[1] **Lina Jia**
[2]**Qiang Yang***

Optic Cup and Optic Disc Segmentation Based on improved TransUnet

Journal of Electrical Systems

Abstract: - Glaucoma ranks as the second leading cause of blindness globally, surpassed only by cataracts. It inflicts irreversible harm to the optic nerve, and once vision is lost, restoration is unattainable for life[1]. Therefore, early detection of glaucoma is imperative.The cup-to-disc ratio serves as a primary diagnostic tool for identifying glaucoma. Generally, an excessively large cup-to-disc ratio in fundus photographs strongly suggests glaucoma. However, human error in diagnosing fundus photographs by clinicians is prevalent, leading to time-consuming, labor-intensive, expensive, and potentially inaccurate diagnoses during extensive screenings.To combat this challenge, an enhanced TransUnet-based method is proposed for optic cup and optic disc segmentation, aiding clinicians in large-scale glaucoma screenings. The model incorporates the Focus structure and CBAM structure. The Focus structure addresses information loss during single downsampling of images, preserving more data without changing image dimensions. Meanwhile, the CBAM structure integrates channel and spatial attention, enhancing the model's ability to extract features.On the ORIGA-650 dataset, this enhanced method achieved a mean Dice coefficient of 0.984 for the disc and 0.947 for the cup, along with a mean IOU value of 0.968 for the disc and 0.902 for the cup. Compared to alternative algorithms, the segmentation results exhibit superior accuracy, showcasing the efficacy of our proposed model.

Keywords: Deep Learning, Fundus Color, Optic Disc and Optic Cup Segmentation, TransUnet.

I. INTRODUCTION

Glaucoma is a common cause of blindness in ophthalmology, ranking as the second leading cause of blindness worldwide, following only cataracts.According to statistics from the World Health Organization (WHO), approximately 120 million people worldwide suffer from glaucoma, with many facing a high risk of vision impairment or blindness. Every year, over 3 million people become blind due to glaucoma.This disease often lacks obvious symptoms in the early stages, but it can lead to permanent damage to the optic nerve, resulting in gradual loss of vision function. Once the condition deteriorates to the point of blindness, patients will no longer be able to regain their sight.On the other hand, glaucoma has a relatively long progressive course. As long as treatment is initiated early in the disease, it can significantly delay the progression of the condition, allowing most patients to maintain a certain level of vision throughout their lives.The optic disc (OD) is the brighter area in the fundus image, approximately elliptical in shape. The optic cup (OC) is the variable-sized central depression present on the optic disc region. Previous studies[2] have shown that the larger the cup-to-disc ratio, the higher the probability of having glaucoma. Therefore, it is crucial to develop an accurate algorithm for segmenting the OC and OD[3].

A. Data set introduction

Fig. 1 illustrates examples of fundus color images from the ORIGA dataset along with their corresponding labels.

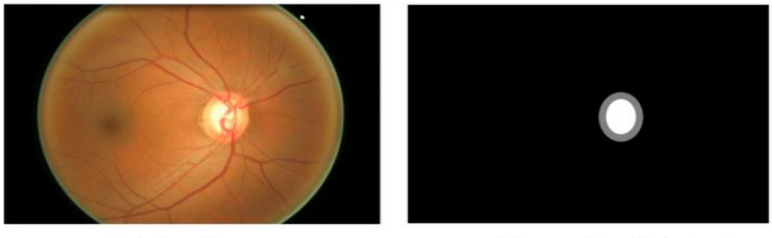

Figure 1: Example images from the ORIGA-650 dataset

Dataset: ORIGA-650

[1] Lina Jia, Affiliation: Academy of Artificial Intelligence,Beijing Institute of Petrochemical Technology, Beijing,102600, China.
[2] *Corresponding author: Qiang Yang, Affiliation: Academy of Artificial Intelligence,Beijing Institute of Petrochemical Technology, Beijing,102600, China. (Email: yangqiang@bipt.edu.cn)

The dataset chosen for this study is the publicly available dataset ORIGA-650. It was compiled by the ophthalmology research team at the National University of Singapore from various glaucoma screening projects, consisting of retinal images. The dataset comprises 650 color retinal images, each typically with a resolution of 2448x2050 pixels. These images contain detailed information on structures such as the optic cup, optic disc, and retinal vessels. Each image comes with corresponding manual annotations, which typically include boundary information for the optic cup and optic disc, as well as other features potentially relevant to glaucoma. The dataset encompasses patients of different ages, races, and case presentations.

