RETINAL LAYER SEGMENTATION FROM OCT IMAGES USING 2D-3D HYBRID NETWORK WITH MULTI-SCALE LOSS AND REFINEMENT MODULE

Tsubasa Konno[⋆], *Takahiro Ninomiya*[†], *Kanta Miura*[⋆], *Koichi Ito*[⋆], *Noriko Himori*[†], *Parmanand Sharma*[†], *Toru Nakazawa*[†], *and Takafumi Aoki*[⋆]

[⋆] Graduate School of Information Sciences, Tohoku University, Japan.
[†] Department of Ophthalmology, Graduate School of Medicine, Tohoku University, Japan.

ABSTRACT

We propose a method of segmenting retinal layers from optical coherence tomography (OCT) images for the diagnosis. The proposed method estimates the pixel-wise labels of each retinal layer and each layer surface position using convolutional neural network (CNN). We introduce CNN to a multi-scale loss and a refinement module to improve the accuracy of pixel-wise labels and layer surface position. Through experiments using a public OCT image dataset, we demonstrate that the proposed method exhibits higher accuracy of segmenting retinal layers than the state-of-the-art methods.

***Index Terms*—** retinal layer, optical coherence tomography, segmentation, CNN

1. INTRODUCTION

Optical coherence tomography (OCT), which can noninvasively obtain high-resolution three-dimensional images of the retina, is widely used for diagnosis in ophthalmology. The retinal layers need to be segmented from OCT images to diagnose glaucoma and age-related macular degeneration (AMD), which changes the thickness of retina. An automatic and accurate segmentation method is required since manual annotation of retinal layers is time-consuming and labor-intensive.

Retinal Layer segmentation network (ReLayNet) is a pioneering method using convolutional neural network (CNN) for retinal layer segmentation [1]. ReLayNet is an encoder-decoder model like U-Net [2] that assigns pixel-wise labels to retinal layers. Pixel-wise labeling does not always produce continuous and smooth retinal layer surfaces, and does not guarantee the order of the retinal layers. To address the above problems, the methods have been proposed to detect the boundaries of the retinal layer simultaneously with pixel-wise labels by dividing the final layer of 2D CNN into two branches [3, 4]. The boundaries of the retinal layers can be used to obtain continuous and smooth retinal layer surfaces and to guarantee the order of the retinal layers. The use of 3D CNN is suitable for considering 3D features of OCT data since OCT data is acquired as 3D data of the retina. On the other hand, the resolution anisotropy and vertical misalignment of a set of 2D images consisting of OCT data prevent direct use of 3D CNN. To consider 3D features of OCT data, Liu et al. proposed simultaneous alignment and surface regression (SASR) using the 2D-3D hybrid network [5]. SASR consists of a 2D encoder to extract features from 2D images, a 3D decoder to align a set of 2D images, and a 3D decoder to obtain boundaries of the retinal layer and pixel-wise labels. SASR estimates boundaries of the retinal layer and pixel-wise labeling by dividing the final layer of 3D CNN into two branches as well as 2D CNN-based methods [3, 4], resulting in no direct impact on each estimation. A method has been proposed to detect the boundaries of the retinal layer after pixel-wise labeling [6], while the accuracy of pixel-wise labeling influences the accuracy of retinal layer boundary detection.

In this paper, we propose a 2D-3D hybrid network with multi-scale loss and refinement module to improve the segmentation accuracy of retinal layers. By introducing multi-scale loss, each layer of 3D decoder is trained to reduce the loss of pixel-wise loss and boundaries of retinal layers. In the refinement module, pixel-wise labels and boundaries of the retinal layers obtained from 3D decoder are combined in the channel direction, and both are estimated by the convolution layers and the fully-connected layer to improve the accuracy of both estimations. Through the experiments using the public OCT dataset [7], we demonstrate the effectiveness of the proposed method compared with the conventional methods.

