise binary annotations to segment the ELM line. The OCT images present high variations in image contrast, motion, brightness, and speckle noise which can affect the robustness of applied algorithms, so we performed an extensive OCT imaging and annotation data quality analysis. Imaging data quality control included noise, blurriness and contrast scores, motion estimation, darkness and average pixel scores, and anomaly detection. Annotation quality was measured using gradient mapping of ELM line annotation confidence, and idiopathic full-thickness macular hole detection. Finally, we compared qualitative and quantitative results with seven state-of-the-art machine learning-based segmentation methods to identify the ELM line with an automated system.

1. Introduction

Idiopathic full-thickness macular holes (MHs) are a common vitreoretinal interface abnormality that affects approximately 1 in 200 people aged 60 years or older and can result in significant visual deterioration [1, 45]. Surgery is the main treatment for MHs, with pars plana vitrectomy and intraocular gas tamponade being the current gold standard intervention, which achieves successful post-operative MH closure in more than 90% of cases [59, 43]. After successful hole closure, visual acuity (VA) usually improves by a mean of three-lines of vision, although final VA achieved can be variable with, for instance, only 35-40% achieving a United Kingdom (UK) driving level of vision after otherwise successful surgery [35, 59].

Ophthalmologists utilize a broad range of imaging-based techniques to aid diagnosis and the monitoring of eye diseases. Optical coherence tomography (OCT) [32] is a non-invasive, high-resolution imaging technique that uses infrared light to image the retinal structure in three-dimensions (3D) [30]. OCT scans are used to qualitatively and quantitatively characterise and monitor many retinal diseases including diabetic retinopathy (DR), age-related macular degeneration (AMD), diabetic macular edema (DMO), and MHs [56, 48, 16]. The accurate identification of eye diseases and their features plays a key role in correctly diagnosing and measuring disease progression.

\textsuperscript{*}Corresponding author. Email: boguslaw.obara@newcastle.ac.uk (B. Obara)

ORCID(s):

Figure 1: Spectral-domain optical coherence tomography central slice through the macula demonstrating an idiopathic full-thickness macular hole. The external limiting membrane is clearly labelled.

The ocular fundus, which comprises the retina, macula, optic disc, fovea and blood vessels, can be visualised in detail using OCT which allows more than ten retinal laminations to be identified. Fig. 1 shows the appearance of an MH when imaged using spectral-domain OCT, and specifically identifies the ELM. Accurate visualization of retinal anatomy allows ophthalmologists to determine which retinal layers are affected in specific disease processes. The retina can be anatomically divided into inner and outer zones. The inner retina defines all structures from the internal limiting membrane (ILM) to the ELM, and the outer retina includes all structures from the photoreceptors to the choroid. The ELM is formed by junctional complexes between photoreceptor and Müller cells, located between the photoreceptor nuclei and their inner segments. The integrity of the ELM is important in many disease states. In this study, we specifically investigate the ELM in MHs.
Idiopathic full-thickness MHs form due to age-related changes at the vitreoretinal interface, with antero-posterior vitreous traction on the inner retinal surface being transmitted by Müller cells to the outer retina, which leads to outer retinal traction and dehiscence. Subsequently, a full-thickness MH forms with the associated movement of the outer retinal layers towards the inner retinal surface [62]. The integrity of the ELM after surgery is associated with post-operative vision. The height of the ELM on the sides of the MH also appears to provide prognostic information, and it is thought that post-operative vision is partly related to the extent of tractional changes in the outer retina when the MH forms, as well as its chronicity [24, 41]. Detecting the ELM in MHs is not only clinically useful, but also a particularly challenging image segmentation task with high applicability to other diseases and biomedical imaging modalities.

There is significant potential for computer-aided diagnosis (CAD) systems to aid OCT image analysis of MHs and assist ophthalmologists in the detection and characterisation of the ELM. Identification and quantification of the ELM can be processed manually, but the many challenges make this a slow and error-prone process. Accurately identifying the ELM is also difficult because the detection or inhibition of partially coherent optical beams causes speckle noise in OCT images, causing difficulty in identifying the ELM line and differentiating it from neighboring retinal layers.

OCT scans require various solutions such as reducing the speckle noise, correctly differentiating the multiple retinal layers and changes of the tissue under imaging, as well as the progress of the disease in the tissue that affects the quality of acquired images with its analysis [6]. Curvelet transforms [36], Contourlet transform [67], and Ripplet transforms [27] were employed for image de-noising to reduce the speckle noise from the OCT dataset. Stefanos et al. [3] suggested a deep neural network-based CNN improving noisy low-quality OCT scans to better contrast high-quality images with decreased speckle noises. Jing et al. [65] developed a CNN-based architecture to classify the OCT images quality based on signal completeness, location, and effectiveness. Various deep neural networks were included such as VGG, ResNet-18, ResNet-50, and Inception-V3 in the classification task through transfer learning.

There are two main types of automated image analysis approaches that have been utilized for CAD system development, which includes classical image analysis techniques and machine learning-based image informatics techniques [18, 20, 64, 42, 54, 51]. Classical image informatics methodologies have several limitations such as threshold methods which struggle with discontinuities and intensity variations across different retinal layers, and they also cannot combine previously obtained information such as retinal layer thicknesses from previous OCT scans. These limitations fail to provide a fully automated solution. [34, 38, 17].

