2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Locate before Segment: Topology-guided Retinal Layer Segmentation in Optical Coherence Tomography Images

Ye Lu, Yutian Shen, Xiaohan Xing, and Max Q.-H. Meng*, *Fellow, IEEE*

***Abstract*— Optical Coherence Tomography (OCT) is a non-invasive imaging technique that is instrumental in retinal disease diagnosis and treatment. Segmentation of retinal layers in OCT is an essential step, but remains challenging for common pixel-wise segmentation methods usually fail to obtain the correct layer topology. To tackle this challenge, we propose a novel Locate-to-Segment (L2S) framework to provide a layer region location guidance for pixel-wise labeling learning so as to obtain better segmentation with the correct topology and smooth boundaries. Specifically, a Structured Boundary Regression Network (SBRNet) is devised to first predict the surface positions. For effective learning on normal-size images, we design two regression branches to regress the top surface and eight layer widths separately in SBRNet to locate each layer region with absolutely correct orderings. Then, we take the prediction of SBRNet as an additional input for a common pixel-wise segmentation network to provide the guidance of correct topology. In this L2S manner, our framework takes merits of regression-based methods and pixel-wise labeling-based methods to obtain accurate segmentation with the correct topology and smooth continuous boundaries. Experimental results on a public retinal OCT dataset demonstrate the effectiveness of our method, outperforming state-of-the-art segmentation methods with the highest average Dice score of 90.29% and the lowest average MAD score of 0.782.**

I. INTRODUCTION

Optical Coherence Tomography (OCT) [1] is a non-invasive imaging technique capable of high-resolution visualization of human retina, which has been instrumental in eye clinics [2]. In OCT images, the retina presents as multiple layers in a strict order, providing micro-structure information near the level of histopathology, as is shown in Fig. 1 (a). Then, the thicknesses and shapes of different layers serve as important indicators for various diseases [3]. For instance, multiple sclerosis (MS) can result in thinning of the retinal nerve fiber layer (RNFL) [4]. Therefore, accurate retinal layer segmentation in OCT is an essential step in the eye disease diagnosis and treatment, as in Fig. 1 (b). However, due to the low contrast and inevitable noises in OCT images, it remains a challenge to automatically identify the multiple layers with correct anatomical orderings and smooth boundaries in retinal OCT.

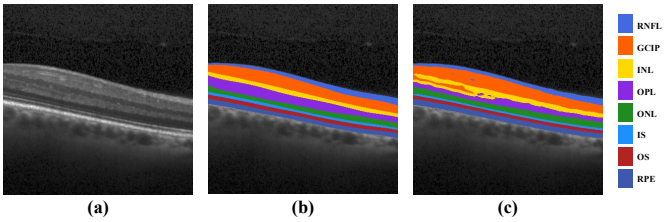

Fig. 1. (a) Retinal OCT image sample. (b) Segmentation ground truth of eight retinal layers, including retinal nerve fiber layer (RNFL), ganglion cell and inner plexiform complex (GCIP), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), Inner photoreceptor segments (IS), Outer photoreceptor segments (OS) and retinal pigment epithelium (RPE). (c) Segmentation result of a conventional pixel-wise labeling model [13], presenting the incorrect topology and discontinuous boundaries.

With the rapid development of artificial intelligence, there have been various automatic retinal OCT segmentation approaches explored based on deep learning models. The conventional and popular methods are with fully convolutional networks and their extensions to perform pixel-wise classification for the whole image [5]–[8]. Although they can achieve some decent results, these methods with pure pixel-wise labeling maps cannot always obtain correct layer topology and continuous boundaries, as shown in Fig. 1 (c), for not considering the global features.

More recently, some researchers have proposed to directly regress the layer boundaries for segmentation [9]–[11], considering the unique characteristics of retinal OCT data. Unlike pixel-wise labeling, the regression methods can globally locate each layer region to guarantee correct anatomical orderings explicitly. However, due to the indirect logical relationship between the boundary positions and image pixels, the performance of those regression networks is limited. In the original literature, the small-size inputs that only keep the regions of interest (ROIs) with leaving out large backgrounds are adopted to decrease the regression difficulty, which requires additional annotations and is infeasible in practical applications. Moreover, we experimentally observe that normal-size inputs with large-scale backgrounds included always lead to poor performance through these methods.

