Foundation Models Permit Retinal Layer Segmentation Across OCT Devices

Olivier Morelle[1,2][0000−0001−6404−2726] and Thomas Schultz[1,3][0000−0002−1200−7248]

[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

Abstract. The segmentation of retinal layers in images from optical coherence tomography (OCT) is an important step in ophthalmological diagnosis and disease monitoring. Current CNN-based models perform well on images from the same OCT scanner on which they have been trained, but their performance can degrade drastically when images are acquired with other devices. We present the first method for OCT layer segmentation that builds on recent Vision Transformer (ViT) foundation models. We demonstrate that, compared to a state-of-the-art CNN approach, doing so significantly improves their ability to generalize to devices for which no training data was available. This highlights the potential of foundation models to enable more robust medical image analysis. We also analyze the effect of using different foundation models. Notably, more generic foundation models from computer vision permitted better generalization than an equally large foundation model that was specifically trained for OCT analysis.

Keywords: Optical Coherence Tomography · Foundation Models · Retinal Layer Segmentation

1 Introduction

Medical image analysis has become an indispensable tool in modern healthcare, providing critical insights for the diagnosis, treatment, and monitoring of various medical conditions. Among the advanced imaging techniques, Optical Coherence Tomography (OCT) stands out due to its non-invasive nature and high-resolution capabilities, making it particularly valuable in ophthalmology for detailed visualization of the retinal architecture. A key challenge in leveraging OCT data lies in the accurate segmentation of retinal layers, which is crucial for diagnosing and tracking diseases such as age related macular degeneration.

One of the major obstacles in developing models for layer segmentation and other tasks is a performance degradation when the test data is from a different distribution than the training data. Such domain shift can be caused by different imaging conditions or patient demographics and influences where a model can be deployed.

In recent years, Vision Transformers (ViTs) have emerged as a powerful tool in the field of computer vision [4]. Unlike traditional convolutional neural networks (CNNs), ViTs leverage self-attention mechanisms to model long-range dependencies in images, offering a more flexible and scalable approach to image analysis. Their ability to capture global context makes them particularly well-suited for tasks requiring detailed spatial understanding, such as layer segmentation in OCT images. On the down side, training these models from scratch is very expensive due to the required large batch sizes. The first ViT for example was trained with a batch size of 4096 [4].

Using pre-trained models hence became a common approach to capitalize on the benefits of ViTs. These foundation models based on Vision Transformers represent a significant advancement in building robust and generalizable models for medical image analysis since they are pre-trained on large-scale datasets, capturing a wide array of visual patterns and features. The extensive pre-training enables foundation models to generalize better to unseen domains, addressing the issue of domain shift effectively [8]. By leveraging the comprehensive feature representations learned during pre-training, these models can be fine-tuned on specific medical imaging tasks with relatively smaller datasets, thereby enhancing their applicability and performance in diverse clinical scenarios.

Our first contribution in this work is a method that leverages foundation models for the segmentation of retinal layers in OCT by combining a Low-Rank Adaptation (LoRA) fine-tuning [7] of the pre-trained encoder with an added decoder built from two Transformer blocks and a task specific Layer Head [12].

For the encoder of our model, we can choose from a growing number of pre-trained ViTs. This will allow us to benefit from even stronger models as they become available, but it also poses the question which current model yields the best results. Our second contribution hence is a performance comparison of layer segmentation models building on different foundations. We evaluate the layer segmentation performance both in the training domain and under domain shift and present qualitative results for all models and datasets. As a baseline and example for a more traditional approach we use the FeatureRefinementNet (FRN), a CNN specifically designed for OCT Layer Segmentation [12].

2 Related Work

In ophthalmology, self-supervised training methodologies already showed some promising results for solving downstream tasks. RETFound [18] provides foundation models for color fundus photography as well as OCT, but has only been evaluated for image classification, not for segmentation. Uni4Eye [3] is another example where a ViT was pretrained on multimodal ophthalmological data as a masked auto-encoder and subsequently finetuned to a classification task. Another work improved data efficiency on retinal datasets compared to traditional

U-Net-based methods through self-supervised learning of a Transformer-U-Net hybrid architecture [16].

Hybrid models incorporating transformers were also explored for segmentation tasks like OCTA vessel segmentation [14] and layer and fluid segmentation [17], but without self-supervised learning.

