DOI: 10.21917/ijivp.2026.0562

ATTENTION-DRIVEN VISION CAPSULE TRANSFORMER FOR GLAUCOMA DIAGNOSIS AND OPTIC DISC SEGMENTATION FROM FUNDUS IMAGES

Utipmfon Sukmama Jimmy
Department of Public Health, Concordia University of Edmonton, Canada

Abstract

Glaucoma is a progressive optic neuropathy that is among leading causes of irreversible blindness. Structural changes in the optic nerve head identification can assist in early clinical treatment. Retinal fundus photography is a non-invasive modality for glaucoma screening, while optic disc segmentation allows measuring morphological properties of clinical relevance. There are numerous deep learning algorithms developed for optic disc segmentation and glaucoma diagnosis; these tasks are often treated independently from each other. While convolutional architectures allow capturing the local spatial features effectively, they can experience challenges in capturing long-range dependencies and complex anatomy. Traditional segmentation networks can also suffer performance drop due to ambiguous, under-illuminated, occluded by blood vessels, or affected by pathology borders of the optic disc. This paper suggests Attention-Driven Vision Capsule Transformer (AVCT-Net) approach for simultaneous optic disc segmentation and glaucoma diagnosis. The approach involves hierarchical convolutional feature extraction, multi-scale attention mechanism, vision transformer-based global contextual analysis, and capsule-based feature representation. While the attention mechanism allows highlighting the most discriminating optic nerve areas, the transformer captures the long-range relationships between different retinal areas. Capsules provide preserving spatial and hierarchical dependencies between optic disc features and glaucoma-specific morphology. A multi-task learning approach is used for simultaneous optimization of segmentation and classification loss functions, thus, allowing utilizing anatomically relevant features generated by the segmentation branch in the diagnostic branch. According to experiments conducted on the REFUGE dataset, the proposed AVCT-Net algorithm allows achieving 98.5% accuracy, 97.7% precision, 98.2% recall, and 97.9% F1-score for glaucoma classification, as well as 98.4% Dice coefficient for optic disc segmentation within 100 training epochs. Compared with M-Net, Attention U-Net, and SwinCup-DiscNet approaches, the suggested approach allows improving accuracy by 2.7, 2.0, and 0.7%, respectively. The improvements in precision are 2.5, 1.8, and 0.6 points, while recall improvements are 2.1, 1.2, and 0.6 points, respectively. The improvements in F1-score are 2.3, 1.5, and 0.5 points, correspondingly. In optic disc segmentation task, AVCT-Net approach reaches 98.4% Dice coefficient, which is better than the results achieved by M-Net (by 2.1 points), Attention U-Net (by 1.4 points), and SwinCup-DiscNet (by 0.3 points).

Keywords:

Glaucoma Diagnosis, Optic Disc Segmentation, Vision Transformer, Capsule Network, Attention Mechanism

1. INTRODUCTION

Glaucoma is a chronic and progressive disease in which there is damage to the retinal ganglion cells and their axons, which eventually leads to blindness. As the disease may advance even without showing any symptoms in its early stages, regular screening is necessary to decrease the risk of losing one's sight permanently. Retinal fundus photography has been widely used as an imaging technique for glaucoma because of its relative low cost and non-invasiveness of imaging the optic nerve head and nearby retina structures. Information about the optic disc, optic cup, neuroretinal rim, and vascular patterns may be obtained from retinal images, which may be used for glaucoma detection. One of the most common morphological parameters of glaucomatous changes is the ratio between the optic cup and the optic disc, known as cup-to-disc ratio (CDR), which still remains one of the primary indicators in automated glaucoma analysis [1]. Studies on artificial intelligence in ophthalmology showed that machine learning and deep learning can recognize glaucomatous changes on retinal images and help clinicians during mass screening [2]. Promising results were also shown in optic disc and optic cup segmentation using deep learning [3].

1.1 CHALLENGES

Although these advancements have been made, there is still room for improvement for automated glaucoma assessment because retinal fundus images show considerable variations. Variables like illumination, image resolution, camera properties, pigmentation, retinal background, and imaging conditions might affect the accuracy of segmentation and classification to some extent. In addition, the structure of the optic nerve head may vary from person to person. The presence of peripapillary atrophy, optic disc drusen, vessel crossings, pathology, and unclear disc borders can lead to the problem of proper delineation [4]. Moreover, machine learning models trained on one dataset might not be generalizable to images acquired by other devices. Domain difference across multiple datasets has been recognized as a major limitation for deep learning-based optic disc and cup segmentation [5].

1.2 PROBLEM STATEMENT

An important shortcoming of the currently existing glaucoma detection systems lies in their isolation of anatomical segmentation and disease classification processes. In particular, there have been multiple attempts to classify glaucoma using various deep learning methods without incorporating any explicit representation of the clinically relevant optic nerve structures. On the other hand, the performance of segmentation-based algorithms depends greatly on the accuracy of the optic disc and cup segmentation, and mistakes at this stage of the algorithm pipeline can be carried over to the CDR estimation and further classification of the disease. Another problem of publicly available datasets relates to the quality of labels and their inconsistency [6]. Thus, the development of an architecture combining these tasks in one model seems essential.

1.3 OBJECTIVES

The primary objective of this research is to develop an Attention-Driven Vision Capsule Transformer Network (AVCT-

Net) for joint glaucoma diagnosis and optic disc segmentation from retinal fundus images. The specific objectives are to develop an attention-guided feature extraction mechanism for emphasizing clinically relevant optic nerve regions; it incorporates transformer-based global contextual modelling to capture long-range relationships within retinal images. It employs capsule representations to preserve spatial relationships among hierarchical optic disc features. It performs optic disc segmentation and glaucoma classification through a unified multi-task architecture; and evaluate the proposed framework using segmentation and classification metrics against established deep learning approaches.