Consider manual segmentation in practice, where clinicians usually first browse the whole image to locate each layer region and then focus on the detailed boundary pixels to delineate the multiple surfaces. Inspired by this work-

Y. Lu and Y. Shen are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China (e-mail: luyyy@link.cuhk.edu.hk).

X. Xing is with the Department of Electrical Engineering, The City University of Hong Kong, Hong Kong SAR, China.

Max Q.-H. Meng is with Shenzhen Key Laboratory of Robotics Perception and Intelligence, and the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China, on leave from the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, and also with the Shenzhen Research Institute of The Chinese University of Hong Kong, Shenzhen 518055, China (e-mail: max.meng@ieee.org).

* Corresponding author.

ing mechanism, we propose a novel Structured Boundary Regression Network (SBRNet) that first locates the layer regions as a topology guidance for conventional pixel-wise segmentation models to tackle the aforementioned issues. Specifically, we devise two branches including top surface regression and layer width regression in a fully convolutional network for SBRNet to locate the regions of each layer in an absolutely correct orderings. Then, a pixel-wise labeling network takes the prediction of SBRNet as an additional input with the original image to make predictions under the guidance of correct topology. With such Locate-to-Segment (L2S) framework, our method can take merits of boundary regression-based models and pixel-wise labeling-based models to obtain more accurate segmentation with the correct topology and smooth boundaries. The main contributions are summarized as follows:

- We propose a L2S framework that first locates the layer regions in correct orderings by SBRNet to provide a topology guidance for a pixel-wise segmentation network, so as to achieve more accurate performance with the correct topology.
- We develop a fully convolutional SBRNet with two regression branches to predict the top surface position and eight layer widths separately, so that effective regression learning is guaranteed for normal-size images.
- Effectiveness of our proposed SBRNet is validated on a public dataset [12]. Extensive experiments demonstrate that our L2S framework outperforms the state-of-the-art segmentation algorithms.

II. RELATED WORK

With the introduction of convolutional neural networks (CNN), deep learning-based algorithms have achieved impressive breakthroughs in image semantic segmentation, which also trigger substantial studies in medical image processing. Among them, UNet [13] is a common and effective model for medical image segmentation, consisting of a symmetric encoder-decoder framework with skip-connections. More recently, its extensions ResUNet [14] and TransUNet [15] have gained popularity and made some improvements. For retinal OCT image segmentation, several deep learning-based methods have been developed and gained promising performance in the past five years. In 2017, Fang et al. [5] presented a CNN-GS framework that combined convolutional neural networks with graph search methods for automatic retinal layer segmentation. Roy et al. [6] proposed a fully convolutional network (FCN) [16] with a joint loss for segmentation of retinal layers and fluid regions. Cao et al. [7] introduced regression of signed distance maps into the pixel-wise segmentation task to provide extra topological supervision for retinal layer segmentation. Li et al. [8] developed a multi-scale graph convolutional network (GCN)-assisted two-stage framework to simultaneously label retinal layers and the optic disc.

However, all these aforementioned approaches perform segmentation in a pixel-wise labeling manner, which cannot always guarantee the correct topology and continuous boundaries. To tackle this problem, regression-based models have been recently proposed to directly predict layer boundary positions to obtain topology-correct segmentation results. Shah et al. [9] firstly developed a deep network to segment three retinal surfaces by directly predict the surface positions. He et al. [10] adopted a second regression network to compute multiple surface positions from the predicted probability maps with prior shape information. He et al. [11] utilized a FCN and a column-wise soft-argmax method to predict the multiple surface positions for segmentation. Although achieved some success, these explicit regression models are performed on small-size images only including ROIs, whose application is limited and infeasible in practice. In our work, we exploit a regression-based model on normal-size images to provide a topology guidance for pixel-wise segmentation models, taking advantage of both kinds of methods to achieve better performance.

III. METHOD

The architecture of our proposed framework is illustrated in Fig. 2. For a given image x, it is first fed into our SBRNet to locate the positions of each retinal layer. Particularly, the SBRNet consists of a shared feature extractor and two different heads, composing two branches for top surface regression and layer width regression separately to jointly predict the boundary position matrices $\widehat{S}$ and $\widehat{W}$. Then the predicted position matrices are transformed to a pixel-wise segmentation mask $\widehat{Y}_r$, which locates the layer regions in a correct anatomical ordering. Next, taking the transformed location mask as a guidance, $\widehat{Y}_r$ and the original image x are separately processed and then fed into a conventional pixel-wise labeling network to achieve a more accurate segmentation result $\widehat{Y}$ with correct topology and finer boundaries. These two networks are trained independently in an end-to-end way. In the following, we will illustrate the details of the top surface regression and layer width regression modules in SBRNet as well as the overall L2S framework training.

A. Top Surface Regression

Since retinal layers in OCT images present a fixed structure with strict anatomical orderings, it is intuitive and more efficient to explicitly regress the multiple surface positions for layer segmentation to guarantee the correct topology and continuous boundaries. However, since the retina has eight layers, it is not easy to perfectly predict all boundaries in normal-size images for the class imbalance problem and limited regression ability. For instance, the performance of those regression models in existing studies [10] and [11] suffers dramatic degradation when trained with images of a larger size than their adopted sizes. To tackle this issue, our SBRNet aim to globally locate the regions of each layer in normal-size images. We propose to predict the top surface position and eight layer widths separately to reduce the regression difficulty and guarantee the correct layer orderings.

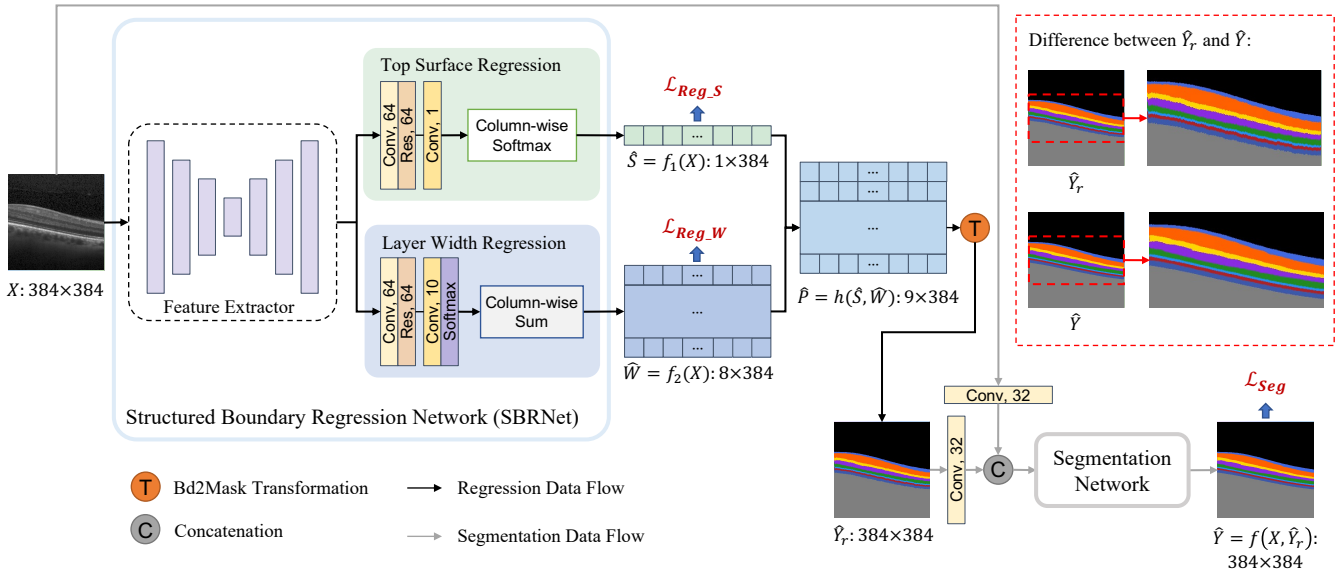

Fig. 2. Overview of our L2S framework. The SBRNet contains a shared feature extractor and two regression heads to predict top surface position matrix $\widehat{S}$ and layer width matrix $\widehat{W}$ separately, which are later converted into the position matrix $\widehat{P}$ for nine boundaries. Then a pixel-wise segmentation mask $\widehat{Y}_r$ with correct layer topology transformed from $\widehat{P}$ is taken as another input for a conventional segmentation network as a topology guidance to achieve final result $\widehat{Y}$. The segmentation boundaries of $\widehat{Y}_r$ and $\widehat{Y}$ are amplified to highlight the difference to illustrate the effectiveness of our L2S framework.

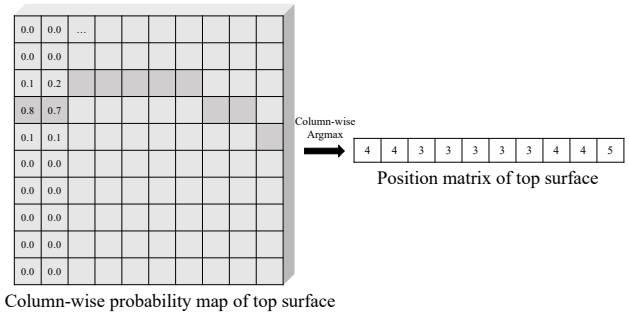

Fig. 3. Toy example for top surface regression module.

