UViT-Transformative Vision: Fusion of UNet and ViT for Precise Optic Disc and Cup Segmentation

Mitesh Sanjay Jalan (115222506) | Smit Kumbhani (114964474),
Mohammad Khakhariyawala (115363539) | Gaurav Kamdar (115919985)

Abstract

Glaucoma, a prominent cause of irreversible blindness, necessitates early detection through optic cup-to-disc ratio (CDR) assessment. Addressing limitations in current Convolutional Neural Network (CNN)-based segmentation approaches, we introduce UViT-Net a novel model seamlessly fusing U-Net and transformer architectures. Our proposed UViT-Net incorporates an attention-gated bilinear fusion scheme and Multi-Head Contextual attention but also surpasses the performance of individual UNet and transformer models, as evidenced by evaluations on the DRISHTI-GS dataset. This innovative approach holds promise for advancing glaucoma risk evaluation and contributing to the field of Computer-Aided Diagnosis (CAD) systems.

1. INTRODUCTION

1.1 Background

Glaucoma, the second-largest cause of blindness globally and the leading cause of irreversible blindness is characterized by increased intraocular pressure, resulting in optic cup deformation and retinal nerve fiber thickening. Given its irreversible nature, early detection through screening is crucial for preserving vision. Traditional optic nerve head (ONH) assessment, a common screening method, involves binary classification between healthy and diseased subjects, often relying on metrics such as cup-to-disc ratio (CDR). Manual assessment is time-consuming and impractical for large-scale screening, necessitating automated and accurate segmentation methods.

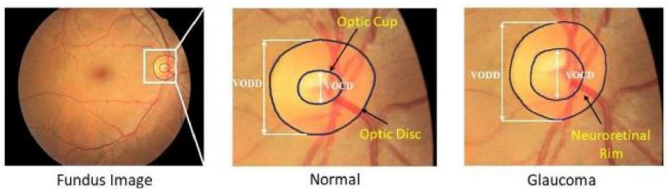


Fig. 1: Structure of retinal fundus: the region enclosed by the outer ring is the optic disc, whereas the region enclosed inside the smaller ring is the optic cup, and the region between them is the neuroretinal rim. The optic cup gets larger in case of glaucoma. The figure is taken from the open-source DRISHTIGS dataset [1].

1.2 Motivation

This project focuses on addressing the challenges associated with glaucoma detection, emphasizing the importance of precise segmentation of the optic disc (OD) and optic cup (OC). Accurate segmentation enables the calculation of CDR, a key indicator of glaucoma risk. While various clinical methods and deep learning approaches have been proposed, there is a need for a robust solution that combines the strengths of Convolutional Neural Networks (CNNs) and transformers, offering global contextual information and efficient feature extraction.

1.3 Objectives

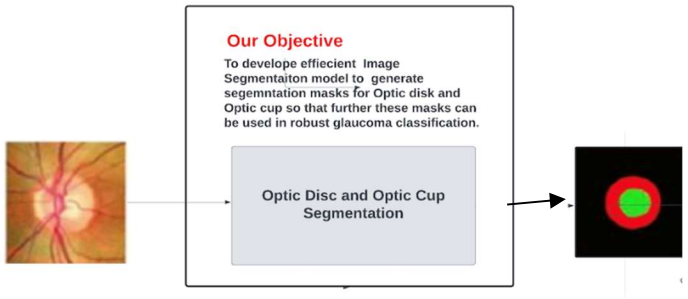


Fig. 2: Full Workflow of our project

Our main objective is to improve Glaucoma detection by developing a novel image segmentation approach. This can be broken down as follows:

1. Combine the U-Net and transformer architectures.
2. Focus on joint segmentation of the Optic Disc (OD) and Optic Cup (OC) while generating accurate masks.
3. Leverage the capacity of transformers to capture global context along with that of U-Net architecture for localization accuracy and feature discrimination.

1.4 Contributions:

Key contributions encompass the introduction of a unique encoder-decoder architecture, the implementation of Multi-Head Contextual Self Attention for transformer enhancement, the incorporation of a sophisticated bilinear fusion module for intelligent feature synthesis, the integration of multiscale context extractors, and the validation of UViT-Net's[6] superior performance on publicly available dataset.

