Video inspired lesion segmentation algorithm for OCT retinal images

Zeyu Zhang[a], Huiyu Duan[a], Guangtao Zhai[a], Bo Jiang[b], and Yuye Ling[a,*]

[a]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
[b]John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Optical Coherence Tomography (OCT) retinal lesion segmentation is critical for ophthalmic diagnosis and treatment. However, current OCT retinal image segmentation methods primarily focus on individual B-scans and neglect the continuity condition of lesions across OCT volume B-scans, which leads to suboptimal segmentation results. To address this issue, we propose an innovative 2.5D segmentation algorithm that treats OCT retinal images as video sequences and leverage advanced video segmentation models. We compared our method with state-of-the-art 2D image-based segmentation methods on a public dataset. Experimental results demonstrate that our method significantly improves both visual quality and quantitative metrics, particularly in terms of structural continuity and robustness of the segmentation results.

Keywords: OCT retinal lesion segmentation, Video sequences, Video segmentation

1. INTRODUCTION

Optical Coherence Tomography (OCT) is an important imaging modality for the diagnosis and monitoring of retinal diseases, which could provide high-resolution cross-sectional images of the retina.[1] These images allow for the detailed visualization of retinal lesions, essential for precise diagnosis and effective treatment planning. Moreover, the segmentation and quantification of these lesions, as imaging biomarkers, are pivotal for clinicians to understand their implication on disease progression.

Recent advancements in deep learning-based segmentation methods have shown promising results in the task of OCT retinal lesion segmentation. Most of these methods use 2D image-based models that process each B-scan independently, focusing on expanding the receptive field[2] or incorporating multi-scale information[3] to improve model performance. However, they tend to concentrate solely on the information from individual frames and, neglect the inherent continuity condition of lesions across adjacent B-scans within an OCT volume. This neglection can limit the effectiveness of lesion detection and segmentation in a clinical setting and further result in suboptimal segmentation outcomes, which is especially critical in the context of reconstructing 3D retinal structures for clinical purposes.

As illustrated in Fig. 1, adjacent frames in a video are separated by minimal temporal intervals, similar to the minimal scanning intervals between adjacent B-scans in an OCT volume. Therefore, we can apply the optical flow constraint equation to the OCT volume, the pixel motion relationship of adjacent B-scans can be represented as:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial z}v + \frac{\partial I}{\partial y} = 0, \tag{1}$$

where y denotes the position of B-scans in an OCT volume instead t which denotes the positions of frames in a video. This similarity emphasizes the inherent continuity present in both video sequences and OCT volumes, highlighting the potential for applying mature video-based segmentation techniques to OCT retinal volume analysis.

Inspired by video processing techniques, we introduce a novel concept for OCT retinal lesion segmentation by treating adjacent B-scans as video sequences in this manuscript. This concept frames the segmentation task

*Send correspondence to: yuye.ling@sjtu.edu.cn

Medical Imaging 2025: Image Processing, edited by Olivier Colliot, Jhimli Mitra, Proc. of SPIE Vol. 13406, 1340628 · © 2025 SPIE
1605-7422 · doi: 10.1117/12.3047184