Despite these advances, most segmentation models for ophthalmological data, including transformer-based ones, were learned in a fully supervised fashion, while domain-specific self-supervised approaches were evaluated only on classification tasks. In our work, we demonstrate the benefits of using foundation models for OCT layer segmentation.

3 Methods

3.1 Datasets

All models were trained on the Duke dataset [5] which consists of OCT scans from Control and Age-related Macular Degeneration (AMD) subjects, with annotations delineating the Inner Limiting Membrane (ILM), the inner boundary of the Retinal Pigment Epithelium (IBRPE) and Bruch's Membrane (BM) within a circular region of 5mm diameter, centered at the Fovea. All data was collected using Bioptigen devices. The dataset includes 115 volumes from Control subjects and 269 volumes from AMD subjects and was split into 164 subjects for training, 20 for validation, and 200 for testing, consistent with previous work [12].

The two-dimensional slices in which the 3D OCT images are acquired are referred to as B-scans, while the individual image columns in a B-scan are called A-scans. B-scans labeled on less than 50% of the full width were removed from the training set, resulting in 8,928 B-scans from 164 subject for training. For in domain evaluation we use all labeled B-scans from the Duke test set. Performance under domain shift is evaluated using two publicly available OCT datasets.

The OCT5k dataset has 60 manually labeled OCT volumes in total assembled from 20 AMD, 20 DME and 20 Control subjects resulting in 1672 B-scans. For each B-scan layer annotations from three graders exist. We compare against the mean of the three graders. The OBRPE Layer in the OCT5k dataset is compared to our BM predictions. We do not use the automatic segmentations provided with the dataset. The data was collected with Heidelberg Spectralis devices [1]. The AROI dataset has 24 subjects with neovascular AMD (nAMD) resulting in 1,136 B-scans with manual layer annotations. The imaging device is Zeiss Cirrus HD OCT 4000 [11].

We also show qualitative results for the RETOUCH dataset [2]. Even though it lacks the layer annotations that would be required for a quantitative analysis, it is still interesting due to the diverse imaging devices used (Spectralis, Cirrus and Topcon), as well as the additional difficulty of showing fluids within the retina.

No new human or animal studies were conducted; the research utilized existing datasets with all necessary ethical approvals and informed consent secured, adhering to ethical standards and regulations.

3.2 Data Preprocessing and Augmentation

All our presented models process B-scans at a resolution of 224 × 224 pixels. B-scans were resized to the respective input resolutions and normalized using the ImageNet channel means and standard deviations. Data augmentation techniques included random horizontal flipping, random vertical shifts (ensuring no layer cutoff while maintaining a uniform vertical distribution of the retina in the training data), and random cropping between 80-100% of the original image dimensions.

3.3 Model Architectures

Our baseline model is the FeatureRefinementNet (FRN), a CNN that was designed to give every output position global context to enable OCT layer segmentation through a Layer Head. The Layer Head is based on the idea that labeling regions above the top-most layer as 1 and those below each layer as 0 allows its position to be obtained as a column-wise sum. Positions of subsequent layers are obtained relative to the previous one via cumulative sums, so that constraining terms to be non-negative ensures the correct ordering [12].

Our proposed models build on pretrained ViT backbones as a feature encoder and add two more Transformer blocks as a decoder. All transformer blocks have embedding dimension 1024, 16 attention heads, and operate on 16 × 16 pixel patches. The token sequence produced by the decoder is projected to the required length and then reshaped to the image shape, similar to the self-supervised pre-training. The final OCT layer output is obtained by feeding the resulting channels in the input shape to the same Layer Head architecture that is used in the baseline FRN [12].

An overview of our chosen encoders is given in Table 1. A natural choice for the encoder is the RETFound model, which provides a ViT-Large encoder that has been pre-trained on an OCT dataset [18]. Although it was only evaluated for classification tasks so far, it has been trained as a masked auto-encoder, so it should maintain sufficiently detailed and localized information about image contents to serve as a foundation of segmentation as well. To investigate the benefit of domain-specific pre-training on OCT data, we compare results to a ViT-Large encoder that is equivalent to the one from RETFound, but has been trained on ImageNet [6]. Since we are targeting a segmentation task, our comparison also includes the encoder from the Segment Anything Model (SAM), which is trained in a supervised fashion on a large segmentation dataset with the goal to enable zero-shot generalization [8]. To investigate the relevance of model size, we evaluate the SAM encoder both in the Large and in the Base configuration. The last model we test is MedSAM, a SAM model in the Base configuration that is initialized with the SAM weights and then further trained on a large and diverse set of medical images [10]. Even though the fraction of OCT images in the training data of MedSAM is rather small, having seen a large number of medical images might still be beneficial due to shared properties such as the predominance of gray-scale images in medical imaging.

Table 1. Model comparison in terms of number of parameters and required time for a forward path with a single B-scan on an A40 GPU. We report the average time over 1000 inferences. FRN is a state of the art CNN for layer segmentation and all other models are based on pretrained Vision Transformers.

Model	**Parameters**	**Trained**	**Forward Path**
FRN [12]	391 555	391 555	$\approx$ 5 ms
MAE-L [6]	331 990 784	27 842 304	$\approx$ 38 ms
RETFound-L [18]	331 990 784	27 842 304	$\approx$ 38 ms
SAM-L [8]	331 990 784	27 842 304	$\approx$ 38 ms
SAM-B [8]	113 201 664	26 767 616	$\approx$ 20 ms
MedSAM-B [10]	113 201 664	26 767 616	$\approx$ 20 ms
SAM-L-noLoRA	330 417 920	26 269 440	$\approx$ 26 ms
SAM-B-noLoRA	112 611 840	26 177 792	$\approx$ 13 ms

We use the SAM implementation that injects additional relative positional encodings to the self attention. Since these weights are not available for the other models, we reset and train them from scratch in all models. All other parameters in the original encoder are frozen during training. Instead, we add low-rank adaptation (LoRA) [7] to all linear layers in the encoder. For this LoRA we use a rank of 4 and an initial α of 1 that decays towards earlier layers. α scales the weight matrix that is learned by the LoRA and is added to the frozen encoder weights. Hence, changing α is roughly the same as changing the learning rate [7]. To obtain comparable results for ViT variants of different sizes, we set α_i for the ith layer as

$$\alpha_i = \alpha \cdot \beta^{\left(\frac{depth-i}{depth}\cdot 24\right)} \tag{1}$$

where $\alpha = 1$ is the base value of the parameter, $\beta = 0.8$ is the effective decay rate for a transformer with 24 blocks, depth is the total number of transformer blocks, and i is the current block index. This matches the range of α to ≈ 0.006 for the initial layer and 1 for the final layer for transformers with varying depths.

3.4 Training Procedure

All models were trained for 50 epochs using the AdamW optimizer [9] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay of 0.01. For each epoch, 50,000 samples were drawn with replacement from the training data using a batch size of 32. While the ViT models were trained with a maximum learning rate of 0.0006 reached after 10 epochs, then decaying following the one-cycle policy [15], the FRN was trained without a warm-up starting at a learning rate of 0.001. All models were trained with a Sum aggregated smooth L1 loss over all layer positions. We use the Sum instead of a Mean aggregation to make a B-scan's contribution to the training proportional to its number of labeled A-scans. Early stopping is applied after 25 epochs without improvements of the validation loss and we choose the best model based on the validation loss.

Table 2. Layer segmentation results for our models building on different foundation models compared to a state of the art CNN (FRN) on the test set of the training domain. We show median and 95% quantile of the absolute layer error over all A-scans. The error is reported in pixel. All models have an input shape of 224^2.

	Control			**AMD**		
	ILM	**IBRPE**	**BM**	**ILM**	**IBRPE**	**BM**
FRN	**0.14** (**0.41**)	**0.26** (1.25)	**0.26** (**0.85**)	**0.14** (**0.43**)	**0.23** (**1.1**)	**0.32** (**1.35**)
MAE-L	0.16 (0.47)	0.28 (1.18)	0.28 (0.9)	0.16 (0.49)	0.27 (1.17)	0.35 (1.43)
RETFound-L	0.16 (0.49)	0.28 (1.2)	0.28 (0.9)	0.16 (0.51)	0.28 (1.23)	0.35 (1.39)
SAM-L	0.15 (0.44)	0.27 (**1.18**)	0.28 (0.89)	0.15 (0.46)	0.26 (1.12)	0.34 (1.45)
SAM-B	0.16 (0.49)	0.32 (1.32)	0.31 (0.98)	0.16 (0.55)	0.28 (1.24)	0.42 (2.26)
MedSAM-B	0.16 (0.51)	0.31 (1.31)	0.31 (0.97)	0.17 (0.55)	0.29 (1.26)	0.42 (2.13)
SAM-L no-LoRA	0.15 (0.47)	0.29 (1.22)	0.29 (0.91)	0.16 (0.48)	0.28 (1.2)	0.37 (1.49)
SAM-B no-LoRA	0.16 (0.5)	0.31 (1.29)	0.31 (0.97)	0.17 (0.56)	0.29 (1.29)	0.43 (2.31)

Table 3. Layer segmentation results for our models building on different foundation models compared to a state of the art CNN (FRN) on the OCT5k and AROI datasets. We show median and 95% quantile of the absolute layer error over all A-scans. The error is reported in pixel. All models have an input shape of 224^2. Inter-Reader errors for OCT5k are computed from the A-scan wise reader spread mapped to our models input resolution.

	AROI			**OCT5k**		
	ILM	**IBRPE**	**BM**	**ILM**	**IBRPE**	**BM**
FRN	0.56 (15.71)	1.44 (18.78)	1.14 (28.4)	0.64 (45.97)	0.8 (44.03)	0.7 (43.94)
MAE-L	0.23 (0.7)	0.59 (**3.4**)	0.46 (7.81)	0.3 (0.88)	**0.39** (1.37)	0.45 (1.24)
RETFound-L	0.35 (1.13)	0.74 (6.71)	0.51 (**7.04**)	0.94 (2.91)	1.01 (2.91)	0.49 (2.11)
SAM-L	**0.21** (0.66)	**0.53** (4.0)	**0.45** (8.85)	0.26 (**0.77**)	0.41 (**1.3**)	**0.38** (**1.09**)
SAM-B	0.35 (2.05)	0.93 (8.89)	0.89 (17.03)	0.45 (2.03)	0.47 (3.09)	0.4 (2.7)
MedSAM-B	0.38 (1.88)	1.07 (7.5)	0.89 (15.75)	0.44 (1.48)	0.47 (2.51)	0.39 (1.96)
SAM-L no-LORA	**0.21** (**0.65**)	0.6 (4.38)	0.49 (7.73)	**0.25** (0.93)	0.4 (1.56)	0.41 (1.21)
SAM-B no-LORA	0.42 (2.1)	0.96 (10.77)	0.99 (19.3)	0.69 (2.78)	0.54 (4.28)	0.42 (3.51)
Inter-Reader Spread	-	-	-	0.44 (1.31)	0.44 (1.75)	0.44 (1.31)

4 Results

We first consider the segmentation accuracy in domain, by computing absolute errors of predicted layer positions on the test set from the Duke data. Due to their non Gaussian distribution, Table 2 summarizes them via the median and the 95% quantile over all A-scans. In this setting, differences between the methods are minor, mostly on the order of fractions of a pixel. The smaller ViT-Base encoders from SAM-B and MedSAM-B slightly reduce the accuracy, especially when localizing Bruch's Membrane (BM) in AMD patients.

Table 3 shows corresponding results when evaluating the same segmentation models, which have been trained on the Duke dataset, on the AROI and OCT5k datasets. Significant differences become apparent between the models' ability to generalize to images that have been acquired with different devices and partly show additional, sometimes strong pathologies. The traditional FRN, which produced the best results in domain (by a small margin), now clearly provides the least robust results. In particular, high values for the 95% quantile indicate that the domain shift leads to a complete collapse of the models' abilities on a significant part of the dataset.

Our foundation model based approaches generalize substantially better, with the best results for SAM-L, closely followed by the MAE that has been trained on ImageNet. Notably, with respect to almost all numbers, RETFound's more specialized pre-training of that same encoder on OCT data reduces the resulting segmentation model's ability to generalize. Again, the smaller SAM-B is outperformed by the larger SAM-L, and differences between SAM-B and MedSAM-B remain minor, and somewhat mixed. SAM models trained without LoRA mostly perform slightly worse compared to their counterparts with LoRA, while the difference between the SAM-L and SAM-B models remains. Finally, we observe that the ranking of RETFound-L relative to the other foundation models differs between the two datasets: On AROI, it still works better than the smaller SAM-B and MedSAM-B encoders, while it has the last place in terms of median accuracies on OCT5k. A grouped analysis of the OCT5k results (Control, DME and AMD) is shown in the supplement, but did not yield further insights.

