Enhancing Glaucoma Detection with Deep Learning: Segmentation of Optic Nerve Features from Low Quality Images

Anjali Goel*, Dan Luettgen*, Trenton Slocum*, Hilde Younce*, Arjun J Dirghangi, MD, MHS†
*School of Data Science
†School of Medicine, Department of Ophthalmology
University of Virginia, Charlottesville, VA, USA
Emails: arg8qqv@virginia.edu, skk8kc@virginia.edu, nuf8ms@virginia.edu, ksg8xy@virginia.edu

***Abstract*—Glaucoma is a leading cause of irreversible blindness worldwide, with early detection critical for effective intervention. However, access to reliable screening remains limited in many under-resourced settings due to reliance on high-quality imaging and specialist interpretation. This study introduces a deep learning-based system built around an enhanced U-Net architecture for segmenting optic nerve head features from both high-resolution fundus photographs and low-quality clinical video ophthalmoscopy frames. The model uses an encoder–decoder architecture to improve robustness under challenging imaging conditions, such as glare and motion blur. Trained on the ORIGA dataset and evaluated on real-world clinical images collected from high-risk patients, the U-Net model demonstrates strong segmentation performance, particularly when optimized with a hybrid loss function combining binary cross-entropy and Dice loss. Comparative evaluation with a promptable Medical SAM-2 model highlights the U-Net's superior reliability in maintaining accurate optic disc and cup boundaries, especially for the more subtle optic cup region. This work underscores the potential of lightweight, interpretable U-Net-based models to support scalable, equitable glaucoma screening in diverse clinical environments, laying the groundwork for future real-time deployment and integration with diagnostic scoring systems like DDLS.**

I. Introduction

Glaucoma remains one of the leading global causes of irreversible blindness, predominantly impacting populations with limited healthcare access. Early and accurate detection is crucial for effective treatment and prevention of vision loss. However, current glaucoma screening practices often require high-quality imaging equipment and specialist interpretation, creating barriers for under-served communities. Consequently, there is an urgent need for innovative, accessible solutions that can function effectively even with lower-quality imaging modalities.

This paper explores the use of deep learning techniques to automate glaucoma screening through the segmentation of optic nerve head features from video ophthalmoscopy images. Our approach aims to democratize access to glaucoma screening by overcoming traditional limitations associated with image quality. By utilizing a combination of publicly available datasets and clinical data provided by the University of Virginia Health System, we develop and validate models based on architectures including U-Net and Medical-SAM2. The project's ultimate goal is to create a robust, generalizable machine learning system capable of providing real-time assessments of glaucoma risk during routine eye examinations, significantly reducing health disparities associated with glaucoma diagnosis and management.

II. Background and Motivation

A. Related Works

An influential study by Sudhan et al. (2022) implemented a U-Net-based deep learning model to segment the optic cup and disc using the ORIGA dataset, achieving strong performance with a Dice coefficient of 98.42% and a test accuracy of 96.90% [1]. Their work incorporated a DenseNet-201 model for feature extraction and a DCNN for classification, demonstrating the viability of combining semantic segmentation with transfer learning to support early diagnosis of glaucoma from high-resolution fundus images. Although the study presents a robust pipeline for automated classification, it is primarily limited to high-quality static images and does not explicitly address generalizability across variable imaging conditions.

Another line of recent work that informed our approach is Medical SAM-2, proposed by Zhu et al. (2024), which builds on the Segment Anything Model 2 (SAM2) to address segmentation challenges in medical imaging [2]. Their model re-frames segmentation as a tracking problem, using a self-sorting memory bank to store embeddings from high-confidence frames and generalize across an image sequence. Notably, Medical SAM-2 introduces a "One-Prompt Segmentation" paradigm, enabling inference across multiple frames based on a single bounding box prompt—ideal for applications with limited labeling capacity. Inspired by this, we adapted the Medical SAM-2 framework to segment fundus video frames by prompting the model with bounding boxes around the optic nerve head. This direction is especially promising in clinical settings where glare, motion blur, and poor contrast frequently degrade image quality.