The majority of image analysis approaches in the past have used privately owned datasets, and there is currently no large publicly available benchmark dataset to use for ELM line segmentation from OCT images. Therefore, the most appropriate image analysis methodology for automated ELM line segmentation is unclear.

The main objective of this study is to address the challenge of acquiring a publicly available retinal OCT dataset by providing a benchmark ELM line OCT dataset and baseline segmentation results with current state-of-the-art deep learning methods. We present a new benchmark for automated detection of the retinal ELM using a 3D spectral-domain OCT (SD-OCT) image dataset of idiopathic full-thickness MHs. The experimental results on this new ELM line dataset serve as the first building block to foster rapid and objective progress in tackling the research problem at hand as a research community. The main contributions of this work include:

1. We introduce a new benchmark dataset for retinal ELM line segmentation from OCT images as shown in Fig. 2. The proposed ELM line dataset employed 3D SD-OCT (i.e., 5243 2D planes) images of 107 patients.
2. To provide a high-quality benchmark dataset, robust and rigorous quality checks performed on OCT images, including noise score, blurriness score, contrast score, motion score, brightness-darkness score, and average pixel width scores. Moreover, annotation quality analysis includes gradient-based ELM line detection and MH detection.
3. We perform extensive experiments, including a broad range of ablation studies using seven state-of-the-art biomedical and semantic image segmentation methods. Quantitative and qualitative outcomes are compared with six evaluation metrics.

The remainder of the article is structured as followed: Section 2 describes current published literature and their retinal layer segmentation methods. Section 3 describes the benchmark OCT imaging dataset used in this study as well as image acquisition, annotation, and data quality analysis methodologies. Section 4 explains the deep learning-based segmentation methods used in the study. Section 5 details study results and limitations. Section 6 concludes the study findings and provides recommendations for future research.

2. Related Works

A number of retinal layer segmentation approaches have been described in published literature. Below, we present and discuss the most common segmentation methods, ranging from classical to machine learning-based image informatics approaches.

2.1. Classical Image Informatics Approaches

Automatic segmentation of retinal layers based on classical approaches can be categorised as follows: (1) threshold-based methods, (2) level set-based techniques, and (3) dynamic programming/graph cut-based methods. Mostly, these techniques have extracted hand-crafted features that utilize the pixel values, texture, colour, and shape for the segmentation process. To accurately determine the border of the
retinal layers, Ishikawa et al. [34] utilized an adaptive thresholding technique. Previous work has also applied intensity-based Markov boundary models [38] and a texture and shape analysis [37] approach to segment different retinal layers from OCT data. Furthermore, Novosel et al. [49] adapted a level-set technique based on Bayesian inference and anatomical information of the retina to delineate surfaces between layers. Carass et al. [12] employed a deformable model method to segment eight retinal layers across the complete macular cube. This suggested mechanism maintains object relationships and topology whilst limiting overlaps.

A fully automated graph theory and dynamic programming-based method was applied to segment three retinal layers from SD-OCT images of an eye with drusen and geographic atrophy (GA) [18]. Another group [20] used a small dataset of OCT images for statistical shape modelling. Specifically, the Mumford-Shah functional method was employed, which allows the development of a parametric illustration of open contours. Later, they formed small bandfs around the open contours, which enables segmentation by incorporating local information. Furthermore, the kernel regression (KR)-based classification approach [17] was adopted to accurately measure abnormal fluid within and under the retina, and the position of the retinal layers. This classification strategy identifies a pattern to delineate retinal layers based on graph theory and dynamic programming techniques. However, this method also has limitations due to its parameter inference on images with noise. To overcome this challenge, Tian et al. [64] recommended the shortest path-based graph search method, which identifies retinal layers by seeking the shortest path among two end nodes using Dijkstra’s algorithm. Moreover, Bai et al. [8] proposed an adaptive-curve detection technique that explores the retinal area with boundary growth. The recommended approach used the simple linear iterative clustering (SLIC) superpixels, and the adjusted active contour, to sequentially delineate the remaining boundaries. The method was tested using 3D OCT images captured from two different OCT systems.

Classical image informatics methods have several limitations, for example (1) the thresholding methods can cause discontinuities and variances of intensity within the same retinal layer, (2) they lack the ability to consider prior information about the retinal layer thicknesses obtained from previous imaging, and (3) their typically slow performance (especially graph-based techniques) means they are inappropriate for real-time clinical practices.

### 2.2. Machine Learning-Based Image Informatics Approaches

Convolutional neural networks (CNNs) are a deep learning-based technique used in many image-based segmentation applications [39]. Most CNNs are applied in areas with large quantities of imaging data with annotations available to be trained on. In the case of medical imaging, there are numerous unique challenges, such as the availability of data, dimensionality (3D or 4D) of images, annotation complexity, and quality. In addition, personal patient data and information must be anonymized, which is a complex and regulated process [21]. As deep learning methods improve, there are new available models for medical image segmentation tasks such as FCN [44], SegNet [5], and U-Net [52].