Our models also compare favorably against the inter-reader spread on the OCT5k dataset. We computed the spread as median and 95% quantile of all pairwise differences between readers. In this dataset, all readers start their annotations from the same automatic segmentation. Moreover, three iterations of outlier removals between the readers are performed. We note that, while this procedure is reasonable to find a good consensus and decrease the chance of annotation errors, it implies that estimates of inter-rater reliability are optimistic.

Qualitative results for the above-described models and datasets are shown in Figure 1. We found differences on typical cases to be small, in line with the good median accuracy of all models. Therefore, we manually selected B-scans that we expected to be challenging based on diverse degenerations. To avoid bias in favor of any particular method, the selection has been made without referring to the segmentation results. In several examples, the output of the FRN is completely unusable (rows 4 and 9), while the foundation model based approaches continue to work well, and differences between them are more subtle. Our selection also includes a case (row 12) with a strong pathology that was unseen during training, in which none of the methods yield a correct result.

Additional qualitative results for the RETOUCH dataset are shown in Figure 2. Again, challenging cases have been selected manually without referring to the segmentation results. Even though no detailed ground truth is available, several failure cases are apparent. In particular, the FRN results clearly diverge from layer positions in several cases to which the MAE-L based method successfully generalizes (rows 2, 9, 10, 11). MAE-L, RETFound-L, and SAM-L produce similar results overall; in one example with substantial differences (row 9), MAE-L yields the preferred result. In many cases (especially prominent in rows 2, 7, 9, 10, 11, 12) the less powerful ViT-Base encoders lead to obvious deviations

