Foundation Models Permit Retinal Layer Segmentation Across OCT Devices

Olivier Morelle[1,2] and Thomas Schultz[1,3]

[1] University of Bonn, B-IT and Department of Computer Science
[2] Department of Ophthalmology, University Hospital Bonn, Bonn, Germany
[3] Lamarr Institute for Machine Learning and Artificial Intelligence

Problem:

▶ Segmentation of retinal layers is critical in the analysis of optical coherence tomography (OCT) data.

▶ The appearance and quality of OCT datasets vary considerably.

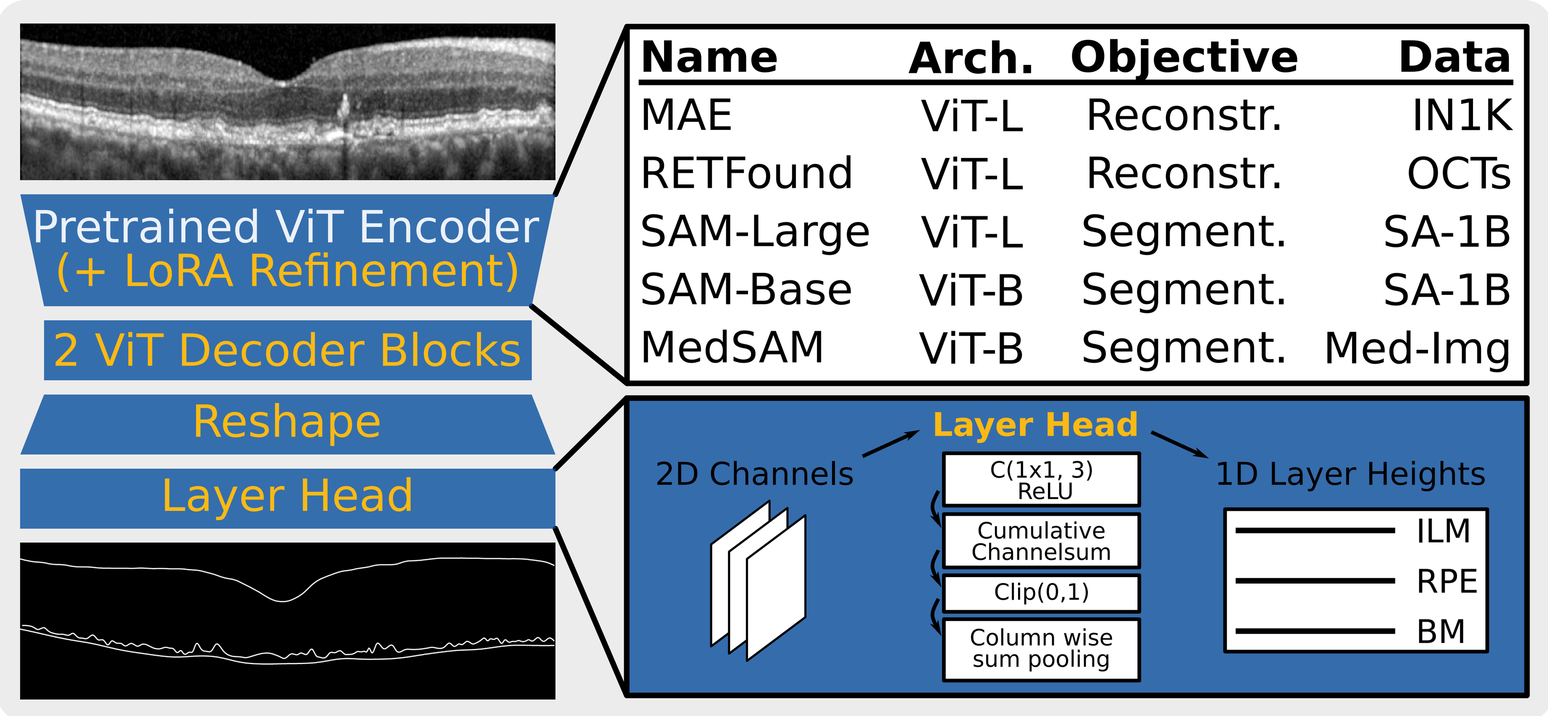
▶ Our model for order-constrained regression[1] provides state-of-the-art results in the training domain but generalises poorly to unseen domains.