B. Glaucoma and the Need for Automation

Glaucoma is a leading cause of irreversible blindness globally, affecting 80 million in 2020 with projected increases to 112 million by 2040 [3]. The disease is often asymptomatic in early stages, resulting in delayed diagnosis and significant vision loss before treatment begins. In many parts of the world - particularly low-resource and rural areas - access to timely and affordable screening remains a major challenge. While technologies like optical coherence tomography and fundus photography are widely used in clinical settings, their cost and reliance on trained personnel have severe limits on their reach. In this context, there is growing interest in leveraging artificial intelligence and deep learning to build scalable, automated screening systems that can enhance early detection and reduce global health disparities.

C. AI-Based Approaches

AI has already made significant progress in ophthalmology, particularly in automating the detection of retinal diseases. In the case of glaucoma, research has explored both one-step and two-step AI frameworks. One-step models use end-to-end deep learning pipelines to classify images directly, often with limited interpretability. In contrast, two-step approaches first segment key anatomical structures (in this case, the optic disk and optic cup) then apply classification rules or machine learning models to determine the presence of disease. Two-step approaches offer several advantages, such as requiring less data, supporting greater model interpretability, and enabling clinicians to visually inspect segmentation outputs, facilitation integration into clinical workflows [4].

D. Optic Disk and Cup Segmentation Using Deep Learning

A large body of research has focused on segmenting the optic nerve head using deep learning techniques. U-Net, a convolutional neural network, has become a popular choice due to its ability to capture contextual information while maintaining high spatial precision. Prastyo and Sumi used U-Net on 650 fundus images from the ORIGA dataset, achieving a Dice coefficient of 98.42%, indicating a nearly perfect agreement between the predictions and the ground truth [5]. Kako et al. developed a multi-label U-Net (MSU-Net and BU-Net) for segmenting the disk and cup, as well as blood vessels and peripapillary atrophy zones. Their model, trained on the HRF dataset, showed strong results with 87% IoU and 86.9% F1 Score, demonstrating the utility of multi-class segmentation [6].

E. E. Automated Classification Systems

In addition to segmentation approaches, several studies have employed deep learning models for direct classification of glaucoma risk. These models typically operate on cropped optic disc regions and output a probability score indicating disease presence. For example, end-to-end CNN architectures such as ResNet, DenseNet, and Inception have been used with varying degrees of success, often requiring large annotated datasets for training. While classification-only methods can offer faster inference, they generally lack the interpretability of segmentation-based pipelines and may be less robust across variable imaging conditions. This motivates the use of two-step systems that combine segmentation and classification to balance accuracy with clinical transparency.

F. Limitations in Current Research

Several limitations remain in the current body of research. Most existing models are trained on high-resolution still fundus photographs captured under ideal clinical conditions. These settings fail to capture the variability introduced by mobile imaging, inconsistent lighting, and motion artifacts. Additionally, many studies lack external validation across device types, raising concerns about generalizability. Few models are designed to operate under the constraints of low-cost clinical workflows, where image quality and available computational resources are limited [4].

G. Our Contribution

In response to these gaps, our research aims to develop a novel AI-based system that automates glaucoma screening using video ophthalmoscopy and an enhanced U-Net segmentation framework. Unlike most prior work, our model is tested on both high-quality fundus photographs and low-quality clinical video frames, enabling broader applicability. By incorporating spatial and channel-wise attention mechanisms, our architecture is designed to remain robust under suboptimal imaging conditions. This project contributes toward scalable, generalizable screening tools that can support earlier glaucoma detection in community and primary care settings, ultimately helping to reduce inequities in eye health outcomes.

III. Methodology

A. Data Description

This study utilizes two complementary datasets: a high-quality open-source dataset, and a lower-quality real-world clinical dataset captured using handheld video ophthalmoscopy.

1) ORIGA Dataset: The first dataset is the publicly available Optic Disc and Rim Image Database for Glaucoma Analysis (ORIGA), which consists of 650 high-resolution fundus images collected as part of the Singapore Chinese Eye Studies (SCES). Each image has been expertly annotated for optic disc and cup segmentation by trained ophthalmologists, providing pixel-level ground truth [5]. The dataset is widely used for glaucoma detection tasks and serves as a benchmark for training and evaluating deep learning-based segmentation models.

2) Clinical Video Ophthalmoscopy Images: The second dataset comprises a series of de-identified fundus video recordings captured from high-risk glaucoma patients at the University of Virginia Health System, using imaging collected via a custom camera developed by Dr. Dirghangi's clinic. The device attaches to an ophthalmologist's binocular indirect ophthalmoscope, enabling low-cost, clinic-friendly

capture of retinal videos. From each video, individual frames were manually extracted when the optic nerve head was clearly visible, allowing us to build a sizable dataset despite a limited number of original recordings. These images exhibit wide variation in resolution, focus, and lighting—conditions that more accurately reflect the environments we aim to optimize our models for in future deployments.

B. Data Preprocessing

Prior to training, both the retinal fundus images and corresponding segmentation masks undergo a series of preprocessing steps to ensure consistency in format and size. Each retinal fundus image is loaded from file, converted to RGB format, and transformed into a PyTorch tensor. The images are then resized to a fixed spatial resolution using bilinear interpolation, to preserve image smoothness and structure during rescaling. The images are resized to 256×256 and 512×512 for the U-Net and Medical SAM-2 models, respectively.

The corresponding ground truth masks are loaded as grayscale arrays. Class labels are encoded based on pixel values, with the optic disc defined as any positive value, and the optic cup as values greater than one. Binary masks are then created for each class (OD and OC), converted to PyTorch tensors with a single channel, and resized using nearest neighbor interpolation to preserve discrete class labels. The masks are resized to 256×256 pixels for the U-Net and 512×512 pixels for Medical SAM-2 model.

C. Model Architectures

We evaluate two segmentation pipelines for identifying the optic disc and cup in fundus images: an enhanced U-Net model, and a promptable Medical-SAM2 model. Both are designed to support downstream computation of clinically relevant metrics such as Cup-to-Disc Ratio (CDR) and Disc Damage Likelihood Scale (DDLS).

1) U-Net: The U-Net architecture is a convolutional neural network widely used for segmentation in medical imaging. We implement a U-Net-based model [Figure 1] to segment the optic disc and optic cup from retinal fundus images. The network takes an RGB image as input and outputs two binary masks—one for the optic disc and one for the optic cup.

The architecture follows the standard U-Net structure, comprising a contracting path (encoder), a bottleneck, and an expanding path (decoder), with skip connections at each resolution level. The encoder consists of three blocks, each with two 3×3 convolutional layers (with padding and ReLU activations) followed by max pooling to downsample spatial dimensions and increase feature depth. Feature maps before pooling are stored for skip connections.

The bottleneck contains two 3×3 convolutional layers with ReLU activations. The decoder mirrors the encoder and consists of three blocks, each beginning with a 2×2 transposed convolution for upsampling, followed by concatenation with the corresponding encoder features, dropout, and two 3×3 convolutions with ReLU activations. A final 1×1 convolution reduces the output to two channels, followed by a sigmoid activation to produce pixel-wise probability maps for the target masks [1].

2) Medical SAM-2 with Bounding Box Prompts.: To address the challenges posed by glare, motion blur, and poor contrast in clinical video ophthalmoscopy, we also explore a parallel segmentation pipeline using Medical SAM-2—a foundation model adapted from the Segment Anything Model 2 and extended for biomedical imaging by Zhu et al. (2024) [2]. The key motivation for integrating Medical SAM-2 into our pipeline is its support for bounding-box-based prompting, which allows us to explicitly guide the model's attention to the optic nerve head region.

In our implementation, we manually draw a bounding box around the optic disc on a single frame of a fundus video. Medical SAM-2 then leverages its "One-Prompt Segmentation" capability to segment subsequent frames without further intervention, using a self-sorting memory mechanism to propagate reliable visual features. This prompting approach enables more robust segmentation in frames with heavy artifacts—where traditional models may struggle to identify structure without guidance.

Unlike our U-Net model, which is trained on labeled data and learns fixed spatial priors, Medical SAM-2 is used in a zero-shot inference setting. It offers a flexible, lightweight alternative for rapid deployment in noisy or low-resource environments. This approach is being investigated as a potential strategy for improving segmentation performance in real-world conditions where traditional models might fail without targeted spatial guidance.

D. Model Training

1) U-Net: Several iterations of the U-Net model were trained over 60 epochs using the Adam optimizer with a learning rate of 1×10^{-4}. Model performance was evaluated using the Dice coefficient [Equation 1], a metric that quantifies the spatial overlap between predicted and ground truth segmentation masks. The Dice coefficient penalizes both false negatives (missed relevant pixels) and false positives (included irrelevant pixels), making it particularly suitable for medical image segmentation tasks where precision is critical. In this context, it ensures that the model accurately segments both the larger optic disc and the comparatively smaller optic cup regions.

$$\text{Dice}(Predicted, Truth) = \frac{2|Predicted \cap Truth|}{|Predicted| + |Truth|} \quad (1)$$

The first model was trained on the ORIGA dataset and optimized using binary cross-entropy (BCE) loss with mean reduction. The second model employed a hybrid loss function consisting of a weighted sum of BCE and Dice loss [Equation 2]. This combination allows BCE to guide the model at the pixel level, while Dice emphasizes the overall shape and boundary accuracy of the segmented regions.

$$\text{Loss} = \alpha \cdot \text{BCE} + (1 - \alpha) \cdot \text{Dice} \quad (2)$$