II. METHODS

A. Previous relevant methods

With the continuous development of deep learning technology, significant progress has been made in the field of medical image segmentation. Many scholars have also conducted in-depth exploration and research on the segmentation of optic disc and optic cup in fundus photographs.

Maninis et al.[4] proposed a fully convolutional neural network based on the VGG-16 network and transfer learning techniques, which accurately performs segmentation of retinal vessels and optic discs.

Edupuganti et al.[5] implemented optic disc and cup segmentation using FCN.

Sevastopolsky et al.[6] applied an improved U-Net[7]for optic disc segmentation, incorporating Dropout operations after each convolutional layer to enhance the model's generalization capability. They further utilized the segmented optic disc results for the task of optic cup segmentation, achieving a more refined analysis of fundus images.

Fu et al.[8] proposed a deep learning network architecture named M-Net by improving the U-Net network. This structure mainly consists of multiscale input layers, a U-shaped convolutional network, lateral output layers, and a multi-label loss function. It achieves the segmentation of both optic cup and optic disc simultaneously in one system.

Shankaranarayana et al.[9] incorporated the idea of residual improvement into the U-Net network, proposing Res-UNet.

Y.Qin et al.[10] combined the deformable convolutional network[11] with the U-Net network, proposing a deformable U-Net network for optic disc and optic cup segmentation.

The deep learning-based segmentation algorithms for OD and OC primarily include Fully Convolutional Networks (FCN)[12] and U-Net, along with their modifications.U-Net demonstrates superior performance in segmenting medical images with large dimensions but small datasets. As a result, many scholars utilize improved versions of U-Net for medical image segmentation research, making U-Net one of the most common methods in the field. Many variants of U-Net have been proposed, such as 3DUNet[13], Res-UNet, U-Net++[14], U-Net3+[15], and others.With the emergence of Vision Transformer (VIT)[16], Transformer's applications in the medical imaging domain are also increasing.

Due to the limitations of traditional CNN convolutional operations, there are deficiencies in capturing distant semantic relationships, while Transformers exhibit excellent performance in global perception.Xu et al. proposed TransUNet[17] in 2021, which combines the strengths of CNN and Transformer, fully leveraging the advantages of both.The encoder part adopts traditional CNN for feature extraction and upsampling, which effectively captures local information and details of the image. Meanwhile, the decoder part utilizes transpose convolution for upsampling, aiding in restoring the spatial resolution of the image. Through skip connections, the encoder and decoder feature maps are fused, providing richer information and better spatial localization capability.On the other hand, TransUNet introduces Transformer at the middle position of the U-shaped structure, providing the model with powerful global perception capability, which helps improve the performance and generalization ability of the model.

B. Proposed Method

1) FOCUS MODULE

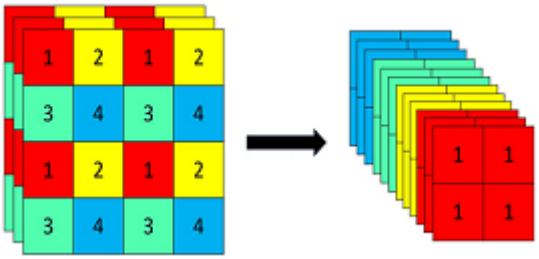

Figure 2: Focus Module

From Figure 2, medical image data is often limited in quantity and large in size. However, large-scale pretrained networks heavily rely on input image size. Resizing images to a fixed size before feeding them into the network model can result in the loss of valuable information. To address this issue, we introduce the Focus module.

In the YOLOv5 network, the Focus module slices the input image before it is passed into the backbone. Specifically, it divides the image into four parts, sampling every other pixel value, similar to nearest downsampling. The resulting four images are complementary in information, with similar sizes but no loss of information.

This process concentrates the width and height information of the image into the channel space, resulting in a fourfold increase in the number of input channels. It is equivalent to performing downsampling without information loss, thus preserving more information.

For example, if the input image size is 448×448×3, then each channel of the 448×448 image will be split into 224×224×4 images. Finally, they are stacked along the channel axis to obtain a tensor of size 224×224×12. Afterwards, it undergoes convolution, normalization, activation function, and is passed to the next layer.

2) CBAM MODULE

Woo et al.[18] proposed the Convolutional Block Attention[19] Module (CBAM), a lightweight and versatile module that seamlessly integrates into any CNN architecture and can be trained end-to-end with basic CNNs.