Research Question:

▶ Do vision foundation models improve the robustness of retinal layer segmentation models?

▶ Which foundation models work best?

Methods:



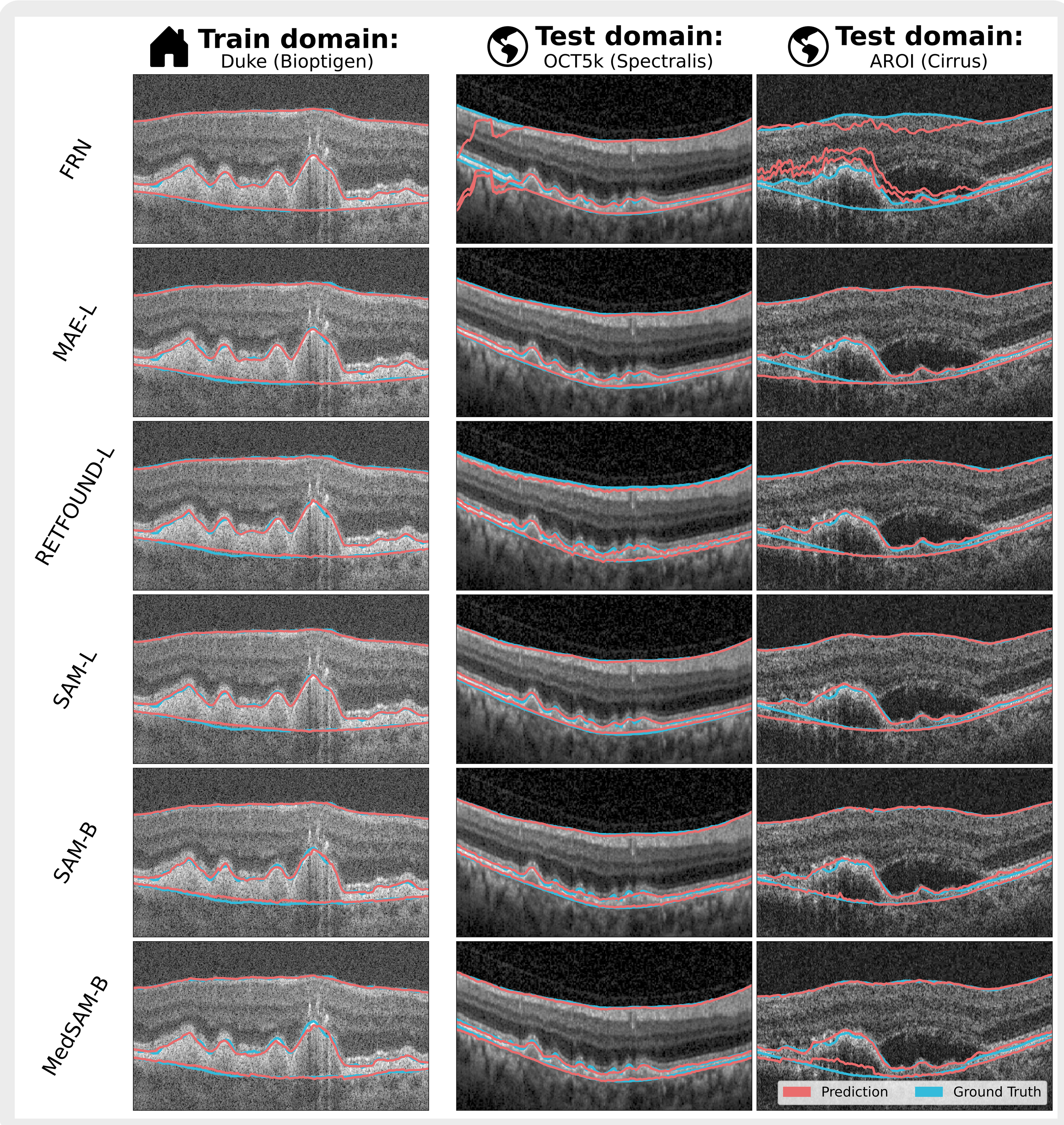
Name	Arch.	Objective	Data
MAE	ViT-L	Reconstr.	IN1K
RETFound	ViT-L	Reconstr.	OCTs
SAM-Large	ViT-L	Segment.	SA-1B
SAM-Base	ViT-B	Segment.	SA-1B
MedSAM	ViT-B	Segment.	Med-Img