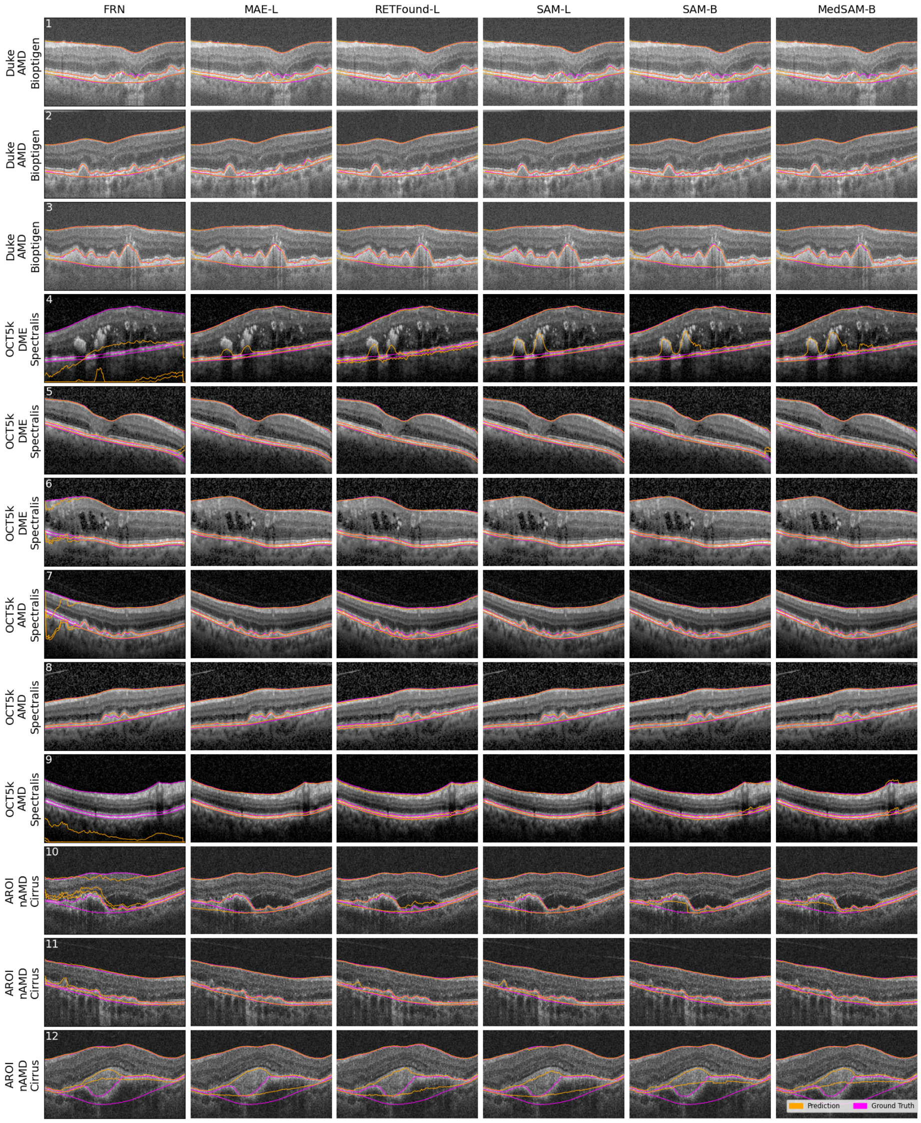

Fig. 1. Qualitative results for all models on the Duke test data (in domain) and the OCT5k and AROI datasets (domain shift). B-scans where selected manually to focus on challenging cases, but without referring to segmentation results. In domain, all models perform almost identical. Under domain shift, the performance of the FRN degrades noticeably, while differences between the other models are more subtle.

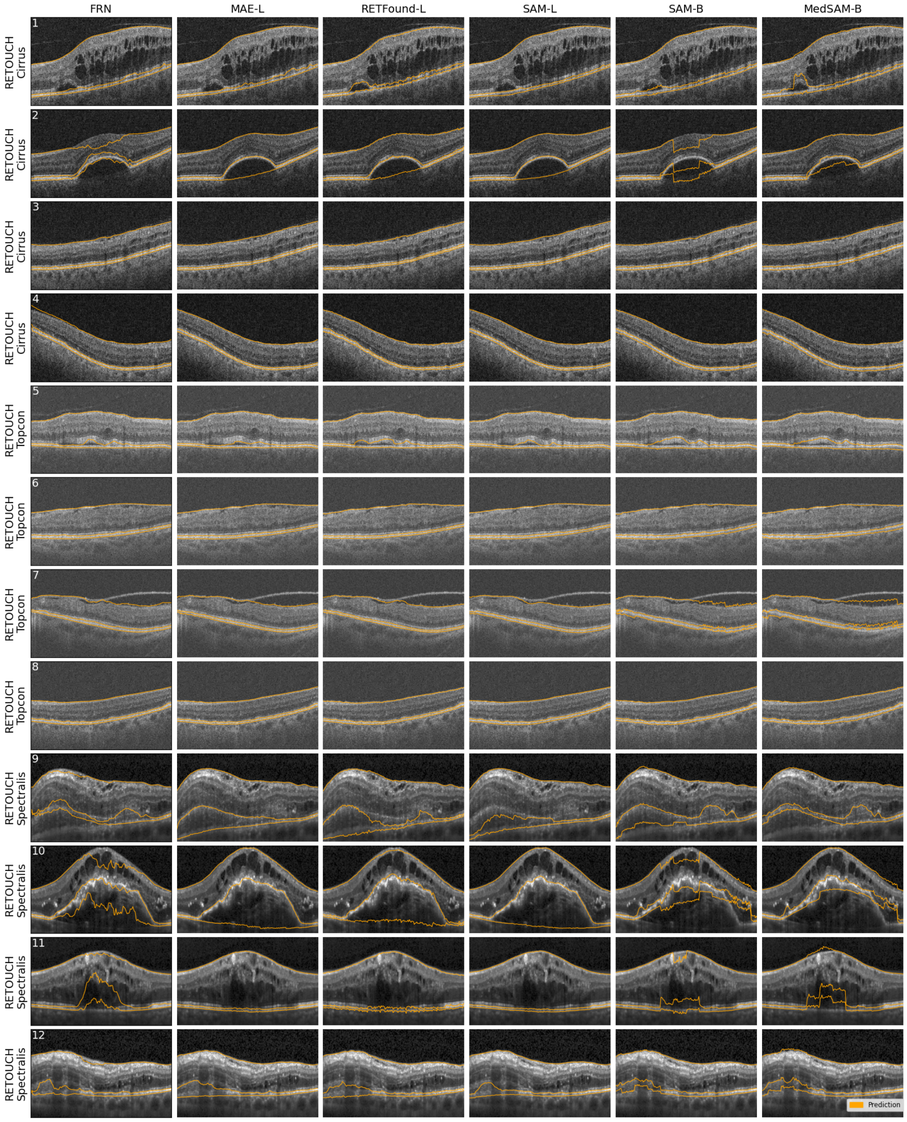

Fig. 2. Additional qualitative results for all models on the RETOUCH dataset (domain shift). Even though no annotations are available, this dataset provides insight on additional acquisition devices and pathologies.

from the true layer positions. Overall, these additional results are in line with the quantitative results on the datasets for which annotations are available.

