IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 February 2024, accepted 6 March 2024, date of publication 18 March 2024, date of current version 4 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3379015

RESEARCH ARTICLE

Interactive Deep Learning-Based Retinal OCT Layer Segmentation Refinement by Regressing Translation Maps

GUILHERME ARESTA, TERESA ARAÚJO, BOTOND FAZEKAS, JULIA MAI, URSULA SCHMIDT-ERFURTH, AND HRVOJE BOGUNOVIĆ

Christian Doppler Laboratory for Artificial Intelligence in Retina, Department of Ophthalmology and Optometry, Medical University of Vienna, 1090 Vienna, Austria

Corresponding author: Guilherme Aresta (guilherme.moreiraaresta@meduniwien.ac.at)

This work was supported in part by the Christian Doppler Research Association; in part by the Austrian Federal Ministry for Digital and Economic Affairs; and in part by the National Foundation for Research, Technology, and Development.

ABSTRACT Retinal layer segmentation in optical coherence tomography (OCT) is essential for the diagnosis and follow-up of several diseases. Despite the success of deep learning approaches for this task, their clinical applicability is limited, since they neither account for pathologies other than those present in the training set nor for the specialists' subjectivity. Thus, we propose an interactive layer segmentation approach that allows to obtain an initial segmentation and, more importantly, to interactively correct those segmentations. Our deep learning-based approach predicts the translation required to correct layer boundary segmentations by regressing pixel-wise translation maps that account for the user input. The method is designed to allow for segmentation correction by interactions with point-clicks or line-scribbles. Additionally, the system outputs a coordinate-wise confidence, allowing to automatically identify regions of possible segmentation failure that may require user attention. We extensively validate our approach on multiple private and public datasets with different pathomorphological complexities, achieving state-of-the-art performance, while allowing for a simple and efficient user interaction.

INDEX TERMS Image segmentation, interactive annotation, deep learning, human-in-the-loop, optical coherence tomography, retina.

I. INTRODUCTION

In recent years, the intersection of computer vision and medical imaging has significantly advanced diagnostic and therapeutic practices in healthcare. Indeed, medical image analysis now plays a pivotal role in interpreting and extracting clinically relevant information from various medical imaging modalities [1]. Within this interdisciplinary domain, retinal imaging emerges as a crucial subset, focusing on the acquisition, processing, and analysis of images captured from the retina. The acquisition and analysis of retinal images allows the non-invasive diagnosis of a myriad of pathologies, including not only ocular diseases but also systemic conditions such as cardiovascular problems and dementia [2].

Retinal optical coherence tomography (OCT) is a 3D non-invasive imaging technique that allows to evaluate and monitor the status of the retina and its layers. Quantifying changes in these layers is essential for the diagnosis and follow-up of eye diseases such as age-related macular degeneration (AMD) and diabetic macular edema (DME), as well as conditions such as multiple sclerosis [3], [4], [5]. However, the manual annotation of these dense volumetric images is time-consuming, and thus automated segmentation tools are being standardly offered by OCT device manufacturers. Retinal layer boundary segmentation methods predict a set of coordinates of the most probable location of the interface between two retinal layers. These approaches usually rely

The associate editor coordinating the review of this manuscript and approving it for publication was Chulhong Kim.

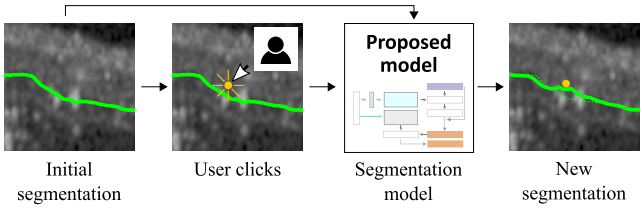


FIGURE 1. Our approach corrects OCT retinal layer boundaries segmentations using user clicks and the previous segmentation.

on finding an initial boundary proposal, which can then be refined with graph- or level set-based approaches [6], [7], [8], [9], [10], [11].

Fully-convolutional networks have become the state-of-the-art approach for retinal boundary segmentation, obtaining high quality results by either segmenting the retinal layers and identifying their interfaces [12], [13] or by directly inferring their boundaries [10], [14]. Recently, [15] proposed a deep model that relies on both layer and boundary segmentations while imposing anatomical constraints, achieving low error without additional post-processing.

Despite the high quality of existing methods, local segmentation failures still occur, in particular for cases with pathologies other than those in the training set. Also, retinal layer segmentation, similarly to other medical segmentation tasks, has an associated subjectivity, with specialists often preferring different but valid solutions in severe pathological regions. Because of this, obtaining a precise and personalized layer boundary still requires physicians to actively correct different regions on multiple layers [16]. Yet, to the best of our knowledge, end-to-end deep learning approaches that allow for interactive retinal layer boundary segmentation on OCT images have yet to be explored. To further reduce the workload of the specialists, these systems should focus on user experience, (i) allowing to perform corrections near the interaction region without degrading the segmentation quality elsewhere, (ii) being capable of correcting multiple boundaries at once while interacting with only one of them.

With this in mind, we developed a novel approach (Fig. 1) that proposes an initial segmentation of the retinal layer boundaries, which can then be interactively corrected with the assistance of the system, reducing the workload of the user.

A. RELATED WORK

Interactive segmentation, i.e. adjusting a previous segmentation result according to the input of an annotator, is an actively researched topic in both computer vision and medical image analysis fields, with both non-deep and deep learning approaches being used in a complementary fashion. In retinal OCT, fully graph-based interactive layer segmentation methods [17], [18] such as LOGISMOS-JEI [19] use intensity and gradient information to construct cost maps, which can be adjusted by the users' interactions, to compute the segmentations. Despite the success of these algorithms, graph-based approaches tend to present a trade-off between segmentation quality and processing time, since they are computationally demanding and reliant on the design and tuning of task-specific constraints. Deep learning approaches, although more data demanding, usually achieve higher quality results within a similar time frame.

As a hybrid approach, in [20] the authors rely on deep learning to obtain an initial estimation, which can then be corrected using different image processing techniques. There, a deep learning system predicts a pixel-wise segmentation of two retinal layers and potential pathological manifestations in OCT B-scans. The inferred pixel-wise probabilities are used for automatically inferring regions of high uncertainty. These can then be semi-automatically corrected by the users. Notably, the proposed method allows to also automatically correct neighboring B-scans, decreasing the overall interaction time. To correct the segmentations, the authors provide a set of tools such as constrained shortest path prediction, B-spline-based corrections and polynomial smoothing. While correcting, users have to identify an appropriate tool and perform the corresponding hyperparameter tuning whenever necessary. In contrast, deep interactive approaches, such as ours, are much simpler from the user standpoint: as shown in Fig. 1, the user is only required to select the region where the boundary is expected, and the system will handle the correction automatically.

In general, interactive deep learning segmentation approaches can be divided into *region-focused*, where the goal is to change the segmentation at a pixel level, or *contour-focused*, where the goal is to solely correct the object boundary. In both approaches, the user input is commonly encoded as an attention map, which can then be used as a feature map and/or a weighting term for the training loss [21], [22], [23], [24].

1) REGION-FOCUSED METHODS

Such methods aim at utilizing the user input to include/exclude pixels from the current segmentation. A common approach is to adjust the model's weights case-wise according to the user interaction [21], [25], [26], [27]. For instance, in [21] they proposed to fine-tune a pre-trained model during the inference stage to a set of manual correction scribbles. These scribbles are encoded as a weight map for the loss function, encouraging the network to include (or exclude) the highlighted pixels. However, requiring test time training is time-consuming, which hinders real-world applications, where a near-instant feedback is expected. This is particularly relevant for retinal OCT images, where the time burden is further increased by the large number of boundaries and the 3-dimensionality of the data.

An alternative is to use networks that account for possible user inputs without the need for additional weight refinement [23], [28], [29], [30], [31], [32]. For example, [23] encodes users' clicks and scribbles as binary maps (in case of clicks) or as binarized distance transform maps (for scribbles), which are used as additional input channels of the image to segment.

The resulting segmentation can be changed by changing these extra input maps. However, previous interaction results are not considered, which may lead to incoherence between the steps.

Indeed, interactive approaches often do not retain memory of their previous state, i.e. there is no guarantee that the network will consider previous outputs and user inputs to produce the new result [25], [27], [28], [33], [34]. Previously corrected regions can become partially or totally lost, increasing the overall time required to achieve a satisfactory result [35], [36]. Because of this, in [35] they proposed an interactive segmentation framework that accounts for both the previous model's inference and the user input. In particular, user interactions are encoded using geodesic image transforms, and then used together with the previous segmentation as extra input channels to a second segmentation model. This model contains a trainable Conditional Random Field to promote a more spatially consistent result and allows using the input scribbles as hard segmentation constraints. To address this lack of memory, in [28] they proposed a two-stream network: one receives as input the image to segment and an encoding of the user clicks, and the other this same encoding together with the previous segmentation prediction. The resulting features are then combined and decoded to produce a new output, allowing to improve the quality of the segmentation without affecting the previously corrected details. Other notable examples are the Segment Anything Model (SAM) and its derivatives [37]. These models use transformer-based architectures trained with a very large number of images from different acquisition settings (either natural or medical images) in order to learn highly general representations that allow to segment objects of interest identified by (usually) clicks and bounding boxes. However, despite being promising approaches, their performance is still subpar when compared to domain-specific models, especially those that account for specific anatomical prior knowledge [38].

Overall, despite the merits of Region-focused approaches, their applicability in retinal OCT boundary segmentation is still limited. First, scribbles for pixel-wise correction are not intuitive for users, as it is not trivial to decide which of the two contiguous structures is under- or over-segmented or where the location of the interaction should be. Instead, it is more intuitive to simply indicate where the segmentation boundary should move to. In addition, usually the segmentation is not a direct correction of the previous result, but instead a novel prediction generated in a black-box fashion with no guarantee of the stability of the result between the interactions. Consequently, it has become of interest to address retinal OCT layer segmentation as a contour-focused problem.