The top surface of the retina decides the whole foreground location and general curve directions, which carries more significance in prediction. Since it is easier to regress the position of single boundary than those of all boundaries, we propose to predict the position matrix of the top surface with an individual branch. However, in normal-size images without cropping the ROIs for retina, there exists severe class imbalance problem as the retinal layers are usually thin. Conventional regression methods with fully connected layers to directly predict the surface positions are unable to achieve satisfying performance and even with large numbers of parameters. Therefore, we devise a fully convolutional network to build a location probability map to predict the positions via column-wise classification, as illustrated in Fig. 3. In particular, considering the top surface lies across the image and intersects each pixel column only once, the ground truth should indicate one at the intersection position and zero at the others in each column. Then viewing the pixel rows as various class channels, we perform classification in the column direction to select the row index with the highest probability for the surface position in each column.

Formally, consider a $M \times N$ input image $X = \{x_{i,j} | 1 \leq i \leq M, 1 \leq j \leq N\}$, where $x_{i,j}$ is the pixel value in the i^{th} row and j^{th} column, with the corresponding top surface ground truth $S = [s_1, \cdots, s_n]$, where s_j $(1 \leq j \leq N)$ is the row index at which the surface intersects the j^{th} column and $1 \leq s_j \leq M$. Denoting the shared feature extractor with the top surface regression head as $f_1(\cdot)$, the supervised loss for top surface regression is formulated as follows:

$$\mathcal{L}_{Reg\_S}(X, S) = \frac{1}{N} \sum_{j=1}^{N} \mathcal{L}_{ce}(f_1(X)_{\cdot,j}, s_j), \qquad (1)$$

where $\mathcal{L}_{ce}(\cdot)$ is the cross-entropy loss [17] for classification. In this manner, our model can utilize global information to absolutely guarantee the surface topology while requiring less learning parameters.

B. Layer Width Regression.

With the top surface position predicted, we propose to simultaneously regress eight layer widths to get positions of the remaining surfaces. In most existing regression-based methods, directly predicting the multiple layer surface positions is adopted, which can cause incorrect layer orderings. In contrast, our layer width regression module estimates the relative distances between the adjacent surfaces, so that the non-negative outputs can guarantee the topological correctness of layer orderings.

Meanwhile, conventional fully connected regression requires much more computations and performs the calculation in a black box manner, where the relevance between the image pixels and boundary positions is not clear, leading to

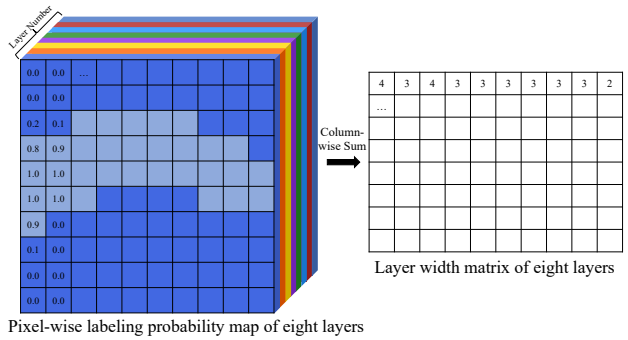

Fig. 4. Toy example for layer width regression module.

the limited regression performance. To solve this problem, we employ the pixel-wise labeling probability maps to model the prediction of multiple layer widths in a fully convolutional manner. Considering the pixel-wise segmentation map for each layer class indicates one in the corresponding layer region and zero outside the region, ideally its sum in each column is equivalent to the layer width. Following this insight, we adopt a convolutional head with the feature extractor to output segmentation probability maps with 10 class channels, including eight retinal layers and two backgrounds. Then, we compute the column-wise sums in the eight layer channels to obtain the corresponding layer widths, as presented in Fig. 4.

Formally, for X with input size $M \times N$, we have the ground truth of layer width matrix $W = \{w_{j,k} | 1 \leq j \leq N, 1 \leq k \leq C{+}2\}$, where $w_{j,k}$ is the width in the j_{th} column for the k_{th} layer and C is the number of layer classes. Here $C = 8$ for our retinal OCT data. Denoting the shared feature extractor with the layer width regression head as $f_2(\cdot)$, the supervised loss for layer width regression is defined as:

$$\mathcal{L}_{Reg\_W}(X, W) = \frac{1}{(C+2) \cdot N} \sum_{k=1}^{C+2} \sum_{j=1}^{N} \left\| \sum_{i=1}^{M} f_2(X)_{i,j,k} - w_{j,k} \right\|^2. \quad (2)$$

In this manner, we build a direct logical relationship between the image pixels and the surface positions, making the boundary position regression feasible and effective for normal-size images.