To overcome the aforementioned challenges, an end-to-end multi-scale nested U-Net [42] based method has been employed to segment seven retinal layers and three types of retinal fluid respectively. The important feature of this model is the utilization of multi-scale input, multi-scale side output, re-designed skip connections from U-Net++ [68], and...
and dual attention technique. Specifically, U-Net++ allows for encoding and decoding layers connected by a series of nested dense skip connections. This allows the semantic gap between encoding and decoding layers to be decreased. Similarly, Kugelman et al. [40] applied the recurrent neural network trained as a patch-based retinal boundary classifier with a graph search (RNN-GS) to segment seven retinal layers in OCT images from healthy children and three retinal layer edges in OCT images from patients with age-related macular degeneration (AMD). However, Pekala et al. [51] suggested a novel method for automated segmentation of OCT which utilized deep learning, and combined fully convolutional networks with Gaussian processes (GP). The result confirms the combined efforts of DenseNet [33] and regression-based post-processing where DenseNet allows each layer of the network to directly process outputs from all previous layers.

An automated retinal layer segmentation network called ReLayNet [54] successfully segmented a retinal OCT B-scan into 7 retinal layers and areas of fluid accumulation. This proposed approach utilizes convolutional encoding blocks to learn a hierarchy of contextual features that follow an expansive path of decoders for semantic segmentation. To enable automated investigation of abnormal maculae, Sun et al. [61] implemented an FCN to recognize retinal areas in OCT images. Gopinath et al. [26] used a combination of a CNN which extract image layers and edges of interest, and long and short-term memory (LSTM) [31] to trace the layer boundaries enabling segmentation of the retinal layers from OCT scans. In [60], the authors adopted a two-stage FCN method which trained sequentially to achieve better segmentation performance. In the first stage, the OCT image was segmented using a trained FCN and the second stage was refined by another trained model with a decision mask to enhance the segmentation result.

However, the aforementioned machine learning-based image informatics methods have limitations relating to data annotation, model robustness when using OCT machines from different manufacturers, and the parameters of deep models required for clinical practice.

### 3. Materials

#### 3.1. OCT Imaging Data

Table 1 shows the detailed description of the ELM line dataset. The images utilized in this study were captured using SD-OCT imaging using the Heidelberg Spectralis (Heidelberg, Germany) as part of routine care, using the same imaging protocol in Sunderland Eye Infirmary, United Kingdom (UK). A high density central horizontal scanning protocol with 29-30 micron line spacing was used in the central 15 degrees. The individual OCT line scans were 768 × 496 pixels, with the scaling varying slightly between datasets but typically equating to 5.47 microns per pixel in the X (horizontal) axis and 3.87 microns per pixel in the Y (vertical) axis. With 29-30 microns spacing between scans (Z-axis), there were 49 scans per dataset. Image dataset and clinical data on 107 eyes from 107 patients were analysed. The mean age was 70 years old (SD 6.6, range 48–84), 88 (82%) were female and 54 (50%) were right eyes. The mean minimum linear diameter of the holes was 384.7 microns, and the median duration of symptoms was 6 months. All scans used a 16 automatic real-time setting, enabling multi-sampling and noise reduction over 16 images.

#### 3.1.1. Imaging Data Quality Control

The OCT scans have large variations in image data quality, related to the patients, operator, and the acquisition machine software. Since image data quality suffers from several artifacts such as speckle noise, contrast changes, and motion, it is necessary to remove them from the benchmark dataset using conventional methods. Furthermore, it serves as a guide to detecting anomalies among all data. Thus, a high-quality benchmark dataset is provided by robust and rigorous quality checks of OCT images. For this purpose, we utilized mathematical models that followed various evaluation measures, which are described below.

- **Noise score:** To represent the optimal noise thresholds of OCT images, noise estimation algorithms have been studied and developed by researchers for many years [19, 28]. This study measured the noise value by using a robust wavelet-based estimator of the Gaussian noise standard deviation, called the wavelet transformation, see Fig. 3.

- **Blurriness score:** Among the existing image blurriness estimation methods, a variance of image \( I(x, y) \) intensity smoothed with a Gaussian filter \( G(x, y) \) has been extensively used [23, 9]. Similarly, we measured the variance of image intensities. If the calculated variance of the image is low, then it is concluded as *sharp*; if variance is high, then it is *blurry* (see Fig. 4, corresponding intensity values are included).

- **Contrast score:** The perceived contrast provides the sharpness of the image, which must be also accounted for in the quality evaluation. It is calculated as the ratio of the mean intensity of the image to the standard deviation of the image intensity. A high contrast score indicates a sharp image, while a low contrast score indicates a blurry image. The contrast score is calculated using the following formula:

\[
\text{Contrast Score} = \frac{\text{Mean Intensity}}{\text{Standard Deviation}}
\]

![Small.](image1.png)  ![Large.](image2.png)

**Figure 3:** Spectral domain optical coherence tomography images demonstrating a small and large noise score.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Illustration of ELM line dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of 3D Volumes</td>
<td>No. of 2D Images</td>
</tr>
<tr>
<td>107</td>
<td>5243</td>
</tr>
</tbody>
</table>

Vivek Kumar et al.: Preprint submitted to Elsevier
for effectiveness in data pre-processing. Local measure methods [57, 66] reveal local contrast maps, which carry important information about the structure of the image. With this approach, we used the gradients of each pixel of the image and then computed the standard deviation (the square root of variance) as a measurement of image contrast.