2) CONTOUR-FOCUSED METHODS

Such methods define the segmentation task as predicting the position of a set of points that define the object of interest [39], [40], [41]. The contour is considered as an ordered point sequence, where each point's position depends on all the previous ones. In [42], they used a recurrent neural network to assist the creation of polygons around objects, where the network predicts the location of the next user click by assessing all the pre-existing points. Users interact by adjusting the position of the predicted vertices, which are then re-assessed by the network. However, this leads to long inference times. Because of this, [43] used a graph convolutional neural network (GCN). At each iteration, a feature vector is extracted from the output of a convolutional neural network (CNN) for each contour vertex. These features are then used in a GCN to predict the required point movements.

Although showing great potential, these methods are not easily translatable into retinal OCT layer segmentation. First, retinal layers are not closed objects, and thus approaching the task as a polygon delineation problem is not desirable. Furthermore, the extraction of accurate biomarkers from these images needs a segmentation as precise as possible, thus requiring a large number of vertices and consequently increasing the complexity of the models and the inference time. An alternative would be to use less control points and then approximate the final prediction. However, fitting a curve to the detected points requires extra hyperparameter tuning and may affect the local quality of the results, especially on pathological regions.

B. CONTRIBUTIONS

In this work, we propose a deep learning-based framework that provides an initial set of retinal layer boundaries in OCT images and allows for their correction via user introduced clicks/scribbles. The approach combines the strengths of pixel- and contour-focused approaches by treating the layer boundary segmentation task as a pixel-wise regression problem where the network has to predict the amount of movement required to properly readjust the coordinates of the current segmentation. Our contributions to the state-of-the-art are:

- a novel deep learning approach for retinal boundary segmentation, that not only achieves a performance similar or better than other state-of-the-art approaches, but importantly, also allows for interactive correction. In particular, we propose a new segmentation and interaction scheme that predicts the translation required to correct segmentations by regressing pixel-wise translation maps that account for the user input. The approach is specifically designed to allow for segmentation correction by simple user interactions with point clicks or line scribbles;
- an interactive system that outputs a coordinate-wise confidence, allowing to automatically identify regions of possible segmentation failure that may require user attention;
- an extensive quantitative and qualitative validation of the approach on private and public datasets spanning

different retinal diseases, and of different pathomorphological complexities.

II. METHODS

The proposed method (Fig. 2) predicts an initial set of retinal layer boundaries, which are then sequentially adjusted according to the provided user input. The model infers for each layer boundary a pixel-wise translation map (Fig. 2, Translation map) with the estimation of the signed vertical distance to the corresponding boundary coordinate. These values are used to move the previously predicted boundaries to new positions. Corrections proposed by the user are encoded as extra input channels (Fig. 2, Inputs) and promote changes to these translation maps and consequent update of the boundary coordinates. Ultimately, at each prediction iteration, the model accounts for user feedback to predict how much a point anywhere on the image would have to move to be in the correct location. Because the segmentation is never predicted from scratch, but is instead moved from a previous location, the evolution of the result is coherent between interactions. In addition, the model outputs a pixel-wise confidence map (Fig. 2, Confidence map) of how likely it is that a specific boundary occurs on that location, easing the identification of regions where the segmentation is likely to need correction.

Specifically, at each iteration with an user input, the model predicts a novel translation map T_l for the retinal layer l, with the same size as the input grayscale B-scan image. Each element of T corresponds to the estimated vertical translation required for the current prediction to reach the new location pointed by the user (Fig. 3). The system also predicts a confidence map C related to the current location of the segmentation.

Let $\mathbf{B}^i = \{b_0, \ldots, b_n\}$ be the set of predicted layer coordinates at the user interaction iteration i, where b are the predicted boundary coordinates and n is the number of layers to segment, and $\mathbf{C}^i = \{c_0, \ldots, c_n\}$ the corresponding prediction confidence $\in [0, 1]$. The overall goal of the network is to adjust the proposed solution from interaction $i-1$, $\mathbf{B}^{i-1}$, so that the segmentation error of $\mathbf{B}^i$ is lower than $\mathbf{B}^{i-1}$ and likewise that the confidence increases, i.e. $\sum \mathbf{C^i} > \sum \mathbf{C^{i-1}}$.

Our approach can use any encoder-decoder network backbone. In this study, we opt for a deep residual U-Net [15], [44], which has already been successfully used for retinal OCT images. Let $H \times W \times (1+n)$ be the input's shape. The first channel of the input is the OCT B-scan to segment, and the n channels encode the user interactions, one for each layer boundary (details in Section II-C). The backbone has residual convolutional blocks with batch normalization and outputs a $H \times W \times 64$ feature map, where H and W are the height and width of the input image, respectively. At the end of the backbone, there are two branches responsible for predicting the column-wise translation (the translation branch) and the corresponding prediction uncertainty (the confidence branch).

A. LAYER-WISE TRANSLATION AND COORDINATE PREDICTION

The translation branch predicts the column-wise shift needed to move a current boundary prediction b^{i-1} to a new location according to the user input. For that, the backbone network's output is convolved with a $(3 \times 3) \times 64$ kernel, followed by n distinct $(3 \times 3) \times 1$ kernels with linear activation, resulting in the preliminary maps $T'_0, \ldots, T'_n$. To promote anatomical coherence, each translation map T_l is computed with reference to the layer immediately above:

$$T_l = T_{l-1} + \mathrm{ReLU}(T'_l) \tag{1}$$

and for the first layer, no rectification is performed. The new boundary is based on the previous prediction iteration:

$$b^i_l = b^{i-1}_l + T_l\left(b^{i-1}_l\right) \tag{2}$$

where $T_l\left(b^{i-1}_l\right)$ is evaluated by bilinear interpolation. Fig. 3 shows a schematic representation of the correction. Anatomically coherent layer ordering is guaranteed as in [15]:

$$b_l = b_{l-1} + \mathrm{ReLU}\left(b'_l - b_{l-1}\right) \tag{3}$$

where b'_l is the result of Eq. 2. Because the network output follows the expected layer ordering, i.e. b_l is the boundary that immediately follows b_{l-1}, Eq. 3 ensures each boundary depth is always greater (or at least the same) as its preceder, thus guaranteeing a proper layer ordering.

Note that for the initial prediction $i = 0$, $\mathbf{B}^0$ is not defined, as it is not a direct output of the network. To circumvent this, the layer coordinates are initialized as:

$$\mathbf{B}^0 = \left\{T^0_0\left(\mathrm{argmin}\left(T^0_0\right)\right), \ldots, T^0_n\left(\mathrm{argmin}\left(T^0_n\right)\right)\right\}, \tag{4}$$

i.e., as a subpixel-refined coordinates of the positive-negative transition of the translation maps, which corresponds to the region where the boundary is most likely to occur. Note that $\mathbf{B}^0$ can also result from any boundary segmentation algorithm.

Predicting $\mathbf{B}^i$ via Eqs. 1, 2 and 3 embeds anatomical knowledge into the model, forcing an anatomically consistent layer ordering $b_l \geq b_{l-1}$. Also, at each user interaction the network always has access to the previous segmentation state b^{i-1}, guaranteeing that b^i is a direct modification of b^{i-1} instead of a new solution predicted from scratch. This leads to a coherent progression of the segmentation without requiring costly and slow parameter tuning via, for instance, retraining.

B. POSITION CONFIDENCE ESTIMATION

For each layer boundary coordinate, the system also infers how confident it is that each point is now on the correct location. The confidence prediction branch has a similar structure to the translation branch. Specifically, the confidence results from a bilinear interpolation of b^i_l with a $W \times H$ map that results from a sigmoid activation, i.e. $c^i_l = C\left(b^i_l\right)$.

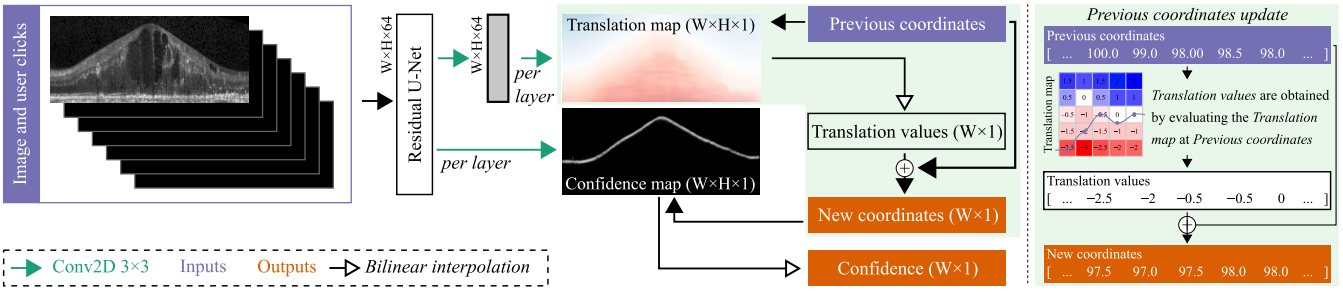


FIGURE 2. Proposed algorithm structure for interactive retinal boundary segmentation in OCT images.

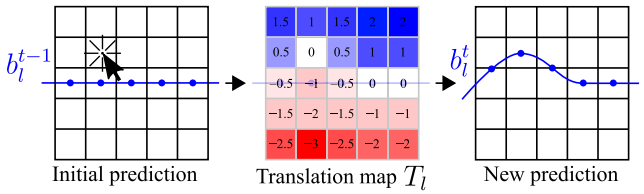


FIGURE 3. Updating a previous prediction b_l^{t-1} by vertically adjusting its coordinates via the inferred translation map T_l.

C. ENCODING THE USER INTERACTION

The model interprets users' clicks and scribbles to adapt the translation maps T so that the boundaries move to the desired new location. This interaction is encoded with maps that are provided to the network together with the input image. We use binary matrices (one per retinal boundary) with value 1 on the coordinates clicked by the user. Although a common approach is to process these inputs with 2D Gaussian kernels of different spreads [27], preliminary experiments showed no meaningful gain on using mappings other than the binary matrix.