2. RELATED WORK

The fusion of Convolutional Neural Networks (CNNs) and transformers has significantly advanced medical image segmentation [4]. While traditional CNNs focus on local features, transformers, originally designed for natural language processing, excel at capturing global contextual information [3]. SETR, a prominent transformer-based model, faces challenges in handling spatial information [3]. Recent works aim to address this by integrating transformers and CNNs [2] [20].

Modifications to U-Net architectures, such as deconvolutions [5] [8], and skip connections [7] [9], enhance feature extraction and resolution but struggle with accurate localization of small, subtle structures like the optic cup in glaucoma detection.

In response, the UVit-Net model is introduced [12], combining transformers and U-Net in parallel branches of an encoder-decoder pipeline. UVit-Net leverages transformers' global context awareness and U-Net's feature extraction capabilities. The model incorporates a multi-head contextual self-attention module and a bilinear fusion module [12].

UVit-Net's evaluation on retinal fundus image dataset demonstrates its superiority over existing methods [12]. This multi-encoder architecture proves promising for accurate segmentation and glaucoma detection, offering valuable support for medical professionals in diagnosis and treatment planning.

3. METHODOLOGY

3.1 UNet-encoder:

The initial segment of the downsampling pathway in the proposed design involves an encoder based on the U-Net [12] architecture, depicted in Figure 3. This pathway comprises 5 stages, each encompassing a 3 × 3 convolutional process, succeeded by batch normalization. To introduce non-linearity to the model, the output of each batch normalization undergoes a Rectified Linear Unit (ReLU) activation, followed by a 2 × 2 MaxPooling operation. The downsampling process systematically extracts features from the images, progressively augmenting the dimensions of these features in each layer. Consequently, a final layer is reached, yielding a high-dimensional feature representation enriched with substantial semantic information. The incorporation of skip connections facilitates the recovery of fine-grained details from the encoder during the upsampling phase of the network.

3.2 Transformer-encoder:

A transformer encoder employs stacks of self-attention, treating query and key-value pairs with point-wise fully connected layers. The result is a weighted sum of values, with weights computed by a compatibility function of the corresponding query. CNNs, when applied sequentially, struggle to capture essential long-range dependency information crucial for semantic segmentation [10]. To address this limitation, a transformer encoder is introduced to effectively capture global contextual information. For self-containment, we provide a brief overview of the transformer encoder network.

Initially, an input image, $Z \in \{H \times W \times C\}$, is evenly divided into $N = H/P \times W/P$ patches. The patch grid is flattened into a sequence that passes through a linear embedding layer with an output dimension of Dl. Learnable positional embeddings of matching dimensions are added to the raw embedding sequence, $eb \in FN \times Dl$, to incorporate spatial priors. The resulting embedding sequence, Zl, serves as the input to the 12-layered transformer encoder, comprising Multi-head Self-Attention (MSA) [13] followed by a multi-layer perceptron (MLP) [14]. Notably, instead of traditional transformers, we employ Multi-Head Contextual Self-Attention (MCSA) in place of MSA.

Multi-Head Contextual Self-Attention: In vision transformers, a self-attention module is typically applied to 2-D feature maps, generating an attention matrix, A, from isolated query-key pairs at a specific spatial location. However, this approach often neglects the global context present in neighbouring keys, particularly problematic for medical datasets due to high cross-image similarity. To address this issue, we introduce the contextual self-attention block, as illustrated in Figure 3.

The self-attention (SA) mechanism within a transformer enhances the status of individual patches through the comprehensive consolidation of features, a formulation exemplified in Equation (1).

$$SA(Z) = Z + softmax(\frac{ZW_q(ZW_k)^T}{\sqrt{d}})(ZW_v) \quad (1)$$

In the self-attention mechanism, denoted as Z, input is represented as a triplet comprising a query (Q), a key (K), and a value (V), where Q = ZWq, K = ZWk, and V = Z. Here, Wv, Wq, and Wk are trainable parameters. In contrast to the 1×1 convolutions utilized in vision transformers for encoding individual keys, MCSA (Multi-Context Self-Attention) employs a group k×k convolution on neighbouring keys. This process constructs a spatial k×k grid, incorporating contextually rich key representations. Consequently, the newly learned keys (K′ ∈ RH×W×C) encompass contextual information from local adjacent keys, serving as a consistent representation of X in space.

