Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

SAMURAI: A Two-Stage Foundation Model Pipeline for Robust Optic Nerve Head Segmentation in Fundus Images

Carlos Perez[†,*], Neeru Gupta[†], Ipek Oruc[†]
[†] Department of Ophthalmology and Visual Sciences, University of British Columbia

Abstract

Accurate segmentation of the optic nerve head (ONH) is essential for automated glaucoma assessment using the Cup-to-Disc Ratio (CDR). However, conventional convolutional neural networks (CNNs) often exhibit performance degradation under domain shift caused by variations in fundus imaging devices and protocols. Foundation models offer a potential solution due to their large-scale pre-training and intrinsic feature invariance. While the Segment Anything Model (SAM) offers a robust alternative, recent adaptations have resorted to complex, task-specific architectural modifications to handle retinal geometry. In this paper, we propose SAMURAI, a two-stage foundation model pipeline that combines a YOLOv12x-based ONH localizer with a minimally adapted MedSAM foundation model. We rigorously evaluate this supervised baseline against exploratory variants incorporating geometric inductive biases (polar transformations) and semi-supervised learning (SSL). On the REFUGE benchmark, our simplified approach establishes a new state-of-the-art, achieving an Optic Cup Dice of 0.920, significantly outperforming specialized models like FunduSAM (0.867). Furthermore, our ablation study reveals that additional architectural complexity does not confer measurable performance gains over the foundation baseline. These findings suggest that large-scale pre-trained foundation models provide sufficient robustness for ONH segmentation without task-specific architectural modifications.

Keywords: Medical Image Analysis, Glaucoma, Segment Anything Model

1. Introduction

Glaucoma is a leading cause of irreversible blindness worldwide [1]. The primary structural indicator of damage is neuroretinal rim thinning, quantified clinically as an increase in the vertical cup-to-disc Ratio (CDR). Automated calculation of CDR from fundus photography requires precise pixel-level segmentation of the optic disc (OD) and the optic cup (OC). State-of-the-art approaches for optic nerve segmentation have predominantly relied on Convolutional Neural Networks (CNNs), particularly U-Net [2] variants, such as M-Net [3] and CE-Net [4]. While these models achieve high accuracy on homogeneous datasets, they exhibit significant performance degradation when deployed in real-world clinical settings. This fragility stems from domain shifts caused by variations in fundus camera hardware, illumination, and patient populations, as demonstrated in the REFUGE challenge [5].

The recent advent of the Segment Anything Model (SAM) [6] offers a potential solution to this generalization bottleneck. In contrast to traditional CNNs, which are typically constrained by the limited size and variability of task-specific medical datasets, SAM utilizes a Vision Transformer (ViT) backbone pre-trained on the massive SA-1B dataset (11 million images, 1.1 billion masks). This exposure to diverse visual features allows the model to learn generalized, class-agnostic representations of object boundaries that are less dependent on local texture statistics [7]. MedSAM [8] adapts this foundation to the medical domain. However, it remains an open question whether standard fine-tuning suffices for the complex geometry of the optic nerve head, or whether specialized domain-adaptive architectures are necessary.

[*] cperez67@student.ubc.ca

We introduce SAMURAI, a modular two-stage pipeline based on MedSAM whose primary contribution is demonstrating that minimal decoder fine-tuning of a foundation model, combined with a purpose-built YOLO localizer, is sufficient to surpass specialized retinal architectures—without complex adapters, polar constraints, or domain-specific modules. Beyond standard fine-tuning, we rigorously evaluate variants incorporating a geometric inductive bias and semi-supervised learning (SSL). Benchmarking against U-Net, M-Net, and FunduSAM reveals that our Supervised Baseline sets a new state-of-the-art. Interestingly, the exploratory extensions did not yield consistent performance gains, suggesting that the intrinsic domain invariance of the foundation model may reduce the need for additional task-specific adaptation techniques.

1.1. **CNNs and Domain Shift in Retinal Segmentation**

Since its inception, the U-Net architecture [2] has firmly established itself as the de facto backbone for medical image segmentation [9]. In ophthalmology, researchers have developed specialized variants to address the nested geometry of the optic nerve head. For instance, M-Net [3] utilizes polar coordinate transformations to linearize the circular optic disc. However, fully supervised CNNs are highly susceptible to covariate shift, where variations in camera spectral sensitivity or illumination cause performance degradation on unseen domains [5]. While Domain Adaptation (DA) techniques like Boundary Entropy Adversarial Learning (BEAL) [10] address this via explicit statistical alignment, they introduce significant architectural complexity. This motivates the search for a modular, generalizable pipeline capable of robust performance without task-specific architectural over-engineering.

1.2. **Foundation Models and Adapters**

MedSAM [8] represents a paradigm shift, adapting the Segment Anything Model (SAM) by fine-tuning on a large-scale dataset of over 1.5 million image-mask pairs spanning 10 imaging modalities and more than 30 cancer types. While this dataset includes diverse targets such as abdominal organs (CT/MRI), cellular structures (microscopy), and endoscopic views, retinal fundus photography represented only a small proportion of the training data. Consequently, while MedSAM possesses strong general semantic understanding, it may lack the specific boundary precision required for the low-contrast optic cup within the optic nerve head.

Recent work such as FunduSAM [11] attempted to bridge this gap by introducing complex, multi-scale adapter modules specifically tailored for retinal features. In contrast, our work investigates the efficacy of a streamlined approach. By coupling a state-of-the-art localizer (YOLOv12x) [12], implemented with the Ultralytics software package [13], with a minimally adapted MedSAM baseline, we aim to establish a universal, two-stage pipeline for fundus segmentation. We posit that a streamlined MedSAM baseline provides a more scalable clinical AI solution: a single adaptable architecture retrained for diverse ophthalmic tasks without task-specific geometric modules.

2. **Methodology**

The proposed pipeline is structured into a two-stage framework: Stage 1 utilizes YOLOv12x to localize the optic nerve head, and Stage 2 employs the SAMURAI segmentation model. We also investigate two exploratory extensions (SAMURAI-A and SAMURAI-B).

2.1. **Stage 1: Optic Disc Localization (YOLOv12x)**

We utilize YOLOv12x [12], via the Ultralytics framework [13], to localize the optic disc. We adapted the model by generating ground-truth bounding boxes from the minimal enclosing rectangle around the union of the optic disc and cup masks.

To ensure the tightest possible crop for the subsequent segmentation stage, we replaced YOLO's standard mean Average Precision (mAP) model selection with a custom fitness stopping criterion based on bounding box Intersection over Union (IoU):

$$F = \text{IoU}(B_p, B_g) = \frac{\text{Area}(B_p \cap B_g)}{\text{Area}(B_p \cup B_g)} \tag{2.1}$$

where B_p and B_g are the predicted and ground-truth boxes. We maximized this fitness score on the validation set for early stopping. At inference, we retain only the single highest-confidence detection per image. To prevent processing severe artifacts, predictions below a 0.10 confidence threshold are discarded, yielding a null mask.

Stage 1 is critical: by providing a tightly cropped region of interest, it reduces the input distribution seen by MedSAM to a near-canonical view of the ONH, substantially lowering the segmentation difficulty and partially explaining why additional domain-adaptation techniques proved unnecessary.

2.2. **Stage 2: SAMURAI Segmentation (Baseline)**

We adapt the MedSAM foundation model [8] by freezing its Vision Transformer (ViT-B) image encoder — a deliberate departure from MedSAM's original protocol, which fine-tunes both the encoder and decoder — to preserve pre-trained feature extraction capabilities while fine-tuning only the lightweight mask decoder. We supervise the segmentation heads using an unweighted combination of Binary Cross-Entropy (BCE) and Dice Loss, matching MedSAM's original loss formulation. This minimal adaptation strategy retains foundation model generalization without task-specific architectural changes.

2.3. **Stage 3: Exploratory Variants**

SAMURAI-A (Semi-Supervised Learning): We implement a self-training framework leveraging unlabeled EyePACS data. The iterative process entails: (1) training the YOLOv12x teacher on labeled source data using the Box IoU criterion (exclusive to the object detector); (2) generating pseudo-labels on unlabeled images using a strict confidence threshold (> 0.50); and (3) fine-tuning a student model on the pseudo-labeled dataset.

SAMURAI-B (Geometric & Attention Components): This variant integrates FunduSAM [11] components to address topological variability. We apply a Linear Polar Transformation post-localization to map Cartesian (x, y) to Polar (r, θ) coordinates, unwrapping the optic nerve into a layered linear structure. Additionally, a CBAM Attention module [14] is inserted into the bottleneck to refine feature maps. To enforce anatomical consistency, we add a Containment Loss term to the objective function:

$$\mathcal{L}_{total} = \lambda_d \mathcal{L}_{disc} + \lambda_c \mathcal{L}_{cup} + \lambda_{topo} \mathcal{L}_{contain} \tag{2.2}$$

where $\mathcal{L}_{contain}$ strictly penalizes cup pixels (P_{cup}) predicted outside the disc (P_{disc}):

$$\mathcal{L}_{contain} = \frac{1}{N} \sum_i P_{cup}^{(i)} (1 - P_{disc}^{(i)}) \tag{2.3}$$

$\lambda_d = 1.0$, $\lambda_c = 2.0$, and $\lambda_{topo} = 1.0$ to prioritize cup boundary and structural validity.