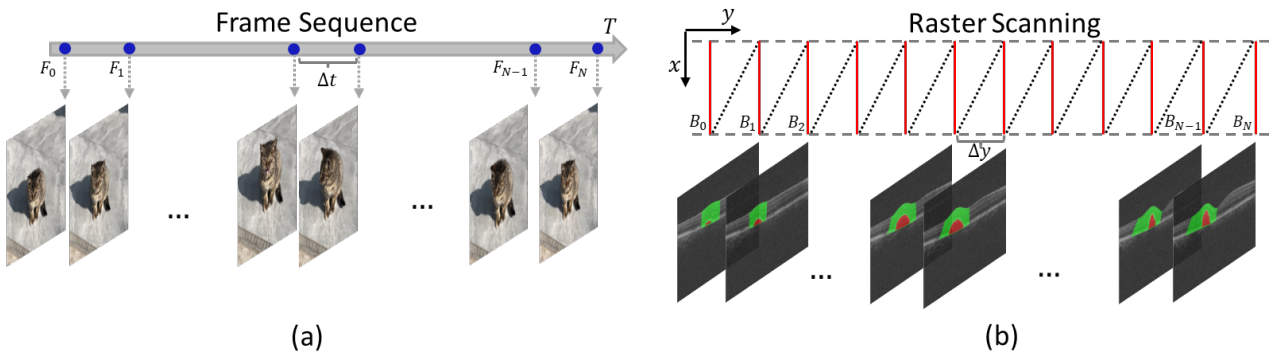

Figure 1. (a) Consecutive Frames in a video. (b) Consecutive B-scans in an OCT retinal volume.

as a 2.5D problem rather than a purely 2D one. By leveraging the OCT raster scanning mechanism and the inherent continuity of retinal lesions, we employ a Mamba-based video method[4] that accurately captures the consistent presence and gradual changes of lesions across consecutive OCT B-scans. This 2.5D approach could bridge the gap between 2D and 3D segmentation, providing a more balanced performance for segmenting retinal pathology. The contributions of our approach can be summarized as follows: (1) We introduce an innovative concept for treating OCT retinal lesion segmentation task as an analogy of video segmentation; (2) We achieve notable results using video segmentation models, surpassing state-of-the-art methods on a public dataset.

2. METHOD

2.1 Framework

In this work, we employ the Mamba-based Vivim model[4] for OCT retinal lesion segmentation, where the model's input is video sequence. The initial step entails transforming consecutive B-scans from an OCT retinal volume into a video sequence. When processing a volume containing N B-scans, our approach involves converting every consecutive set of T B-scans into a video sequence. By setting the stride to 1, each new video sequence overlaps with the preceding one by $T-1$ B-scans. This methodology ensures continuity across sequences and enhances the robustness of the dataset. These preprocessed video sequences are then used as input for the Vivim model, as demonstrated in Fig. 2. Vivim primarily comprises two components: a hierarchical encoder featuring stacked Temporal Mamba Blocks designed to extract multi-scale feature sequences, and a lightweight CNN-based decoder head that integrates multi-scale feature sequences to generate segmentation masks.

2.2 Loss Function

We use a combined approach of Binary Cross-Entropy Loss ℓ_{bce} and Dice Loss ℓ_{dice} to supervise the learning of our model. This hybrid loss function is designed to optimize the model's performance in handling imbalanced data and improving the precision of segmentation in target areas.

2.3 Dataset and Implementation

We evaluated our innovative concept for OCT retinal lesion segmentation on the public Retinal Edema Segmentation Challenge (RESC) dataset,[5] containing subretinal fluid (SRF) and pigment epithelial detachments (PED) lesion images. The dataset comprises training, validation, and test splits, consisting of 70, 15, and 15 cases respectively, each case includes 128 OCT slices at a resolution of 512 $\times$ 1024 pixels. However, the dataset only provides pixel-level annotations for training and validation splits, therefore, we implement 5-fold cross-validation on the training split on the case level in our experiments and test on the validation split. The proposed framework was trained on one NVIDIA RTX 3090 GPU. During each iteration of the training process, video frames are resized to 256×256 pixels. Each video sequence contains 5 B-scans, and each batch comprises 8 such sequences. We conducted data augmentation strategies on these sequences by random flipping and rotation. The initial learning rate is set at 1×10^{-4} and is gradually reduced to 1×10^{-4} using a cosine annealing strategy. The entire training extends over 100 epochs, with the Adam optimizer utilized to optimize model parameters.

