Segmentation of Retinal Fluid Using Foundation Models

JOY Y CHENG[1,2], Anthony Wu[3], Jeffrey N Chiang[3,4]

[1]Bruins in Genomics (BIG) Summer Program, Institute for Quantitative and Computational Biosciences, University of California, Los Angeles (UCLA). [2]Department of Computer Science, Samueli School of Engineering, UCLA.
[3]Department of Computational Medicine, David Geffen School of Medicine, UCLA. [4]Department of Neurosurgery, David Geffen School of Medicine, UCLA.

Abstract

Neovascular age-related macular degeneration (nAMD) represents a leading cause of vision loss worldwide. Optical coherence tomography (OCT) B-scans can be observed for the presence of retinal fluid to assess disease progression. The time-consuming and subjective nature of manual OCT fluid segmentation demands clinically applicable computational segmentation methods; however, the development of these methods is hampered by a lack of high-quality annotated data. We approached automatic segmentation with deep learning by adapting and fine-tuning the vision transformer-based foundation model MedSAM using natural language prompts and benchmarked it against U-Net, a convolutional neural network model. We experimented with different sets of text prompts and unfrozen model components in our training. Overall, the U-Net outperformed the foundation model in our experiments, but the foundation model offers an innovative approach to automatic segmentation with limited data and additional research is required to determine its clinical applicability. Our work may be adapted for clinical uses such as predicting visual acuity and quantifying the effectiveness of nAMD treatments.

Background

- Optical coherence tomography (OCT) is a widely used non-invasive imaging modality that allows clinicians to view cross sections of the retina.
- Neovascular age-related macular degeneration (nAMD) is characterized by three types of retinal fluid: intraretinal cystoid fluid (IRF), subretinal fluid (SRF), and fluid in pigment epithelial detachments (PED)[1].

Figure 1. OCT machine by Topcon Healthcare.

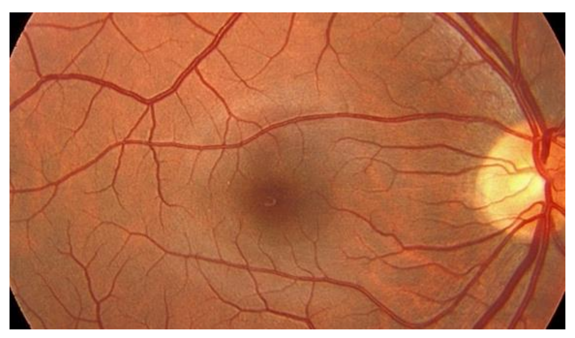

Figure 2. Fundus photograph of normal retina[2].

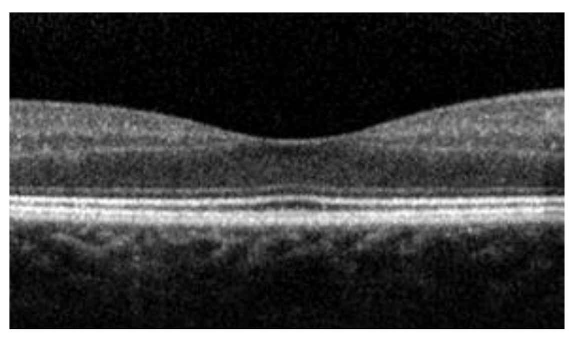

Figure 3. OCT scan of normal retina[3].

Figure 4. Fundus photograph of retina affected by nAMD[4]. Exudative fluid can be seen in the macular region.

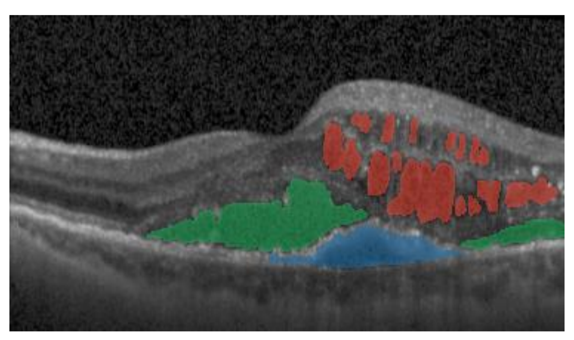

Figure 5. OCT scan of retina affected by nAMD from RETOUCH dataset. IRF is shown in red, SRF in green, and PED in blue.