The concatenation of the key (K′) and query (Q) leads to the generation of an attention matrix through two sequential convolutional operations with learnable weights Wθ and Wϕ, where Wθ has ReLU activation, and Wϕ lacks activation. The attention matrix is formulated as per Equation (2).

$$A(Z) = [QK'^T]W_\theta W_\phi \quad (2)$$

In each head, the attention matrix at every spatial location of A is learned by utilizing the contextual query-key pair instead of the localized one. This enriches the self-attention output by incorporating statically mined K′. The attentive feature map (K″) is obtained by aggregating values, following the regular self-attention framework outlined in Equation (3).

$$K'' = ZW_v \odot A(Z) \quad (3)$$

The MCSA involves m such self-attention operations to project a concatenated output, which is then subjected to a residual skip operation, as expressed in Equation (4).

$$output = MCSA(Z) + MLP(MCSA(Z)) \quad (4)$$

Subsequently, layer normalization is applied, and the result is forwarded to the bilinear fusion block. The overall transformer encoder architecture is illustrated in Figure 3.

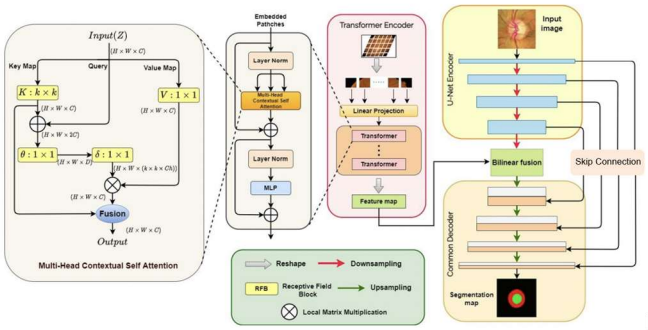


Fig. 3: Structure of the UViT-Net, comprising two distinct encoder branches: U-Net and Transformer. The Transformer processes patch-based images, while the U-Net handles the entire image. Extracted features from both encoders undergo fusion through a bilinear fusion block, followed by a shared decoder path, ultimately generating the final segmentation map.

3.3 Bilinear Fusion:

To achieve effective integration of features derived from the dual encoder branches, we introduce a bilinear fusion block illustrated in [6] and Figure 4. This block incorporates a self-attentive multi-modal fusion mechanism. The fusion process involves a sequence of linear operations applied to features extracted from the two encoders operating in parallel branches, culminating in their merging. Global information from the transformer branch is enhanced using channel attention. Meanwhile, to eliminate potential noise in the low-level features derived from the U-Net branch, a spatial attention operation is employed. Ultimately, a Hadamard product is computed using the weights from the respective branches, facilitating a nuanced interaction between these two feature sets.

Channel Attention[6][16]: The filters in each channel operate on a limited local receptive field, leading to suboptimal exploration of features beyond this local area. To address this issue, we alleviate the problem by incorporating global spatial information from the transformer branch. This is achieved by compressing the spatial information into a channel descriptor through average pooling, thereby creating statistics oriented towards each channel.

Initially, the input feature F ∈ RH×W×C undergoes a reshaping process to produce intermediate features F1 ∈ R(H∗W)×C and F2 ∈ RC×(H∗W). Subsequently, these intermediate features are multiplied and divided by the square root of a factor C. Ultimately, the influence of the b^{th} channel on the a^{th} channel, referred to as the channel attention map, is determined through the Softmax operation:

$$\mathcal{M}_c^{a,b} = Softmax\left(\frac{f(F_a, F_b)}{\sqrt{C}}\right) \quad (5)$$