- Motionscore: Accurate estimation of patient motion in medical imaging data is an important task that has been extensively researched [46]. In this study, the Horn-Schunck optical flow motion estimation method has been performed to perceive distortions in the smooth flow of information along the Z-axis of every 3D image. The method uses first-order derivatives as seen in the Equation 1:

\[
\min_{u,v} \left[ \nabla_x I u_j + \nabla_y I v_i + \nabla_z I \right]^2 ,
\]

where \( u, v \) represent the horizontal and vertical optical flow vectors, and \( \nabla_x I, \nabla_y I, \) and \( \nabla_z I \) represent the derivatives of image across \( x, y, t \) dimensions (\( t \)-time, which would be \( z \) in our case). To illustrate the motion estimation score, an RGB colour map was used, as shown in Fig. 5. Here, grey colour indicates low motion and red or blue colour indicates high motion between any two neighbouring 2D planes in a 3D image.

- Brightness-Darkness score: To research image quality control, the level of brightness/darkness in an image can help identify specific image characteristics. This measurement achieves the greatest difference in terms of luminance. To reveal pixel brightness values, we performed a brightness editing algorithm [10].

- Average pixel width score: To calculate the average pixel score, first Gaussian smoothing was applied to decrease the noise impact, followed by an image edge detection with John Canny’s method [11]. Following this, pixels that appeared far from potential edges were removed, named non-maximum suppression. In addition, a double threshold was performed, to reveal the essential edges.

3.1.2. Anomaly Detection

A common requirement, when rapidly and efficiently processing real-world data in deep learning applications, is to determine which data samples differ from others [53, 25]. The use of anomaly detection methods enables the identification of cases that are unusual in real-world datasets before being used by machine learning models.

According to the research conducted by Gopinath et al. [25], classification, statistics and density based anomaly detection methods were proposed as an optimal solution, when prior knowledge about the datasets is unavailable.

In this study, we used the statistics based robust covariance approach. Robust covariance estimation is a technique that assumes the entire dataset is of Gaussian distribution [4, 22]. The elliptic envelope method, which is a robust covariance estimation method, was chosen in accordance with our dataset distribution as it reveals an elliptic boundary. This boundary regularizes the covariance matrix and identifies abnormal and normal data with prediction scores.

The above-defined imaging data quality scores, estimated for each 2D image slice in every 3D volume, are used by the robust covariance method to detect anomalies in our imaging dataset. Algorithm 1 presents the proposed image quality approach to determine anomaly candidates \( a \), with anomaly scores \( d \), based on image data quality measurements \( f_z \) calculated for each 2D image slice \( I_z \) in a set of 3D volume \( I \).

**Algorithm 1: 2D Image Anomaly Candidates and Anomaly Scores Calculation**

**Require**: \( \{I\} \) //Set of 3D images  
**Ensure**: \( \{a\}, \{d\} \) //Sets of anomaly candidates and anomaly scores for all 2D image slices of 3D images  
\( p \leftarrow |\{I\}| \) //Number of 3D images  
for \( i \leftarrow 1 \) to \( p \) do  
\( s_{x,y,z} \leftarrow |I| \) //Image size for \( x, y, z \)  
for \( z \leftarrow 1 \) to \( s_z \) do  
\( \{f_z^i\} \leftarrow quality(I_z^i) \) //2D image quality scores  
end for  
end for  
\( \{f_z^i\} \leftarrow normalise(\{f_z^i\}) \)  
\( \{a_z^i\}, \{d_z^i\} \leftarrow anomaly(\{f_z^i\}) \)
3.2. OCT Imaging Annotation Data

Annotation of the ELM line in MHs was performed by a data science researcher and two experienced ophthalmologists. Specifically, one academic ophthalmology trainee doctor with 2 years of experience, and one consultant ophthalmologist with over 20 years of clinical experience in their field. Firstly, the data science researcher manually annotated each slice of the OCT image. All annotations were then checked by the junior ophthalmologist. In challenging cases, the experienced ophthalmologist was consulted, and annotations were adjusted accordingly. We exported the slice-wise volumetric resultant annotations as Tagged Image File Format (TIFF) files. These annotated binary masks correspond to [0, 255] for black and white pixels into background and foreground classes, respectively.

3.2.1. Annotation criteria

Annotation data quality plays an important role in building accurate segmentation methods and facilitating the diagnosis and monitoring of disease. Therefore, we measured relevant annotation quality indices, including gradient-based ELM line detection and MH detection, in the annotated OCT image slices. Fig. 8 shows an example of a gradient-based method that measures the intensity variation of the ELM line from a 3D OCT image. Yellow or red corresponds to the intensity variation of pixel change, whereas blue represents no change in pixel intensity. We used a Gaussian filter, which...
Figure 9: Spectral domain optical coherence tomography (OCT) slices from a single idiopathic full-thickness macular hole (MH) from our dataset, with the external limiting membrane (ELM) shown in red colour.

calculates a multi-dimensional maximum filter on the edge of the ELM line. To measure the changes in the neighbouring pixels, connected components with five pixels compute the gradient local maxima and measure the variance. The main purpose of this method is to estimate the intensity variation that presents any uncertainty that could affect the annotation of the ELM line.

Idiopathic full-thickness macular holes were detected using the following three key steps: 1) we used a 3D annotated mask by estimating the maximum of an array or maximum along an axis towards the direction of depth; 2) we set the background label to zero to explicitly focus on the foreground ELM line; and 3) we separated the connected components of defined labels by taking the maximum from them. This helped to identify the continuous line (surface in 3D) representing the ELM line, and the one discontinuity corresponding to the presence of a macular hole. Each patient had only one macular hole present in their annotated images. Fig. 9, shows multiple OCT slices of a single patient from our benchmark dataset, with its ELM line marked with red colour.
4. Methods

In this section, we present the architectural details of seven segmentation methods including fully convolutional networks (FCN) [44], U-Net [52], SegNet [5], Attention U-Net [50], recurrent residual convolutional neural networks (R2U-Net) [2], Efficient U-Net [7], and DeepLabv3+ [14] employed to segment the ELM line from OCT image.