Workflow

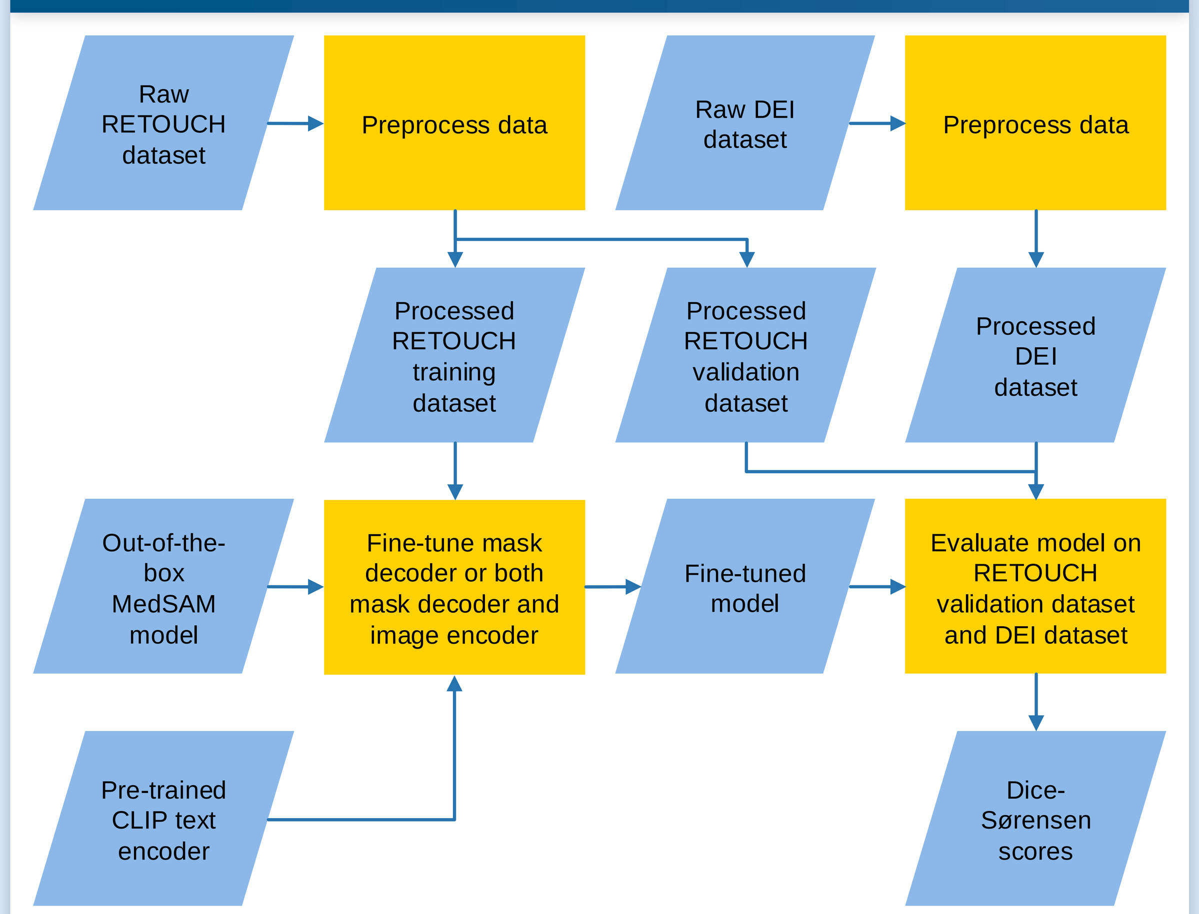

Figure 6. Diagram of workflow including data preprocessing, fine-tuning, and evaluation.