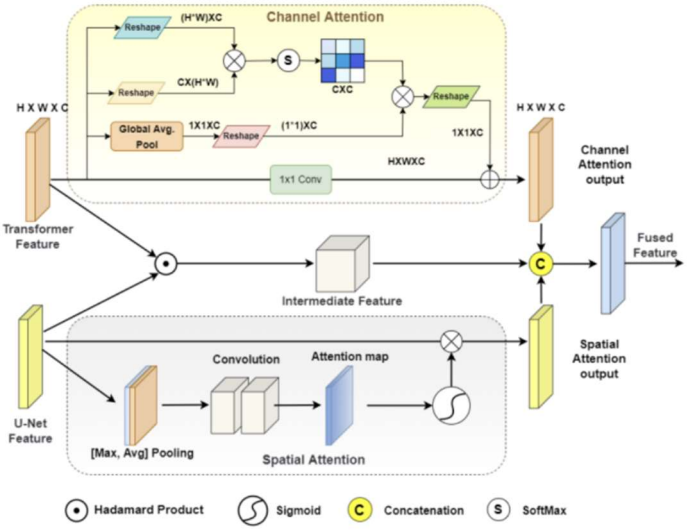


Fig. 4: Bilinear fusion scheme of features extracted from two different encoder branches.

The function f is utilized for computing the mutual reliance among channels. Furthermore, we utilize global average pooling to capture comprehensive discriminatory features across channels, incorporating high-level semantic information. The output of the pooling process undergoes a 1 × 1 convolution, gets multiplied by Mc, and is then reshaped into a dimension of 1 × 1 × C.

Spatial Attention [6]: The U-Net encoder's output undergoes processing through the spatial attention segment within the bilinear fusion block. After combining the channel attention and spatial attention map through concatenation, a feature fusion is performed between the transformer and the U-Net encoder. Ultimately, by employing weights W1 and W2 for the corresponding branches in the bilinear fusion block, the Hadamard product of the input features is obtained, as expressed in Equation (7).

$$B_i = Conv(Z_i W_{1i} \odot U_i W_{2i}) \quad (7)$$

In Equation (7), ⊙ denotes the Hadamard operation, where Zi and Wi represent input features from the two branches. Subsequently, the resulting Bi, obtained through the Hadamard product, is concatenated with the outputs from the spatial and channel attention branches. This concatenation forms the conclusive output of the bilinear fusion block. The spatial attention block's representation is illustrated in Figure 4.

3.4 LOSS FUNCTION:

This study employs a combined use of the Dice coefficient, Intersection over Union (IoU), and Binary Cross Entropy (BCE) loss to train the network. The segmentation pipeline relies on ground truth for supervision throughout the segmentation process. The BCE loss evaluates individual pixels, ensuring equal learning for each pixel without gradient explosion. The Dice coefficient gauges the overlap between ground truth and a segmentation map. These BCE [11], Dice, and IoU losses are represented by Equations (8), (9), and (10), respectively.

$$L_{BCE}(GT,S) = -(GT_i \cdot \log S_i + (1-GT_i)\cdot \log(1-S_i)) \quad (8)$$

$$\mathcal{L}_{Dice}(GT,S) = 1 - \sum_{k=1}^{c} \frac{2\omega_k \sum_{j=1}^{n} S(k,j)GT(k,j)}{\sum_{j=1}^{n} S(k,j)^2 + \sum_{j=1}^{N} GT(k,j)^2} \quad (9)$$

$$\mathcal{L}_{IoU}(GT,S) = 1 - IoU(GT,S) - \frac{|X - GT \cup S|}{|X|} \quad (10)$$

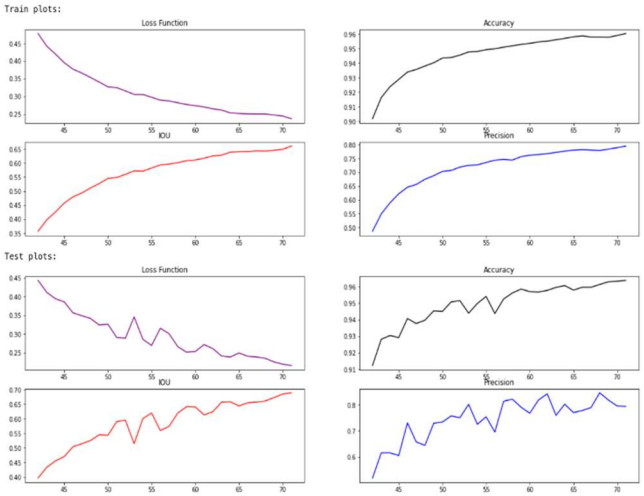
The segmentation task involves defining ground truth (GT) and prediction map (S) with individual images GTi and Si for N available image data, where i ∈ {1,2, 3,..., N} and k ∈ {1,2,3,...,c}. The pixel number is denoted as n, and ωk represents the weight of the kth class. Variable X signifies the smallest bounding box covering the segmentation map and ground truth. The overall loss function is given in Equation 11.

$$L(GT,S) = L_{MAE}(GT,S) + \lambda_1 L_{Dice}(GT,S) + \lambda_2 L_{IoU}(GT,S) + \lambda_3 L_{BCE}(GT,S) \quad (11)$$

Experimentally set hyperparameters λ1, λ2, λ3, and λ4 are 0.15, 0.4, 0.3, and 0.15, respectively. LMAE denotes mean absolute error loss, detailed in Section IV of the supplementary material.