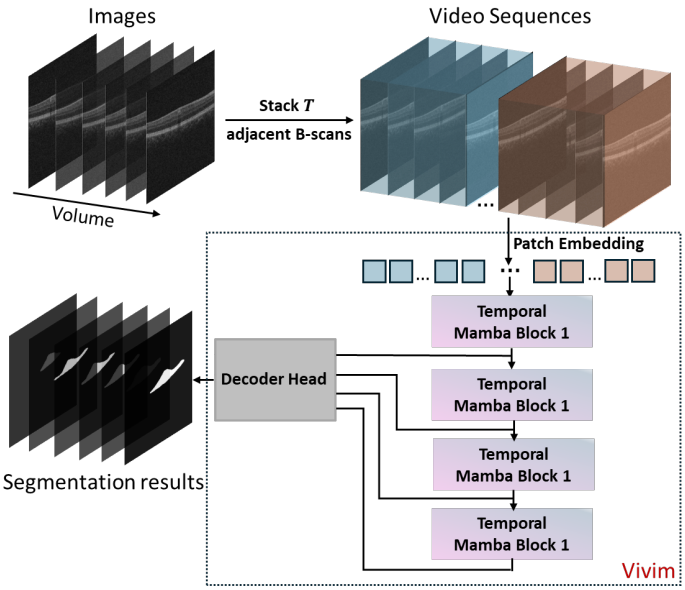

Figure 2. Overview of the framework.

3. RESULTS

We employed five segmentation evaluation metrics, including Dice, Precision, Recall, S-measure (S_α)[6] and E-measure (E_ϕ),[7] and conducted comparisons with current popular medical image segmentation baseline models, including Unet,[8] Unet++,[9] and TransUnet.[10] All experimental settings were kept consistent, and The quantitative results are presented in Tab. 1, tested on the validation set with 5-fold cross-validation on the training set.

Table 1. Quantitative comparison with state-of-the-art methods on the RESC validation set. Mean(Std) of 5-fold cross-validation. The best results in this table are labeled in **bold**.

Dataset	Lesions	Model	Metric				
			Dice(%)	Precision(%)	Recall(%)	E_ϕ(%)	S_α(%)
RESC	SRF	Vivim	**80.84(2.59)**	83.44(2.08)	**80.92(3.20)**	**95.49(1.36)**	**91.56(1.50)**
		Unet	76.08(1.57)	84.45(0.92)	74.13(1.89)	90.13(0.8)	87.12(1.20)
		Unet++	74.90(2.51)	83.69(1.96)	72.59(2.94)	87.87(2.41)	86.10(1.63)
		TransUnet	77.26(2.45)	**84.55(1.67)**	75.25(2.61)	91.22(2.33)	89.19(1.66)
	PED	Vivim	30.55(1.03)	**35.83(1.79)**	28.71(1.74)	**77.60(3.59)**	**73.50(3.39)**
		Unet	30.79(3.74)	35.04(6.48)	29.23(3.39)	68.31(3.90)	70.77(2.93)
		Unet++	17.96(2.65)	20.63(3.05)	16.74(2.17)	63.12(0.84)	70.40(0.92)
		TransUnet	**36.30(11.92)**	35.49(12.13)	**30.32(10.99)**	71.31(8.92)	69.54(4.21)
	Average	Vivim	**75.70**	78.57	**75.58**	**93.66**	**89.71**
		Unet	71.45	79.40	69.54	87.90	85.45
		Unet++	69.08	77.24	66.88	85.34	84.50
		TransUnet	73.07	**79.53**	70.66	89.19	86.29