Data

- We used datasets from the Retinal OCT Fluid Challenge (RETOUCH)[5] and the Doheny Eye Institute (DEI).
- We used 90% of the RETOUCH dataset for training and 10% for validation. We evaluated our models on both the RETOUCH validation dataset and the entire DEI dataset.

Table 1. Characteristics of RETOUCH training, RETOUCH validation, and DEI datasets.

	RETOUCH training	RETOUCH validation	DEI
Image count	6,040	896	3,472 (1,371 with PED label)
Volume count	63	7	50 (27 with PED label)
Mask overlap	False	False	True

Methods

- We fine-tuned the foundation model MedSAM[6] in four experiments with varying text prompts and unfrozen model components.
- We employed early stopping with a patience of 25. Due to noise in validation curves, we trained models for at least 100 epochs.

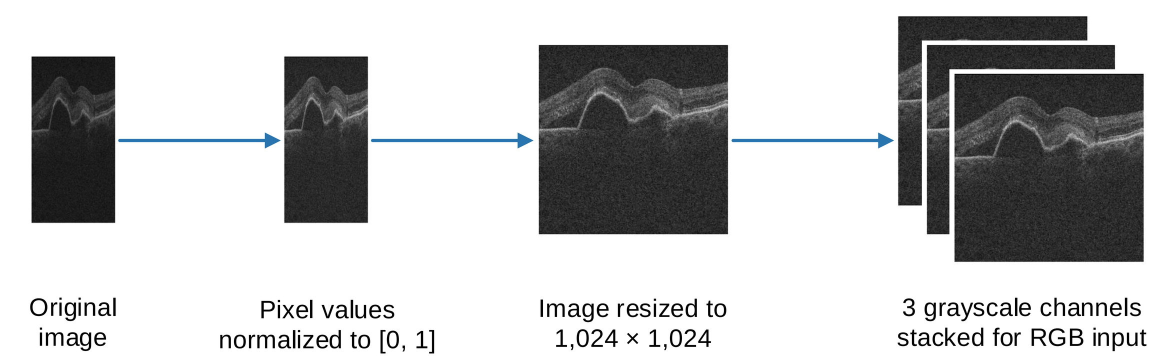

Figure 7. Data preprocessing pipeline.

Table 2. Training parameters for foundation model experiments and U-Net.

	Foundation	U-Net
Training examples	Only examples containing fluid from RETOUCH training dataset	All examples from RETOUCH training dataset
Validation examples	All examples from RETOUCH validation dataset	All examples from RETOUCH validation dataset
Training protocol (per epoch)	Train on all image-mask pairs resulting from choosing a random mask with fluid for each image	Train on all image-mask pairs
Validation protocol (per epoch)	Test on all image-mask pairs	Test on all image-mask pairs
Augmentations (50% chance per example)	Horizontal flip	Horizontal flip
Loss function	Dice loss + cross-entropy loss	Dice loss + focal loss