The proposed method efficiently segments optic disc and optic cup from retinal images, utilizing CNN and transformer network advantages. Accurate segmentation aids in detecting Glaucoma onset by measuring the Cup-to-Disc Ratio (CDR). Experimental results and comparisons with state-of-the-art methods are discussed in the following section.

4. TRAINING



Graph I: Test vs Train results for loss, Accuracy, IOU Precision

In training the UViT Net, a convolutional neural network architecture, with a batch size of 1, a learning rate of 0.0002, and an Adam optimizer over 75 epochs, remarkable performance was achieved. The model demonstrated high training accuracy (96.07%), precision (79.44%), and Intersection over Union (IOU) of 66.29%. Additionally, on the test set, the UViT Net exhibited a competitive test accuracy of 95.97%, precision of 85.94%, and IOU of 65.26%. The consistent learning rate of 0.0002 contributed to stable convergence. The results underscore the efficacy of the UViT-Net in learning complex representations, as evidenced by its superior performance in image classification and segmentation tasks. This suggests its potential for practical applications in computer vision domains, affirming its robustness and generalization capabilities.

5. EXPERIMENTAL RESULTS

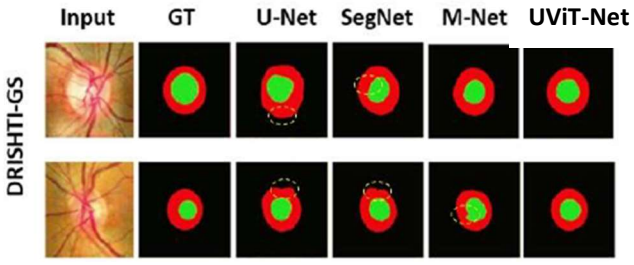


Fig. 5: Visual comparison of the results obtained by the proposed UViT-Net on the dataset with the existing state-of-the- art methods. The red and green regions represent the optic disc and optic cup respectively. The highlighted regions represent the wrongly segmented parts

The UViT-Net model was evaluated on the publicly available dataset—DRISHTI-GS using three metrics: Intersection over Union (IoU), Sensitivity, and Accuracy. The results, detailed in Table I, demonstrate strong performance in optic disc (OD) and optic cup (OC) segmentation on the dataset for all the models. Visual comparisons in Figure 5 illustrate the UViT-Net's effectiveness, with red and green regions representing the segmented optic disc and optic cup, respectively. The proposed model consistently outperforms some of the existing methods in terms of IoU, sensitivity, and accuracy. Notably, UViT-Net excels in OD and OC segmentation compared to other methods, showcasing its robustness. Despite the challenging nature of the segmentation due to its obscure boundary within the optic disc and cup, UViT-Net achieves commendable results, further emphasizing its efficacy.

Model Name	Metrics		
	IoU	Sensitivity	Accuracy
SegNet	73.86	87.11	99.91
Transformer-SegNet	63.38	70.62	95.01
M-Net	62.95	69.98	94.77
UNet	62.43	69.76	94.03
UViT (proposed)	***65.26***	***71.76***	***95.97***

TABLE I: Quantitative results of the proposed UViT-Net on the publicly available dataset for OD and OC segmentation. IoU: Intersection over Union.

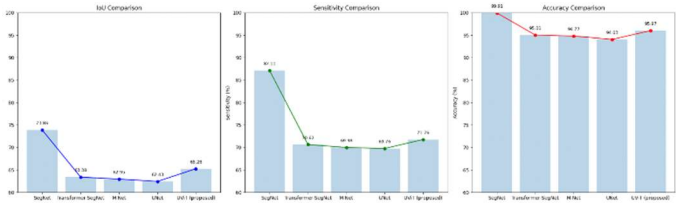


Fig. 6: Evaluation of different models where each bar in the histogram represents a specific model, and the height of the bar corresponds to the prevalence of IoU, Sensitivity, and Accuracy scores.