4.1. Fully Convolutional Networks (FCN)

To perform semantic segmentation, Long et al. [44] introduced the first deep learning-based fully convolutional network (FCN) method. We adopted FCN to provide an accurate ELM line segmentation. During training, the encoder starts with an initial block which performs convolution on the input OCT image with a kernel of size 7 × 7 and a stride of 2. Then, we employed the first four layers from a pre-trained ResNet-18 (residual networks) [29] in an encoder. Specifically, we utilize the pre-trained ImageNet weights. The early residual layers are used to extract the most relevant salient features (i.e., edge, texture, and intensity, etc.) from OCT images that lead to the precise ELM line segmentation. We upsampled the encoded features by applying bilinear interpolation to predict the segmentation mask. In addition, employing the skip connection to combine low and high layer features in the final output layer for fine-grained ELM line segmentation.

4.2. U-Net

We used U-Net [52] which includes five encoder and decoder layers. The encoder layer repeated two 3 × 3 convolutions and strides, 1 × 1 followed by batch normalization and the non-linear ReLU activation function. These convolutional filters directly learn intrinsic retinal features (gray-level, texture, gradients, edges, etc.) from the OCT images. To down sample the feature maps, we applied a 2 × 2 max-pool operation with stride 2 × 2 after each encoder layer. Each decoder layer consists of an upsampling of the feature maps followed by a 2 × 2 up-convolution that divides the number of feature channels, a skip connection ( concatenation) with the correspondingly cropped feature map from the encoder layers, and two 3 × 3 convolutions, each followed by a ReLU. This skip connection between encoding and decoding layers were added to preserve relevant information from the input features. At the final layer, a 1 × 1 convolution was adopted to map each 64 element feature vector into the binary segmentation task of the ELM line.

4.3. SegNet

We employed the SegNet [5] method to segment the ELM line. This network is followed by encoder and decoder blocks. Specifically, the encoder utilized a 13 convolution layer of the VGG16 [58] network. During the training process, convolution filters with size 3 × 3 extracted the set of feature maps from OCT images. Moreover, due to variance in the feature maps, batch normalization was applied to avoid overfitting. Non-linear ReLU activation functions were applied. To evade overfitting and reduce the input representation (feature dimensionality), max-pooling with size 2 × 2 and non-overlapping stride were 2 × 2 employed to the resultant feature maps. Likewise, the decoder has 13 deconvolutional layers that were used to up-sample the extracted feature maps of the ELM line. The high dimensional feature description at the output of the final decoder was served to a sigmoid classifier to generate class probabilities for each pixel individually.

4.4. Attention Gates in U-Net (Attention U-Net)

Attention U-Net [50] was employed to provide ELM line segmentation from the OCT image dataset. It is the modified version of the very popular biomedical image segmentation model U-Net to advance model sensitivity to foreground pixels without needing complex design. This network architecture represents a similar structure of standard U-Net with five encoding and decoding layers. After each encoding layer, the attention module was applied. This module is employed by introducing a grid-based gating mechanism. Each encoder layer of attention-based features was passed to the corresponding decoder layer through the skip connection, which evades disambiguating inappropriate and noisy features. This model helps to highlight more relevant ELM line features by ignoring other background regions and generate a precise output segmentation map.

4.5. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net)

The R2U-Net [2] is a medical image segmentation model that provides an advancement from the classical U-Net method. It consists of five encoding and decoding layers. We used the recurrent residual convolution block in each encoding and decoding layer. Especially, this recurrent convolution was applied before the features downsampling and upsampling process in encoding and decoding layers respectively. Finally, this recurrent convolutional neural network (RCNN) block was also added before the final segmentation result. Similar to U-Net, R2U-Net utilized a skip connection between the encoding and decoding layers. The sigmoid activation function was used to calculate the final pixel probability of ELM line output.

4.6. Efficient U-Net

The EfficientNet [63] network was utilized as an encoder for feature extraction from OCT images and the decoder was used to provide the segmentation map of the ELM line. This EfficientNet has eight variations that include EfficientNetB0 to EfficientNetB7. Due to computational restriction, we applied EfficientNetB3 [63] to perform feature extraction inside the encoder layers of U-Net. It consists of mobile inverted bottleneck convolution (MBConv) [55] that combines the squeeze and excitation mechanism to highlight the more relevant features of the ELM line. Similarly, the decoder has kept a similar architecture to standard U-Net. Skip connection with element-wise feature concatenation was employed to preserve the high-level features and spatial information by filling the semantic gaps during the reconstruction process.
4.7. DeepLabv3+

To provide a precise semantic segmentation, we employed DeepLabv3+ [14], which is an extension of the previous method called DeepLabv3 [13]. During the training process, a pre-trained ResNet-18 model as a network backbone was used to extract rich feature information from the OCT dataset. Without adding any computational complexity, the extracted feature maps are inputted into the atrous convolution block that is used to control the resolution of obtained features from CNNs. Furthermore, it allows low and high-level features to be combined (ELM line, and its background). The atrous convolutional block also enables increments to capture multi-scale contextual information from OCT images. We used an output stride of 16 which enables a $1 \times 1$ convolution and three $3 \times 3$ dilated convolutions of varying dilation rates of 1, 6, 12, and 18 respectively. The last convolutional layer has a kernel size of $1 \times 1$ served by a global average pooling operation. To obtain high-level feature maps, an upsampling operation was applied after each atrous convolution layer, and then all extracted feature maps were concatenated. In the decoder, to recover the high-level and low-level feature maps, bi-linear interpolation was used to upsample the features twice. To obtain rich information, we combined both high and low-level features. Ultimately, the bi-linear interpolation technique was employed for the final segmentation of ELM line output by upsampling the feature map by 2.