D. TRAINING SCHEME

During training, we simulate user inputs by selecting for each layer boundary the reference standard coordinate that has the highest absolute distance from the respective prediction. Each batch is interacted several times in the same training step [28] to better mimic the behavior at test time. This scheme also promotes a segmentation error reduction on the location pointed by the user without degrading the result elsewhere.

Let $\mathcal{L}'$ be an adapted Huber loss [45]:

$$\mathcal{L}'\left(y,\hat{y},w,c,\delta\right)=\begin{cases}\frac{1}{2}\left(\left(y-\hat{y}\right)w\right)^2 c & \text{for } |y-\hat{y}|\le\delta\\ \delta\left(|y-\hat{y}|\,w\,c-\frac{1}{2}\delta\right) & \text{otherwise}\end{cases} \tag{5}$$

where $\delta = 7$ defines where $\mathcal{L}'$ is quadratic or linear, y and $\hat{y}$ are the reference standard and predictions, respectively, w is a weighting factor and c is the prediction confidence.

The network is trained to minimize, for each boundary, the translation and coordinate predictions as well as to maximize the confidence estimation on the correct regions:

$$\mathcal{L}_l=\mathcal{L}'\left(T_l,\hat{T}_l,w_{T,l},1,\delta\right)+\mathcal{L}'\left(b_l,\hat{b}_l,w_l,c_l,\delta\right)+\mathcal{L}'\left(1,c_l,w_l,1,1\right) \tag{6}$$

with the terms focusing on the error reduction of the translation map, the error of the boundary and on ensuring that the network does not learn the trivial solution of predicting zero confidence everywhere, respectively; $\hat{T}$, $\hat{b}$ and c are the predicted translation maps, boundary coordinates and confidence, respectively, and T and b the corresponding reference standard; $w_T = 1 - \frac{T}{\max(\text{abs}(T))}$ is a pixel-wise factor that prioritizes correct translation values near the true boundary; w is a weighting factor that prioritizes the correctness of the segmentation near the user clicks. We set $w_l = 10$ for the 51 A-scans around the user click and $w_l = 1$ elsewhere. Hyperparameters were selected based on the performance on the validation sets. Initial experiments also showed that very high values of w_l near the correction regions lead the network to overly focus on those locations, degrading the segmentation performance elsewhere.

III. EXPERIMENTS

In our experiments, user input is simulated similarly to the training scheme (Section II-D), i.e., for each layer the A-scan with the highest absolute error from the previous prediction is selected. Unless stated otherwise, each iteration has one simulated click per boundary before inferring a new result.

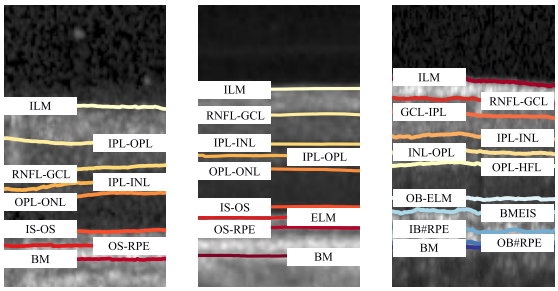
A. DATASETS

The proposed system is evaluated on two public datasets, HC/MS [46] and DME [6] and on an internal dataset consisting of scans of eyes with neovascular AMD (nAMD) (Fig. 4).

1) PUBLIC HC/MS DATASET

Contains scans from 14 healthy controls (HC) and 21 patients with multiple sclerosis (MS), acquired with a Spectralis (Heidelberg Engineering, Germany) scanner[1] [46]. Nine surfaces were manually delineated for all B-scans. MS subjects show mild thinning of retinal layers and possible microcystic

[1] https://iacl.ece.jhu.edu/index.php?title=Resources



(a) DME (b) HC/MS (c) nAMD

FIGURE 4. Retinal boundary nomenclature for the used datasets.

macular edema (MME), but the retinal layer structure of all the images is intact. Each OCT has 49 B-scans (496 $\times$1024 pixels) with an axial resolution of 3.87μm. Following [15], the last 6 HCs and 9 MS are used for training (735 B-scans) and the other 20 subjects are used for testing (979 B-scans, with one B-scan removed due to low contrast between the background and the retina). Within the 15 training subjects, the first HC and first two MS are used for validation. To ease comparison with other methods, we used the B-scans flattened by the Bruch's membrane (BM), with a size of 128 $\times$ 1024 pixels.

2) PUBLIC DME DATASET

Contains scans of 10 diabetic macular edema (DME) patients acquired with a Spectralis scanner, with 11 B-scans per patient annotated[2] [6]. The axial resolution is 3.87μm, and the B-scan size is 496 $\times$ 768 pixels. Reference segmentations by two annotators are available for eight retinal surfaces, as well as intraretinal cystoid fluid regions in the central portion of the images. Similarly to other studies, the last 55 B-scans are used for training the model, and the other 55 scans of highly pathological cases are used for testing. We only consider the annotations of the first reader. Due to the highly pathological nature of some of the examples, there are segments for which the manual annotations of some retinal layer boundaries do not exist, as the layers got disrupted by the fluid compartments. These segments were ignored both during model training and evaluation. To ease comparison with other methods, all B-scans were flattened by BM and have a size of 224 $\times$ 512 pixels.

3) INTERNAL nAMD DATASET

Contains 68 scans from 68 eyes of as many patients with nAMD, undergoing treatment at the eye clinic of Medical University of Vienna, Austria. The analysis adhered to the tenets of the Declaration of Helsinki, and was approved by the Ethics Committee of the Medical University of Vienna (EK Nr: 1246/2016). The scans were acquired with a Spectralis scanner covering a 6 $\times$ 6 $\times$ 2mm^3 volume, with an axial resolution of 3.87μm; the B-scan size is 496 $\times$ 768 $-$ 1024 pixels. Seven B-scans uniformly spaced across the volume had 12 layer boundaries manually annotated by experts (476 annotated B-scans). The 68 volumes were randomly split into training, validation and test sets with 282, 71 and 77 B-scans, respectively. All B-scans were flattened by BM and have a size of 256 $\times$ 512 pixels. Axial resolution was kept by cropping to the smallest possible power of 2 height, and width was subsampled by 2 due to memory constraints.

[2]https://people.duke.edu/ sf59/Chiu_BOE_2014_dataset.htm

B. COMPARISON TO NON-INTERACTIVE STATE-OF-THE-ART SEGMENTATION METHODS

We compare the initial segmentation of our approach (i.e. prior to user interaction) to other state-of-the-art methods to confirm that it is competitive even as a standalone application. We particularly aim at verifying that the initial segmentation error of our approach is in a similar range of other methods (not necessarily lower), thus showing that it can perform similarly to existing solutions while in addition advantageously allowing for subsequent segmentation correction. We focus on state-of-the-art deep learning segmentation networks proved to work on medical images. (i) ReLayNet [12], a U-Net-based model that performs retinal layer segmentation; (ii) MGU-Net [47], a graph convolutional neural network specifically designed for retinal layer segmentation; (iii) UNeXt [48], a Convolutional multilayer perceptron (MLP) based network dedicated to medical image segmentation and (iv) the method from He et al. [15], which uses a deep residual U-Net to predict anatomical coherent layer boundaries. Except for [15], for which the results are directly retrieved from the publication, all methods were trained having as base the available source code.

C. COMPARISON TO INTERACTIVE SEGMENTATION BASELINE

We compare our method with a fully region-based approach to assess whether our translation-based interaction scheme is harder to optimize than a pure region-level interaction in which both layer delineation and corrections are based on a pixel-wise segmentation. We are particularly interested in understating the influence of our proposed translation-based approach on the interactive segmentation performance. In an ablation fashion, we opt for comparing it to an end-to-end fully deep learning approach that uses the same backbone and overall interaction scheme and training pipeline as the proposed method. There, the user interacts with the network by selecting over- or under-segmented regions on an inferred pixel-wise segmentation of the same size as the input B-scan. We refer to this approach as our *baseline* model.

The baseline model uses the same backbone as ours. The output is a $H \times W \times (n+1)$ matrix with the same size as the input, and each channel represents the probability of a pixel belonging to one of the retinal layers or the background. We adapt our training scheme (Section II-D) with changes to the loss function and the user interaction encoding. Namely,

as loss, we use a multiclass intersection over union loss:

$$\mathcal{L}_{\text{IoU}} = \sum_{l}^{n+1} \frac{\sum (s_l \times \hat{s}_l)}{\sum (s_l + \hat{s}_l)} \tag{7}$$

where s is the reference segmentation, $\hat{s}$ is the pixel-wise prediction and l is each of the layers (and background) to segment. To promote focus on more complex regions of the image, all correct pixels scored above 0.7 are ignored [49].

Interaction is simulated by, for each of the $n+1$ labels, clicking the maximum of the distance transform of the largest under- or over-segmented region. This strategy ensures that the click stays inside the region of interest, independently of its shape. Clicks on over- and under-segmented regions are encoded with -1 and 1, respectively. We also provide as input channels the interaction maps smoothed with Gaussian kernels of spread 8 and 12. For evaluation, the final segmentation is the pixel-wise argmax of the predictions, and, for each layer, the boundaries are composed of the corresponding segmented pixels with the lowest depth per A-scan, following the definition shown in Fig. 4. The layer boundary coordinates for columns with no segmentations are linearly interpolated from the nearest existing neighbors.

D. INTERACTIVE SEGMENTATION EVALUATION

1) SEGMENTATION BEHAVIOR WITH INTERACTION

The behavior of the segmentation with user interaction is assessed with both simulated and manual annotations. Namely, we compare the evolution of the segmentation error for simulated interactions with and without the assistance of our system.

2) MANUAL ANNOTATIONS

Complementarily, we evaluate on 5 randomly selected cases from the DME dataset with real user interactions to verify that the system does not require perfectly positioned user-clicks. To assess the consistency of the results with different clicks and users, two different volunteers performed the experiment. This was performed on a custom graphical user interface that showed the user the B-scan, the system's prediction as full lines and the reference standard as dashed lines. The interface allows to correct, for each model prediction, any subset of boundaries with an arbitrary number of interaction points per layer. Interaction was performed by selecting the layer of interest with a right mouse click on the current prediction, providing the correct locations with left clicks, repeating the procedure for the desired layers and finally requesting a new set of locations by pressing *Enter*. For a fair comparison with the simulated interaction, the user was requested to provide for each boundary a point on a region where there was a need for a correction (not necessarily the highest error), predict a new result and repeat the process for 10 times.

3) EXTERNAL INITIALIZATION

Our method doesn't directly output a boundary, but instead translates it to new positions. Thus, it is expected that the system can also be used to correct segmentations from an external method. To test this hypothesis, we evaluate the performance of the model on the HC/MS test set using as initial segmentations the predictions from [15] instead of the argmin of the translation maps.

4) QUALITATIVE ASSESSMENT

We conduct qualitative experiments with manual interactions to explore the behavior of the segmentation in different scenarios.

5) CORRECTION LENGTH

In practice, corrections are commonly performed with both strokes and single clicks. Therefore, we simulate interactions with different stroke lengths to determine the length from which manual corrections are better than having the assistance of the system. For that, we assess the performance ratio between the non- and the system-assisted corrections, $\text{error}_{\text{user}}/\text{error}_{\text{system}}$. At each interaction, the A-scan with the highest segmentation error is selected, and used as the center of the correction stroke.

6) SEGMENTATION CONFIDENCE

The prediction confidence of the location of the boundary for each layer potentially allows to automatically identify cases of possible failure, which can then be revisited by the retinal specialist. With this in mind, we study the behavior of the segmentation error as a function of the predicted confidence and the number of clicks. In particular, for each simulated user interaction, we bin the predicted confidence in intervals of 0.25 and measure the segmentation error on the locations corresponding to each bin.

E. EVALUATION METRICS

The performance of the system is evaluated using the segmentation metrics mean absolute distance (MAD) and root-mean-square error (RMSE) to allow for the comparison with other state-of-the-art methods:

$$\text{MAD} = \frac{1}{W} \sum |y - \hat{y}| \tag{8}$$

$$\text{RMSE} = \sqrt{\frac{1}{W} \sum (y - \hat{y})^2} \tag{9}$$

where y and $\hat{y}$ are the reference and predicted coordinates of the boundaries, respectively. Experiments on the HC/MS dataset are reported using both MAD and RMSE and those on the DME dataset consider only MAD, following the performance metrics reported previously in the literature for those two datasets. For experiments with simulated interaction, we assess RMSE as a function of the number of clicks per layer. Each click is simulated by selecting the reference position of the A-scan with the highest current error.

F. TRAINING DETAILS

Training was performed with an Adam optimizer with learning rate 10^{-4} and a batch size of 4. Training data

TABLE 1. Mean absolute distance / root mean square error (standard deviation) in μm for the HC/MS dataset.

Boundary	ReLayNet [12]	UNeXt [48]	MGU-Net [47]	He [15]	Ours
ILM	$4.75_{3.02}$ / $36.24_{59.61}$	$7.26_{3.11}$ / $27.09_{33.74}$	$3.19_{1.76}$ / $17.98_{28.21}$	$\mathbf{2.41}_{0.81}$ / $\mathbf{3.10}_{1.35}$	$2.45_{0.98}$ / $3.11_{2.06}$
RNFL-GCL	$5.27_{4.98}$ / $39.44_{64.25}$	$5.66_{2.92}$ / $22.65_{34.09}$	$4.72_{2.29}$ / $19.07_{30.00}$	$\mathbf{2.96}_{1.70}$ / $\mathbf{4.02}_{2.59}$	$3.14_{1.54}$ / $4.27_{2.59}$
IPL-INL	$5.85_{5.36}$ / $41.21_{66.78}$	$5.10_{2.95}$ / $23.07_{37.64}$	$3.79_{1.33}$ / $14.04_{24.24}$	$\mathbf{2.87}_{1.69}$ / $\mathbf{3.80}_{2.74}$	$3.22_{1.34}$ / $4.09_{1.81}$
INL-OPL	$6.03_{5.34}$ / $42.00_{67.91}$	$5.35_{3.13}$ / $21.74_{39.53}$	$4.11_{1.09}$ / $14.16_{24.72}$	$\mathbf{3.19}_{1.49}$ / $4.12_{2.46}$	$3.20_{1.06}$ / $\mathbf{3.99}_{2.20}$
OPL-ONL	$5.37_{5.81}$ / $42.87_{69.47}$	$4.16_{1.20}$ / $9.25_{12.90}$	$3.38_{1.21}$ / $14.04_{25.34}$	$\mathbf{2.72}_{1.70}$ / $3.69_{2.75}$	$3.00_{1.32}$ / $\mathbf{3.59}_{1.72}$
ELM	$5.52_{6.43}$ / $44.51_{72.69}$	$3.77_{1.01}$ / $8.69_{11.39}$	$3.59_{1.77}$ / $15.01_{27.20}$	$\mathbf{2.65}_{1.14}$ / $3.29_{1.62}$	$2.86_{1.20}$ / $\mathbf{2.82}_{1.77}$
IS-OS	$4.42_{5.44}$ / $40.57_{67.45}$	$2.99_{0.96}$ / $6.14_{8.11}$	$3.27_{1.31}$ / $14.10_{26.85}$	$\mathbf{2.01}_{0.88}$ / $\mathbf{2.51}_{1.32}$	$3.76_{2.18}$ / $4.74_{2.82}$
OS-RPE	$4.97_{5.09}$ / $34.83_{62.04}$	$4.01_{1.60}$ / $6.94_{8.49}$	$3.86_{1.43}$ / $11.85_{21.46}$	$3.55_{1.73}$ / $4.34_{1.92}$	$\mathbf{3.21}_{2.18}$ / $\mathbf{3.85}_{2.41}$
BM	$4.07_{3.60}$ / $19.24_{33.21}$	$3.70_{1.37}$ / $7.54_{10.93}$	$3.55_{1.66}$ / $10.77_{16.58}$	$3.10_{2.21}$ / $\mathbf{3.66}_{2.28}$	$\mathbf{3.01}_{1.44}$ / $3.88_{2.30}$
Average	$5.14_{4.83}$ / $38.56_{62.81}$	$4.67_{1.60}$ / $16.85_{24.61}$	$3.72_{1.21}$ / $14.77_{23.87}$	$\mathbf{2.83}_{1.48}$ / $\mathbf{3.60}_{2.11}$	$3.09_{1.47}$ / $3.82_{2.6}$

was artificially augmented by performing, with 50% probability, random horizontal flips, vertical scaling (up to the point where ILM may occur on the upper region of the image), and vertical and horizontal translations of up to 10% of the image's height and width, respectively. On the HC/MS dataset and nAMD datasets, training was performed until there was no improvement on the validation loss for 100 epochs. For DME dataset, which has no validation set, training was stopped when there was no improvement on the non-augmented training set.

To incentivize the network to use the user input encoded in the prior maps, training started by, at each user interaction, simulating adjacent clicks (i.e. strokes) that cover 20% of the image width. This percentage exponentially decayed over the epochs so that after 50 epochs 1 pixel was selected per iteration. No post-processing of the predicted boundaries was applied. Experiments were performed using Python 3.8 and Keras 2.5.0, on a workstation with an Intel(R) Core(TM) i7-10700K CPU and NVIDIA RTX3080 GPU.

IV. RESULTS

Performance evaluation considered the central 98% portion of the scans, as for some cases our network does not properly handle the lack of information on the image boundaries. For the DME dataset, evaluation is performed on the same regions as [15], i.e. we ignore regions without reference standard and those where [6] does not report results.

A. COMPARISON TO STATE-OF-THE-ART SEGMENTATION METHODS

The performance of the method without user interaction (i.e. considering the initial segmentation, see Section II-A) on the public datasets is shown in Tables 1 and 2. Larger segmentation errors usually occurred on pathological regions of the images, in particular if the pathologies affect the definition of the layer boundaries, as well as regions where the signal/noise ratio is low, as depicted in Figs. 6b and 7b. From the tables, we can observe that our method achieved a performance similar or better than other state-of-the-art approaches on both datasets. For the HC/MS dataset, specifically, we had the best score on 4 out of the 9 boundaries (RMSE metric). The system is a viable alternative for existing segmentation methods, while additionally providing a framework that enables user interaction and real time segmentation correction.

TABLE 2. Mean absolute distance (standard deviation) in μm for the DME dataset. Bold indicates the best performance. Comparison with ReLayNet [12], UNeXt [48], MGU-Net [47] and He et al. [15]. Standard deviation for [15] is not reported in the original work.

Boundary	ReLayNet	UNeXt	MGU-Net	He	Ours
ILM	$15.72_{19.46}$	$7.89_{4.06}$	$4.78_{1.33}$	**4.51**	$4.78_{1.49}$
RNFL-GCL	$13.96_{8.70}$	$8.65_{1.52}$	$8.85_{2.75}$	**6.71**	$7.09_{3.25}$
IPL-INL	$12.47_{9.45}$	$9.37_{2.14}$	$8.64_{1.902}$	8.29	$\mathbf{7.14}_{4.52}$
INL-OPL	$13.36_{8.59}$	$9.50_{2.06}$	$11.09_{3.48}$	10.71	$8.78_{5.95}$
OPL-ONL	$12.49_{3.98}$	$10.44_{3.20}$	$11.18_{3.36}$	**9.88**	$10.53_{9.87}$
IS-OS	$8.38_{3.34}$	$5.22_{1.15}$	$4.83_{1.09}$	**4.41**	$5.09_{2.74}$
OS-RPE	$7.55_{1.56}$	$\mathbf{4.30}_{0.45}$	$4.83_{1.06}$	4.52	$5.39_{2.81}$
BM	$12.24_{8.99}$	$\mathbf{4.35}_{0.86}$	$5.74_{1.22}$	4.61	$5.13_{2.17}$
Average	$12.01_{6.29}$	$7.37_{0.93}$	$7.34_{1.08}$	**6.70**	$6.74_{4.10}$

B. COMPARISON TO INTERACTIVE BASELINE

A comparison of our solution with the interactive baseline as a function of the number of user clicks is shown in Fig. 5(a-c). The behavior of the baseline indicates that the model has learned to account for the spatial location of the user clicks to improve the segmentation. However, the average error is still much larger than that of the proposed solution. Indeed, pixel-wise segmentation approaches tend to under-perform on retinal boundary segmentation in OCT images as they do not properly encode anatomical priors such as layer ordering, allowing segmented layer to occur at unreasonable locations [15]. Furthermore, as shown in Figs. 5 a, b and c, the error reduction with interactions tends to be less stable. We hypothesize that this is due to the extra correction effort, since several clicks may be needed to fill/delete the same incorrectly segmented regions. Instead, our contour-focused approach allows to operate on the final predictions' coordinates in an end-to-end fashion, easing the addition of anatomical constraints that improve the local segmentation behavior. In a practical use-case, interaction with this type of approach is simpler for the user, as it only requires to click on the desired location without needing to additionally indicate if that region is under- or over-segmented. Note also that the proposed interactive interaction model does not introduce significant computational overhead, increasing the number of parameters of the baseline backbone from

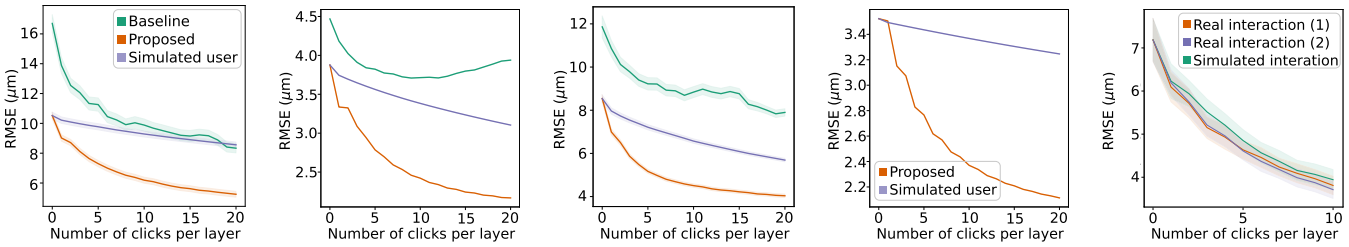


FIGURE 5. Average segmentation error (and 95% confidence interval) as function of the number of clicks per layer. (a, b and c): DME, HC/MS and internal (nAMD) datasets, respectively; (d) with initial segmentation from [15] for the HC/MS dataset; (e) performance on 5 random cases from the DME dataset for simulated and real user interactions using the proposed method.

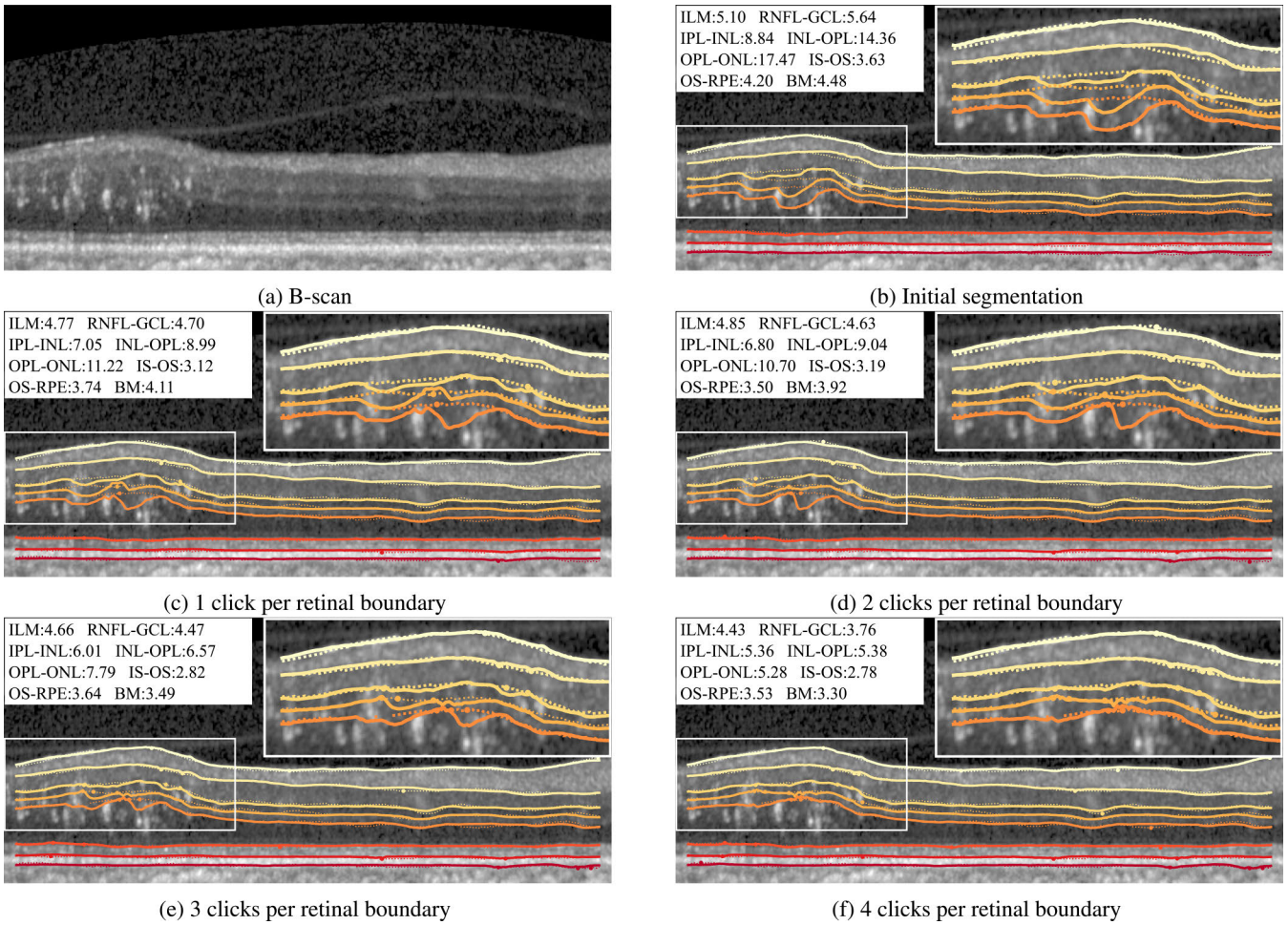


FIGURE 6. Behavior of the proposed interactive segmentation model on a case from the DME dataset without user interaction and after $1, \ldots, 4$ clicks per boundary. The displayed values are the root-mean-square error (μm) between the reference segmentation (dotted line) and prediction (full line). Colored circles are the clicks for the respective boundaries. Best viewed in color using the digital version of the manuscript.

45 166 208 to 45 215 830, i.e. a mere 0.1% parameter increase.

C. INTERACTIVE SEGMENTATION

1) SEGMENTATION BEHAVIOR WITH INTERACTION

The mean of the layer-wise average error for all test samples is depicted on Fig. 5, with Figs. 5a–c showing the performance of the model using a segmentation initialized with the argmin of the predicted inferred maps. Using our approach allows a better mitigation of existing segmentation errors in comparison to manual corrections. For instance, for the HC/MS dataset (Fig 5a) a single click per layer already achieves a performance better than the state-of-the-art [15], whereas with the manual correction more

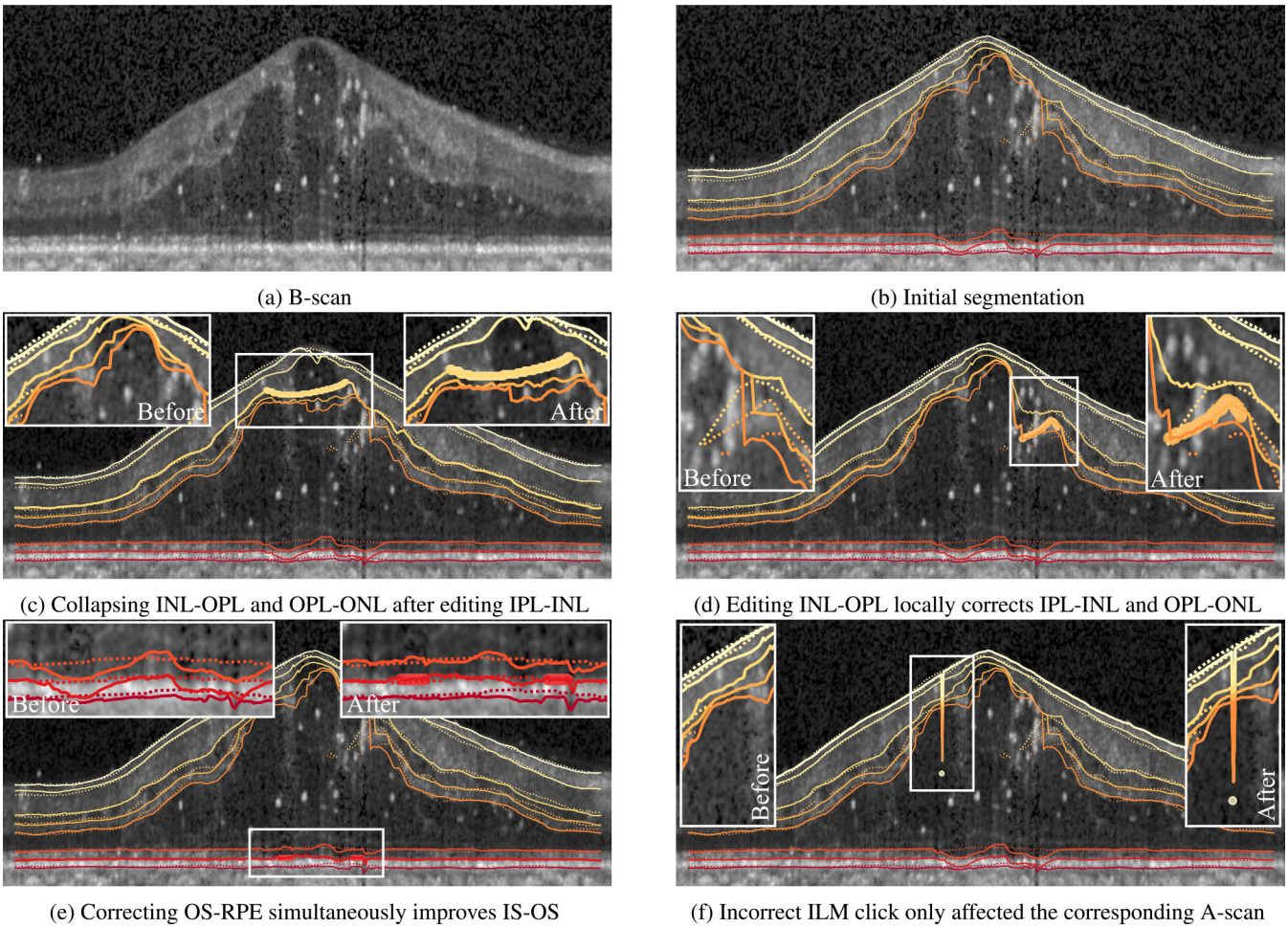


FIGURE 7. Segmentation behavior after different types of interaction (thick stroke lines) on a test case from the DME dataset. Best viewed in color using the digital version of the manuscript.

than 15 clicks per layer would be required. This is also quantitatively shown in Figs. 6b and 6c, where a single click per layer improves the segmentation in a pathological region. In particular, the INL-POL and OPL-ONL errors are reduced by more than 30%. In addition, the behavior of the network is similar across datasets with different types of retinal diseases. The low inference time per prediction iteration of approx. 0.25s/2.5s (GPU/CPU, independent of the number of interactions performed), further facilitates the translation of the system to the clinic.

2) MANUAL ANNOTATIONS

Fig. 5e shows the segmentation error for 5 cases from the DME dataset with real interactions assisted by our system in comparison to the simulated interaction. These results validate our findings using the simulated interactions. In particular, we show that the error reduction with both simulated and real interactions is very similar. Also, the similar performance of both manual annotations suggests that the system is robust to different types/styles of interactions. In fact, it appears that the simulated interaction, which follows the training scheme by always selecting the highest error coordinate, may not even be the best interaction strategy. Indeed, both real users were capable of achieving an overall better performance than the simulation. We believe that this is due to the nature of the interaction, since the real users had a tendency to interact with a larger portion of the image where errors occurred, regardless of whether these were indeed the locations with the largest errors. This provides more context to the network, thus allowing for more efficient corrections.

3) EXTERNAL INITIALIZATION

Fig. 5d shows the evolution of the segmentation error using as starting point the boundaries predicted by [15] for the HC/MS dataset. These results suggest that our method can be used to correct other segmentation models, with the behavior of the segmentation error being similar to our initialization.

4) QUALITATIVE ASSESSMENT

Fig. 6 depicts a potential use case of our method, where a poor segmentation on a pathological region is corrected, and Fig. 7 shows examples of different interactions and

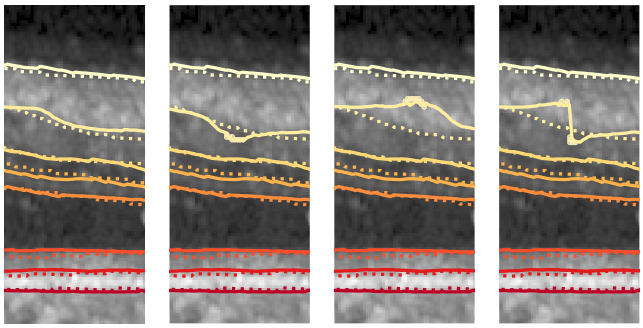


FIGURE 8. Behavior of the segmentation for the IPL-OPL layer for different user interactions (DME dataset). First image is the initial segmentation, followed by 3 different results after 3 different interactions. Dotted lines indicate the reference standard. Best viewed in color using the digital version of the manuscript.

the corresponding output on a highly pathological case. As shown, interacting with a single boundary affects neighboring segmentations. For example, on Fig. 7c, increasing the IPL-INL depth leads to a collapse of both INL-OPL and OPL-ONL while maintaining the relative distances between the 3 boundaries. This suggests that the network has learned not only the expected anatomical order of the layers, but also that layer thickness is a relevant parameter for correctly identifying the different boundaries. Likewise, Fig. 7d and 7e show scenarios where correcting one of the boundaries locally improved the quality of the neighboring segmentations. This behavior is generally desirable, allowing to reduce the workload of the specialist by reducing the total number of interactions required to correct the result. Finally, Fig. 7f depicts the prediction of the model after a single poorly located click. Unlike the cases shown in Fig. 6, where clicks made on plausible locations affect the correction of the adjacent portions of the boundaries, in this case the click only affects its respective column. This hints that the model identifies this noisy interaction as an outlier and minimizes the impact of the manual correction on the final result. Even though it would be better to completely ignore the click, the model was trained assuming that the user input would lead to a correction/improvement of the segmentation. Thus, changes on the interacted A-scan are always expected.

Fig. 8 illustrates how the system can account for user subjectivity. For simplicity, we force the non-interacted layers to maintain their initial position. Notice that none of the interactions matches the reference segmentation. In a scenario where user interaction was not properly accounted for, one would expect the model to always guide the boundaries to the correct location. Instead, the system is adjusting to the user's opinion, moving the boundary towards the requested locations.

5) CORRECTION LENGTH

Fig 9 presents the stroke lengths for which the deep learning assisted correction is better than the manual delineation. If all layers are to be corrected, the assisted correction is particularly efficient in comparison to the manual annotations

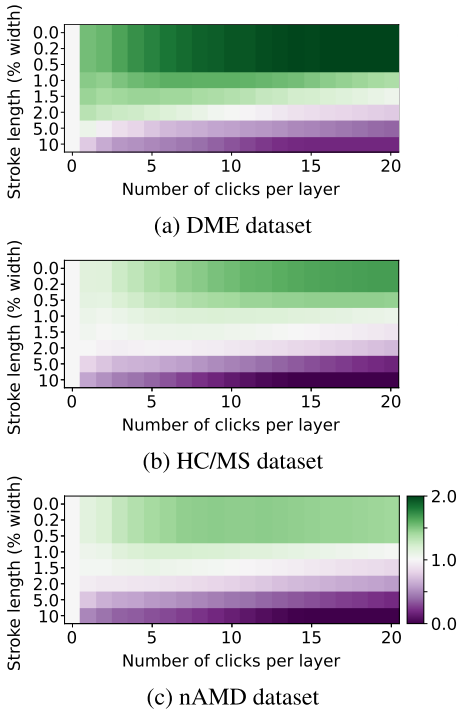


FIGURE 9. Ratio of the segmentation root-mean-square error between the simulated manual user interaction and the deep learning assisted interaction as function of the number of interactions and length of the correction stroke (percentage of image width). Green indicates that the deep learning is better than the manual user correction, and purple the opposite.

when very small (single clicks and small strokes) are provided. Specifically, the model stops actively contributing to a better correction when the input strokes are larger than approximately 2% of the images' width (approx. 10 to 20 pixels, depending on the dataset). This behaviour is to be expected, as the model was specifically trained to perform corrections based on clicks and not large strokes, with the latter being characteristic of fully manual corrections. Also, the assisted interaction allows to obtain smooth transitions between the corrected and neighbour regions of the boundary, resulting in a more natural segmentation than the manual correction, as exemplified in Fig. 7e, where the length of the stroke is smaller than the corrected depressions of the OS-RPE layer. Click-based interactions are of particular interest for regions where multiple boundaries have an initial poor estimation, as they significantly reduce the amount of manual effort required for correction (Fig. 6). Note how larger segmentation errors, in particular for the IPL-INL, INL-OPL and OPL-ONL (see Fig 6b, left half), are corrected with a few clicks (Fig. 6f). Internally, these clicks lead to updates mainly on the same neighborhood region of the translation map, allowing to maintain the quality of the segmentation on the remaining portions of the image.

6) SEGMENTATION CONFIDENCE

The average segmentation error as a function of the model's confidence and the number of clicks per layer is presented

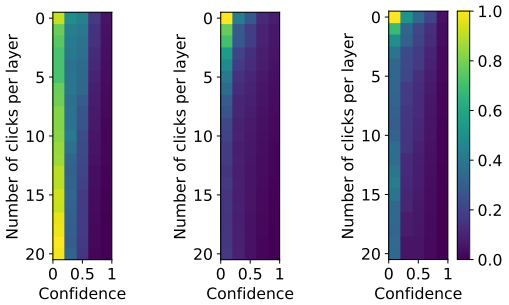


FIGURE 10. Max-min normalized average error as a function of the number of interactions and inferred confidence for the datasets in the study. Color codes the normalized error.

in Fig. 10. Results indicate that the model has learned to leverage the confidence as a weighting factor on the loss function to reduce the influence of high error locations during training. Indeed, for all datasets the upper left region of the plot has high error, which is to be expected as it corresponds to segmentation regions with low confidence and low number of interactions. Likewise, increasing the number of interactions and/or the predicted confidence results in lower errors. Note, however, that in the DME dataset (Fig. 10(a)), there is a set of cases for which increasing the number of clicks degrades the quality of the segmentation (first column of plot (a), corresponding to 0-0.25 confidence). These correspond to highly pathological images not meaningfully represented in the training data, leading to improper behavior during test time. Despite this, the confidence for those cases is still low, and thus are properly highlighted by the system as being likely erroneous regions requiring manual correction.

V. DISCUSSION AND CONCLUSION

This paper presents a novel interactive segmentation method for retinal boundary segmentation in OCT images. Our approach, based on predicting translation maps, allows obtaining an initial segmentation performance in line with other state-of-the-art methods (Tables 1 and 2) and, more importantly, to correct those segmentations. To the best of our knowledge, such end-to-end deep learning approach that allows for efficient and coherent retinal boundary segmentation while respecting anatomic constraints has never been explored in the literature before. Indeed, the system was shown capable of correcting layer boundary segmentations in scans covering three different retinal diseases with minimal annotation effort (Fig. 5).

We showed that a few user provided clicks in the regions with large error allow to substantially improve the initial segmentation (Fig. 6). Particularly, the sparse interaction encoding and the extra loss weight near the interaction regions promote a local segmentation correction without degrading the performance elsewhere. In addition, training to account for an expected layer ordering allows correcting multiple layers while interacting with just one, as exemplified for example in Fig. 7e. The system also learned to rely on the user input to produce different outputs according to the opinion of the annotator, thus accounting for the subjectivity inherent to the segmentation process (Fig. 8). Overall, this type of effective interaction is essential in clinical practice, where the time availability of retinal specialists is severely limited and the amount of work is high, hence fully manual correction to measure retinal layer thickness is not feasible. Furthermore, our system has an associated uncertainty estimation (Fig. 10), which can potentially be used for identifying regions of failure in an automated manner (e.g. via a threshold) and guide the user, further streamlining the segmentation correction process.

Of note, a segmentation proposed by a different system can be used as the starting point for interaction, because the correction method is based on the translation of existing boundaries instead of their direct prediction. This increases the potential of clinical applicability, where often segmentation software is distributed and locked together with the image acquisition hardware. As long as access to the initially predicted coordinates is possible, our approach allows training an interactive correction method that can be used by clinicians to refine segmentations according to their needs.

As a limitation, there is a risk that the method does not generalize to datasets that have a domain shift such as a novel pathology or a different acquisition setting, similarly to other deep learning approaches. In particular, in this study we opted by not studying the performance of the system in a multi-pathology setting (i.e. combining multiple datasets) as each dataset has a different number of layers annotated using distinct annotation/initialization schemes. This makes their combination and management of possible domain-shits non-trivial. Because of this, future efforts should focus on the collection of large uniform and highly representative datasets, as well as on the development of techniques that increase the robustness of these types of systems to outlier cases and domain shifts. Also, the system was trained assuming that user interactions are always meaningful, i.e. would always lead to a better segmentation. For extreme outlier clicks (see Fig. 7f), the segmentation behaviour is suboptimal. As a consequence, further efforts to increase the model robustness to such noisy interactions should be done. In addition, at this stage, the proposed system only predicts layers' positions from a single B-scan at a time, which may be imposing a performance limit compared to predictions from 3D volumes. With sufficient computational resources, however, this limitation would be solved as our algorithm should be directly extendable to 3D.

In conclusion, a novel interactive segmentation approach for retinal boundary segmentation in OCT images is presented. We demonstrate the effectiveness of the approach by conducting extensive quantitative and qualitative validations in both private and public datasets. In particular, user interaction allows the method to achieve the segmentation performance beyond the current state-of-the-art. Also, the simplicity of the interaction required to correct the

TABLE 3. Relevant parameters used for describing the proposed method.

Parameter	Description
l	Retinal layer
n	Number of retinal layers to segment
i	Current interaction
b_l	Coordinates of the boundary of retinal layer l
c_l	Prediction confidence of the boundary of retinal layer l
T^l	Translation map associated with retinal layer l
C_l	Predicted confidence map associated with retinal layer l
$\mathbf{B}$	Set of predicted layer coordinates
$\mathbf{C}$	Set of prediction confidences for the retinal layers
w_T	Pixel-wise weighting factor for the optimization of the translation maps
w	Weighting factor for the optimization of the boundary location near the user interaction
H	Input height
W	Input width

TABLE 4. Summary of relevant state-of-the-art approaches reviewed in this work. DL: deep learning; Gr: graph-based; RF: random fields; FzC: fuzzy clustering; R: region-focused method; C: contour-focused method. Coherent: whether the method has properly in account previous interactions. Confidence: whether the method provides a confidence metric associated with the output.

Work	Focus	Interactive	Fast	Coherent	Confidence
Ours (DL)	R+C	Y	Y	Y	Y
[17] (Gr)	C	Y	N	Y	N
[18] (Gr)	C	Y	N	Y	N
[20] (Gr+DL)	R+C	Y	Y	Y	Y
[21] (DL)	R	Y	N	N	N
[22] (DL)	R	Y	Y	N	N
[23] (DL)	R	Y	Y	N	N
[24] (DL)	R	Y	Y	N	N
[25] (DL)	R	Y	Y	Y	N
[27] (DL)	R	Y	Y	N	N
[23] (DL)	R	Y	Y	N	N
[28] (DL)	R	Y	Y	Y	N
[30] (Gr+DL)	R	Y	Y	Y	N
[31] (RF)	R	Y	Y	Y	N
[33] (DL)	R	Y	Y	N	N
[25] (DL)	R	Y	Y	Y	N
[35] (DL)	R	Y	Y	Y	N
[36] (DL)	R	Y	Y	Y	N
[37] (DL)	R	Y	Y	Y	Y
[39] (DL)	R+C	N	Y	N	N
[40] (DL)	R+C	Y	Y	N	N
[41] (FzC)	C	Y	Y	Y	N
[42] (DL)	C	Y	N	Y	N
[43] (DL)	C	Y	Y	Y	N
[12] (DL)	R	N	Y	N	N
[47] (DL)	R	N	Y	N	N
[48] (DL)	R	N	Y	N	N

segmentation makes it an attractive tool for clinical practice, and thus it is our hope that this line of work will enable efficient and accurate retinal layer thickness measurements from OCT images.

DATA AVAILABILITY

The Internal nAMD dataset used to train and evaluate our method cannot be shared at the current time due to data confidentiality agreements and privacy constraints. The public HC/MS dataset is available athttps://iacl.ece.jhu.edu/index.php?title=Resources. The public DME dataset is available at https://people.duke.edu/ sf59/Chiu_BOE_2014_dataset.htm.

APPENDIX A

See Table 3.

APPENDIX B

See Table 4.

REFERENCES

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Jan. 2017.

[2] S. K. Wagner, D. J. Fu, L. Faes, X. Liu, J. Huemer, H. Khalid, D. Ferraz, E. Korot, C. Kelly, K. Balaskas, A. K. Denniston, and P. A. Keane, "Insights into systemic disease through retinal imaging-based oculomics," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 6, Jan. 2020.

[3] U. Schmidt-Erfurth, S. Klimscha, S. M. Waldstein, and H. Bogunović, "A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration," *Eye*, vol. 31, no. 1, pp. 26–44, Jan. 2017.

[4] U. Schmidt-Erfurth, J. Garcia-Arumi, F. Bandello, K. Berg, U. Chakravarthy, B. S. Gerendas, J. Jonas, M. Larsen, R. Tadayoni, and A. Loewenstein, "Guidelines for the management of diabetic macular edema by the European society of retina specialists (EURETINA)," *Bulgarian Rev. Ophthalmol.*, vol. 66, no. 1, p. 3, Feb. 2022.

[5] S. Saidha, S. B. Syc, M. A. Ibrahim, C. Eckstein, C. V. Warner, S. K. Farrell, J. D. Oakley, M. K. Durbin, S. A. Meyer, L. J. Balcer, E. M. Frohman, J. M. Rosenzweig, S. D. Newsome, J. N. Ratchford, Q. D. Nguyen, and P. A. Calabresi, "Primary retinal pathology in multiple sclerosis as detected by optical coherence tomography," *Brain*, vol. 134, no. 2, pp. 518–533, 2011.

[6] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomed. Opt. Exp.*, vol. 6, no. 4, p. 1172, 2015.

[7] S. P. K. Karri, D. Chakraborthi, and J. Chatterjee, "Learning layer-specific edges for segmenting retinal layers with large deformations," *Biomed. Opt. Exp.*, vol. 7, no. 7, p. 2888, 2016.

[8] A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunović, "Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context," *Biomed. Opt. Exp.*, vol. 8, no. 3, p. 1874, 2017.

[9] F. Rathke, M. Desana, and C. Schnörr, "Locally adaptive probabilistic models for global segmentation of pathological OCT scans," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10433, 2017, pp. 177–184.

[10] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomed. Opt. Exp.*, vol. 8, no. 5, p. 2732, 2017.

[11] J. Novosel, K. A. Vermeer, J. H. de Jong, Z. Wang, and L. J. van Vliet, "Joint segmentation of retinal layers and focal lesions in 3-D OCT data of topologically disrupted retinas," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1276–1286, Jun. 2017.

[12] A. Guha Roy, S. Conjeti, S. Phani Krishna Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional network," 2017, *arXiv:1704.02161*.

[13] S. Borkovkina, A. Camino, W. Janpongsri, M. V. Sarunic, and Y. Jian, "Real-time retinal layer segmentation of OCT volumes with GPU accelerated inferencing using a compressed, low-latency neural network," *Biomed. Opt. Exp.*, vol. 11, no. 7, p. 3968, 2020.

[14] J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, "Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search," *Biomed. Opt. Exp.*, vol. 9, no. 11, p. 5759, 2018.

[15] Y. He, A. Carass, Y. Liu, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Structured layer surface segmentation for retina OCT using fully convolutional regression networks," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101856, doi: 10.1016/j.media.2020.101856.

[16] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101693.

[17] M. Montazerin, Z. Sajjadifar, E. Khalili Pour, H. Riazi-Esfahani, T. Mahmoudi, H. Rabbani, H. Movahedian, A. Dehghani, M. Akhlaghi, and R. Kafieh, "Livelayer: A semi-automatic software program for segmentation of layers and diabetic macular edema in optical coherence tomography images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Jul. 2021, doi: 10.1038/s41598-021-92713-y.

[18] K. Lee, H. Zhang, A. Wahle, M. D. Abràmoff, and M. Sonka, "Multi-layer 3D simultaneous retinal OCT layer segmentation: Just-enough interaction for routine clinical use," in *Proc. VipIMAGE*, J. M. R. S. Tavares and R. M. Natal Jorge, Eds. Cham, Switzerland: Springer, 1007, pp. 862–871.

[19] H. Zhang, K. Lee, Z. Chen, S. Kashyap, and M. Sonka, *LOGISMOS-JEI: Segmentation Using Optimal Graph Search and Just-Enough Interaction*. Amsterdam, The Netherlands: Elsevier, 2019, doi: 10.1016/B978-0-12-816176-0.00016-8.

[20] S. Gorgi Zadeh, M. W. M. Wintergerst, and T. Schultz, "Intelligent interaction and uncertainty visualization for efficient drusen and retinal layer segmentation in optical coherence tomography," *Comput. Graph.*, vol. 83, pp. 51–61, Oct. 2019, doi: 10.1016/j.cag.2019.07.001.

[21] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[22] G. Aresta, C. Jacobs, T. Araújo, A. Cunha, I. Ramos, B. van Ginneken, and A. Campilho, "iW-Net: An automatic and minimalistic interactive lung nodule segmentation deep network," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Aug. 2019.

[23] N. A. Koohbanani, M. Jahanifar, N. Z. Tajadin, and N. Rajpoot, "NuClick: A deep learning framework for interactive segmentation of microscopic images," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101771.

[24] S. Zhang, J. H. Liew, Y. Wei, S. Wei, and Y. Zhao, "Interactive object segmentation with inside-outside guidance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2020, pp. 12231–12241.

[25] X. Liao, W. Li, Q. Xu, X. Wang, B. Jin, X. Zhang, Y. Wang, and Y. Zhang, "Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9391–9399.

[26] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin, "F-BRS: Rethinking backpropagating refinement for interactive segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8620–8629.

[27] W.-D. Jang and C.-S. Kim, "Interactive image segmentation via backpropagating refinement scheme," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5292–5301.

[28] M. Forte, B. Price, S. Cohen, N. Xu, and F. Pitié, "Getting to 99% accuracy in interactive segmentation," 2020, *arXiv:2003.07932*.

[29] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102062.

[30] X. Luo, G. Wang, T. Song, J. Zhang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102102.

[31] R. Li and X. Chen, "An efficient interactive multi-label segmentation tool for 2D and 3D medical images using fully connected conditional random field," *Comput. Methods Programs Biomed.*, vol. 213, Jan. 2022, Art. no. 106534.

[32] X. Li, M. Xia, J. Jiao, S. Zhou, C. Chang, Y. Wang, and Y. Guo, "HAL-IA: A hybrid active learning framework using interactive annotation for medical image segmentation," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102862.

[33] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2746–2754.

[34] X. Zhao, L. Zhang, and H. Lu, "Automatic polyp segmentation via multi-scale subtraction network," 2021, *arXiv:2108.05082*.

[35] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "DeepIGeoS: A deep interactive geodesic framework for medical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1559–1572, Jul. 2019.

[36] I. Mikhailov, B. Chauveau, N. Bourdel, and A. Bartoli, "A deep learning-based interactive medical image segmentation framework with sequential memory," *Comput. Methods Programs Biomed.*, vol. 245, Jan. 2024, Art. no. 108038. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260724000348

[37] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.

[38] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, no. 1, p. 654, Jan. 2024. [Online]. Available: https://www.nature.com/articles/s41467-024-44824-z

[39] H. He, X. Li, Y. Yang, G. Cheng, Y. Tong, L. Weng, Z. Lin, and S. Xiang, "BoundarySqueeze: Image segmentation as boundary squeezing," 2021, *arXiv:2105.11668*.

[40] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, "Deep snake for real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8530–8539.

[41] C. Militello, L. Rundo, M. Dimarco, A. Orlando, V. Conti, R. Woitek, I. D'Angelo, T. V. Bartolotta, and G. Russo, "Semi-automated and interactive segmentation of contrast-enhancing masses on breast DCE-MRI using spatial fuzzy clustering," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103113.

[42] L. Castrejón, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4485–4493.

[43] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, "Fast interactive object annotation with curve-GCN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5252–5261.

[44] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[45] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[46] Y. He, A. Carass, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls," *Data Brief*, vol. 22, pp. 601–604, Feb. 2019, doi: 10.1016/j.dib.2018.12.073.

[47] J. Li, P. Jin, J. Zhu, H. Zou, X. Xu, M. Tang, M. Zhou, Y. Gan, J. He, Y. Ling, and Y. Su, "Multi-scale GCN-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary OCT images," *Biomed. Opt. Exp.*, vol. 12, no. 4, p. 2204, Apr. 2021. [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-12-4-2204

[48] J. M. J. Valanarasu and V. M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022* (Lecture Notes in Computer Science), L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham, Switzerland: Springer, 2022, pp. 23–33.

[49] T. Pissas, C. S. Ravasio, L. Da Cruz, and C. Bergeles, "Effective semantic segmentation in cataract surgery: What matters most?" in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 1–8.

GUILHERME ARESTA received the M.Sc. degree in bioengineering–biomedical engineering from the Faculty of Engineering, University of Porto, Porto, Portugal, in 2016, and the Ph.D. degree in electrical and computer engineering from the Faculty of Engineering, University of Porto, focused on the development of deep learning approaches for medical image analysis and especially lung cancer screening. He has been collaborating as a Researcher with the Biomedical Imaging Laboratory, Center for Biomedical Engineering Research (C-BER), Institute for Systems and Computer Engineering, Technology and Science (INESC-TEC), since 2014. Since 2021, he has been with the Christian Doppler Laboratory for Artificial Intelligence in Retina as a Postdoctoral Researcher.

TERESA ARAÚJO received the M.Sc. degree in bioengineering and biomedical engineering and the Ph.D. degree in electrical and computer engineering from the Faculty of the Engineering, University of Porto (FEUP), Porto, Portugal, in 2016 and 2021, respectively. For her thesis, she worked on diabetic retinopathy grading in color eye fundus images. Since 2014, she has been collaborating as a Researcher with the Biomedical Imaging Laboratory, Center for Biomedical Engineering Research (C-BER), Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto. In 2021, she joined the Christian Doppler Laboratory for Artificial Intelligence in Retina, as a Postdoctoral Researcher.

BOTOND FAZEKAS received the B.Sc. degree in software engineering and the M.Sc. degree in computational intelligence from Vienna University of Technology. He is currently pursuing the Ph.D. degree. During and after his graduation, he worked for several years as a freelancer contractor for different Austrian and German companies, leading large-scale software development projects as a software architect, besides working on research projects focusing on predictive maintenance for locomotives and high-speed computer vision solutions for train pantograph certification and safety evaluation. He joined the research group as a research engineer, in 2020.

JULIA MAI received the M.D. degree from Ludwig-Maximilians-University (LMU) Munich, Germany in May 2018. She is currently pursuing the Ph.D. degree in medical imaging with the Medical University of Vienna. She joined the Laboratory for Ophthalmic Image Analysis (OPTIMA), in December 2019. Besides the Ph.D. study, she works in the outpatient clinic with the Department of Ophthalmology and Optometry to gain clinical knowledge and experience. She currently focuses her research on OCT biomarkers in atrophic age-related macular degeneration (AMD).

URSULA SCHMIDT-ERFURTH has completed a medical training with Ludwig Maximilians-University, Munich, Germany. She began a career as a Research Fellow with the Harvard Medical School, Boston, where she pioneered in the development of photodynamic therapy, a breakthrough in retinal therapy. She is a Professor and the Chair of the Department of Ophthalmology, University Eye Hospital, Vienna, Austria, one of the largest academic institutions in ophthalmology in Europe. She is also an Adjunct Professor of ophthalmology with Northwestern University, Chicago. She has founded the Vienna Study Center (VSC), which serves as the principal investigator site for multi-center clinical trials; and the Vienna Reading Center (VRC), an institution for digital retinal imaging performing image analysis for clinical trials connected with over 400 clinical centers worldwide. In 2013, she founded the OPTIMA Project, an interdisciplinary laboratory, including a computer scientists, a physicists, and a retina experts introducing artificial intelligence into ophthalmic image analysis. She is an inventor on several patents on retinal imaging and therapeutic methods. She is the author of over 360 original articles. Her clinical activities include both surgical and medical retina. Her scientific research focuses on the development of novel diagnostic techniques, e.g., retinal imaging and novel treatment strategies, such as intravitreal pharmacotherapy.

She is a Founding Member of the Medical Imaging Cluster (MIC) and the speaker elect. She is a Board Member of EURETINA, the European Retina Specialists, and served as the president of the society. She is a member of many professional organizations, including the Association for Research in Vision and Ophthalmology (ARVO), the Macula Society, the Retina Society, the Gonin Club, the European Academy of Ophthalmologists, and the American Academy of Ophthalmology. She has received numerous grants and awards, such as the Research Award by the German Ophthalmological Society, the Achievement Award of the American Academy of Ophthalmology, the Roger Johnson Award by the University of Washington, the Donald Gass Award of the Retina Society, and the Donald Gass Medal of the Macula Society. She serves as a reviewer for the European Commission, the Welcome Trust, the German Research Foundation (DFG), and other funding organizations. She is active in the board of the Austrian Research Foundation, OEFG, and Vicepresident of the European Forum Alpbach, an interdisciplinary platform for science, politics, business, and culture, established in 1945, addressing the relevant socio-political questions of modern time. She serves on the editorial board for the *British Journal of Ophthalmology* (BJO), *Investigative Ophthalmology and Visual Science* (IOVS), *Acta Ophthalmologica*, and *European Journal of Ophthalmology*.

HRVOJE BOGUNOVIĆ received the B.Sc. and M.Sc. degrees in computer science from the University of Zagreb, Croatia, and the Ph.D. degree from Universitat Pompeu Fabra (UPF), Barcelona, Spain, in 2012. For his thesis, he worked on medical image segmentation and shape analysis applied to blood vessels of the brain imaged with various angiographic modalities.

After graduation, he did a postdoctoral research with Iowa Institute for Biomedical Imaging (IIBI), University of Iowa, USA, specializing in medical image analysis for applications in ophthalmology. He moved to the Medical University of Vienna, Austria, in 2015, to join Christian Doppler Laboratory for Ophthalmic Image Analysis, as a Lead of computational imaging methods and machine learning. In 2018, he was a tenure-track Faculty Member with the Medical University of Vienna. His general research interests include medical image analysis, imaging genetics, computer vision, and machine learning, with applications in healthcare. He is particularly interested in machine learning for predicting disease progression and in knowledge discovery from large clinical longitudinal imaging and genetic datasets.