The Output analysis further underscores the superior performance of UVit-Net [6] compared to some other models, including M-Net, Transformer-SegNet, and U-Net. Notably, UViT-Net [6] consistently exhibits higher scores across all metrics. This vividly illustrates the robustness and efficacy of UViT-Net [6], reaffirming its position as one of the

leading models in terms of segmentation accuracy, sensitivity, and overall performance.

In the evaluation of segmentation models within the context of medical imaging, our proposed UViT-Net[6] model distinguishes itself as a preeminent performer across pivotal metrics, substantiating its efficacy in precise object boundary delineation and pixel classification. The Intersection over Union (IoU), a critical gauge of segmentation accuracy, positions UViT[6] as the foremost contender, exhibiting a notably high IoU of 65.26. This surpasses the performance of competing models such as Transformer-SegNet[15], M-Net, and UNet, with SegNet showing a marginally higher IoU at 73.86.

The sensitivity metric, elucidating the model's discernment of positive instances, fortifies UViT's superiority. Recording a sensitivity score of 71.76, UViT surpasses its counterparts, attesting to its proficiency in accurately identifying true positive regions. While SegNet closely leads with a sensitivity of 87.11, UViT sustains a competitive score. Conversely, Transformer-SegNet[15], M-Net, and UNet manifest lower sensitivities at 70.62, 69.98, and 69.76, respectively, indicating a relatively higher rate of false negatives in comparison to UViT.

Concerning overall accuracy, UViT again emerges as the model boasting one of the highest accuracy scores of 95.97, underscoring its robust performance in accurate pixel classification. While SegNet maintains a commendable accuracy of 99.91, UViT-Net still outperforms it in the overarching accuracy metric. Transformer-SegNet, with an accuracy of 95.01, demonstrates effective pixel classification but falls short of UViT's performance. M-Net and UNet report accuracy scores of 94.77 and 94.03, respectively, further highlighting UViT's efficacy in achieving superior accuracy.

CONCLUSION

In conclusion, our UViT[6] model exhibits heightened performance relative to some of the existing models, establishing its prominence in the domain of medical image segmentation. With high IoU, sensitivity, and accuracy, UViT[6] stands as a robust and effective solution for precise object delineation and pixel classification in the realm of medical imaging applications.

Reference

[1] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[2] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[3] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[4] A. Sevastopolsky, "Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network," *Pattern Recognition and Image Analysis*, vol. 27, no. 3, pp. 618–624, 2017.

[5] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J., "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE Transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[6] Rukhshanda Hussain, Hritam Basak, "UT-Net: Combining U-Net and Transformer for Joint Optic Disc and Cup Segmentation and Glaucoma Detection", *arXiv:2303.04939*

[7] M. K. Hasan, M. A. Alam, M. T. E. Elahi, S. Roy, and R. Mart´ı, "Drnet: Segmentation and localization of optic disc and fovea from diabetic retinopathy image," *Artificial Intelligence in Medicine*, vol. 111, p. 102001, 2021.

[8] S. M. Shankaranarayana, K. Ram, K. Mitra, and M. Sivaprakasam, "Fully convolutional networks for monocular retinal depth estimation and optic disc-cup segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1417–1426, 2019.

[9] A. Tulsani, P. Kumar, and S. Pathan, "Automated segmentation of optic disc and optic cup for glaucoma assessment using improved unet++ architecture," *Biocybernetics and Biomedical Engineering*, 2021

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[11] Shruti Jadon, "A survey of loss functions for semantic segmentation"

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multihead self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *arXiv preprint arXiv:1905.09418*, 2019.

[14] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[15] Robin Strudel, Ricardo Garcia, van Laptev, Cordelia Schmid, "Segmenter: Transformer for Semantic Segmentation", 2021 *IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 7242-7252, doi: 10.1109/ICCV48922.2021.00717*

[16] Nitin Kisan Ahire, R. N. Awale, Abhay Wagh, "Attention module-based fused deep cnn for learning disabilities identification using EEG signal" *Multimed Tools Appl ,2023.*