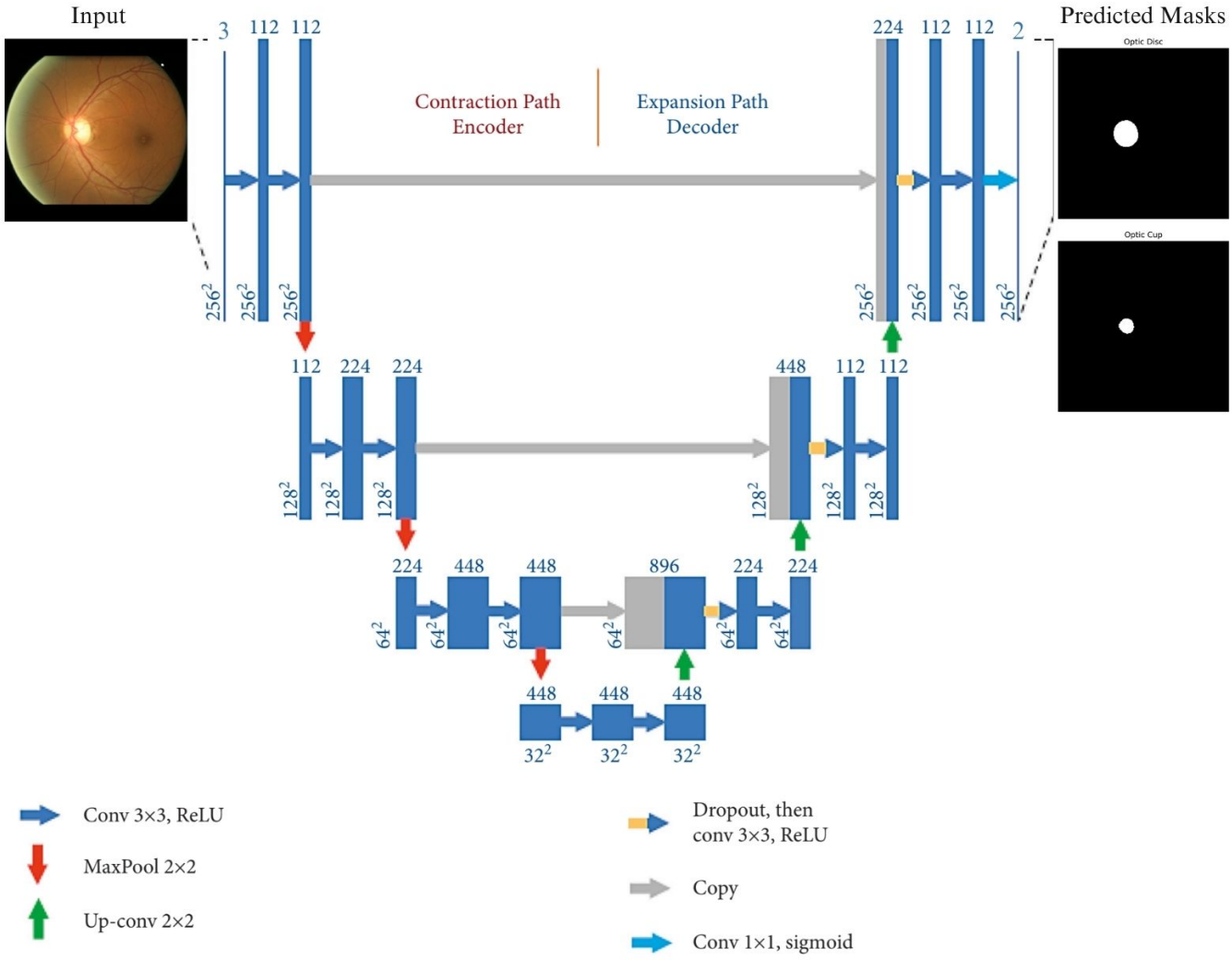

Fig. 1: The proposed U-Net architecture [1].

E. Medical SAM-2

The Medical SAM-2 model was trained on the ORIGA dataset using a weighted BCE loss. The model architecture was altered to force dual-mask prediction. For a single image, the model predicts two masks. In training, the first mask is always considered the optic disc and the second the optic cup. Loss is calculated separately against each mask and combined using a weighted sum with λ set to 0.3:

$$\text{Loss}_{\text{total}} = \lambda \cdot \text{Loss}_{\text{disc}} + (1 - \lambda) \cdot \text{Loss}_{\text{cup}} \tag{3}$$

Using pre-trained weights from the SAM-2 small model, fine-tuning was performed against all layers of our altered Medical SAM-2 model. Training for a maximum of 50 epochs using the Adam optimizer with a learning rate of 1×10^{-4}, using early stopping with a patience of 5, yields a model with a test loss of .012 and a Dice score of .8219.

IV. Conclusions

A. Results

1) Performance Analysis: Model performance was evaluated using the Dice coefficient. Two loss functions were compared: standard BCE and a hybrid loss function that combines BCE with Dice loss. As shown in Table I, the model trained with BCE loss achieved a training loss of 0.0043 and a test loss of 0.0046, with corresponding Dice scores of 0.8718 and 0.8745. In comparison, the model trained with a hybrid loss function (a weighted sum of BCE and Dice) showed slightly higher training (0.0621) and test (0.0547) losses but improved Dice scores of 0.8848 and 0.8989. These results indicate that while BCE provides stable optimization, the hybrid loss enhances segmentation accuracy, particularly in capturing finer structures such as the optic cup.

2) Predicted Masks: Figure 2 presents a visual comparison between the ground truth segmentation masks and

Metric	Model	Training	Test
BCE Loss	BCE-only	0.0043	0.0046
	Hybrid	0.0621	0.0547
Dice Coefficient	BCE-only	0.8718	0.8745
	Hybrid	0.8848	0.8989

TABLE I: U-Net training and test results for BCE-only and Hybrid loss models.

the predictions generated by the U-Net model trained with the hybrid loss function. The model demonstrates strong performance in delineating both the optic disc and optic cup. Quantitatively, the optic disc segmentation achieved a lower loss of 0.0226, while the optic cup segmentation, which is more challenging due to its smaller size and less defined boundaries, yielded a higher loss of 0.0765.

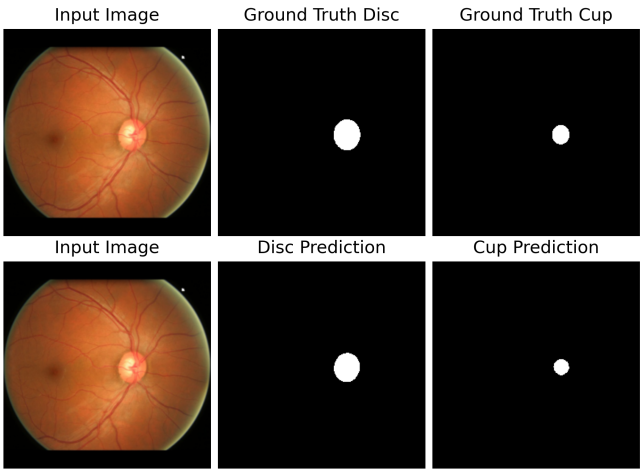

Fig. 2: Ground truth and predicted masks of an ORIGA image.

Figure 3 displays three stills from a clinical video of Subject 1's eye, with overlaid contours around the optic disc and optic cup generated from the U-Net model's segmentation predictions. Despite the lower resolution and variable lighting typical of clinic-acquired images, the model is able to produce contours that closely resemble those drawn by ophthalmologists, suggesting its potential for practical application in real-world settings. These qualitative results highlight the model's robustness and generalizability beyond curated datasets. However, formal validation by clinicians is necessary, and further refinement is needed to address subtle inaccuracies and ensure consistent performance across diverse clinical conditions.

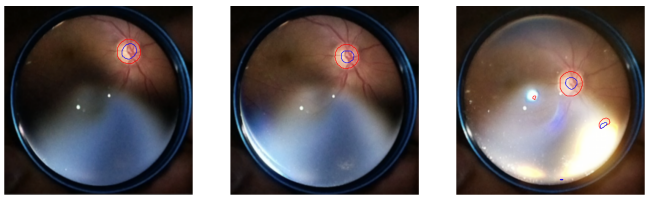

Fig. 3: Predicted mask contours of clinic Subject 1.

Figure 4 shows a test example from the fine-tuned Medical SAM-2 model. While the model segments the optic disc well, it struggles to segment the optic cup in a relatively continuous manner. This results in an optic cup mask that does not represent the approximately ovular shape that is expected for the optic cup. This compares to the U-Net model which consistently generates an approximately ovular mask.

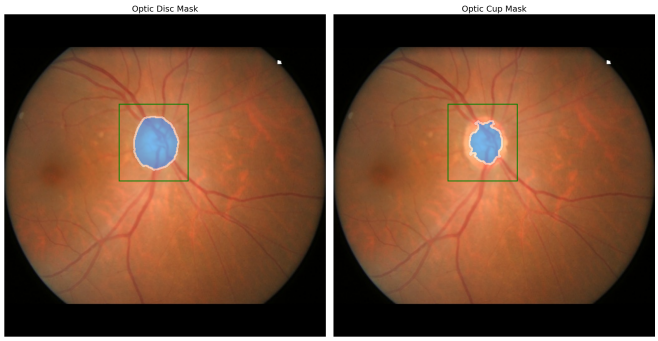

Fig. 4: Predicted masks of an ORIGA image.

Figure 5 exhibits the prompting power of the altered Medical SAM-2 model. Despite the strong presence of glare and artifacts in the clinic image, the bounding box prompting is successful in guiding the model to predict highly plausible locations for the optic disc and optic cup. However, the optic cup mask still lacks continuous edges.

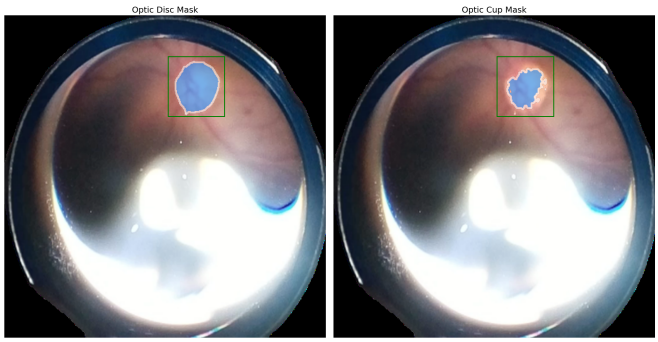

Fig. 5: Predicted masks of a clinic Subject 1 image.

B. Future Work

While the U-Net model demonstrates promising performance on low-quality clinical images, it struggles to accurately segment regions affected by large glare artifacts, which are often misclassified. To address this limitation, we are exploring several potential improvements. First, we plan to collaborate with ophthalmologists to obtain ground truth annotations for clinical images, enabling the inclusion of these real-world samples in the training set. This would help the model learn to better generalize to clinical conditions. Additionally, we are investigating the integration of prompting techniques from the Medical SAM-2 model, such as defining bounding boxes that exclude glare regions and guide the model to predict segmentation masks within a refined

region of interest (ROI). These strategies aim to improve segmentation accuracy in challenging clinical scenarios.

In conjunction with modifications to the U-Net model, we will continue to explore modifications to the Medical SAM-2 model to increase performance on images, specifically with optic cup segmentation. Proposed modifications include further manipulation of the loss function used for training and alternate prompting techniques. Specifically, we will investigate a combined U-Net and Medical SAM-2 model where predicted masks from the U-Net model are used as prompts for Medical SAM-2 model. Until the Medical SAM-2 model reliably generates optic cup masks with significantly continuous edges, the U-Net model is preferred as it provides masks that more strongly reflect real world expectations of optic cup shape.

In the next phase, we aim to calculate DDLS scores directly from the segmented disc and cup masks. DDLS is a clinically validated metric used by ophthalmologists to assess glaucoma related damage, and automating this process would enable rapid, standardized evaluation without manual measurement. We plan to develop a pipeline that integrates vertical cup-to-disc ratio, rim width, and other anatomical features derived from the segmentation outputs to generate DDLS values for each frame.

Our long-term vision is to implement this system in a mobile application designed for use in real-time clinical settings. The app would allow ophthalmologists to record fundus videos using a binocular indirect ophthalmoscope with an attached smartphone camera. The model would then process video frames on-device or via cloud inference, generate segmentation masks, and compute DDLS scores to assist in clinical decision-making during the exam. This would significantly streamline the glaucoma screening workflow and expand access to early detection tools in under-resourced settings.

Additional future work includes incorporating temporal analysis across video frames for greater segmentation consistency, fine-tuning the model on a larger and more diverse clinical dataset, and validating DDLS predictions against expert annotations to ensure alignment with clinical standards.

ACKNOWLEDGMENTS

The authors would like to our wonderful sponsor Dr. Arjun Dirghangi and mentor Dr. Aiying Zhang for providing immense industry and technical knowledge throughout the project.

REFERENCES

[1] M. B. Sudhan, M. Sinthuja, S. P. Raja, J. Amutharaj, G. C. P. Latha, S. S. Rachel, T. Anitha, T. Rajendran, and Y. A. Waji, "Segmentation and classification of glaucoma using u-net with deep learning model," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–10, 2022, article ID 1601354.

[2] Z. Zhu, Y. Xia, J. Wang, J. Liu, A. Yuille, and Y. Zhou, "Medical sam-2: Towards one-prompt segmentation for any medical image," arXiv preprint arXiv:2408.00874, 2024. [Online]. Available: https://arxiv.org/abs/2408.00874

[3] H. A. Rahsheed, T. Davis, E. Morales, Z. Fei, L. Grassi, A. De Gainza, K. Nouri-Mahdavi, and J. Caprioli, "Ddlsnet: A novel deep learning-based system for grading funduscopic images for glaucomatous damage," in *Ophthalmology Science*, 2023.

[4] L. J. Coan, B. M. Williams, V. K. Adithya, S. Upadhyaya, A. Alkafri, S. Czanner, R. Venkatesh, C. E. Willoughby, S. Kavitha, and G. Czanner, "Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review," *Survey of Ophthalmology 68*, 2023.

[5] P. H. Prastyo and A. S. Sumi, "Optic cup segmentation using u-net architecture on retinal fundus image," *Journag of Information Technology and Computer Engineering*, 2020.

[6] N. A. Kako, A. M. Abdulazeez, and D. N. Abdulqader, "Multi-label deep learning for comprehensive optic nerve head segmentation through data of fundus images," *Heliyon 10*, 2024.