4.8. Cost Function

In this work, we used the sum of binary cross-entropy (BCE) and the Dice loss as a utilized cost function. Here, Dice loss $L^{\text{Dice}}$ is the Dice coefficient $\text{Dice}$ that can be expressed as follows:

$$L^{\text{Dice}}(G, P) = 1 - \text{dice}(G, P) = 1 - \frac{2|G| \cdot |P|}{|G|^2 + |P|^2}$$

where $G$ is the ground truth image mask comprising an ELM line and $P$ is the predicted mask from the image segmentation output.

The overall loss function is formulated as follows:

$$L^{\text{OL}}(G, P) = \alpha(-(G \cdot \log(P) + (1 - G) \cdot \log(1 - P))) + (1 - \alpha)L^{\text{Dice}}(G, P).$$

Where, $\alpha$ is an empirical weighting factor.

5. Experimental Design and Results

5.1. Parameter Selection and Training

We implemented our network using Python 3.6, CUDA 10.2, cuDNN 7.0, PyTorch 1.8.1 running on a 64-bit Ubuntu operating system using a 3.4 GHz Intel Core-i9 with 32 GB of RAM and NVIDIA RTX 2080 Super GPU with a memory of 8 GB. We have trained our model by using the input resolution size of 256 $\times$ 256. During training, we used ADAM optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate = 0.0002 with a batch size of two. We set the loss weighting factor $\alpha$ to 0.5. Finally, we trained the model with 100 epochs and evaluated it on the best checkpoint saved by the highest Dice coefficient score. In order to generate the final segmentation result, we set the threshold value to 0.5.

The dataset was divided patient-wise into training, validation, and testing sets by using five-fold cross-validation with a ratio of 70%, 10% and, 20% respectively. Specifically, each cross-validated fold included the 75, 11, and 21 patients into training, validation, and testing, respectively. Further, after the image quality control and anomaly detection, we excluded 152 OCT images from the network training. Finally, we utilized 3675, 387, and 980 OCT images for training, validation, and testing respectively. We evaluate the model performance on the independent test set. Note, we confirm that there are no patients shared within the splits and the folds of the cross-validation study. To provide variability and artificially increase the size of the ELM line dataset, various data augmentation techniques were applied. We scaled the images by changing the scaling variable from 0.5 to 1.5 with a step size of 0.25. Next, we applied gamma correction to the images by varying the gamma scaling constant from 0.5 to 1.5 with a step size of 0.5. Finally, we horizontally and vertically flipped the images, and rotated them with different angles of 15 degrees.

5.2. Quantitative Evaluation

In this study, we employ six evaluation metrics to evaluate the performance of each segmentation method, including Dice coefficient score (DSC), intersection over union (IoU), root-mean-square-error (RMSE), Hausdorff distance (HD), sensitivity (SEN), and false-positive error (FPR). The segmentation results analysis relied on a confusion matrix that encompasses four different variables: true positive (TP), false positive (FP), true negative (TN), and false-negative (FN). Here, we present the mathematical formulations of the six metrics: DSC, IoU, SEN, FPR, RMSE, and HD.

**Dice Coefficient:** The Dice coefficient was used to measure the similarity between ground truth and the predicted mask.

$$DSC = \frac{2|P \cap G|}{|P| + |G|} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

where $G$ is the ground-truth mask and $P$ is the predicted mask.

**Intersection Over Union:** The intersection over union calculates the percentage overlap between the ELM line ground truth and predicted mask.

$$IoU = \frac{TP}{TP + FP + FN}$$

**Sensitivity:** Sensitivity calculates the portion of positive pixels in the ground truth that is also identified as positive by the algorithm being evaluated. Sensitivity can be obtained by the Equation (6):

$$SEN = \frac{TP}{TP + FN}$$
Table 2
Quantitative comparison of the seven segmentation methods with and without the effect of data augmentation on the ELM line test set by using six evaluation metrics i.e., DSC, IoU, RMSE, HD, SEN, and FPR. Statistically notable results are highlighted in bold font.