3. **Datasets**

We aggregated 9 diverse public datasets (Table 1) covering varied fundus cameras, resolutions, and pathological severities, partitioned at the image level into training (80%),

validation (10%), and testing (10%) subsets. REFUGE test set images were strictly held out to provide a zero-shot standardized benchmark against prior methods.

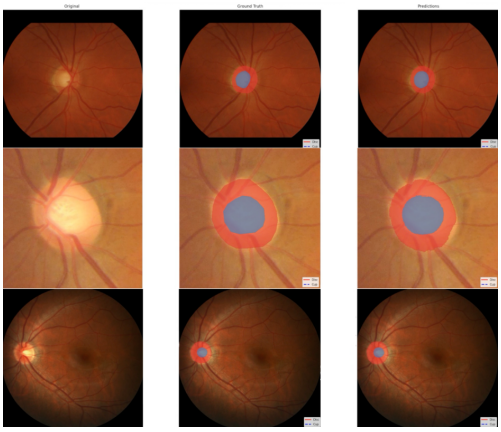

Figure 1. Original, GT, Pred (left, center, right columns). REFUGE, GRAPE, and PAPILA datasets (top, middle, bottom rows).

Table 1. Summary of the annotated datasets used in this study.

Dataset	Images
REFUGE / REFUGE2 [5, 15]	2000
CHAKSU [16]	1343
G1020 [17]	1020
RIGA [18]	749
ORIGA [19]	650
RIM-ONE DL [20]	485
DRISHTI [21]	101
GRAPE [22]	630
PAPILA [23]	488

4. **Results**

We evaluated our pipeline on the REFUGE test set to provide a standardized comparison against state-of-the-art methods. Table 2 summarizes the Dice coefficients for optic disc (OD) and optic cup (OC) segmentation.

We benchmarked our approach against the comprehensive evaluation recently reported by Yu et al. [11], which includes both Convolutional (ResUNet [24], nnU-Net [25]) and Transformer-based (TransUNet [26], Swin-UNETR [27]) architectures; baseline Dice values for these methods are as re-implemented and reported by Yu et al. Our Supervised Baseline achieved an Optic Cup Dice of 0.920. This performance represents an improvement over the recently published FunduSAM (0.867) and the nnU-Net baseline (0.849) under a comparable evaluation protocol. By leveraging the large-scale pre-training of the MedSAM foundation model, our approach surpasses these task-specific architectures without requiring complex attention modules or adapter tuning.

Table 2. Comparison with REFUGE Challenge Winners and Recent SOTA.

Method	Disc Dice	Cup Dice
REFUGE Challenge Winners [5]		
Team CUHKMED (DenseNet)	0.960	0.883
Team Masker (ResNet)	0.946	0.884
Alternative Architectures		
Mask R-CNN (Wu et al.) [28]	0.962	0.887
Deep Level Set (Liu et al.) [29]	0.966	0.891
Foundation Models		
FunduSAM [11]	0.961	0.867
SAMURAI	**0.966**	**0.920**

Table 3. Ablation study of SAMURAI components with Semi-Supervised Learning (A) and Polar Transformation (B).

Method	REFUGE TEST Disc	Cup
SAMURAI	**0.966**	**0.920**
+ [A] SSL (Teacher-Student)	0.958	0.910
+ [B] Polar Transform	0.930	0.820

4.1. **Component Ablation Study**

To assess the contribution of our specific architectural modifications, we conducted a rigorous ablation study evaluating the impact of Semi-Supervised Learning (SSL) and Geometric Polar Transformations. We compared the performance of SAMURAI against variants incorporating these modules individually and in combination.

As shown in Table 3, neither SSL nor the polar transformation improved upon the baseline; the polar variant degraded performance substantially (OD Dice 0.930, OC Dice 0.820). SSL likely failed because pseudo-labels introduced noise that compounded uncertainty at the cup boundary. These results suggest that the foundation model's large-scale pre-training already provides sufficient domain invariance, with limited benefit from additional task-specific adaptation, though generalization to pathological images remains an open question.

Acknowledgements

This work was supported by the Department of Ophthalmology and Visual Sciences and through computational resources and services provided by Advanced Research Computing at the University of British Columbia. The authors thank Simrat Binning, Parsa Delavari, and Athavan Gananathan. We gratefully acknowledge the creators and maintainers of the public datasets used in this study. AI tools were used for limited research and editorial support.

References

[1] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng. "Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis". In: *Ophthalmology* 121.11 (2014), pp. 2081–2090.

[2] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer. 2015, pp. 234–241.

[3] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao. "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation". In: *IEEE Transactions on Medical Imaging* 37.7 (2018), pp. 1597–1605.

[4] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu. "CE-Net: Context encoder network for 2d medical image segmentation". In: *IEEE Transactions on Medical Imaging* 38.10 (2019), pp. 2281–2292.