2. METHODS

Fig. 1 shows an overview of the proposed method. The proposed method consists of SASR [5] with multi-scale loss and refinement modules. First, 2D encoder is used to extract features from OCT volume data. Next, the displacement field, which indicates the displacement between 2D images, is obtained from the features output by each layer of 2D encoder using 3D decoder. The features obtained by 2D encoder and the displacement field obtained by 3D decoder are input to spatial transformer module (STM) [8] to cancel the vertical misalignment in the feature maps. Then, the corrected feature maps are input to 3D decoder to obtain layer maps and surface distributions. The layer map represents pixel-wise labels of a background and each layer. The surface distribution represents the existence probability of boundaries at each pixel. The surface position is the height of the boundary in each column, which is calculated from the surface distribution. In the proposed method, a module for calculating multi-scale loss is connected to 3D decoder. For each decoder block of 3D decoder, we add two branches that perform 1×1 convolution, outputs layer map, surface distribution, and surface position at each resolution of the feature map. By training to minimize these losses, this module improves the accuracy of layer map and surface distribution output by 3D decoder. In the proposed method, a refinement module is also added after 3D decoder. This module refines layer map and surface distribution in light of their characteristics. We describe the details of the refinement module and multi-scale loss in the following.

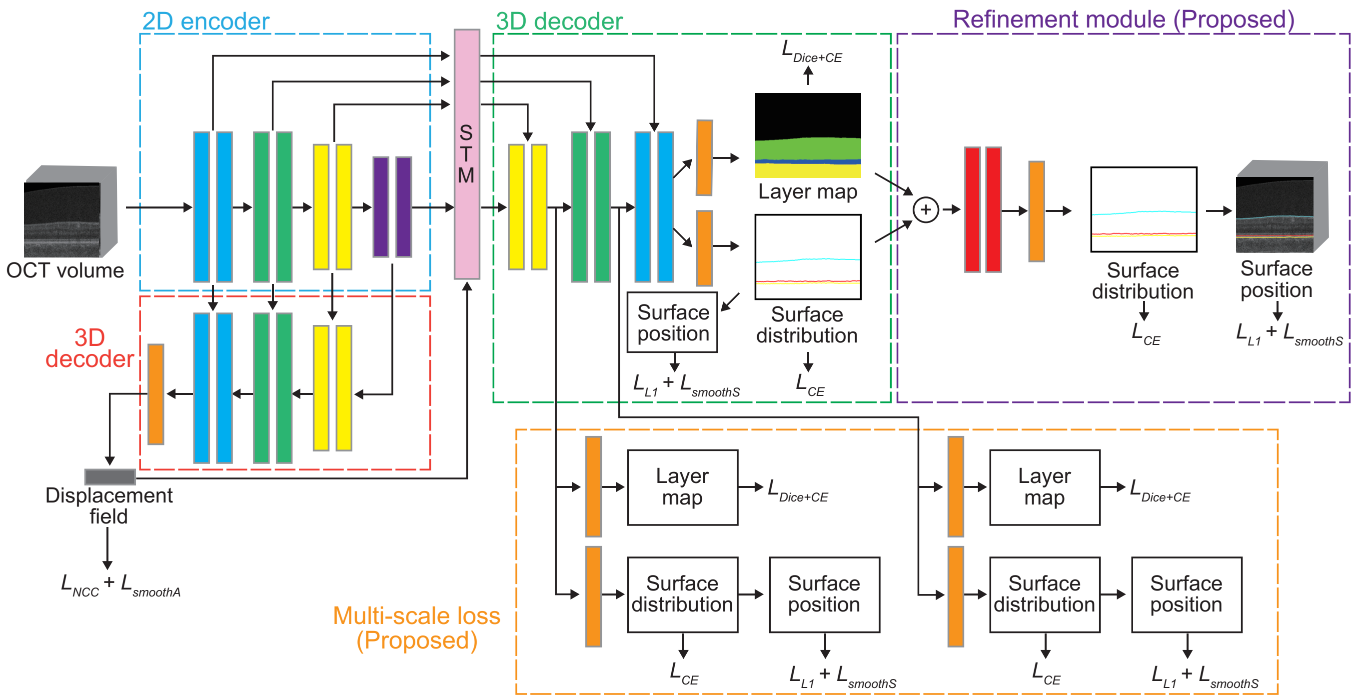

Fig. 1. Overview of the proposed method consisting of SASR [5] with multi-scale loss and refinement module.