5 Discussion

On independent test data from the same domain as the training data, all models achieved small median errors and 95% quantiles, with minor differences in segmentation accuracy between the ViT-based models and the traditional CNN baseline, the FeatureRefinementNet. This suggests that with sufficient training data, and in the absence of a domain shift, no extensive pre-training is required for the task of retinal layer segmentation in OCT.

However, an advantage of our foundation model based approach becomes evident when evaluating the methods under domain shift. On the AROI and OCT5k datasets, the ViT-based models significantly outperformed the FRN, which exhibited noticeable degradation in performance. The FRN's high values of the 95% error quantile indicate that the CNN-based model fails completely on a substantial portion of the dataset to which the ViT-based models successfully generalized. This is also apparent in the qualitative results.

Among the ViT-based models, the SAM-L and the MAE-L encoders, which have both been pre-trained on large and diverse natural image datasets, provided the best generalization. Since the image contents and characteristics of retinal OCT images are quite different from the color photographs that are used to train general foundation models in computer vision, we expected that the RETFound encoder, which has been pre-trained specifically on a large OCT dataset, would be even better suited for our task. Because limited information is available on its training data, we can only speculate why the RETFound-based model turned out to generalize less well than those based on SAM-L and MAE-L.

We believe that our results might reflect biases in the RETFound training data. Even though it contains images from different types of OCT devices [18], it might be that certain device types dominated, and others might not have been included at all. This might explain why features from a generic computer vision model which is completely agnostic to OCT empirically provided the best generalization across devices. None of the datasets that we used in our study (Duke, OCT5k, AROI) were included in the training of the RETFound OCT model [18] or MedSAM [10]. However, it is likely that there are overlaps in terms of device types, and differences in such overlaps between the training data of RETFound and the AROI or OCT5k datasets, respectively, might explain the different ranking of RETFound with respect to the other ViT-based approaches in those two cases.

In addition to differences in the training data, we note that there are also differences between the training regimes of the foundation models we compared. For example, RETFound was trained with a batch-size of 1792 and a base learning rate of 1×10^{-3}, while the MAE used a batch-size of 4096 and a base learning rate of 1×10^{-4}. Since prior work demonstrated that differences in optimizers and training hyperparameters of segmentation CNNs that make little difference when evaluating in domain can have a significant impact on across-scanner generalization [13], it is possible that such factors also play a role here.

In our current investigation, the larger SAM-L encoder consistently outperformed the smaller SAM-B, especially under domain shift. Our comparison between models trained with and without LoRA indicates that the better performance of SAM-L cannot be attributed to the larger number of LoRA parameters for the SAM-L model since the gap in performance between SAM-L and SAM-B remains also without LoRA finetuning of the encoder.

The comparison between SAM-B and MedSAM-B models yielded mixed results, suggesting that MedSAM's additional pre-training on medical images did not confer a consistent advantage for our task. This might be explained by the relatively small fraction of OCT images in MedSAM's training data.

Finally, even though the use of foundation models greatly improved generalization across scanners, there was still a noticeable decrease in accuracy compared to the evaluation in domain. Based on inspecting qualitative results, it is our impression that, in addition to small differences in annotation protocols, and potentially small remaining effects of scanner changes, pathologies that were not seen during training continue to pose an important problem that the use of currently available foundation models does not resolve.

6 Conclusion

In this study, we present a novel method for OCT layer segmentation leveraging ViT foundation models. Our approach demonstrates significant improvements in robustness and generalization compared to traditional convolutional neural networks (CNNs), particularly when faced with domain shifts due to data from devices not present during training.

Our study underscores the potential of ViT foundation models to enhance the robustness and generalizability of learning-based medical image analysis, but also highlights problems that remain despite their use, especially when facing pathologies that were not seen during training. It also illustrates that the specific choice of foundation model matters and that, somewhat surprisingly, generic models from computer vision can sometimes produce better results than more specialized foundation models whose training more closely matches the images and task at hand.