[5] J. I. Orlando, H. Fu, J. B. Breda, et al. "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs". In: *Medical Image Analysis* 59 (2020), p. 101570.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[7] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Kocak, B. P. Kelley, H. Chen, and P.-H. Lo. "Segment anything model for medical image analysis: an experimental study". In: *Medical Image Analysis* 89 (2023), p. 102918.

[8] J. Ma, Y. He, F. Li, et al. "Segment anything in medical images". In: *Nature Communications* 15.1 (2024), p. 654.

[9] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. "U-Net and its variants for medical image segmentation: A review of theory and applications". In: *IEEE Access* 9 (2021), pp. 82031–82057.

[10] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng. "BEAL: Boundary entropy adversarial learning for domain adaptation in semantic segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 74–82.

[11] J. Yu et al. "FunduSAM: A Specialized Deep Learning Model for Enhanced Optic Disc and Cup Segmentation in Fundus Images". In: *arXiv preprint arXiv:2502.06220* (2025).

[12] Y. Tian, Q. Ye, and D. Doermann. "YOLOv12: Attention-Centric Real-Time Object Detectors". In: *arXiv preprint arXiv:2502.12524* (2025).

[13] G. Jocher, A. Chaurasia, and J. Qiu. *Ultralytics YOLO*. Version 8.0.0. Accessed: 2026-02-08. 2023. URL: https://github.com/ultralytics/ultralytics.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. "CBAM: Convolutional block attention module". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.

[15] H. Fu, J. I. Orlando, J. Barbosa Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, et al. "REFUGE2: A Large-scale Retinal Fundus Image Database for Glaucoma Screening". In: *Proceedings of MICCAI 2020 Ophthalmic Medical Image Analysis Workshop*. 2020, pp. 1–8.

[16] J. R. H. Kumar, C. S. Seelamantula, J. H. Gagan, Y. S. Kamath, N. I. R. Kuzhuppilly, U Vivekanand, P. Gupta, and S. Patil. "Chákṣu: A glaucoma specific fundus image database". In: *Scientific Data* 10.1 (2023), p. 70.

[17] M. N. Bajwa, G. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed. "G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–7.

[18] A. Almazroa, S. Alodhayb, E. A. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan. "Retinal fundus images for glaucoma analysis: the RIGA dataset". In: *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 10579. SPIE. 2018, 105790B.

[19] Z. Zhang, F. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong. "ORIGA-light: An online retinal fundus image database for glaucoma analysis and research". In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE. 2010, pp. 3065–3068.

[20] F. J. Fumero, J. Sigut, S. Alayón, M. Gonzalez-Hernandez, and M. Gonzalez de la Rosa. "RIM-ONE DL: A Unified Retinal Image Database for Assessing Glaucoma Using Deep Learning". In: *Image Analysis & Stereology* 39.3 (2020), pp. 161–167. DOI: 10.5566/ias.2346.

[21] J. Sivaswamy, S. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish. "Drishti-GS: Retinal image dataset for optic nerve head (ONH) segmentation". In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI 2014)*. IEEE. 2014, pp. 53–56.

[22] X. Huang, X. Kong, Z. Shen, J. Ouyang, Y. Li, K. Jin, and J. Ye. "GRAPE: A multi-modal dataset of longitudinal follow-up visual field and fundus images for glaucoma management". In: *Scientific Data* 10.1 (2023), p. 520.

[23] O. Kovalyk, J. Morales-Sánchez, R. Verdú-Monedero, I. Sellés-Navarro, A. Palazón-Cabanes, and J.-L. Sancho-Gómez. "PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment". In: *Scientific Data* 9.1 (2022), p. 291.

[24] X. Xiao, S. Lian, Z. Luo, and S. Li. "Deep ResU-Net for segmentation of retinal blood vessels". In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2018, pp. 3371–3376.

[25] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (2021), pp. 203–211.

[26] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[27] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu. "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images". In: *International MICCAI Brainlesion Workshop*. Springer. 2022, pp. 272–284.

[28] J. Wu and L. Chiariglione. "Automated Optic Disc and Cup Segmentation for Glaucoma Detection from Fundus Images Using the Detectron2's Mask R-CNN". In: *Proceedings of the IEEE International Conference on E-health Networking, Application & Services (HealthCom)* (2022), pp. 1–6.

[29] B. Liu et al. "Deep level set method for optic disc and cup segmentation on fundus images". In: *Biomedical Optics Express* 12.11 (2021), pp. 6969–6983.

Appendix A. Hardware and Software

Experiments ran on UBC Sockeye (NVIDIA A100, 40 GB) using Python 3.10, PyTorch 2.1, Ultralytics 8.x, and the official MedSAM repository.