C. Locate-to-Segment Framework Training

With the predicted top surface position and all layer widths, we can obtain the position matrix of all layers by adding the layer widths column-wise to the top surface position in a correct layer order. Specifically, given the predicted top surface position matrix $\widehat{S} \in \mathbb{R}^{1 \times N}$ and layer width matrix $\widehat{W} \in \mathbb{R}^{C \times N}$, where $C = 8$ is the layer class number, the position matrix of all layer boundaries can be computed as follows:

$$\widehat{P} = h(\widehat{S}, \widehat{W}) = cumsum(\begin{bmatrix} \widehat{S} \\ \widehat{W} \end{bmatrix}, \dim = 0), \quad (3)$$

where $\widehat{P} \in \mathbb{R}^{(C+1) \times N}$ and $cumsum(\cdot)$ is a column-wise cumulative sum.

Then we can transform $\widehat{P}$ into a pixel-wise segmentation mask $\widehat{Y}_r$ with absolutely correct topology and continuous boundaries. Though it can effectively locate the regions of each retinal layer, the transformed segmentation mask inevitably has jagged and coarse boundaries. Consider conventional pixel-wise labeling-based methods can focus more on local and fine features but have trouble guaranteeing the correct topology. We propose a locate-to-segmentation (L2S) mechanism, firstly locating the layer regions and then performing pixel-wise segmentation locally to combine two types of methods to make up their shortcomings for better segmentation. Particularly, we take the prediction of SBRNet as another input for a second pixel-wise labeling network to provide a topology guidance for finer segmentation.

As shown in Fig. 2, for any effective pixel-wise segmentation network, the original image x and the transformed segmentation mask $\widehat{Y}_r$ from the regression result are first fed into two convolutional blocks separately. Then, the obtained low-level feature maps are concatenated and passed to the original segmentation network for pixel-wise classification. Denoting the segmentation ground truth as $Y = \{y_{i,j} | 0 \leq y_{i,j} \leq C{+}1, 1 \leq i \leq M, 1 \leq j \leq N\}$, where C layer classes and 2 background classes are counted, the supervised loss for pixel-wise segmentation is formulated as:

$$\mathcal{L}_{Seg}(X, \widehat{Y}_r, Y) = \mathcal{L}_{ce}(f(X, \widehat{Y}_r), Y) + \mathcal{L}_{dice}(f(X, \widehat{Y}_r), Y), \quad (4)$$

where $f(\cdot)$ refers to the segmentation network, $\mathcal{L}_{ce}(\cdot)$ is the cross-entropy loss [17] and $\mathcal{L}_{dice}(\cdot)$ is the dice loss [18].

In our L2S framework, the two networks are trained individually. Firstly, the SBR-Net is trained with the total regression loss $\mathcal{L}_{Reg} = \mathcal{L}_{Reg\_S} + \mathcal{L}_{Reg\_W}$ until convergence. Then, the regression results are transformed to pixel-wise segmentation masks and input with original images into a pixel-wise segmentation network, which is trained with the segmentation loss $\mathcal{L}_{Seg}$ until convergence. As for inference, the test image also follows this procedure to obtain the final segmentation results.

IV. EXPERIMENTS

A. Dataset

We evaluated the proposed framework on a public retinal OCT dataset [12], which consists of retinal OCT volumes of 35 subjects with 14 healthy controls (HC) and 21 patients of multiple sclerosis (MS). Each OCT volume contains 49 B-scans with the size of 496×1024. Each B-scan has 8 retinal layers delineated, including RNFL, GCIP, INL, OPL, ONL, IS, OS and RPE, as presented in Fig. 1 (b). For data preparation, we cropped each B-scan height-wise from 496 into 384 pixels and then divided it width-wise into two non-overlapping slices, to obtain 3430 retinal OCT scans with the size of 384×512. Then, we split them into the training set of 2940 scans from random 30 subjects, the validation set of 196 scans from random 2 subjects, and the testing set of 294 scans from the left 3 subjects. For pre-processing,

TABLE I

COMPARISON WITH SOTAS FOR RETINAL LAYER SEGMENTATION ON DICE SCORE (%) ↑. THE RETINAL LAYERS INCLUDE RNFL, GCIP, INL, OPL, ONL, IS, OS AND RPE. BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED AND UNDERLINED.