<table>
<thead>
<tr>
<th>Models</th>
<th>With data augmentation</th>
<th>Without data augmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSC ↑</td>
<td>IoU ↑</td>
</tr>
<tr>
<td>FCN [44]</td>
<td>77.307</td>
<td>63.562</td>
</tr>
<tr>
<td>U-Net [52]</td>
<td>78.026</td>
<td>64.358</td>
</tr>
<tr>
<td>SegNet [5]</td>
<td>76.501</td>
<td>62.435</td>
</tr>
<tr>
<td>Attention U-Net [50]</td>
<td>78.270</td>
<td>64.711</td>
</tr>
<tr>
<td>R2U-Net [2]</td>
<td>78.299</td>
<td>64.179</td>
</tr>
<tr>
<td>Efficient U-Net [7]</td>
<td>77.583</td>
<td>63.832</td>
</tr>
<tr>
<td>DeepLabv3+ [14]</td>
<td>76.328</td>
<td>62.277</td>
</tr>
</tbody>
</table>

False Positive Rate: This is the ratio of false positives to the sum of false positives and true negatives.

\[ FPR = \frac{FP}{FP + TN} \] (7)

Root Mean Square Error: RMSE measures the difference between the ground truth and the predicted mask of the ELM line (units in pixels). Here, a smaller value represents better segmentation output.

\[ RMSE = \sqrt{\frac{1}{n} \sum_{i \in T} (P(i) - G(i))^2} \] (8)

where \( G(i) \) is the ground-truth of pixel \( i \), \( P(i) \) corresponding predicted mask, \( T \) is the set of valid pixels (i.e., both the ground-truth and predicted mask pixels that do not have mask values equal to zero or non-black regions), \( n \) is the number of elements in \( T \).

Hausdorff Distance: This is a symmetric measure of distance between two masks (smaller is better) and is defined as [15]. This distance can be calculated in pixels. Here, The point sets \( S_G \) and \( S_P \) belong to the pixels of the ground truth mask and the prediction, respectively.

\[ HD = \max \left\{ \max_{S_G \in S(G)} d(S_G, S(P)), \max_{S_P \in S(P)} d(S_P, S(G)) \right\} \] (9)

Where, \( d \) refers to the distance between the two points.

5.3. State-Of-The-Art Method Comparisons
In Table 2, we present the quantitative results of seven segmentation methods on the ELM line dataset. These methods include the (FCN) [44], U-Net [52], SegNet [5], Attention U-Net [50], recurrent residual convolutional neural networks (R2U-Net) [2], Efficient U-Net [7], and DeepLabv3+[14]. Initially, the experiments were performed without using any data augmentation techniques, but later performed data augmentation on the segmentation results (see subsection 5.5). Experimental results demonstrated that R2U-Net shows effectiveness and superiority in DSC, RMSE, HD, and FPR scores of 78.03%, 0.069, 5.99, and 0.141%, respectively compared with other methodologies. The recurrent convolutional neural network (RCNN) leads to precise segmentation due to its temporal dependencies from OCT images. It extracts rich low-level features (i.e., edge, texture, and intensities, etc.) and high-level features. In turn, U-Net and Attention U-Net performed similarly and yielded 63.63% and 63.53% IoU scores, respectively. The attention mechanism allows capturing the most relevant foreground ELM line features and discards the irrelevant ones. SegNet scored comparatively lower segmentation results compared to the rest methods. Overall, we observed that the U-Net family models (i.e., U-Net, Attention U-Net, R2U-Net, and Efficient U-Net) generate similar ELM line segmentation results.

In Fig. 10a, 10b we have provided descriptive statistics of box-plot analysis for the most relevant DSC, and IoU scores. These measurements constitute each of the employed segmentation methods. We used different colour boxes to indicate the score range of several methods; the white line inside each box represents the median value, box limits include interquartile ranges Q2 and Q3 (from 25% to 75% of samples), upper and lower whiskers are computed as 1.5 × the distance of upper and lower limits of the box, and all values outside the whiskers are considered outliers, which marked with the (+) symbol. As clearly shown, R2U-Net outperforms the other six methods with a lower standard deviation. Descriptive statistics also demonstrated that all the ELM line segmentation methods produced few outliers. These outliers may affect segmentation results for certain cases.

5.4. Qualitative Evaluation
Fig. 11 presents the three qualitative examples to compare the seven state-of-the-art segmentation methods (i.e., FCN, SegNet, U-Net, Attention U-Net, R2U-Net, Efficient U-Net, and DeepLabv3+) used to segment the ELM line. The white rectangular box provides a zoom-in visualization of the specific region, where all the utilized methods failed to provide precise ELM line segmentation. It is evident that most of the compared ELM line segmentation methods performed similarly by generating only very few false positives due to the fuzzy boundaries and high variation in pixel intensity across the ELM line, as highlighted in the zoom-in box. Moreover, the neighboring retinal layers create a challenge for correctly identifying the ELM line. Conclusively, all deep neural network-based methods used are susceptible
to intensity variation and achieve better performance in the presence of consistent pixel intensity regions. To visually examine the segmentation results, we used color coding in which the orange color represents true positives, red color denotes the false negatives and green corresponds to false positives.

In Fig. 11, the last row includes the pixel intensity gradient map generated by the ground truth image where the blue colour corresponds to no variations in intensity, whilst certain areas, shown as a yellow or red colour, depict increasing changes in intensity. The generated gradient map clearly confirms that there is a specific region where the intensity of an image differs from its neighboring pixel region. In other words, we analyzed the most in-homogeneous regions that infer the segmentation performance of deep neural networks. And we found that all the employed deep learning models performed nicely with homogeneous intensity regions and failed in the presence of ambiguous boundaries or changes in pixel intensity shown in the zoom-in white box.

5.5. Ablation Study

In recent years, ablation studies are involved in various relevant scientific publications to assess deep learning performance [47]. We provide a broad range of ablation studies that involve the effect of data augmentation, the effect of the loss function, and the computational complexity of each segmentation method.