▶ We compare ViT encoders of varying sizes, pre-trained with different data and objectives.

▶ Encoder parameters are fixed except for a LoRA[9] refinement

▶ 2 Transformer blocks and a layer head are trained to map ViT embeddings to layer heights.

Results:



Comparison of the absolute segmentation error in pixel for the tested models on the training domain (Duke) and two seperate datasets. We report the Median and 95% Quantile over all layers withing each dataset.

Model	Duke[2]	OCT5k[3]	AROI[4]
CNN[1]	**0.21 (0.97)**	0.71 (44.62)	0.99 (21.75)
MAE-L[5]	0.23 (1.01)	0.37 (1.2)	0.37 (**3.58**)
RETFound-L[6]	0.24 (1.02)	0.8 (2.73)	0.5 (4.68)
SAM-L[7]	0.22 (0.99)	**0.34 (1.1)**	**0.35** (3.88)
SAM-B[7]	0.25 (1.18)	0.44 (2.63)	0.64 (10.37)
MedSAM-B[8]	0.25 (1.17)	0.43 (2.01)	0.7 (9.19)

In Domain:

▶ The CNN trained from scratch is the best by a small margin.

Under Domain Shift:

▶ All tested pre-trained models do better on new domains than the baseline CNN. SAM-L is the best.

▶ RETFound-L domain specific pre-training reduces the performance compared to the more general MAE-L

▶ RETFound-L performance varies between test datasets. Is there a bias in the training data?

Conclusion

▶ ViT-based foundation models can increase the robustness of OCT layer segmentation compared to CNNs trained from scratch.

▶ The choice of the right foundation model matters.

▶ Within the foundation models, bigger models are better.

▶ Sometimes generic models outperform specialized models.

▶ LoRA refinement of the encoder helps all models, but does not change their relative performance.

References

1 Morelle et al. (2023) Scientific Reports 13 (1): 8162. https://doi.org/10.1038/s41598-023-35230-4.
2 Farsiu et al. (2014) Ophthalmology 121 (1): 162–72. https://doi.org/10.1016/j.ophtha.2013.07.013.
3 Arikan et al. (2023) bioRxiv. https://doi.org/10.1101/2023.03.29.534704.
4 Melinščak,et al. (2021) Automatika 62 (3–4): 375–85. https://doi.org/10.1080/00051144.2021.1973298.
5 He et al. (2022) CVPR 15979–88. https://doi.org/10.1109/CVPR52688.2022.01553.
6 Zhou et al. (2023) Nature 622 (7981): 156–63. https://doi.org/10.1038/s41586-023-06555-x.
7 Kirillov et al. (2023) ICCV 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371.
8 Ma et al. (2024) Nature Communications 15 (1): 654. https://doi.org/10.1038/s41467-024-44824-z.
9 Hu et al. (2022) ICLR https://doi.org/10.48550/arXiv.2106.09685

UNIVERSITÄT BONN INFORMATIK b-it ukb LAMARR INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE