arXiv:2608.18516v1 [cs.CV] 19 Aug 2026

OptiModNet: A UNet-Transformer Hybrid with Grouped-Query and Channel Attention for Optic Disc and Cup Segmentation

Soumili Ghosh[1 []], Debapriya Roy[2 [0000−0001−8657−3130]], Aryan Das[3 [0000−0003−0439−7351]], and Bikash Santra[4 [0000−0002−6833−140X]]

[1] IEM Saltlake,India ghoshsoumili03@gmail.com
[2] Techno Main Salt Lake, India debapriyakundu1@gmail.com
[3] VIT Bhopal University, India aryan.das2021@vitbhopal.ac.in
[4] IIT Jodhpur, India bikash@iitj.ac.in

Abstract. Precise segmentation of the optic disc and cup is critical for the early detection and diagnosis of glaucoma. However, achieving consistently high performance across datasets while maintaining low computational requirements remains a significant challenge. In glaucoma detection, low-computation methods are crucial for enabling rapid, large-scale screening and facilitating deployment in resource-limited clinical environments. While deep learning models such as UNets, Vision Transformers (ViTs), and Diffusion models have demonstrated strong segmentation performance but these methods often come with substantial computational overhead. UNets are efficient at capturing local features but are limited in modeling global contextual information. Conversely, ViTs excel at long-range dependency modeling but are computationally intensive. Hybrid architectures, such as UNetR, which combine transformer-based encoders with UNet-style decoders, have shown improved performance but while incurring additional complexity. Considering these, in this work, we propose OptiModNet, a light weight novel hybrid architecture tailored for optic disc and cup segmentation. The model integrates diverse attention mechanisms at multiple stages of the network to enhance both local and global feature representation. We include an Aggregated Pyramid Loss that supervises predictions at multiple decoder depths, to promote better gradient flow and structural consistency. We evaluate OptiModNet on the REFUGE2 dataset for both optic disc and cup segmentation tasks. Our method achieves state-of-the-art performance, exceeding existing approaches by over 2.5%, while maintaining high efficiency with only 3.73 GFLOPs and 1.93M parameters. The code is available at https://github.com/SG1947/OptiModNet.

1 Introduction

Glaucoma is a major cause of irreversible blindness worldwide and it's early stages are often symptomless. Accurate analysis of retinal fundus images plays a vital role in its detection and monitoring. In particular, segmentation of the

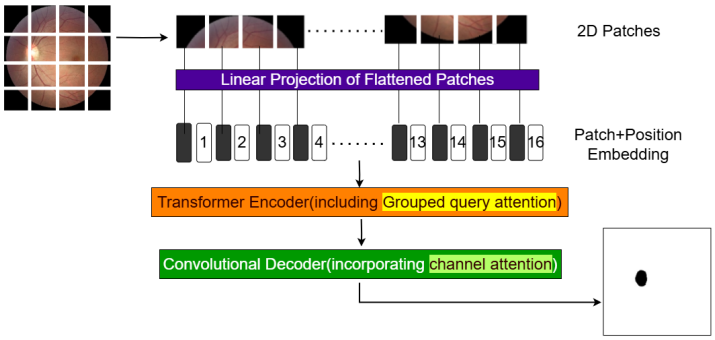

Fig. 1: Overview of the proposed OptiModNet.

optic disc (OD) and optic cup (OC) is a critical step in computing the cup-to-disc ratio (CDR), which is an important biomarker for glaucoma screening and progression assessment. Automated OD/OC segmentation facilitates large-scale screening programs and reduces inter-observer variability in clinical settings. However, OD/OC segmentation is challenging due to factors such as low contrast between cup and disc boundaries, peripapillary atrophy (PPA), blood vessel occlusion at the optic nerve head, and variability in image quality across acquisition devices. Over the years, significant advances have been made in retinal image segmentation, leveraging deep learning architectures to address these challenges. Approaches based on UNet, Vision Transformers (ViTs), and generative models have achieved promising results OD/OC segmentation.

In the last decade, UNet-based models have been widely adopted for OD/OC segmentation due to their strong local feature learning capabilities. Yet, their intrinsic limitation lies in modeling long-range dependencies [20], which are essential for capturing the global anatomical context of the optic nerve head. Conversely, ViTs effectively capture global relationships but require large annotated datasets and incur high computational costs, as computational complexity of the self-attention mechanism increases quadratically with the number of image patches. [7]. This makes their deployment in high-resolution fundus imaging resource-intensive. Hybrid architectures like TransUNet [4] and UNetr [9] combine UNet's local representation learning with transformer-based global context modeling, showing improved performance for OD/OC segmentation. Nevertheless, they inherit the computational inefficiencies of vanilla Transformers. SwinBTS [13] and Swin-UNetr [8] address some of these challenges with shifted window mechanisms, yet remain memory-intensive for full-resolution fundus images. Boundary-aware models such as BEAL [23] and BAT [22] have enhanced delineation of OD/OC boundaries, but at the cost of additional computational complexity. In addition, diffusion-based approaches like MedSegDiff [28] and EnsemDiff [26] have recently demonstrated competitive results for medical image segmentation, but their high training and inference costs make them less suited for real-time clinical workflows.

To overcome these segmentation limitations, we introduce a hybrid UNetR-based architecture that incorporates grouped-query attention in the encoder [1] and channel attention in the decoder [10]. While UNetR has shown strong performance in 3D medical segmentation, our 2D adaptation is tailored for fundus images, enabling efficient extraction of both global context and fine boundary details. We also propose a new loss function, Aggregated Pyramid Loss, that enforces supervision at multiple decoder stages, improving gradient flow and enhancing the segmentation of fine OD/OC boundaries. Through extensive experiments on public OD/OC segmentation benchmarks, we demonstrate that our method achieves state-of-the-art accuracy while significantly reducing computational overhead. Our contributions can be summarized as the following:

- We incorporate grouped query attention into the encoder, reducing computational complexity while effectively modeling long-range dependencies.
- We incorporate channel attention in the decoder to focus on more informative feature channels, improving boundary delineation between the optic disc and cup.
- We introduce a novel loss function, Aggregated Pyramid Loss, supervising multiple decoder stages to enhance gradient flow and fine detail segmentation.
- The proposed model builds upon the complementary strengths of UNet, Transformers, and attention mechanisms to achieve improved OD/OC segmentation accuracy with fewer parameters.

2 Literature Survey

Medical image segmentation has advanced significantly with encoder–decoder architectures, notably UNet and its numerous variants. For example, ResUNet [6] enhances feature reuse via residual connections, while ARU-GD [16] integrates attention mechanisms to improve localization precision. nnUNet [11] offers a self-configuring framework that adapts network architecture, training schemes, and preprocessing steps automatically. BAT [22] introduces boundary-aware modules to enhance edge precision, and DAGAN [31] leverages adversarial training to improve robustness and segmentation quality.

Vision Transformers (ViTs) have recently emerged as a promising paradigm for segmentation tasks, capable of capturing global contextual relationships through self-attention, which CNNs—focused primarily on local receptive fields—often miss. The Vision Transformer (ViT)[7] processes images as sequences of patches, proving effective in medical imaging applications. The Swin Transformer[15] addresses computational constraints through hierarchical feature extraction with shifted windows. Hybrid architectures such as TransUNet [4] integrate transformer layers into a UNet encoder, while UNetr [9] and Swin-UNetr [8] adopt Swin Transformer designs. Other models like FAT-Net [27] and Swin-UNet [3] focus on fusing local CNN features with global transformer cues. For volumetric segmentation, TransBTS [25] and SwinBTS [13] combine 3D CNNs with transformer encoders.

Diffusion-based approaches have recently gained traction. EnsemDiff [26] and SegDiff [2] apply score-based models for structured prediction. MedSegDiff [28] and MedSegDiff-V2 [29] leverage denoising diffusion processes to generate sharper, more consistent outputs. These models show strong performance but often increase computational cost. Combining Vision Transformers, UNet, and attention methods [1] creates a powerful tool for medical imaging.

Within the specialized field of OD and OC segmentation, accurate boundary delineation is crucial for computing the CDR, an important biomarker for glaucoma diagnosis. Early works like Sevastopolsky [21] adapted UNet for fundus images, while JointRCNN [12] introduced region-based convolutional neural networks for simultaneous OD and OC segmentation. Robustness improvements were demonstrated in Yu et al.[33] with a deep learning framework capable of handling variable image qualities, and Liu et al.[14] proposed a spatial-aware network for improved combined segmentation. Later, attention-based strategies such as Zhao et al.[35] incorporated transfer learning into Attention U-Net for better generalization, while Pachade et al.[19] leveraged a lightweight Nested EfficientNet with adversarial learning to reduce computational cost without sacrificing accuracy. Transformers have also entered the OC/OD segmentation space, with works like Yi et al.[32] introducing coarse-to-fine transformer networks for high-precision results and Yan et al.[30] combining large-kernel residual convolution with self-attention for consistent segmentation across datasets. Hybrid architectures combining Vision Transformers (ViTs) and UNet have been explored to address the limitations of each while leveraging their complementary strengths [4, 3]. Our method falls within this category but emphasizes a lighter design with improved performance across diverse OC/OD datasets.

3 Method

The proposed methodology, depicted in Fig. 2, addresses the challenges of semantic segmentation by integrating transformer-based encoders with convolutional decoders. This architecture incorporates two novel components to enhance both efficiency and accuracy. Firstly, Grouped Query Attention replaces the conventional multihead self-attention mechanism within the encoder, significantly reducing computational complexity without compromising performance on high-resolution medical images. Secondly, Channel Attention is employed in the decoder to adaptively emphasize salient feature channels hence attenuating less informative ones, this increasing segmentation accuracy.

Patch and Positional Embedding: Let the input image be $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, where H and W denote the height and width of the image, and C represents the number of input channels. The input image is divided into flattened, uniform, non-overlapping patches $\mathbf{I}_{patch} \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where $P \times P$ is the resolution of each patch and $N = \frac{H \cdot W}{P^2}$ is the number of patches i.e., the length of the sequence.

Each patch is projected into a embedding space of dimension K using a learnable linear layer. To conserve spatial characteristics, a learnable $1D$ po-

sitional embedding $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times K}$ is added to the projected patch embedding $\mathbf{E}_{\text{patch}} \in \mathbb{R}^{(P^2 \cdot C) \times K}$. This ensures that the model can leverage spatial relationships across patches. The final embedding, $\mathbf{E}_f$, is computed as:

$$\mathbf{E}_f = [\mathbf{x}^1_{patch} \cdot \mathbf{E}_{\text{patch}}; \mathbf{x}^2_{patch} \cdot \mathbf{E}_{\text{patch}}; \ldots; \mathbf{x}^N_{patch} \cdot \mathbf{E}_{\text{patch}}] + \mathbf{E}_{\text{pos}}, \tag{1}$$

where, $\mathbf{E}_f \in \mathbb{R}^{N \times K}$ is the output of the combined patch and positional embedding. It is sent as the input to the transformer encoder.

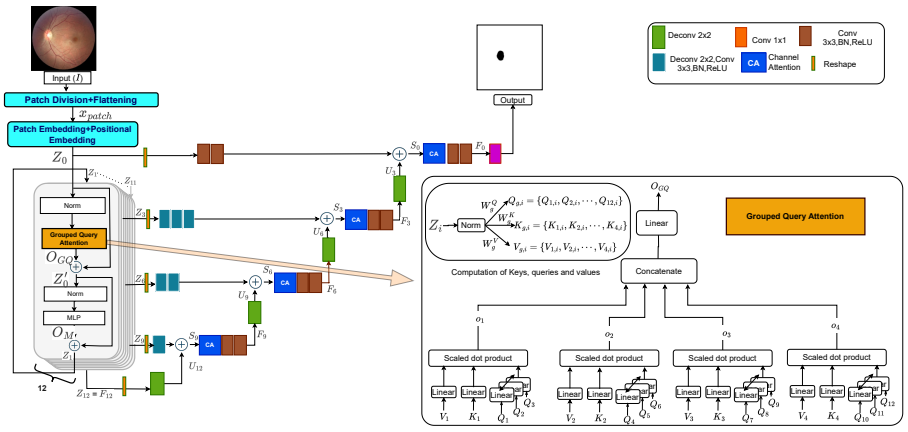

Fig. 2: (Left) Block diagram of the proposed model OptiModNet. (Right) Details of the grouped query attention module.

3.1 Transformer Encoder

The encoder employs multiple stacked transformer blocks to effectively learn long-range interactions and contextual information from the sequence of patch and positional embeddings. Each transformer block consists of a grouped query attention mechanism followed by a feedforward network (MLP) and residual connections to ensure stable training.

In standard transformer models, Multi-head Self-Attention (MSA) computes interactions between all input tokens. This can be expensive, especially for high-resolution inputs. To optimize this, we use Grouped Query Attention (GQA). This method reduces the computational load by splitting the queries into G groups, while associating each group with a single key-value pair, thereby streamlining the attention mechanism.

Given an input $\mathbf{E}_{i-1}$, we first normalize it and then compute the Queries ($\mathbf{Q}_{g,i}$), Keys ($\mathbf{K}_{g,i}$), and Values ($\mathbf{V}_{g,i}$) for each group $g \in \{1, 2, \ldots, G\}$, as follows :

$$\begin{aligned} \mathbf{Q}_{g,i} &= \text{Norm}(\mathbf{E}_{i-1}) \times \mathbf{W}^Q_g, \\ \mathbf{K}_{g,i} &= \text{Norm}(\mathbf{E}_{i-1}) \times \mathbf{W}^K_g, \\ \mathbf{V}_{g,i} &= \text{Norm}(\mathbf{E}_{i-1}) \times \mathbf{W}^V_g \end{aligned} \tag{2}$$

where $\mathbf{W}_g^Q \in \mathbb{R}^{D\times 3q_g}, \mathbf{W}_g^K \in \mathbb{R}^{D\times k_g}, \mathbf{W}_g^V \in \mathbb{R}^{D\times v_g}$ are trainable weight matrices, and q_g, k_g, and v_g denote the number of queries, keys, and values per group, respectively. For example, if there are 12 queries, 4 keys, and 4 values distributed over 4 groups, each group will contain 3 queries, 1 key, and 1 value. Attention for each group is computed as:

$$\mathbf{A}_{g,i} = \text{softmax}\left(\frac{\mathbf{Q}_{g,i}\mathbf{K}_{g,i}^\top}{\sqrt{k_g}}\right), \mathbf{O}_{g,i} = \mathbf{A}_{g,i}\mathbf{V}_{g,i}. \tag{3}$$

Once the attention outputs for all groups are computed, they are concatenated and projected back into the original embedding space:

$$\mathbf{O}_{GQ,i} = \text{Concat}(\mathbf{O}_{1,i}, \mathbf{O}_{2,i}, \ldots, \mathbf{O}_{G,i})\mathbf{W}^O, \tag{4}$$

where $\mathbf{W}^O \in \mathbb{R}^{3q_gG\times D}$ is the output projection matrix. This method significantly reduces the computational complexity of attention while maintaining the ability to model global dependencies in high-resolution inputs.

After grouped query attention, residual connections are added to get $\mathbf{E}_{i-1}^{'} = \mathbf{E}_{i-1} + \mathbf{O}_{GQ,i}$. This result is normalized, and a feedforward network (MLP) is added that refines the output features. This network consists of two linear layers with a Gaussian Error Linear Units (GELU) activation function and a dropout for regularization:

$$\mathbf{O}_{M,i} = \text{Dropout}(\text{GELU}(Norm(\mathbf{E}_{i-1}^{'})\mathbf{W}_1 + \mathbf{b}_1))\mathbf{W}_2 + \mathbf{b}_2 \tag{5}$$

where $\mathbf{W}_1 \in \mathbb{R}^{D\times D'}$ and $\mathbf{W}_2 \in \mathbb{R}^{D'\times D}$, with D' being the intermediate dimension. Adding the residual connection gives: $\mathbf{F}_i = \mathbf{E}_{i-1}^{'} + \mathbf{O}_{M,i}$. The i^{th} transformer block outputs an embeddings $\mathbf{F}_i$, which is passed to the next $(i{+}1)^{th}$ block (shown in Fig. 2). This process continues till the last T^{th} transformer block, producing a final set of embeddings $\{F_1, F_2, \cdots, F_T\}$. In the proposed model we consider $T = 12$.

3.2 Convolutional Decoder

The decoder incrementally upsamples the feature maps while incorporating multi-resolution skip connections. Furthermore, the key contribution of the proposed architecture in this component is the incorporation of a channel attention layer.

The output of the transformer encoder consists of features from various encoder resolutions, we extract a sequence representation $\mathbf{F}_i$, corresponding to specific layers $i \in \{12, 9, 6, 3\}$. Each $\mathbf{F}_i$ is a 1D sequence with dimensions $\frac{H\times W}{P^2} \times C$, where H and W are the input image's spatial dimensions, P is the patch size, and C is the embedding dimension.

We include the encoder's outputs as skip connection within the decoder. To match the spatial dimension they are reshaped into 2D feature maps as the following, $\mathbf{F}_i = \text{Reshape}(\mathbf{F}_i)$, $\mathbf{F}_i \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times C}$. The reshaped features are passed

through 2×2 deconvolutional layers and 3×3 convolutional layers followed by batch normalization (BN) and ReLU activation:

$$\mathbf{C}_i = \text{ReLu}(\text{BN}(\text{Conv2D}(\text{Conv2DTranspose}(\mathbf{F}_i)))) \tag{6}$$

This ensures that the features are better aligned for subsequent processing in the decoder.

Hierarchical Decoder Workflow The decoder operates in a stepwise manner, starting from the highest-resolution feature map $\mathbf{F}_{12}$ and progressively reconstructing the input resolution. Each step of the decoder involves four key operations: Upsampling, Skip Connection Integration, Channel Attention, and Convolutional Refinement.

Upsampling with Deconvolution: At each stage, the spatial resolution of the feature map is doubled using a 2×2 transposed convolution: $\mathbf{U}_i = \text{Conv2DTranspose}(\mathbf{F}_i), \quad \mathbf{U}_i \in \mathbb{R}^{2H' \times 2W' \times C}$. This operation enlarges the feature map while preserving its learned semantic information.

Integration of Multi-Resolution Features: The upsampled feature map $\mathbf{U}_i$ is combined with the corresponding encoder feature $\mathbf{F}_{i-3}$ through an element-wise addition: $\mathbf{S}_{i-3} = \mathbf{U}_i + \mathbf{F}_{i-3}$, This step enriches the upsampled features with contextual information from the encoder.

Channel Attention Mechanism (CA): The combined feature map $\mathbf{S}_{i-3}$ is passed through a channel attention module to recalibrate feature importance. The channel attention weights are computed as follows, first, a Global Average Pooling (GAP) operation is applied to $\mathbf{S}_{i-3}$ to reduce its spatial dimensions: $\mathbf{C}_{\text{avg}} = \text{GlobalAvgPool}(\mathbf{S}_{i-3}), \quad \mathbf{C}_{\text{avg}} \in \mathbb{R}^C$. The pooled vector $\mathbf{C}_{\text{avg}}$ is then passed through two fully connected layers with ReLU and sigmoid activations to compute the attention weights, such as $\mathbf{C}_{\text{attn}} = \sigma\left(\text{Dense}\left(\text{ReLU}\left(\text{Dense}(\mathbf{C}_{\text{avg}})\right)\right)\right)$, $\mathbf{C}_{\text{attn}} \in \mathbb{R}^C$. The weights $\mathbf{C}_{\text{attn}}$ are reshaped and applied to $\mathbf{S}_{i-3}$ through element-wise multiplication as $\mathbf{S}'_{i-3} = \mathbf{S}_{i-3} \odot \text{Reshape}(\mathbf{C}_{\text{attn}})$. This recalibration emphasizes informative channels while reducing the impact of less relevant ones.

Convolutional Refinement: The channel-attended feature map $\mathbf{S}'_{i-3}$ is processed through a sequence of convolutional layers to enhance its spatial and semantic representation: $\mathbf{F}_{i-3} = \text{ReLU}(\text{BN}(\text{Conv2D}(\mathbf{S}'_{i-3})))$. The refined feature map $\mathbf{F}_{i-3}$ serves as the input for the next decoder stage, continuing the upsampling process.

Final Decoding: At the final stage of the decoder, the reconstructed feature map $\mathbf{F}_0$ is passed through a 1×1 convolution to reduce the number of channels to the target classes. The final pixel-wise outputs are obtained by applying a sigmoid activation function:$\mathbf{Y} = \sigma(\text{Conv2D}(\mathbf{F}_0)), \quad \mathbf{Y} \in \mathbb{R}^{H \times W}$. Here, $\mathbf{Y}$ represents the segmentation map, with pixel values indicating the predicted probabilities for each class.

3.3 Aggregated Pyramid Loss

To effectively optimize both region-level overlap and pixel-wise classification accuracy, we adopt a hybrid loss function termed as aggregated pyramid loss that combines the soft Dice loss and Binary Cross-Entropy (BCE) loss. This combined loss is computed in a pixel-wise manner and is defined as:

$$\begin{aligned}\mathcal{L}(\mathrm{GT}, \mathrm{Y}) = 1 - \frac{2\sum_{i=1}^{M} \mathrm{GT}_i \cdot \mathrm{Y}_i + \epsilon}{\sum_{i=1}^{M} \mathrm{GT}_i + \sum_{i=1}^{M} \mathrm{Y}_i + \epsilon} \\ - \frac{1}{M} \sum_{i=1}^{M} \Big[\mathrm{GT}_i \cdot \log(\mathrm{Y}_i) + (1 - \mathrm{GT}_i) \cdot \log(1 - \mathrm{Y}_i) \Big]\end{aligned} \tag{7}$$

where GT_i denotes the ground truth label, Y_i denotes the predicted probability from Final Decoding, M is the total number of pixels (i), ϵ is a small constant for numerical stability.

Given the presence of multiple output stages in the network, we compute the loss at each stage and aggregate them using a weighted sum - hence the name Aggregated Pyramid Loss. Let $\mathcal{L}_k$ denote the combined loss at the k-th output, and λ_k be the corresponding weight. The total loss used for optimization is, $\mathcal{L}_{\text{total}} = \sum_{k=1}^{K} \lambda_k \cdot \mathcal{L}_k$, where K is the number of output branches (e.g., intermediate and final outputs), and λ_k controls the contribution of each output to the total loss. In our implementation we have 3 intermediate output for which we set $\lambda = \{0.2, 0.3, 0.5\}$ and for the final output $\lambda = 1$. The intuition is to ensure supervision at multiple stages while prioritizing the final prediction. An ablation study corresponding to the different values of λ are given the Sec. 4.4.

3.4 Implementation details

This section deals with the detailed implementation parameters and hyperparameters utilized for our model. We resized all the images to 256×256 pixels and normalized to the range [0, 1]. Each image was divided into 16×16 pixel patches to preserve spatial features essential for segmentation. The corresponding masks were resized, normalized. The model processes input images with 3 RGB channels and follows a 12-layer encoder-decoder structure. It employs a hidden dimension of 128 for feature representation and an MLP dimension of 32 for the multi-layer perceptron layers. The transformer architecture includes 12 query heads and 4 key-value heads for multi-head attention, ensuring effective feature extraction and segmentation performance. The model was trained using the proposed Aggregated Pyramid Loss, which supervises each intermediate and final output with a combined BCE and Dice loss. We employed the Adam optimizer with a learning rate of 1×10^{-3}. Training was conducted for 50 epochs with a batch size of 8.

4 Experiments

In this section, we conduct quantitative experimental analyses along with an ablation study. We compare with some benchmark methods on two standard

Table 1: Comparative study using Dice and IoU on REFUGE2 dataset.

Method	REFUGE2-Disc		REFUGE2-Cup	
	Dice	IoU	Dice	IoU
ResUNet [6]	92.90	85.50	80.10	72.30
BEAL [23]	93.70	86.10	83.50	74.10
TransBTS [25]	92.30	85.80	83.70	75.90
SwinBTS [13]	95.20	87.70	86.30	75.70
BAT [22]	92.30	85.80	82.00	73.20
nnUNet [11]	94.50	87.70	83.60	75.40
TransUNet [4]	95.00	87.70	85.40	77.20
UNetr [9]	94.60	87.80	83.50	75.70
Swin-UNetr [8]	95.30	87.90	85.40	77.30
EnsemDiff [26]	92.60	85.20	83.00	73.40
SegDiff [2]	92.10	85.20	83.00	73.40
MedsegDiff [28]+TransUNet	91.80	84.50	82.00	72.60
MedSegDiff-V2 [29]	96.70	88.90	87.90	80.30
OptiModNet (Ours)	**99.45**	**98.16**	**93.23**	**89.41**

segmentation metrics, Dice Coefficient(Dice) and Intersection Over Union(IOU), on two optic disc (OD) and optic cup (OC) segmentation benchmark datasets - REFUGE2 (Disc and cup) [18] and ORIGA [34]. Table 1 and Table. 2 presents the comprehensive comparison of segmentation methods across these datasets.

Dataset details - The REFUGE2 dataset contains a total of 2,000 RGB fundus images, each accompanied by binary masks representing the optic disc and cup areas. The dataset is used with the standard splits into training, validation, and test sets. The ORIGA dataset contains 650 retinal fundus images, including 168 from glaucomatous patients. It provides precise optic disc (OD) and optic cup (OC) boundary annotations, along with cup-to-disc ratio (CDR) values, making it suitable for segmentation-based glaucoma analysis. The images typically have low brightness and noticeable noise, posing additional challenges for automated segmentation methods.

Table 2: Segmentation performance on ORIGA dataset for 3 class classification, optic cup, optic disc and background.

Model	**IoU**	**Dice**
U-Net [20]	91.58	92.76
Attention U-Net [17]	92.39	95.39
DeepLabV3+ [5]	95.10	96.29
TransUNet [4]	96.29	96.29
Swin-UNet [15]	96.00	98.79
Wang et al. [24]	97.60	98.79
OptiModNet(Ours)	**98.92**	**99.31**

4.1 Quantitative Experiments

In Table 1 we can observe, the proposed OptiModNet model outperforms all other methods on the REFUGE2 dataset for both optic cup and disc segmentation. Our proposed approach, OptiModNet, achieves the highest performance across all metrics. Specifically, for the disc segmentation task, OptiModNet attains a Dice score of 99.45% and an IoU of 98.16%, surpassing the next-best method (MedSegDiff-V2) by 2.75% and 2.26%, respectively. For the cup segmentation task, OptiModNet attains a Dice score of 93.23% and an IoU of 89.41%, outperforming the second-best method MedSegDiff-V2 by 2.23% and 1.11%, respectively.

Table. 2 reports the segmentation performance of various methods on the ORIGA dataset for the three-class classification task (optic cup, optic disc, and background). Our proposed method, OptiModNet, achieves the highest scores across both metrics, with an IoU of 98.92% and a Dice score of 99.31%. Compared to the next-best method by Wang et al., OptiModNet improves IoU by 1.32% and Dice by 0.52%.

The improvement margin is more pronounced compared to the classical architectures such as U-Net and Attention U-Net, where the proposed method demonstrates an IoU gain of 7.34% and 6.53%, respectively, and a Dice gain of 6.55% and 6.92%, respectively. These results indicate the efficacy of OptiModNet compared to modern transformer-based models like Swin-UNet and TransUNet as well as traditional convolutional baselines. The superior performance on ORIGA, especially in IoU, shows the model's precise boundary delineation capabilities, which are essential for accurate localization of optic disc and cup regions in clinical settings.

Fig. 3 shows the performance metrics and loss of our method over 50 epochs on the REFUGE2 (optic disc) dataset. The Dice and IoU curves exhibit rapid improvement in both training and validation, reflecting strong alignment with ground truth.

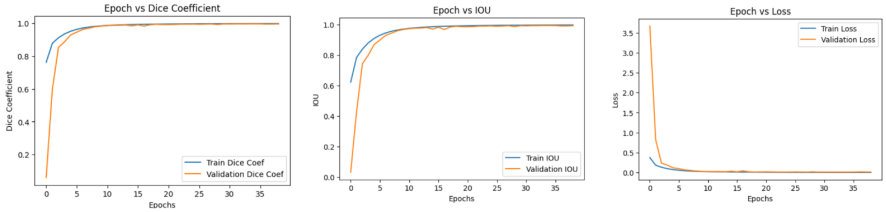

Fig. 3: Training and validation curves for Dice, IoU, and loss on the REFUGE2 (optic disc) dataset show increasing metric values over epochs, and decreasing loss indicating convergence.

4.2 Qualitative Experiments

Fig. 4, Fig. 5 shows our results of optic cup and disc segmentations on the REFUGE2 dataset. In most examples, the predicted masks align closely with the ground truth boundaries, indicating better localization and shape preservation of the optic disc. In general across the results the predictions exhibit smooth and well-defined contours, with only minor deviations near the boundaries, including cases where the disc margin is less distinct due to illumination variation or vessel occlusion. It is observable that these visual observations are consistent with the Dice and IoU scores in Table. 1, indicating that the qualitative performance aligns well with the reported quantitative results.

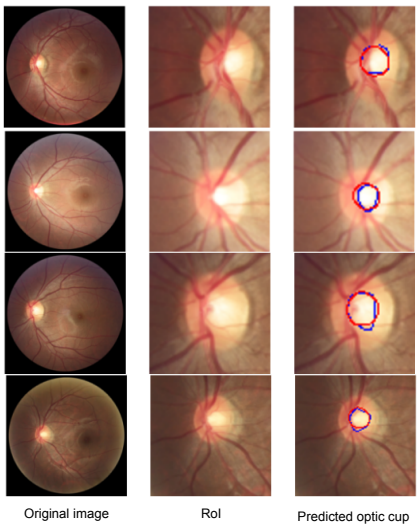

Fig. 4: Qualitative results of optic cup segmentation on the REFUGE2 testset.

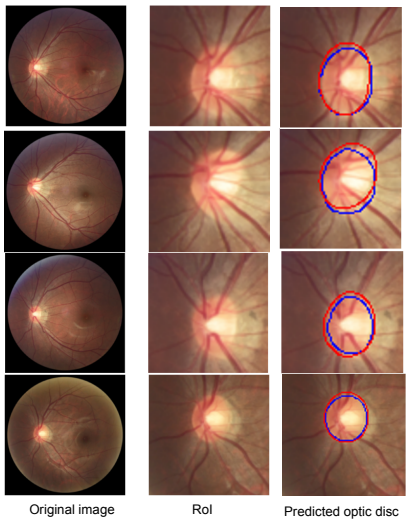

Fig. 5: Qualitative results of optic disc segmentation on the REFUGE2 testset.

4.3 Parameters

Table 3 presents a comparison of various models based on their parameter counts and computational complexity (measured in Gflops). OptiModNet stands out with the lowest parameter count (1.9M), significantly smaller than others (96% parameters efficient compared to MedSegdiff v2) highlighting its efficiency while potentially maintaining competitive performance. A related reduction plot is shown in Figure 6, highlighting that the proposed model retains nearly 96% fewer parameters compared to the highest-performing state-of-the-art models. In addition to parameter efficiency, the proposed OptiModNet achieves the second-best performance while requiring only 0.2% of the GFlops used by MedSegDiff-V2.

Table 3: Comparison of models on Params and Gflops

Model	Params (M)	Gflops
Swin-UNet	21	34
EnsemDiff	23	2203
SegDiff	23	2399
MedSegDiff	25	1770
MedSegDiff-V2	46	983
OptiModNet	**1.93**	**3.73**

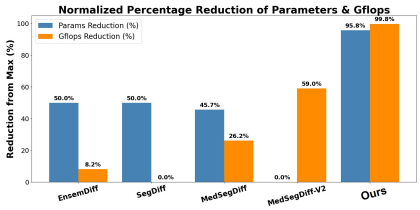

Fig. 6: Reduction plot

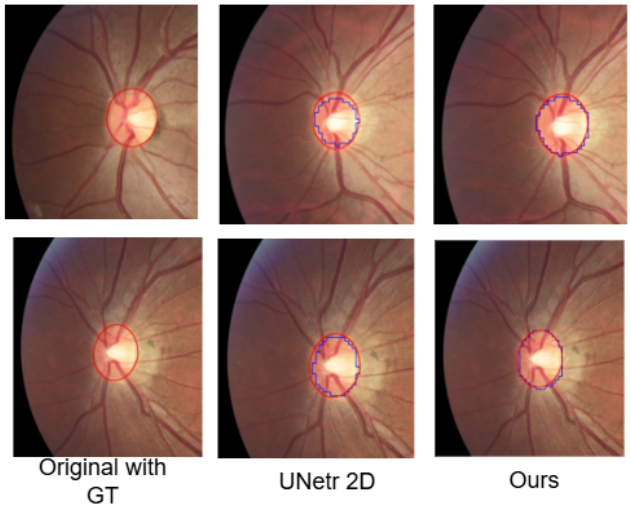

Fig. 7: Comparative study of our result with that of UNETR 2D model on REFUGE-2 disc dataset. For better visual comparability, we have extracted the boundary of the mask and overlaid on the input image. It can be observed that the boundary of our predicted mask (blue) aligns comparatively better with the ground truth (red).

4.4 Ablation Study

To understand the effect of each attention module an ablation study is presented in Table. 4. The incremental improvement in scores confirms that each attention module contributes meaningfully to performance improvement. Table. 5 presents the ablation over different λ values corresponding to the proposed loss function on REFUGE2-disc dataset. A visual ablation presented in Figure. 7 compares our performance with the UNetr 2D exhibiting comparatively that OptiModNet shows improved boundary adherence and captures better fine details.

Table 4: Ablation study: Comparison of OptiModNet with various modifications on the baseline UNETR 2D. Note. CA - Channel Attention and GQA - Grouped Query Attention.

Configuration	REFUGE2-Disc		REFUGE2-Cup	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)
UNETR 2D	94.21	93.83	85.27	80.80
UNETR 2D + CA	95.60	94.28	86.23	81.30
UNETR 2D + GQA	97.23	96.75	89.89	87.56
UNETR 2D + CA + GQA (Ours)	**99.45**	**98.16**	**93.23**	**89.41**

Table 5: Ablation study of our approach with different λ values on the REFUGE 2-disc dataset.

λ_1	λ_2	λ_3	λ_4	Dice (%)	IoU (%)
1	1	1	1	99.03	97.56
0.5	0.5	0.5	0.5	99.41	98.02
0.25	0.25	0.25	0.25	99.38	98.09
0.2	0.3	0.5	1	**99.45**	**98.16**

5 Conclusion and Future Work

In this paper, we propose OptiModNet, a novel framework for optic disc and cup segmentation that extends the UNetR (UNet with Transformers) architecture by integrating Grouped Query Attention and a newly introduced Aggregated Pyramid Loss function. OptiModNet demonstrates state-of-the-art performance on the REFUGE2 and ORIGA datasets. On REFUGE2, it achieves Dice scores of 99.45% for disc and 93.23% for cup, surpassing existing approaches by over 2.5%. On ORIGA, it further attains the highest reported segmentation accuracy, despite requiring nearly 96% fewer parameters than the most complex state-of-the-art model. These results highlight the effectiveness and efficiency of OptiModNet in advancing OC/OD segmentation.

6 Acknowledgements

We thank Dr. Swalpa Kumar Roy for his insightful suggestions and guidance. This work was partially supported by the Anusandhan National Research Foundation (ANRF) through the Prime Minister Early Career Research Grant.

References

1. Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., Sanghai, S.: Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245 (2023)
2. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)

6. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words. arXiv preprint arXiv:2010.11929 **7** (2020)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Jiang, Y., Duan, L., Cheng, J., Gu, Z., Xia, H., Fu, H., Li, C., Liu, J.: Jointrcnn: a region-based convolutional neural network for optic disc and cup segmentation. IEEE Transactions on Biomedical Engineering **67**(2), 335–343 (2019)
13. Jiang, Y., Zhang, Y., Lin, X., Dong, J., Cheng, T., Liang, J.: Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. Brain sciences **12**(6), 797 (2022)
14. Liu, Q., Hong, X., Li, S., Chen, Z., Zhao, G., Zou, B.: A spatial-aware joint optic disc and cup segmentation method. Neurocomputing **359**, 285–297 (2019)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
16. Maji, D., Sigedar, P., Singh, M.: Attention res-unet with guided decoder for semantic segmentation of brain tumors. Biomedical Signal Processing and Control **71**, 103077 (2022)
17. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
18. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
19. Pachade, S., Porwal, P., Kokare, M., Giancardo, L., Meriaudeau, F.: Nenet: Nested efficientnet and adversarial learning for joint optic disc and cup segmentation. Medical Image Analysis **74**, 102253 (2021)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
21. Sevastopolsky, A.: Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. Pattern Recognition and Image Analysis **27**(3), 618–624 (2017)

22. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 206–216. Springer (2021)
23. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: Boundary and entropy-driven adversarial learning for fundus image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22. pp. 102–110. Springer (2019)
24. Wang, S., Kim, B., Eom, D.S.: Boundary-aware transformer for optic cup and disc segmentation in fundus images. Applied Sciences **15**(9), 5165 (2025)
25. Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 109–119 (2021)
26. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: International Conference on Medical Imaging with Deep Learning. pp. 1336–1348. PMLR (2022)
27. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: Fat-net: Feature adaptive transformers for automated skin lesion segmentation. Medical image analysis **76**, 102327 (2022)
28. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning. pp. 1623–1639. PMLR (2024)
29. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6030–6038 (2024)
30. Yan, S., Pan, X., Wang, Y.: Mrsnet: Joint consistent optic disc and cup segmentation based on large kernel residual convolutional attention and self-attention. Digital Signal Processing **145**, 104308 (2024)
31. Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P.L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., et al.: Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. IEEE transactions on medical imaging **37**(6), 1310–1321 (2017)
32. Yi, Y., Jiang, Y., Zhou, B., Zhang, N., Dai, J., Huang, X., Zeng, Q., Zhou, W.: C2ftfnet: Coarse-to-fine transformer network for joint optic disc and cup segmentation. Computers in Biology and Medicine **164**, 107215 (2023)
33. Yu, S., Xiao, D., Frost, S., Kanagasingam, Y.: Robust optic disc and cup segmentation with deep learning for glaucoma detection. Computerized Medical Imaging and Graphics **74**, 61–71 (2019)
34. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: 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. pp. 3065–3068. IEEE (2010)
35. Zhao, X., Wang, S., Zhao, J., Wei, H., Xiao, M., Ta, N.: Application of an attention u-net incorporating transfer learning for optic disc and cup segmentation. Signal, Image and Video Processing **15**(5), 913–921 (2021)