2.1. Refinement Module

This module improves the estimation accuracy of surface distribution and surface position by considering the features of both layer map and surface distribution. The layer map and surface distribution output from the 3D decoder are combined in the channel direction and input to this module. This module consists of two convolution blocks and a 1×1 convolution. The convolution block consists of 3D convolution layer, batch normalization layer, and ReLU, where the number of channels of 3D convolution is 32. Then, the refined surface distribution is obtained and the surface position is calculated based on it. The loss functions are the cross entropy loss for the surface distribution and the L1 loss and smoothS loss for the surface position.

2.2. Multi-Scale Loss Function

We improve the accuracy of layer map and surface distribution output by 3D decoder using multi-scale loss. We obtain the layer map, surface distribution, and surface position from the features extracted by each decoder block of 3D decoder, and train the loss for each of them. For each of the three decoder blocks of 3D decoder, we calculate the loss function at 1/4, 1/2, and 1 resolution, respectively. The loss functions are Dice loss and cross entropy loss for layer map, cross entropy loss for surface distribution, and L1 loss and smoothS loss for surface position. Note that when the resolution is 1/4 and 1/2, we changed the following in calculating the loss functions. The scaling down of the ground truth of the surface distribution is set to 1 if the area before scaling down contains 1, otherwise it is set to 0. The scale down of the ground truth of the layer map is the average of the pixel values of the region before scaling down. Since a resolution reduction by 1/2 results in half the misalignment between the estimated boundary and the ground truth, the inverse of the scale factor is used as the weight for the L1 loss. The smoothS loss calculates the vertical displacement so that the boundary estimates are continuous. The inverse of the scale factor is used as the weights for smoothS loss since the vertical displacement becomes smaller as the resolution decreases.

3. EXPERIMENTS

In this section, we evaluate the accuracy of the proposed method for retinal layer segmentation using the public dataset.

3.1. Dataset

In this paper, we use the SD-OCT public dataset [7]. The dataset consists of OCT data from 269 AMD patients and 115 normal subjects. Since only the region centered on the fovea is manually labeled, $512 \times 400 \times 40$ voxels around the fovea are extracted and used in the experiment. To reduce memory consumption, the retinal B-scan image is flattened with the estimated Bruch's membrane (BM) using the intensity gradient method [9] as in FCBR [3] and SASR [5]. Since BM estimation may fail due to AMD, the BM estimate is shifted to be 64 pixels from the bottom of the image.

3.2. Experimental Condition

In this experiment, we evaluate accuracy by 5-hold cross-validation without changing the ratio of AMD and normal subjects. In each subset, 70% is used as training data, 10% as validation data, and 20% as test data. There are three manually annotated ground truth labels: the inner aspect of the inner limiting membrane (ILM), the inner aspect of the retinal pigment epithelium drusen complex (IRPE), and the outer aspect of Bruch's membrane (OBM). OCT is cropped into $320 \times 400 \times 40$-voxel patches as input. Adam is used as the optimizer, and training is performed for 120 epochs with a mini-batch

Table 1. Mean absolute distance [μm] of boundary position of retinal layers.

Methods	FCBR [3] (experiment in [5])	SASR [5] (experiment in [5])	SASR [5] (our experiment)	Proposed w/o multi-scale loss	Proposed before refinement	Proposed after refinement
ILM (AMD)	1.73 ± 2.50	1.76 ± 2.39	1.281 ± 1.163	1.276 ± 1.228	$\mathbf{1.231 \pm 1.250}$	1.301 ± 1.266
ILM (Normal)	1.24 ± 0.51	1.26 ± 0.47	1.166 ± 0.839	$\mathbf{1.144 \pm 0.900}$	1.148 ± 0.887	1.154 ± 0.876
IRPE (AMD)	3.09 ± 2.09	3.04 ± 1.79	2.710 ± 2.716	2.722 ± 2.676	2.721 ± 2.517	$\mathbf{2.707 \pm 2.601}$
IRPE (Normal)	2.06 ± 1.51	2.10 ± 1.36	1.860 ± 1.303	$\mathbf{1.850 \pm 1.317}$	1.892 ± 1.342	1.885 ± 1.358
OBM (AMD)	4.94 ± 5.35	4.43 ± 2.68	4.149 ± 3.532	4.129 ± 3.679	4.029 ± 3.400	$\mathbf{4.022 \pm 3.416}$
OBM (Normal)	$\mathbf{2.28 \pm 0.36}$	2.40 ± 0.39	2.660 ± 2.111	2.619 ± 1.989	2.594 ± 1.832	2.578 ± 1.768
Overall	2.78 ± 3.31	2.71 ± 2.25	2.458 ± 2.696	2.447 ± 2.734	2.428 ± 2.391	$\mathbf{2.425 \pm 2.610}$

size of 9 patches. The initial learning rate is set to 0.001 and reduced by half if the loss does not improve after 10 consecutive epochs. In the same way as in SASR [5], the weights of smoothS loss are set to 0, 0.3, and 0.5 for ILM, IRPE, and OBM, respectively. As data augmentation, horizontal flipping is performed with a probability of 0.5. The mean absolute distance (MAD) between the estimated boundary position and the ground truth in each column is used as the evaluation metric. The smaller the value of MAD, the more accurately the boundary position is estimated.

3.3. Experimental Results

In this experiment, we compare the accuracy of the state-of-the-art segmentation methods FCBR [3] and SASR [5]. FCBR refers to the results in [5] since [3] used the different dataset and the code is not publicly available. SASR refers to the results in [5] and shows the result of experiments performed under the same condition as in this paper using the publicly available code. Note that the paper [5] did not perform cross-validation, which differs from the experimental conditions in this paper. The proposed method is evaluated before and after refinement and without multi-scale loss. The accuracy of estimating the boundaries of the retinal layers for each method is summarized in Table 1. The proposed method has higher estimation accuracy than FCBR and SASR, and also has higher estimation accuracy after refinement than before refinement. Fig. 2 shows the results of boundary estimation of SASR and the proposed method with/without multi-scale loss The proposed method can estimate the detailed shape of the boundary, resulting in the shape close to the ground truth. Fig. 3 shows the boundary estimation results before and after refinement, and the layer map. The refinement module improves the accuracy of boundary estimation by making the pixel-wise labels consistent with the boundaries. These results indicate that the proposed method can estimate the boundaries of retinal layers with high accuracy by considering pixel-wise labels.

4. CONCLUSION

We proposed a 2D-3D hybrid network with multi-scale loss and refinement module to improve the segmentation accuracy of retinal layers. Through the experiments using the public OCT dataset, we demonstrate the effectiveness of multi-scale loss and refinement module in retinal layer segmentation compared with the conventional methods. In the future, we plan to investigate methods for myopic eyes, where segmentation of retinal layers from OCT data is extremely difficult, and the application of the proposed method to ophthalmic diagnostic support.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

This work was supported in part by JSPS KAKENHI Grant Numbers 21H03457.

7. REFERENCES

[1] 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, pp. 3627–3642, Aug. 2017.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, vol. 9351, pp. 234–241, Oct. 2015.

[3] Y. He, A. Carass, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Fully convolutional boundary regression for retina OCT segmentation," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, vol. 11764, pp. 120–128, Oct. 2019.

[4] 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, pp. 101856, Feb. 2021.

[5] H. Liu, D. Wei, D. Lu, Y. Li, K. Ma, L. Wang, and Y. Zheng, "Simultaneous alignment and surface regression using hybrid 2D-3D networks for 3D coherent layer segmentation of retina OCT images," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention*, vol. 12908, pp. 108–118, Sept. 2021.

[6] Y. He, A. Carass, Y. Liu, B. M. Jedynak, S. D. Solomon, S. Saidha, P. A. Calabresi, and J. L. Prince, "Deep learning based topology guaranteed surface and MME segmentation of multiple sclerosis subjects from retinal OCT," *Biomedical Optics Express*, vol. 10, pp. 5042–5058, Oct. 2019.

[7] S. Farsiu, S. J. Chiu, R. V. O'Connell, F. A. Folgar, E. Yuan, J. A. Izatt, and C. A. Toth, "Quantitative classification of eyes

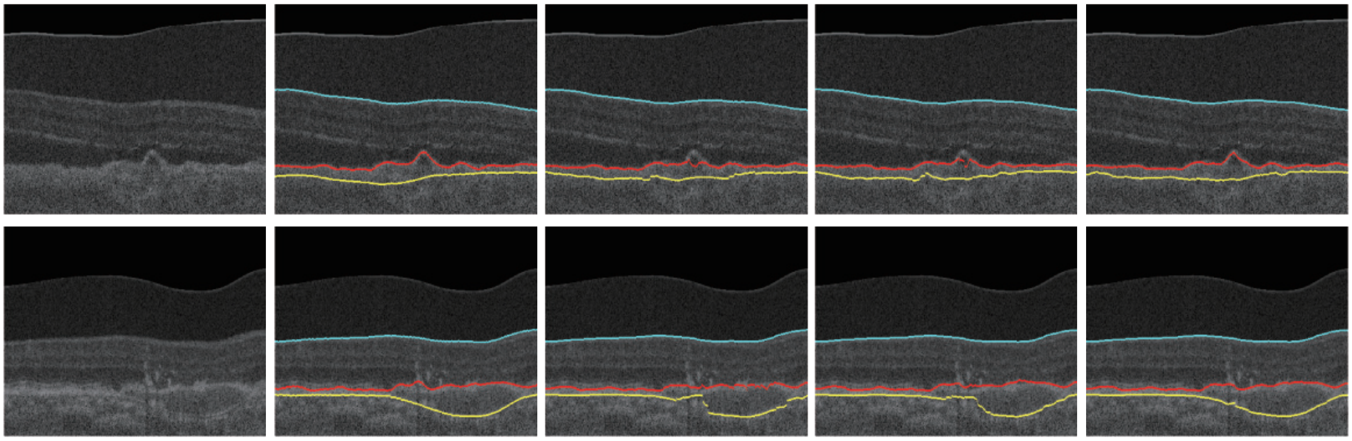

Fig. 2. Examples of estimated boundaries for each method.

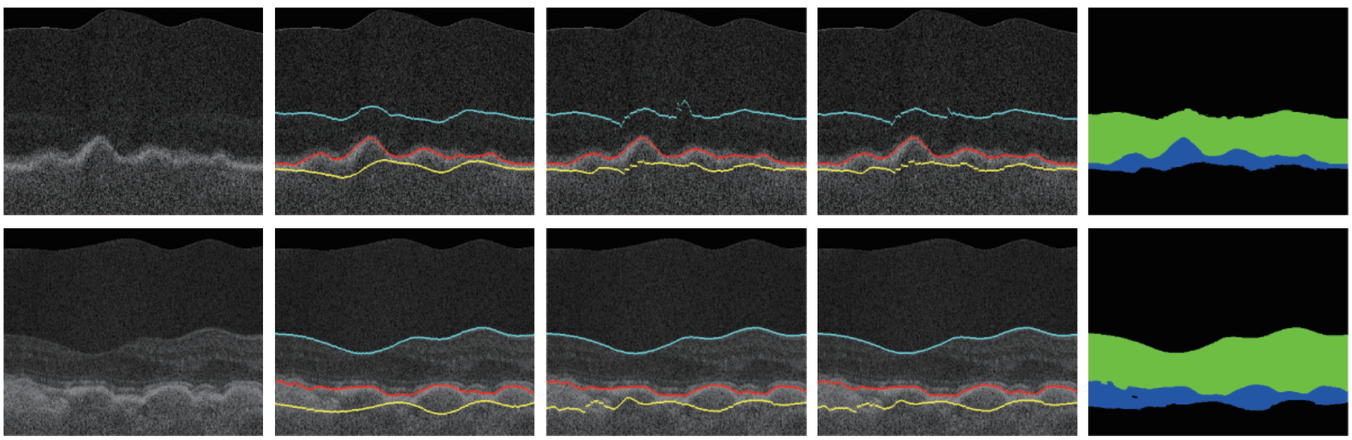

Fig. 3. Examples of estimated boundaries for each method and estimated layer map of the proposed method.

with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162–172, Jan. 2014.

[8] Hongming Li and Yong Fan, "Non-rigid image registration using fully convolutional networks with deep self-supervision," *CoRR*, vol. abs/1709.00799, pp. 1–8, Sept. 2017.

[9] A. Lang, A. Carass, M. Hauser, E. S. Sotirchos, P. A. Calabresi, H. S. Ying, and J. L. Prince, "Retinal layer segmentation of macular OCT images using boundary classification," *Biomedical Optics Express*, vol. 4, pp. 1133–1152, July 2013.