Effect of Data Augmentation: Table 2 shows the effect of data augmentation with each of the employed methods. As can be seen, adding the additional diversity in feature representation leads to a significant improvement in segmentation outcomes. These varying imaging features (e.g., transformation, rotation, contrast changes, and flipping operations) helps to fill the semantic gap and make network training more generalized. This strategy served to improve the model performance by around 1.5% - 2.5%. Moreover, extensive testing confirms rotation of the OCT images by more than 15 degrees did not result in any significant deterioration in its performance.

Effect of Loss Function: Fig. 12 presents the effects of various loss functions (i.e., BCE, Dice loss, IoU loss, BCE+IoU, and BCE+Dice loss) on ELM line test set. We evaluated all the seven segmentation methods underlying different loss functions. Experimental results confirm that the R2U-Net has outperformed other methods with a combination of BCE and Dice loss functions. This combination of loss functions allows faster loss convergence during training and achieves more precise ELM line segmentation. Further, it helps in reducing the number of false positives from the segmented mask remarkably. This ablation study verifies the utilized loss function (BCE + Dice loss) yields a better increment at pixel-level in the ELM line segmentation results.

Computational Complexity: In Table 3, we present the computational complexity of each method in terms of its parameters in million (M) and multiply–accumulate operation (MACs). FCN follows a large number of convolution layers that contained 135.53M parameters with 50.13 MACs. However, DeepLabv3+ showed the second-highest complexity with 59.34M parameters compared to the rest methods. Further, U-Net and Attention U-Net have very comparable trainable parameters and achieved similar segmentation results. But, R2U-Net has 39.09M parameters with the highest 152.94 MACs that increased by the recurrent neural network operation. Besides, Efficient U-Net only contained 21.44M parameters with the lowest 9.69 MACs. This allows the
Figure 11: Qualitative comparison of seven state-of-the-art segmentation methods evaluated on the ELM line test set. In a row, we present three examples used to compare all the segmentation methods. Here, we map the results into three colors: orange color refers to true positive, red color a false negative, and green color presents false positives. Further, gradient-based color maps depict the dark blue color that highlights no variation in the pixel intensities. In addition, a yellow or red color indicates increasing changes in intensity. The white box provides a zoom-in visualization of the specific region, where the compared segmentation methods failed to provide precise ELM line detection.
Figure 12: Illustration of various loss functions performance (i.e., BCE, Dice, IoU, BCE+IoU, and BCE+Dice) comparison by the best results achieved by R2U-Net method. The dice coefficient and IoU scores were adapted to measure the quantitative changes.

Table 3

<table>
<thead>
<tr>
<th>Models</th>
<th>Parameters (M)</th>
<th>MACs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>135.53</td>
<td>50.13</td>
</tr>
<tr>
<td>SegNet</td>
<td>29.44</td>
<td>40.10</td>
</tr>
<tr>
<td>U-Net</td>
<td>34.53</td>
<td>65.53</td>
</tr>
<tr>
<td>Attention U-Net</td>
<td>34.88</td>
<td>66.64</td>
</tr>
<tr>
<td>R2U-Net</td>
<td>39.09</td>
<td>152.94</td>
</tr>
<tr>
<td>Efficient U-Net</td>
<td>21.44</td>
<td>9.69</td>
</tr>
</tbody>
</table>

network to train faster and achieved comparable results with other state-of-the-art segmentation methods.

5.6. Limitations

Accurately identifying the ELM line using OCT images is a challenging assignment due to artifacts and variation in pixel intensities leads to more false-positive results. We accept that our study has limitations. First, we performed an analysis on a dataset compiled from one hospital with a single manufacturers’ OCT device on a predominantly white population. Therefore, the experimental results may not be generalizable to other populations or OCT appliances. Second, the benchmark dataset only includes images from patients with idiopathic full-thickness macular holes, a particularly challenging scenario for ELM line detection. Lastly, in this study, we have not calculated any clinically applicable measurements from the images and compared them to human-derived ones, which is our next aim.

6. Conclusion

In this article, we have presented an OCT image dataset as a new benchmark for ELM line segmentation. Specifically, the dataset contains images from one eye of 107 patients with idiopathic full-thickness macular holes, with precise ELM line annotations. The OCT images have variation in image contrast, motion, brightness, and speckle noise, typical of routinely collected clinical data. We performed a detailed analysis of image and annotation quality, using a range of standard scores, and gradient mapping of the ELM line. We also performed an extensive ablation study that included the effect of the loss function, the effect of data augmentation, and computational complexity and compared the results of seven state-of-the-art image segmentation methods. All the segmentation methods achieved greater than 75% Dice coefficient score, however, there remains potential to improve the methodology with further study. The ability to detect and measure the ELM is an important part of retinal assessment in a number of blinding diseases. In particular in eyes with macular holes, where the ELM can be challenging to evaluate, it can be used to select the optimal surgical techniques and assess prognosis. In future work, we will assess the use of automated ELM detection to derive clinically relevant measures to predict outcome in patients undergoing surgery.

Acknowledgment

This work was supported by a Northern Accelerator grant (OSR/0260/DSTE/HEIF) and was supported by additional funding from Newcastle University.

Data Availability

The benchmark ELM line OCT dataset is available on request from the corresponding author for noncommercial use. The source code is publicly available at https://github.com/vivek231/Benchmark-ELM-Line-OCT-Dataset.

References

Bench-marking Automated Detection Of The Retinal ELM In 3D SD-OCT