Methods	RNFL	GCIP	INL	OPL	ONL	IS	OS	RPE	AVG
UNet [13]	91.90	92.73	85.78	88.09	93.39	85.97	87.52	88.60	89.25
ResUNet [14]	92.79	93.64	86.44	89.04	93.82	86.18	87.38	89.36	89.83
TransUNet [15]	92.60	93.83	86.71	89.42	**94.15**	85.18	86.89	89.50	89.78
RelayNet [6]	91.33	93.25	85.62	88.47	93.39	86.50	87.65	89.34	89.44
MTF [7]	92.66	93.54	86.32	89.34	93.74	86.41	87.83	89.03	89.86
MGU-Net [8]	92.49	93.53	85.95	88.51	93.14	86.01	87.30	89.03	89.49
CRNet [11]	89.01	90.58	81.12	85.15	90.37	78.53	82.79	87.05	85.57
L2S (Ours)	**93.15**	**94.15**	**86.98**	**89.92**	93.99	**86.73**	**87.85**	**89.55**	**90.29**

TABLE II

COMPARISON WITH SOTAS FOR PREDICTED BOUNDARY POSITIONS ON MAD SCORE ↓. THE LAYER BOUNDARIES INCLUDE B1 TO B9 FROM THE TOP TO BOTTOM. BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED AND UNDERLINED.

Methods	B1	B2	B3	B4	B5	B6	B7	B8	B9	AVG
UNet [13]	1.110	1.535	1.356	1.455	1.381	1.049	1.000	1.410	1.629	1.325
ResUNet [14]	0.806	1.103	1.014	1.069	1.071	0.719	0.802	1.071	1.256	0.990
TransUNet [15]	0.667	0.944	0.820	0.945	**0.804**	0.567	0.690	**0.854**	1.119	0.823
RelayNet [6]	6.298	6.299	6.367	6.341	6.540	6.281	6.380	6.595	6.901	6.445
MTF [7]	2.787	3.087	3.087	3.111	3.130	2.888	2.889	3.108	3.459	3.061
MGU-Net [8]	0.664	0.981	0.846	0.969	0.985	0.643	0.596	0.960	1.118	0.862
CRNet [11]	0.957	1.440	1.229	1.310	1.205	1.030	0.801	1.297	1.172	1.160
L2S (Ours)	**0.630**	**0.846**	**0.814**	**0.819**	0.838	**0.558**	**0.579**	0.874	**1.082**	**0.782**

the training data were normalized, randomly cropped into 384×384 pixels and horizontally flipped with a probability of 0.5. The validation and test images were normalized and cropped into 384×384 at the center.

B. Implementation Details

For the network backbone, we employed a U-shaped convolutional network as the shared feature extractor in SBRNet, which contains four residual encoders and decoders. The residual blocks in the feature extractor and two regression heads are similar to that proposed in [14]. As for the segmentation network, any effective pixel-wise labeling-based network can be adopted here, where the input channel should be changed as 64. In our work, we employed the ResUNet [14] as the segmentation network in our basic L2S framework.

The whole framework was implemented with PyTorch using one NVIDIA RTX 2080 Ti. The SBRNet was trained form scratch using Adam optimizer with a step-decay learning rate, initialized with $1e^{-4}$ and decayed by 0.5 every 1000 iterations. The network was totally trained for 5000 iterations. For the second segmentation network, we trained it with a step-decay learning rate for 2000 iterations, initialized with $1e^{-4}$ and decayed by 0.5 every 250 iterations. The parameters of SBRNet were frozen in this process. The batch sizes for the two networks were both set as 4. We made validation every 10 iterations to select the best model for testing. As for the evaluation metrics, we adopted the Dice score on the pixel-wise segmentation masks and the mean absolute distance (MAD) on the boundary positions, which was computed by summing up the corresponding layer maps in each column.

C. Comparison with State-of-the-art Methods

We conducted a series of experiments to compare our framework with state-of-the-art (SOTA) models, including popular medical segmentation baselines (i.e., UNet [13], ResUNet [14] and TransUNet [15]) as well as current algorithms specially developed for retinal OCT segmentation (i.e., RelayNet [6], MTF [7] and MGU-Net [8]). Moreover, CRNet [11] was also employed, which is a latest and feasible boundary regression-based model for retinal layer segmentation. As illustrated in IV-B, we adopted the ResUNet as the pixel-wise segmentation network for our main L2S framework. For a fair comparison, all models were trained with the same experimental settings, while the hyper-parameters were set following their original work.