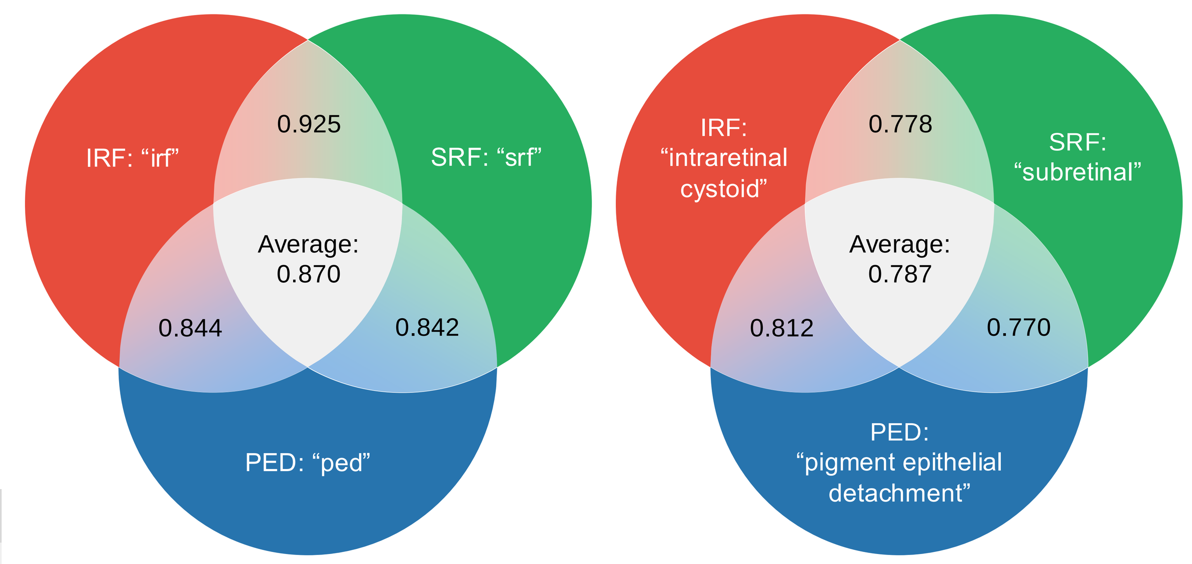

Figure 8. Two sets of three text prompts used in our experiments. Cosine similarity values normalized to [0, 1] are shown for pairs of CLIP text embeddings for text prompts.

Table 3. Text prompts and unfrozen components for foundation model experiments.

	Text prompts	Unfrozen components
Foundation 1	"irf", "srf", "ped"	Mask decoder
Foundation 2	"irf", "srf", "ped"	Image encoder, mask decoder
Foundation 3	"intraretinal cystoid", "subretinal", "pigment epithelial detachment"	Mask decoder
Foundation 4	"intraretinal cystoid", "subretinal", "pigment epithelial detachment"	Image encoder, mask decoder

Results

Table 4. Dice-Sørensen coefficients for IRF, SRF, and PED labels for fine-tuned foundation models and U-Net.

	RETOUCH validation dataset				DEI dataset			
	IRF	SRF	PED	Average	IRF	SRF	PED	Average
Foundation 1	0.564	0.247	0.523	0.445	0.060	0.190	0.209	0.153
Foundation 2	0.540	0.324	0.528	0.464	0.087	0.199	0.194	0.160
Foundation 3	0.528	0.300	0.448	0.425	0.087	0.223	0.138	0.149
Foundation 4	0.520	0.261	0.510	0.430	0.078	0.210	0.229	0.172
U-Net	0.609	0.587	0.621	0.606	0.335	0.675	0.419	0.476

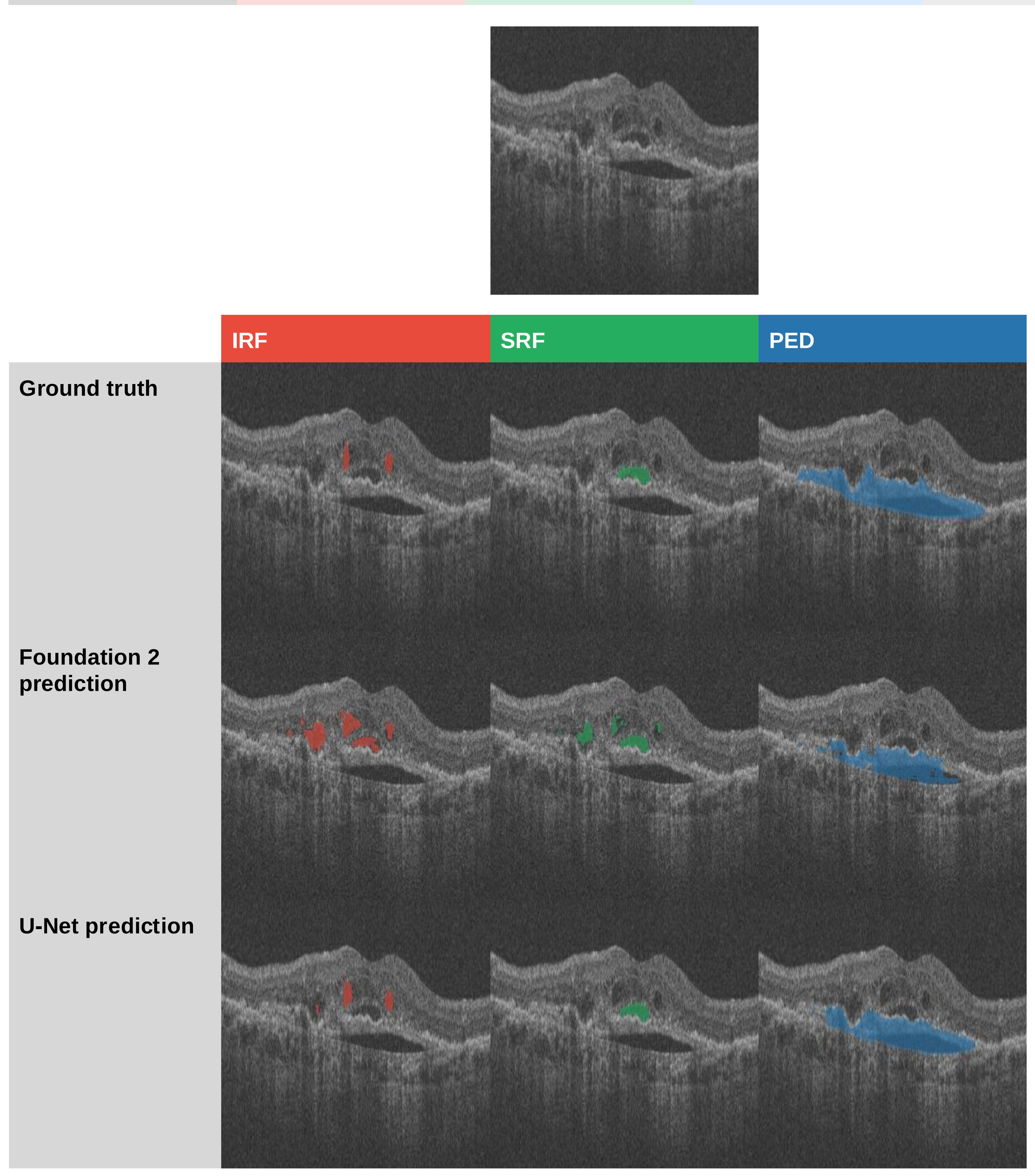

Figure 9. Predictions of fine-tuned foundation model 2 and U-Net for RETOUCH validation image.

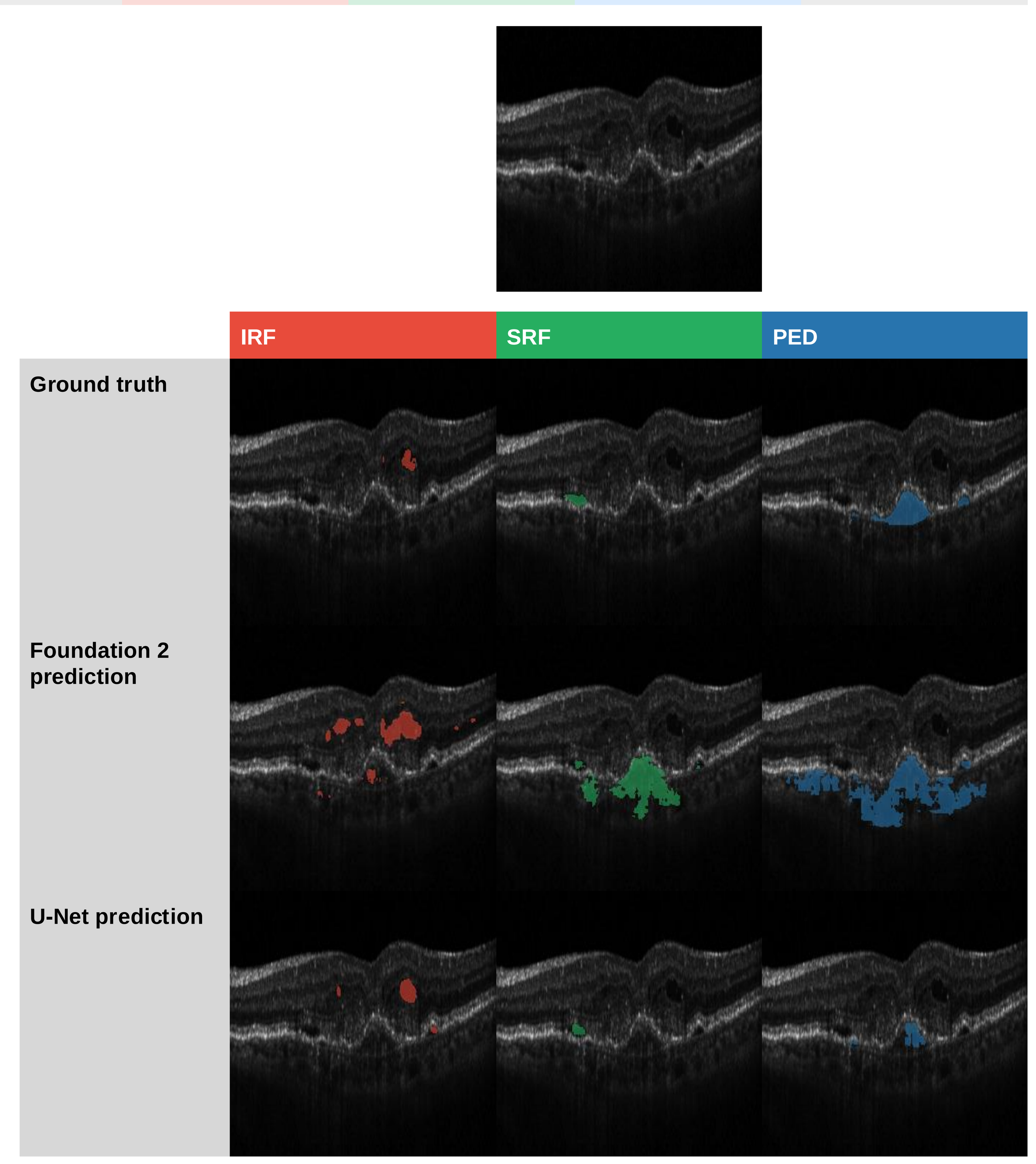

Figure 10. Predictions of fine-tuned foundation model 2 and U-Net for DEI image.

Model Architecture

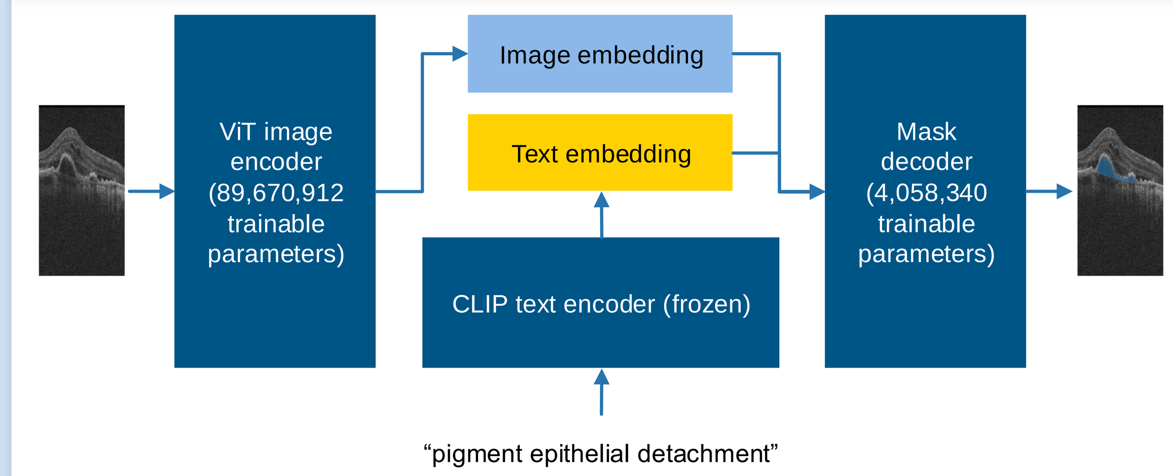

Figure 11. Modified MedSAM model architecture with CLIP text encoder as prompt encoder.

Discussion

Conclusions

- For each dataset, the U-Net achieved a higher average Dice-Sørensen coefficient than all the MedSAM-based models.
- Fine-tuning the image encoder and the mask decoder yielded better results than fine-tuning only the mask decoder.
- Neither of the two sets of text prompts consistently demonstrated an advantage over the other.

Limitations

- We fine-tuned MedSAM only on images with fluid, so the models tended to over-predict masks on images without fluid.
- SAM[7] was fine-tuned on 1,570,263 medical image-mask pairs to obtain MedSAM[6]. Only 803 (0.0005%) of the images were OCT scans[6]. Fine-tuning a model trained on a dataset with greater OCT representation could yield better results.

Future Directions

Model Development

- Explore segmentation with Segment Anything Model 2 (SAM 2) by Meta. Video segmentation abilities could be used for continuous segmentation of three-dimensional volumes, potentially leading to improved performance.
- Modify the architecture by replacing the pre-trained CLIP text encoder with a custom embedding layer.
- Modify training by incorporating some scans without fluid, trying other augmentations such as random cropping and elastic deformation, and training on all masks for each image.

Clinical Applications

- Build a web interface for automatic fluid segmentation to increase annotation efficiency.
- Quantify differences in fluid volume before and after treatment with anti-vascular endothelial growth factor drugs.
- Correlate fluid volume with visual acuity.

References and Acknowledgements

1. Bogunović, H., Vogl, W.-D., Waldstein, S. M., & Schmidt-Erfurth, U. (2019). OCT fluid detection and quantification. *Computational Retinal Image Analysis*, 273–298.
2. Creel, D. J. Clinical electrophysiology. *Webvision: The Organization of the Retina and Visual System [Internet].* (2007). Available at: https://www.ncbi.nlm.nih.gov/books/NBK11553/
3. Retouch - Grand Challenge. *RETOUCH* Available at: https://retouch.grand-challenge.org/.
4. Hobbs, S. D. Wet age-related macular degeneration (AMD). *StatPearls [Internet].* (2024). Available at: https://www.ncbi.nlm.nih.gov/books/NBK572147/
5. Bogunović, H. *et al*. RETOUCH: The Retinal Oct Fluid Detection and segmentation benchmark and challenge. *IEEE Transactions on Medical Imaging* **38**, 1858–1874 (2019).
6. Ma, J., He, Y., Li, F. *et al*. Segment anything in medical images. *Nature Communications* **15**, 654 (2024).
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., & Girshick, R.B. (2023). Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992-4003.

Acknowledgements

Special thanks to Dr. Jeffrey Chiang for his wonderful mentorship and guidance and for the opportunity to explore this fascinating topic; Anthony Wu for his countless invaluable insights and help with onboarding; Alex Chen for his companionship throughout the BIG Summer Program; and Simon Lee, Yanai Halperin, and Helio Halperin for all the awesome lunchtime discussions. Additionally, I would like to thank Dr. Alexander Hoffman, Caroline Baron, and the rest of the BIG Summer Program team for their hard work in cultivating a lively and supportive research community which I will miss dearly!