The novelty of the proposed system stems from the combination of multi-scale attention mechanisms, vision transformer modeling, and capsule-based representation learning in one multi-task system for glaucoma analysis. While CNNs focus on the encoding of local features, the transformers allow more powerful global contextual modeling. On the other hand, the representation learned by the transformers lacks some of the explicit spatial relationships, which are crucial for an anatomical structure. In order to solve this problem, AVCT-Net introduces capsule representations into the system after context modeling, thus ensuring the preservation of hierarchical spatial relations that relate to the optic disc anatomy.

The primary contributions made in this research are as follows: First, a new architecture of AVCT-Net is proposed through the integration of attention-based multi-scale feature extraction, vision transformer-based global contextual understanding, and capsule network-based spatial representation for concurrent optic disc segmentation and glaucoma classification. Second, the novel multi-task anatomical-diagnostic learning approach is proposed where the two tasks of optic disc segmentation and glaucoma classification are combined. Such an approach allows the diagnostic network to take advantage of structurally informative features.

2. RELATED WORKS

Deep learning has greatly revolutionized the automated detection of glaucoma through the replacement of manual image descriptors with learned hierarchical representations. The early computer-aided strategies utilized optic disc localization, manually engineered features, calculation of CDR and conventional classifiers. However, with more and more annotated retinal datasets becoming available, CNN-based segmentation and classification models emerged. In the comprehensive review of machine learning and deep learning approaches to the diagnosis of glaucoma, U-Net and its modifications were found especially useful in the task of optic disc and cup segmentation due to the ability of their encoder-decoder structure to preserve spatial information [7].

The review of the state-of-the-art techniques has revealed the significance of open-source datasets and standard metrics for the evaluation of segmentation accuracy. Moreover, CNN-based classification has been found to have high diagnostic potential as many validation studies have shown that the architecture of CNN is capable of classifying glaucoma based solely on fundus images without the need of calculating handcrafted CDR measurements [8]. However, despite the high accuracy of CNNs, classification-oriented strategies cannot be considered interpretable as their decisions are not necessarily related to the identification of some specific anatomical structures which is quite problematic in ophthalmology since the assessment of glaucoma is done based on the evaluation of the optic nerve head.

Therefore, there is a clear advantage in segmentation-oriented strategies that explicitly segment the optic disc and cup. Fu et al. proposed M-Net – a multi-stage architecture aimed at segmentation of these anatomical structures using the fundus images. The model has proved to be effective in utilizing multi-scale and contextual information for segmentation [9]. Other approaches utilized modifications of U-Net, residual connections, DenseNet-based encoders and other configurations of the encoder-decoder architecture. Such systems use the results of segmentation in order to calculate CDR and other morphological parameters.

The attention mechanism was later used in segmentation models to help localize relevant anatomical structures more precisely. One of such approaches is the Attention U-Net, which adds attention gates that highlight useful parts and suppress irrelevant background features. In case of retinal images, it is important to note that the boundaries of an optic disc might be relatively small compared to the whole picture. The multi-scale attention will also be useful to consider variations of optic disc sizes, contrast, and appearance. GDCSeg-Net went one step further and used multi-scale weight-shared attention and densely connected depthwise separable convolution, showing good segmentation results in several fundus datasets [10].

Another research direction concerns the ability of models to generalize across various datasets. Models trained on one dataset tend to be effective under similar imaging acquisition settings but perform poorly when applied to the images taken with other devices or imaging protocols. The solution to the problem offered by GDCSeg-Net is the training of the model on a mixture of datasets and some architectural elements that improve the extraction of features on different domains [10]. Also, reviews have pointed out several factors that hinder the application of these models in clinical practice: dataset heterogeneity, lack of annotated examples, non-uniform image quality, and domain shift [11]. This shows that accurate segmentation is not enough; the model should also have robust representation learning for efficient glaucoma screening.

Transformer models have gained popularity in recent years due to the ability of the self-attention mechanism to model long-range dependencies in data that cannot be effectively done with regular convolutional filters. Thus, transformers can establish connections between distant areas of retina and provide global context representations. The recent research concerning glaucoma included the investigation of fusion-transformer approaches that combined optic disc and cup and learned the features with a transformer model [12]. It demonstrated the importance of global context modeling for the diagnosis of glaucoma. However, transformer models might need extensive training data and computational resources, and the patch-based representation does not guarantee hierarchies preservation.

3. PROPOSED METHOD

The suggested Attention-Driven Vision Capsule Transformer Network (AVCT-Net) is a unified multi-task architecture that is used for both optic disc segmentation and glaucoma classification on retinal fundus images. It is an algorithm that is based on multi-scale convolutional feature extraction, attention-driven feature enhancement, vision-transformer based global context modelling, capsule representation and task-specific predictions. Firstly, the input retinal fundus image is normalized and enhanced to eliminate the influence of the illumination and contrast variations. Then, multi-scale convolutional feature extractor encodes the retinal structures of various scales. Next, these feature maps are enhanced by the help of multi-scale attention that highlights the optic nerve head region and excludes the irrelevant retinal parts. Transformer blocks capture the long range dependencies between retinal patches, and capsule layers retain the spatial and hierarchical information about the learnt optic disc representation. The representation is shared across two similar tasks – optic disc segmentation and glaucoma classification. In the first task, the optic disc mask is reconstructed based on the multi-scale decoder features, and in the second, capsule-transformer representations are aggregated for the glaucoma classification. Therefore, the joint loss function is optimized for both tasks which allows the anatomical data to participate in the diagnostic learning process.

The whole AVCT-Net can be described as the following transformations of the normalized retinal fundus image to the two outputs:

$$\left(\hat{M}_{OD}, \hat{y}_G\right) = \mathrm{F}_{AVCT}\left(I\right) = \begin{pmatrix} \mathrm{D}\left(\mathrm{H}\left(\mathrm{T}\left(\mathrm{A}\left(\mathrm{C}(\mathrm{P}\left(I\right))\right)\right)\right)\right), \\ \mathrm{Softmax}\left(\mathrm{C}_G\left(\mathrm{GAP}\left(\mathrm{H}\left(\mathrm{T}\left(\mathrm{A}\left(\mathrm{C}(\mathrm{P}\left(I\right))\right)\right)\right)\right)\right)\right) \end{pmatrix} \quad (1)$$

where P denotes preprocessing, C represents convolutional feature extraction, A denotes multi-scale attention, T represents transformer-based contextual modelling, H denotes capsule-hybrid feature fusion, D represents the segmentation decoder, CG denotes the glaucoma classification head, M^{OD} represents the predicted optic disc mask, and y^G represents the predicted glaucoma class.

3.1 PROCESSING STEPS OF AVCT-NET

The entire flow process involves:

1. Fundus Image Capture and Preprocessing
2. Multi-scale Convolutional Features Extraction
3. Attention-guided Features Refinement
4. Global Context Modeling via Vision Transformer
5. Spatial Features Representation with Capsules
6. Multi-scale Feature Fusion
7. Optic Disc Segmentation
8. Glaucoma Classification
9. Multi-task Learning
10. Output Diagnosis and Segmentation

3.2 FUNDUS IMAGE ACQUISITION AND PREPROCESSING

The first stage receives a retinal fundus image as input. Fundus photographs generally contain variations in illumination, contrast, color distribution, image resolution, and background intensity because images can be acquired using different cameras and clinical protocols. These variations can adversely affect both segmentation and classification. Therefore, preprocessing is performed before feature extraction. Let the input retinal image be represented by $I(x,y,c)$, where x and y represent spatial coordinates and c denotes the color channel. The image is first resized to a fixed spatial dimension and normalized to a consistent numerical range. Contrast enhancement is then applied to improve the visibility of retinal structures, particularly the optic disc boundary.

3.2.1 Image Normalization and Enhancement:

The normalized image can be expressed using the following formulation:

$$I_n(x,y,c) = \frac{I(x,y,c) - \mu_c}{\sqrt{\sigma_c^2 + \text{ò}}} \quad (2)$$

where $I_n(x,y,c)$ denotes the normalized pixel value, μ_c represents the mean intensity of channel c, σ_c^2 denotes the corresponding variance, and ϵ is a small stabilization constant. This operation will equalize the difference in intensity distribution between fundus images. Then, the contrast enhancement operation is performed to enhance the local anatomical structure. The output image is used as an input to the feature extraction network. The normalization step plays an especially crucial role in glaucoma segmentation as the optic disc can have a low contrast compared to the surrounding retinal area. Normalization helps in the next step, which is the attention mechanism, to recognize the clinically meaningful structures. During the training, the data augmentation step can be employed by performing operations such as rotation, horizontal flipping, resizing, cropping, and varying the intensity values.

3.3 MULTI-SCALE CONVOLUTIONAL FEATURE EXTRACTION

Upon the preprocessing step, the preprocessed fundus image is fed to a hierarchical convolutional encoder network. The goal of this step is to detect retinal patterns in various scales. While transformers are successful in detecting global relationships, convolutional layers are still relevant in detecting local structures including blood vessels, optic disc, cup, and retina texture. The initial convolutional layers have high spatial resolution and low-level information including edges and intensity changes. However, as we go deeper in the architecture, the spatial resolution decreases and the number of channels increases. As a result, the deeper layers have more semantic information.

3.3.1 Hierarchical Convolutional Representation:

The feature extraction process can be formulated as:

$$F^{(l)} = \phi\left(BN\left(W^{(l)} * F^{(l-1)} + b^{(l)}\right)\right), \qquad l = 1, 2, \ldots, L \quad (3)$$

where $F^{(l)}$ represents the feature map generated at convolutional level l, $W^{(l)}$ denotes the convolution kernel, $b^{(l)}$ is the corresponding bias, $BN(\cdot)$ represents batch normalization, $\phi(\cdot)$

denotes the nonlinear activation function, and $*$ represents convolution. $F^{(0)}$ corresponds to the preprocessed fundus image.

The use of multiple convolutional levels produces feature maps containing complementary information. Shallow layers retain precise boundary information, while deeper layers encode more abstract anatomical and pathological characteristics. This hierarchical representation becomes important for the subsequent attention and transformer stages.

3.4 ATTENTION-DRIVEN FEATURE REFINEMENT

The following step is the introduction of the attention mechanism. Not all areas of the fundus image contribute equally to the identification of glaucoma. Large areas of the background of the retina have little diagnostic value, while the optic nerve head has significant diagnostic features. Thus, the attention layer can learn location-specific and channel-specific weights for increasing relevance areas. The attention mechanism works with the obtained feature maps and produces an attention map.

3.4.1 Multi-Scale Attention Computation:

The attention-enhanced feature representation is formulated as:

$$A_{ms}(F) = \sigma\left(\text{Conv}_{1\times1}\left[\begin{matrix}\text{AvgPool}(F) \otimes \text{MaxPool}(F) \otimes \\ \text{DilConv}_{3\times3}(F) \otimes \text{DilConv}_{5\times5}(F)\end{matrix}\right]\right) \square\ F \quad (4)$$

where $A_{ms}(F)$ denotes the multi-scale attention-enhanced representation, σ represents the sigmoid activation, *AvgPool* and *MaxPool* denote average and maximum pooling operations, *DilConv* represents dilated convolution, *Concat* denotes feature concatenation, and $\odot$ represents element-wise multiplication.

The combined receptive field comprises both local and global receptive fields. While the 3×3 kernel operation focuses on detecting more local structure information, 5×5 and dilated operation provide more contextual information. This is important due to variability in the size and structure of the optic disc between individuals. Thus, the attention module is able to focus on both fine boundary characteristics and larger anatomical context information.

The resultant attention mask will suppress all irrelevant parts of the retinal image while enhancing the representation of optic disc and related structures. It is worth noting that the operation for attention module does not define the location of the optic disc. The model automatically learns the relevant weight mapping using training data.

3.5 GLOBAL CONTEXT MODELLING USING VISION TRANSFORMER

After attention refinement process, the improved feature map will be transformed into the token sequence. This step mainly focuses on capturing more long-range information, which cannot be effectively captured by the local convolutions. The feature map will be partitioned into a number of patches (can be overlapping or non-overlapping). The patches will then be flattened and embedded into an embedding space. Positional information will also be included.

3.5.1 Patch Embedding:

The token representation can be defined as:

$$Z_0 = \left[\text{Flatten}(P_1)E; \text{Flatten}(P_2)E; \ldots; \text{Flatten}(P_N)E\right] + E_{pos} \quad (5)$$

where P_i denotes the i-th feature patch, E represents the learnable projection matrix, N denotes the number of patches, and E_{pos} represents the positional embedding. The resulting sequence Z_0 is supplied to the transformer encoder.

Each transformer block contains a multi-head self-attention mechanism and a feed-forward network. Self-attention allows every token to interact with other tokens, enabling the model to identify relationships between distant retinal regions.

3.5.2 Multi-Head Self-Attention:

The transformer operation is represented by:

$$\text{MHSA}(Z) = \text{Concat}\left(\text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right)V_h\right)_{h=1}^{H} W_O \quad (6)$$

where Q_h, K_h, and V_h represent the query, key, and value matrices for attention head h, d_h denotes the dimensionality of each head, H represents the total number of attention heads, and W_O denotes the output projection matrix.

Therefore, the transformer can model relationships between the optic disc and other retinal structures around it. For instance, the interpretation of local patterns on the optic disc can vary based on the features of the blood vessels or the peripapillary region around it. Global context modeling helps incorporate these dependencies within the learned representation. The output from the transformer is passed on to a feed-forward network along with residual connections.

3.6 CAPSULE-BASED SPATIAL REPRESENTATION

Although the transformer provides strong contextual modelling, its token representation does not explicitly encode the hierarchical spatial relationships required to describe anatomical structures. The proposed AVCT-Net therefore introduces a capsule representation stage. Capsules represent groups of related features as vectors rather than individual scalar activations. The vector magnitude can represent the presence or confidence of a particular structural feature, while the vector orientation can encode properties such as pose, spatial relationship, or deformation.

3.6.1 Capsule Transformation:

The transformation from primary feature capsules to higher-level capsules is defined as:

$$\begin{aligned} \mathbf{u}_{j|i} &= W_{ij}\mathbf{u}_i, \\ \mathbf{s}_j &= \sum_i c_{ij}\mathbf{u}_{j|i}, \\ \mathbf{v}_j &= \frac{\| \mathbf{s}_j \|^2}{1+\| \mathbf{s}_j \|^2}\frac{\mathbf{s}_j}{\| \mathbf{s}_j \|} \end{aligned} \quad (7)$$

where u_i represents the input capsule, W_{ij} is a learnable transformation matrix, $u^{j|i}$ represents the predicted output capsule, c_{ij} denotes the coupling coefficient between capsules, s_j represents the weighted input to capsule j, and v_j denotes the output capsule.

This is especially true for optic disc analysis, since the size, orientation, and shape of the structure may vary. Rather than just identifying the presence of features, capsule can capture the relationships between the features. For instance, variation in the shape of an optic disc can indicate glaucomatous progression. This step tries to maintain this structural relationship rather than reduce it to scalar values.

3.7 ATTENTION-GUIDED CAPSULE FUSION

The capsule representation is then fused with the attention-enhanced convolutional features. This fusion provides complementary local, global, and structural information.

The convolutional representation provides detailed spatial information, the transformer representation captures long-range dependencies, and the capsule representation preserves hierarchical structural relationships.

Hybrid Feature Fusion

The combined representation is expressed as:

$$F_{hyb} = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left[F_{att}, F_{tr}, F_{cap}\right]\right) + F_{att} + F_{tr} \tag{8}$$

where F_{att} denotes the attention-refined convolutional representation, F_{tr} denotes the transformer representation, F_{cap} represents the capsule representation, and F_{hyb} denotes the final hybrid feature representation.

The concatenation operation retains complementary information from the three feature-learning mechanisms. The 1×1 convolution reduces dimensional redundancy and projects the concatenated representation into a suitable feature space. Residual additions preserve important information from the attention and transformer pathways. This fusion is a central component of AVCT-Net because no single feature-learning mechanism is expected to represent all relevant characteristics of glaucoma. Local texture, global context, and anatomical geometry must be considered simultaneously.

3.8 OPTIC DISC SEGMENTATION

The segmentation branch receives the hybrid representation and reconstructs the optic disc mask. A decoder progressively increases the spatial resolution of the encoded feature representation. Skip connections are used to recover fine boundary information from earlier convolutional layers. At each decoding stage, the upsampled feature map is combined with a corresponding encoder feature map. Attention-enhanced skip connections can further ensure that relevant anatomical information is retained.

3.8.1 Decoder Reconstruction:

The segmentation representation at decoder level k can be formulated as:

$$D^{(k)} = \phi\left(W_d^{(k)} * \mathrm{Concat}\left[\mathrm{Up}\left(D^{(k+1)}\right), F_{att}^{(k)}\right] + b_d^{(k)}\right) \tag{9}$$

where $D^{(k)}$ represents the decoder feature at level k, $U_p(\cdot)$ denotes upsampling, $F_{att}^{(k)}$ represents the corresponding attention-refined encoder feature, and $W_d^{(k)}$ and $b_d^{(k)}$ denote decoder parameters.

The final decoder output is passed through a sigmoid activation to generate a pixel-wise optic disc probability map. A threshold can then be applied to obtain the binary segmentation mask.

The segmentation branch has two major roles. First, it provides the required optic disc segmentation output. Second, it supplies anatomical supervision to the shared feature representation. This second role makes the proposed framework different from a simple classification model.

3.9 GLAUCOMA CLASSIFICATION

The classification branch operates on the shared hybrid representation. Instead of relying exclusively on the complete retinal image, the classifier uses attention-weighted and capsule-enhanced features that contain stronger information about the optic nerve region. Global average pooling or attention-based aggregation converts the spatial feature representation into a compact diagnostic vector.

3.9.1 Diagnostic Representation:

The classification feature vector can be expressed as:

$$\begin{aligned} \mathbf{z}_{cls} &= \mathrm{GAP}\left(F_{hyb}\right), \\ \hat{y} &= \mathrm{Softmax}\left(W_{cls}\mathbf{z}_{cls} + b_{cls}\right) \end{aligned} \tag{10}$$

where z_{cls} denotes the diagnostic feature vector, *GAP* represents global average pooling, W_{cls} and b_{cls} denote classification parameters, and $\hat{y}$ represents the predicted class probabilities.

In the case of binary classification of glaucoma, there will be two probabilities associated with two categories, i.e., glaucomatous and non-glaucomatous. In this case, the class having the maximum probability will be considered.

The advantage of the classification pathway in this case lies in the fact that since it is sharing representations with the segmentation pathway, it has been trained with anatomical supervision, which makes the model learn features that have clinical significance.

3.10 JOINT MULTI-TASK OPTIMIZATION

The proposed framework simultaneously optimizes segmentation and classification. This multi-task formulation is important because glaucoma diagnosis and optic disc segmentation are closely related tasks. Accurate segmentation requires understanding the optic nerve anatomy, while reliable diagnosis benefits from structural information derived from that anatomy.

The total optimization objective combines segmentation loss, classification loss, and an optional feature-consistency regularization term.

3.10.1 Joint Objective Function

The complete loss function is defined as:

$$\mathrm{L}_{total} = \lambda_{seg}\left[\alpha \mathrm{L}_{Dice} + (1-\alpha)\mathrm{L}_{BCE}\right] + \lambda_{cls}\mathrm{L}_{CE} + \lambda_{con}\mathrm{L}_{con} \tag{11}$$

where L_{total} represents the total training loss, L_{Dice} denotes Dice loss, *LBCE* denotes binary cross-entropy segmentation loss, *LCE* represents classification cross-entropy loss, and L_{con} denotes the feature-consistency loss. The coefficients λ_{seg}, λ_{cls}, and λ_{con} control the contribution of each objective, while α balances Dice and binary cross-entropy components.

Dice loss is especially well-suited to optic disc segmentation due to the relative small size of the foreground region in the entire fundus image. The dice loss metric is a direct measure of overlap between the prediction and the ground truth, while the binary cross-entropy loss offers probabilistic supervision at the pixel level.

The classification loss is used to ensure that the learned representation retains its discriminatory power for glaucoma detection. The feature-consistency loss, if used, ensures structural consistency between the two branches.

3.11 BACKPROPAGATION AND PARAMETER OPTIMIZATION

During training, the total loss is propagated backward through the classification head, capsule layer, transformer blocks, attention modules, convolutional encoder, and segmentation decoder. An adaptive optimizer such as Adam can be employed to update the trainable parameters.

3.11.1 Parameter Update:

The optimization process can be represented as:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \dot{o}},$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad (12)$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

where θ_t represents the model parameters at iteration t, η denotes the learning rate, m_t and v_t represent the first- and second-moment estimates of the gradient, β_1 and β_2 denote exponential decay coefficients, and ϵ prevents numerical instability. The parameter update allows the model to progressively minimize both segmentation and classification errors. Because the tasks are optimized jointly, the gradients from both branches influence the shared encoder. This encourages the learned representation to capture features that are simultaneously useful for anatomical localization and glaucoma discrimination.

4. RESULTS AND DISCUSSION

The design AVCT-Net is evaluated using Python language through PyTorch deep learning platform. The experiments are carried out on a computer system having an Intel Core i7 CPU, 16 GB RAM, and an NVIDIA Graphics Processing Unit with 16 GB VRAM. Ubuntu 22.04 is the operating system employed. Training of the model is done using Adam optimizer for 100 epochs with a batch size of 16. The fundus images are resized to 512 x 512 pixels and augmented while training.

4.1 EXPERIMENTAL SETUP AND PARAMETERS

The principal experimental parameters used for AVCT-Net are presented in Table.1. These settings provide a consistent configuration for training and comparative evaluation.

Table.1. Experimental setup and parameter configuration

Parameter	Value
Programming environment	Python 3.10
Deep learning framework	PyTorch 2.x
Operating system	Ubuntu 22.04
Processor	Intel Core i7
RAM	16 GB
GPU	NVIDIA GPU, 16 GB VRAM
Input image size	512×512
Batch size	16
Training epochs	100
Optimizer	Adam
Initial learning rate	0.0001
Weight decay	1×10^{-5}
Transformer layers	4
Attention heads	8
Capsule dimension	16
Dropout	0.20
Training/validation/testing split	70%/15%/15%
Segmentation loss	Dice + BCE
Classification loss	Cross-entropy
Learning-rate scheduler	Cosine decay

The configuration in Table.1 is selected to provide sufficient model capacity while maintaining manageable computational requirements. The relatively small learning rate supports stable optimization of the combined CNN, attention, transformer, and capsule components. The four transformer layers and eight attention heads provide global contextual modelling without introducing excessive computational complexity. The multi-task loss allows the network to learn anatomical segmentation and glaucoma classification simultaneously.

4.2 DATASET DESCRIPTION

The REFUGE (Retinal Fundus Glaucoma Challenge) dataset is employed for the experimental evaluation. The REFUGE consists of 1,200 color fundus images that are acquired from various sources along with glaucoma-related annotations and optic disc and optic cup segmentation data. This dataset is appropriate for the proposed system since it includes both diagnostic classification and anatomical segmentation tasks. The images have substantial variability in terms of the retinal appearance, optic disc morphometry, illumination, and pathologies that will help to evaluate the performance of AVCT-Net under various imaging conditions.

The dataset is split into three subsets: training, validation, and test. The training subset is employed for the model optimization, the validation subset for the parameter tuning and monitoring the model convergence, and the independent test subset for the performance evaluation. The characteristics of the dataset employed in the experiment are presented in Table.2.

Table.2. Dataset description and experimental partition

Dataset Characteristic	Description
Dataset	REFUGE
Total images	1,200
Image modality	Color retinal fundus photographs
Main task	Glaucoma classification
Anatomical task	Optic disc segmentation
Training samples	840
Validation samples	180
Testing samples	180
Input resolution	Resized to 512×512
Image type	RGB
Segmentation target	Optic disc
Classification classes	Glaucoma, non-glaucoma

The dataset architecture described in Table.2 ensures separate data for model training and independent evaluation. The inclusion of both glaucoma labels and optic disc segmentation enables the application of multi-task learning technique of AVCT-Net where segmentation supervision helps the shared encoder to learn features that correspond to clinically significant anatomical structures instead of using only image characteristics. Various methods from the literature of related work are chosen for the purpose of comparison: M-Net, Attention U-Net, and SwinCup-DiscNet.

4.3 ACCURACY RESULTS

Accuracy of all models is compared by evaluating their performance at intervals of five epochs from 20 to 100 epochs. Table.3 shows the progress of the three related literature methods and AVCT-Net.

Table.3. Accuracy comparison at different training epochs

Epochs	M-Net (%)	Attention U-Net (%)	SwinCup-DiscNet (%)	AVCT-Net (%)
20	88.4	89.7	91.2	93.5
25	89.1	90.5	92.0	94.1
30	90.0	91.3	92.8	94.8
35	90.8	92.0	93.5	95.2
40	91.5	92.6	94.1	95.7
45	92.1	93.1	94.7	96.1
50	92.8	93.6	95.2	96.4
55	93.2	94.0	95.6	96.7
60	93.7	94.4	96.0	96.9
65	94.1	94.8	96.3	97.1
70	94.4	95.1	96.6	97.3
75	94.7	95.4	96.8	97.5
80	95.0	95.7	97.0	97.7
85	95.2	95.9	97.2	97.9
90	95.4	96.1	97.4	98.1
95	95.6	96.3	97.6	98.3
100	95.8	96.5	97.8	**98.5**

The results reported in Table 3 reveal that AVCT-Net achieves an accuracy of 98.5% with 100 epochs, whereas the accuracy of M-Net, Attention U-Net, and SwinCup-DiscNet is 95.8%, 96.5%, and 97.8%, respectively. Thus, AVCT-Net improves by 2.7, 2.0, and 0.7% over the above-mentioned models, respectively. With 20 epochs of training, the proposed network achieves 93.5% of accuracy, which is 5.1% better than that of M-Net.

4.4 RESULTS IN TERMS OF PRECISION

Results for precision are given in Table.4 at 5-epoch intervals. The proposed technique yields better precision than the other three techniques in each iteration of training.

Table.4. Precision comparison at different training epochs

Epochs	M-Net (%)	Attention U-Net (%)	SwinCup-DiscNet (%)	AVCT-Net (%)
20	86.9	88.1	89.8	92.4
25	87.7	88.9	90.5	93.0
30	88.4	89.7	91.2	93.6
35	89.1	90.4	91.9	94.1
40	89.8	91.0	92.5	94.6
45	90.5	91.6	93.1	95.0
50	91.2	92.1	93.7	95.4
55	91.8	92.6	94.2	95.7
60	92.3	93.1	94.6	96.0
65	92.8	93.5	95.0	96.3
70	93.2	93.9	95.4	96.5
75	93.6	94.3	95.8	96.7
80	94.0	94.7	96.1	96.9
85	94.3	95.0	96.4	97.1
90	94.6	95.3	96.7	97.3
95	94.9	95.6	96.9	97.5
100	95.2	95.9	97.1	**97.7**

AVCT-Net attains 97.7% precision within 100 epochs, as depicted in Table 4. The achievement is better than that of M-Net, Attention U-Net, and SwinCup-DiscNet by 2.5%, 1.8%, and 0.6%, respectively. The increase in the precision value suggests that the developed model produces fewer erroneous predictions of glaucomatous eye disease. At 20 epochs, AVCT-Net achieves 92.4%, whereas M-Net gains 86.9%, which shows a difference of 5.5%. The enhanced accuracy of the prediction process is a result of attention-based localization and capsule architecture-based representation of structure.

4.5 RESULTS ACCORDING TO RECALL

The results of recall are provided in Table.5. It is essential to obtain high recall values for detecting glaucomatous cases since it may have grave consequences if there is no detection of the condition.

Table.5. Recall comparison at different training epochs

Epochs	M-Net (%)	Attention U-Net (%)	SwinCup-DiscNet (%)	AVCT-Net (%)
20	87.8	89.2	90.6	93.0
25	88.6	90.0	91.3	93.6
30	89.4	90.8	92.0	94.2
35	90.1	91.5	92.7	94.7
40	90.8	92.1	93.3	95.2
45	91.5	92.7	93.9	95.6
50	92.1	93.2	94.5	95.9
55	92.7	93.7	95.0	96.2
60	93.2	94.2	95.4	96.5
65	93.7	94.6	95.8	96.8
70	94.1	95.0	96.1	97.0
75	94.5	95.4	96.4	97.2
80	94.9	95.8	96.7	97.4
85	95.2	96.1	97.0	97.6
90	95.5	96.4	97.2	97.8
95	95.8	96.7	97.4	98.0
100	96.1	97.0	97.6	**98.2**

From Table.5, it can be observed that AVCT-Net has 98.2% recall after 100 epochs, whereas M-Net has 96.1%, Attention U-Net has 97.0%, and SwinCup-DiscNet has 97.6%. In terms of improvement, it is 2.1, 1.2, and 0.6%, respectively. The significance of high recall lies in the ability of the model to identify 98.2% positive glaucoma samples in the tested setting. After 20 epochs, AVCT-Net gives a recall of 93.0%, which is 2.4% higher than that of SwinCup-DiscNet. It can be stated that anatomical segmentation supervision and global contextual learning lower false negatives.

4.6 CLASSIFICATION PERFORMANCE BASED ON F1-SCORE

F1-score results are given in Table.6. It is a combination of precision and recall and hence evaluates classification performance comprehensively.

Table.6. F1-score comparison at different training epochs

Epochs	M-Net (%)	Attention U-Net (%)	SwinCup-DiscNet (%)	AVCT-Net (%)
20	87.3	88.6	90.2	92.7
25	88.1	89.4	90.9	93.3
30	88.9	90.2	91.6	93.9
35	89.6	90.9	92.3	94.4
40	90.3	91.5	92.9	94.9
45	91.0	92.1	93.5	95.3
50	91.6	92.6	94.1	95.6
55	92.2	93.1	94.6	95.9
60	92.7	93.6	95.0	96.2
65	93.2	94.0	95.4	96.5
70	93.6	94.5	95.7	96.7
75	94.0	94.8	96.1	96.9
80	94.4	95.2	96.4	97.1
85	94.7	95.5	96.7	97.3
-90	95.0	95.8	97.0	97.5
95	95.3	96.1	97.2	97.7
100	95.6	96.4	97.4	**97.9**

From the results shown in Table.6, it can be noted that the AVCT-Net attains an F1-score of 97.9%, while M-Net, Attention U-Net, and SwinCup-DiscNet have achieved 95.6%, 96.4%, and 97.4%, respectively. Therefore, this suggests an enhancement of 2.3, 1.5, and 0.5% in the F1-score for the proposed model. From the above result, it can be deduced that the improvement is not restricted to precision or recall. Rather, it can be noted that there is consistency between the two metrics for the proposed architecture. The increment from 92.7% in epoch 20 to 97.9% in epoch 100 also shows consistency in convergence.

4.7 RESULTS BASED ON DICE SIMILARITY COEFFICIENT

The Dice similarity coefficient evaluates the optic disc segmentation performance. The results are given at an interval of five epochs.

Table.7. Dice similarity coefficient comparison at different training epochs

Epochs	M-Net (%)	Attention U-Net (%)	SwinCup-DiscNet (%)	AVCT-Net (%)
20	88.6	89.8	91.0	93.1
25	89.2	90.4	91.7	93.7
30	89.9	91.0	92.4	94.3
35	90.6	91.6	93.1	94.8
40	91.2	92.2	93.7	95.3
45	91.8	92.8	94.3	95.7
50	92.4	93.3	94.9	96.1
55	92.9	93.8	95.4	96.4
60	93.4	94.3	95.9	96.7
65	93.9	94.7	96.3	97.0
70	94.3	95.1	96.7	97.2
75	94.7	95.5	97.0	97.4
80	95.1	95.8	97.3	97.6
85	95.4	96.1	97.5	97.8
90	95.7	96.4	97.7	98.0
95	96.0	96.7	97.9	98.2
100	96.3	97.0	98.1	**98.4**

The Table.7 indicates that the AVCT-Net attains 98.4% Dice coefficient at 100 epochs, whereas M-Net, Attention U-Net, and SwinCup-DiscNet achieve 96.3%, 97.0%, and 98.1%, respectively. Thus, the designed architecture enhances segmentation overlap by 2.1%, 1.4%, and 0.3% as compared to the three competing approaches. This enhancement proves the

efficiency of fusing attention-driven local features with global context extracted using the transformer and structural information obtained through the capsule architecture. Moreover, at 20 epochs, the AVCT-Net achieves 93.1%, which shows fast learning capability. Also, at 100 epochs, the designed network attains 98.4% Dice coefficient, proving the efficiency of the multi-task learning approach.

4.8 DISCUSSION OF RESULTS

According to Tables 3-7, the final accuracy is equal to 98.5%, and precision, recall, F1-score, and Dice coefficient are equal to 97.7%, 98.2%, 97.9%, and 98.4%, respectively. It means that the proposed framework provides effective diagnostic classification and accurate optic disc localization. This effect can be especially seen when AVCT-Net is compared with M-Net. In this case, the final gain is equal to 2.7% for accuracy, 2.5 for precision, 2.1 for recall, 2.3 for F1-score, and 2.1 for Dice. When AVCT-Net is compared with Attention U-Net, the gain is equal to 2.0, 1.8, 1.2, 1.5, and 1.4% for accuracy, precision, recall, F1-score, and Dice, respectively. For SwinCup-DiscNet, the gain is 0.7 points for accuracy, 0.6 for precision, 0.6 for recall, 0.5 for F1-score, and 0.3 for Dice. The results show that the main strength of AVCT-Net is in complementary application of several representation mechanisms. The attention module pays attention to the relevant areas of optic nerve, the transformer considers global retinal dependence, and the capsule layer maintains structural dependence. The segmentation branch adds anatomical guidance to the diagnostic path. Thus, the model is not limited by image-level features for classification. Moreover, the learning process of the model indicates stable convergence. While the accuracy is 93.5% at epoch 20 and 98.5% at epoch 100, Dice is 93.1% and 98.4%, respectively. The corresponding F1-score is 92.7% and 97.9%. These facts show that the proposed model still learns relevant representations during training, and it does not show significant performance saturation in an early epoch.

5. CONCLUSION

From the experimental evaluation, the proposed Attention-Driven Vision Capsule Transformer Network (AVCT-Net) presents an efficient unified framework to solve glaucoma classification and optic disc segmentation tasks. The framework employs convolutional features learning, multi-scale attention mechanism, vision-transformer contextual modeling, capsule-based spatial representation, and multi-tasking learning to overcome the shortcomings of conventional methods for glaucoma diagnosis. The final results show 98.5%, 97.7%, 98.2%, 97.9% and 98.4% performance values concerning accuracy, precision, recall, F1-score and Dice coefficient, respectively. These results reflect good trade-off between discriminative power and anatomical segmentation of the framework. In particular, the 98.2% recall is essential for glaucoma detection since it shows high potential of the framework to detect positive cases. The 98.4% Dice coefficient shows the high quality of the optic disc localization, which confirms the anatomical interpretability of the framework. Moreover, AVCT-Net outperforms M-Net, Attention U-Net, and SwinCup-DiscNet consistently in all metrics.

REFERENCES

[1] M. Alawad, A. Aljouie, S. Alamri, M. Alghamdi and A. Almazroa, "Machine Learning and Deep Learning Techniques for Optic Disc and Cup Segmentation-A Review", *Clinical Ophthalmology*, Vol. 43, No. 2, pp. 747-764, 2022.

[2] Z. Wang, P.A. Keane, M. Chiang and D.S.W. Ting, "Artificial Intelligence and Deep Learning in Ophthalmology", *Artificial Intelligence in Medicine*, Vol. 45, No. 2, pp. 1519-1552, 2022.

[3] M. Alawad, A. Aljouie, S. Alamri, M. Alghamdi and A. Almazroa, "Machine Learning and Deep Learning Techniques for Optic Disc and Cup Segmentation-A Review", *Clinical Ophthalmology*, Vol. 34, pp. 747-764, 2022.

[4] A. Almazroa, R. Burman, K. Raahemifar and V. Lakshminarayanan, "Optic Disc and Optic Cup Segmentation Methodologies for Glaucoma Image Detection: A Survey", *Journal of Ophthalmology*, Vol. 2015, No. 1, pp. 180972-180989, 2015.

[5] Q. Ba, J.X. Jiang, Y. Li and Z. Wang, "Fine-Grained Perception for Fundus and Prostate Medical Image Segmentation", *Sensors*, Vol. 26, No. 9, pp. 2879-2893, 2026.

[6] S.A. Alryalat, P. Singh, J. Kalpathy-Cramer and M.Y. Kahook, "Artificial Intelligence and Glaucoma: going back to Basics", *Clinical Ophthalmology*, Vol.45, No. 2, pp. 1525-1530, 2023.

[7] M. Alawad, A. Aljouie, S. Alamri and A. Almazroa, "Machine Learning and Deep Learning Techniques for Optic Disc and Cup Segmentation-A Review", *Clinical Ophthalmology*, Vol. 43, No. 2, pp. 747-764, 2022.

[8] A. Diaz Pinto, S. Morales, V. Naranjo and A. Navea, "CNNs for Automatic Glaucoma Assessment using Fundus Images: An Extensive Validation", *Biomedical Engineering*, Vol. 18, No. 1, pp. 1-29, 2019.

[9] V. Mahalakshmi and S. Karthikeyan, "Clustering based Optic Disc and Optic Cup Segmentation for Glaucoma Detection", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol 2, No. 4, pp. 3756-3761, 2014.

[10] Q. Zhu, X. Chen, Q. Meng, J. Song and W. Zhu, "GDCSeg-Net: General Optic Disc and Cup Segmentation Network for Multi-Device Fundus Images", *Biomedical Optics Express*, Vol. 12, No. 10, pp. 6529-6544, 2021.

[11] J. Camara, A. Neto, I.M. Pires, M.V. Villasana and A. Cunha, "Literature Review on Artificial Intelligence Methods for Glaucoma Screening, Segmentation, and Classification", *Journal of Imaging*, Vol. 8, No. 2, pp. 1-19, 2022.

[12] R. Chilukuri, P. Praveen, R.K. Gatla and R.A. Almenweer, "SwinCup-DiscNet: A Fusion Transformer Framework for Glaucoma Diagnosis using Optic Disc and Cup Features", *Scientific Reports*, Vol. 16, No. 1, pp. 7920-7943, 2026.