References

1. Arikan, M., Willoughby, J., Ongun, S., Sallo, F., Montesel, A., Ahmed, H., Hagag, A., Book, M., Faatz, H., Cicinelli, M.V., Fawzi, A.A., Podkowinski, D., Cilkova, M., de Almeida, D., Zouache, M., Ramsamy, G., Lilaonitkul, W., Dubis, A.M.: OCT5k: A dataset of multi-disease and multi-graded annotations for retinal layers (Mar 2023). https://doi.org/10.1101/2023.03.29.534704
2. Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., Gopinath, K., Gostar,

A.K., Jeon, K., Ji, Z., Kang, S.H., Koozekanani, D.D., Lu, D., Morley, D., Parhi, K.K., Park, H.S., Rashno, A., Sarunic, M., Shaikh, S., Sivaswamy, J., Tennakoon, R., Yadav, S., De Zanet, S., Waldstein, S.M., Gerendas, B.S., Klaver, C., Sánchez, C.I., Schmidt-Erfurth, U.: RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge. IEEE Transactions on Medical Imaging **38**(8), 1858–1874 (Aug 2019). https://doi.org/10.1109/TMI.2019.2901398

3. Cai, Z., Lin, L., He, H., Tang, X.: Uni4Eye: Unified 2D and 3D Self-supervised Pre-training via Masked Image Modeling Transformer for Ophthalmic Image Classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 88–98. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1$_9$
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (Jan 2021). https://doi.org/10.48550/arXiv.2010.11929
5. Farsiu, S., Chiu, S.J., O'Connell, R.V., Folgar, F.A., Yuan, E., Izatt, J.A., Toth, C.A.: Quantitative Classification of Eyes with and without Intermediate Age-related Macular Degeneration Using Optical Coherence Tomography. Ophthalmology **121**(1), 162–172 (Jan 2014). https://doi.org/10.1016/j.ophtha.2013.07.013
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15979–15988 (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01553
7. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. ArXiv (Jun 2021). https://doi.org/10.48550/arXiv.2106.09685
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3992–4003. IEEE, Paris, France (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00371
9. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: International Conference on Learning Representations (2019). https://doi.org/10.48550/arXiv.1711.05101
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (Jan 2024). https://doi.org/10.1038/s41467-024-44824-z
11. Melinščak, M., Radmilović, M., Vatavuk, Z., Lončarić, S.: Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation. Automatika **62**(3-4), 375–385 (Oct 2021). https://doi.org/10.1080/00051144.2021.1973298
12. Morelle, O., Wintergerst, M.W.M., Finger, R.P., Schultz, T.: Accurate drusen segmentation in optical coherence tomography via order-constrained regression of retinal layer heights. Scientific Reports **13**(1), 8162 (May 2023). https://doi.org/10.1038/s41598-023-35230-4
13. Sheikh, R., Klasen, M., Schultz, T.: Adaptive Optimization with Fewer Epochs Improves Across-Scanner Generalization of U-Net Based Medical Image Segmentation. In: Kamnitsas, K., Koch, L., Islam, M., Xu, Z., Cardoso, J., Dou, Q., Rieke,

N., Tsaftaris, S. (eds.) Domain Adaptation and Representation Transfer. pp. 119–128. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16852-9$_1$2

14. Shi, Z., Li, Y., Zou, H., Zhang, X.: TCU-Net: Transformer Embedded in Convolutional U-Shaped Network for Retinal Vessel Segmentation. Sensors **23**(10), 4897 (Jan 2023). https://doi.org/10.3390/s23104897
15. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay (Apr 2018). https://doi.org/10.48550/arXiv.1803.09820
16. Zhang, H., Yang, J., Zheng, C., Zhao, S., Zhang, A.: Annotation-efficient learning for OCT segmentation. Biomedical Optics Express **14**(7), 3294–3307 (Jul 2023). https://doi.org/10.1364/BOE.486276
17. Zhang, Y., Li, Z., Nan, N., Wang, X.: TranSegNet: Hybrid CNN-Vision Transformers Encoder for Retina Segmentation of Optical Coherence Tomography. Life **13**(4), 976 (Apr 2023). https://doi.org/10.3390/life13040976
18. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A.Y., Topol, E.J., Denniston, A.K., Alexander, D.C., Keane, P.A.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (Oct 2023). https://doi.org/10.1038/s41586-023-06555-x