Table I lists the Dice scores for the eight retinal layers and the average of the above-mentioned methods. We can observe that our proposed framework achieved the best performance on most of the retinal layers, with the highest average score of 90.29%. Notably, the state-of-the-art surface regression method had a much worse performance of 85.57% in Dice compared with other baselines, which illustrates its limited regression ability when trained on normal-size images with severe class imbalance. Table II lists the MAD results for nine layer boundaries (denoted as B1 to B9 from the top to bottom) and the average. Although most baselines obtained decent scores in Dice, several pixel-wise segmentation methods (i.e., RelayNet and MTF) demonstrated poor performance in terms of MAD with average scores larger than 3. We surprisingly find that CRNet with

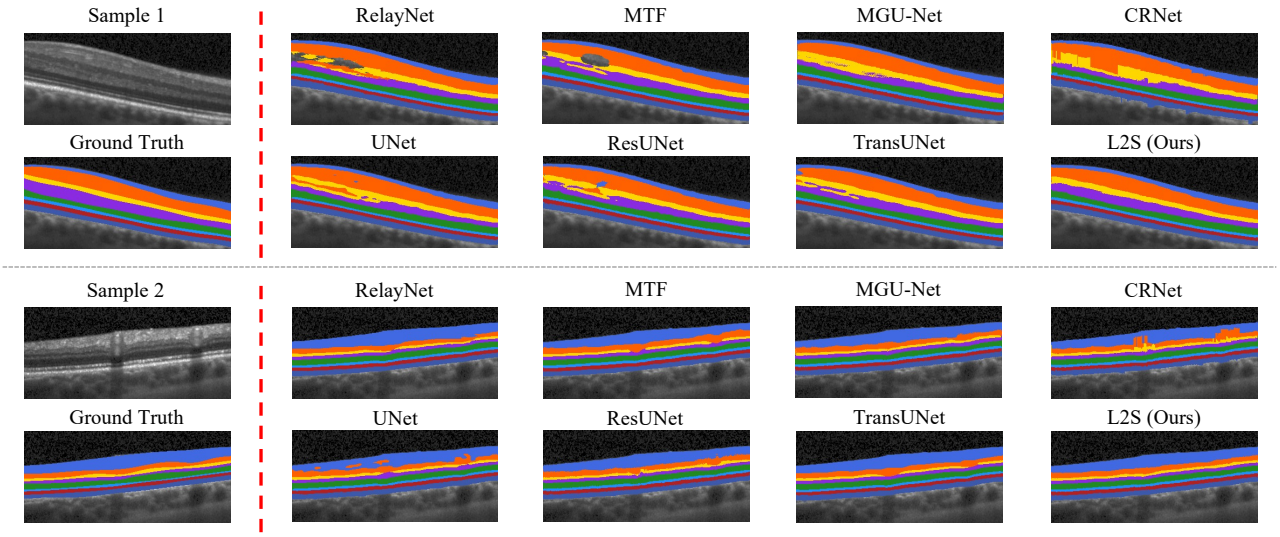

Fig. 5. Examples of retinal layer segmentation results. Due to space limit, we only present the regions of retinal layers for those predictions.

TABLE III

ABLATION STUDY RESULTS ON DICE SCORE (%) ↑. "S", "R" AND "L2S" REFER TO PIXEL-WISE SEGMENTATION, BOUNDARY REGRESSION AND LOCATE-TO-SEGMENTATION TYPES RESPECTIVELY. BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED AND UNDERLINED.

Types	Methods	RNFL	GCIP	INL	OPL	ONL	IS	OS	RPE	AVG
S	UNet [13]	91.90	92.73	85.78	88.09	93.39	85.97	87.52	88.60	89.25
	ResUNet [14]	92.79	93.64	86.44	89.04	93.82	86.18	87.38	89.36	89.83
	TransUNet [15]	92.60	93.83	86.71	89.42	**94.15**	85.18	86.89	89.50	89.78
R	SBRNet	92.58	93.52	84.28	87.57	91.13	79.07	83.48	87.04	87.33
L2S (Ours)	SBR-UNet	<u>93.16</u>	<u>94.18</u>	<u>87.04</u>	<u>89.68</u>	93.75	<u>86.62</u>	87.63	88.55	90.08
	SBR-ResUNet	93.15	94.15	86.98	**89.92**	<u>93.99</u>	**86.73**	<u>87.85</u>	<u>89.55</u>	<u>90.29</u>
	SBR-TransUNet	**93.38**	**94.33**	**87.22**	89.56	93.86	86.54	**88.12**	**89.88**	**90.36**

the worst dice performance even had a lower average MAD score, showing the advantage of boundary regression. On the contrary, our model again achieved the lowest boundary errors with an average MAD score of 0.782. These results prove the superiority of our L2S framework, for taking merits of boundary regression and pixel-wise segmentation. Fig. 5 visualizes the results on two samples with our framework and SOTAs. It shows that our method has the best performance with separable layers and continuous, smooth boundaries.