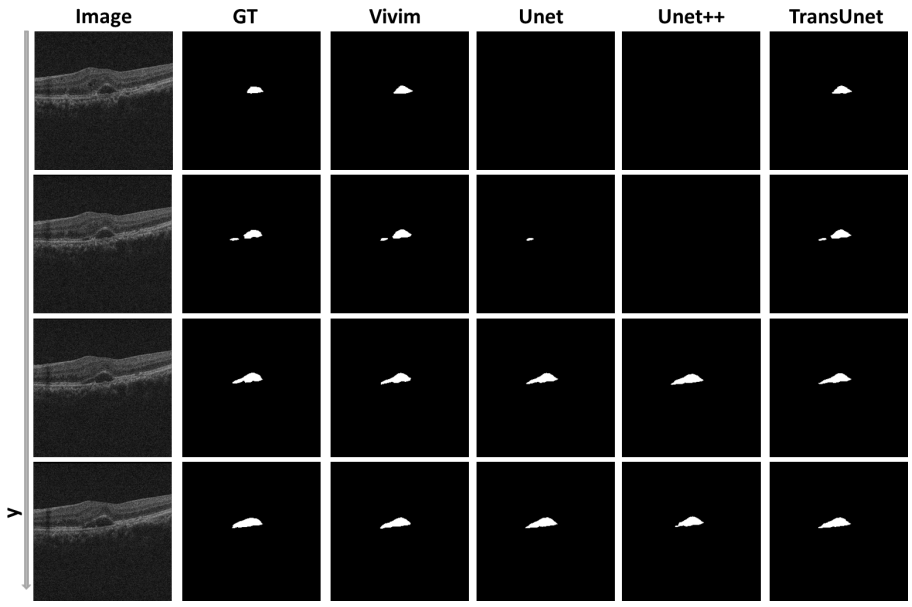

Figure 3. Qualitative results of SRF on the consecutive B-scans along y axis.

We can observe that video-based methods outperform image-based methods on the entire dataset, surpassing the state-of-the-art in SRF segmentation while performing slightly inferior in PED segmentation. However, our method achieves the best results in structured metrics (S_α) and (E_ϕ). This indicates that in the task of OCT retinal lesion segmentation, video-based methods explicitly consider the continuity of adjacent frames, making the segmentation results structurally closer to the ground truth. This ensures continuity of the segmentation results in volume, facilitating subsequent 3D reconstruction. We presented the visualization results of SRF segmentation in Fig. 3. Our method maintains better structural consistency and continuity across consecutive B-scan segmentation results.

4. CONCLUSION

In conlusion, we propose an innovative concept that transforms OCT retinal lesion segmentation into a video segmentation task, leveraging the continuity between B-scans in OCT retinal volumes. By employing mature video segmentation models on a public dataset, we achieve promising segmentation results that exhibit better inter-frame continuity and intra-frame structural robustness, surpassing the performance of image-based state-of-the-art segmentation methods.

REFERENCES

[1] Goebel, W. and Kretzchmar-Gross, T., "Retinal thickness in diabetic retinopathy: a study using optical coherence tomography (oct)," *Retina* **22**(6), 759–767 (2002).

[2] Chen, M., Ma, W., Shi, L., Li, M., Wang, C., and Zheng, G., "Multiscale dual attention mechanism for fluid segmentation of optical coherence tomography images," *Applied Optics* **60**(23), 6761–6768 (2021).

[3] Pappu, G. P., Tankala, S., Talabhaktula, K., and Biswal, B., "Eanet: Multiscale autoencoder based edge attention network for fluid segmentation from sd-oct images," *International Journal of Imaging Systems and Technology* **33**(3), 909–927 (2023).

[4] Yang, Y., Xing, Z., and Zhu, L., "Vivim: a video vision mamba for medical video object segmentation," *arXiv preprint arXiv:2401.14168* (2024).

[5] Hu, J., Chen, Y., and Yi, Z., "Automated segmentation of macular edema in oct using deep neural networks," *Medical image analysis* **55**, 216–227 (2019).

[6] Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., and Borji, A., "Structure-measure: A new way to evaluate foreground maps," in [*Proceedings of the IEEE international conference on computer vision*], 4548–4557 (2017).

[7] Fan, D.-P., Ji, G.-P., Qin, X., and Cheng, M.-M., "Cognitive vision inspired object segmentation metric and loss function," *Scientia Sinica Informationis* **6**(6), 5 (2021).

[8] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*], 234–241, Springer (2015).

[9] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J., "Unet++: A nested u-net architecture for medical image segmentation," in [*Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*], 3–11, Springer (2018).

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