D. Ablation Analysis

To evaluate the effectiveness of our proposed framework, we adopted UNet, ResUNet and TransUNet as the segmentation network in our L2S framework, constituting three L2S models denoted as SBR-UNet, SBR-ResUNet and SBR-TransUNet respectively. Meanwhile, we compare them with the original pixel-wise segmentation networks (UNet, ResUNet and TransUNet), denoted as type "S", and our novel regression-based SBRNet, denoted as type "R". The Dice scores for eight retinal layers and the average are presented in Table III. Our SBRNet obtained a much higher average score 87.33% than CRNet with 85.57%, illustrating the effectiveness of our proposed boundary regression network for normal-size images. Meanwhile, regardless of the segmentation network employed, the three L2S models all achieved the average Dice scores higher than 90%, outperforming their original pixel-wise segmentation models. It demonstrates the superiority and robustness of our "locate before segment" mechanism, which can effectively boost the performance of pixel-wise segmentation networks.

V. CONCLUSIONS

In this work, we propose a novel L2S framework for retinal layer segmentation in OCT images of normal sizes. The main idea is to locate layer regions with correct orderings before pixel-wise segmentation to guarantee the correct topology. Our proposed SBRNet can predict surface positions in correct orderings by regressing top surface position and eight layer widths separately. Then, the regression result provides a topology guidance for the conventional pixel-wise segmentation network to obtain more accurate segmentation with correct topology. Experimental results on a public retinal OCT dataset demonstrate the effectiveness and superiority of our proposed method.

ACKNOWLEDGMENT

The work is supported by National Key R&D program of China with Grant No. 2019YFB1312400 and Hong Kong RGC CRF grant C4063-18G.

REFERENCES

[1] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito *et al.*, "Optical coherence tomography," *science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[2] J. G. Fujimoto, C. Pitris, S. A. Boppart, and M. E. Brezinski, "Optical coherence tomography: an emerging technology for biomedical imaging and optical biopsy," *Neoplasia*, vol. 2, no. 1-2, pp. 9–25, 2000.

[3] A. M. Zysk, F. T. Nguyen, A. L. Oldenburg, D. L. Marks, and S. A. Boppart, "Optical coherence tomography: a review of clinical development from bench to bedside," *Journal of biomedical optics*, vol. 12, no. 5, p. 051403, 2007.

[4] M. Pekala, N. Joshi, T. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, "Oct segmentation via deep learning: A review of recent work," in *ACCV*. Springer, 2018, pp. 316–322.

[5] 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 Express*, vol. 8, no. 5, pp. 2732–2744, 2017.

[6] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical optics express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[7] J. Cao, X. Liu, Y. Zhang, and M. Wang, "A multi-task framework for topology-guaranteed retinal layer segmentation in oct images," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 3142–3147.

[8] J. Li, P. Jin, J. Zhu, H. Zou, X. Xu, M. Tang, M. Zhou, Y. Gan, J. He, Y. Ling *et al.*, "Multi-scale gcn-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary oct images," *Biomedical Optics Express*, vol. 12, no. 4, pp. 2204–2220, 2021.

[9] A. Shah, L. Zhou, M. D. Abrámoff, and X. Wu, "Multiple surface segmentation using convolution neural nets: application to retinal layer segmentation in oct images," *Biomedical optics express*, vol. 9, no. 9, pp. 4509–4526, 2018.

[10] Y. He, A. Carass, Y. Liu, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and Prince, "Deep learning based topology guaranteed surface and mme segmentation of multiple sclerosis subjects from retinal oct," *Biomedical optics express*, vol. 10, no. 10, pp. 5042–5058, 2019.

[11] 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," *Medical image analysis*, vol. 68, p. 101856, 2021.

[12] 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 in brief*, vol. 22, pp. 601–604, 2019.

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

[14] L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[15] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

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

[17] S. Mannor, D. Peleg, and R. Rubinstein, "The cross entropy method for classification," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 561–568.

[18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.