The CBAM moduleintegrates channel attention and spatial attention. Channel attention is used to adaptively adjust the attention weights of different channels in the feature map, while spatial attention is used to adaptively adjust the attention weights of different positions in the feature map. This provides a better attention mechanism to the model, enabling the encoder to more effectively extract image features during downsampling. As shown in Figure 3.

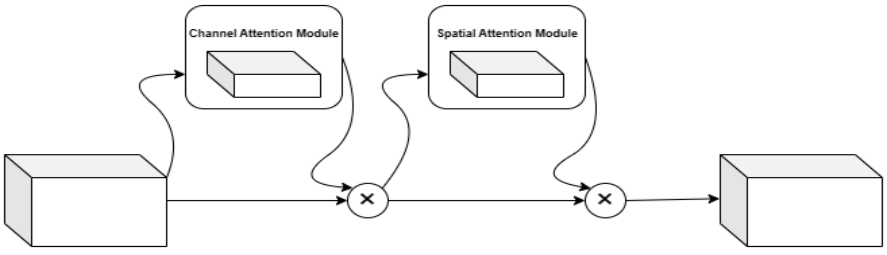

Figure 3: CBAM Module

3) THE PROPOSED MODEL ARCHITECTURE

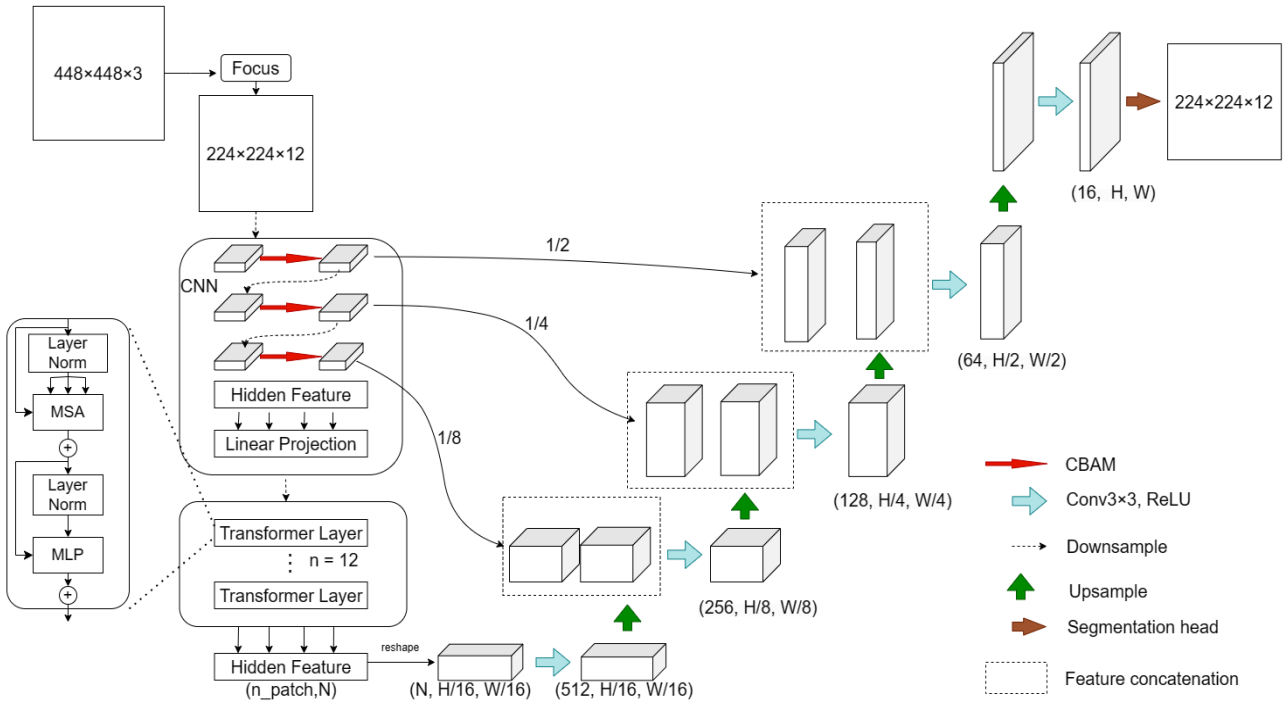

Figure 4: The proposed model architecture

As shown in Figure 4. A pre-trained ViT model was utilized, pre-trained on ImageNet[20].The overall algorithmic process is as follows: Firstly, the input image is downscaled using the Focus module, which employs interleaved sampling and concatenation to reduce the image size. The resulting feature maps are then stacked along the channel dimension and fed into the model for training. The image undergoes feature extraction by the CNN encoder, with the CNN component adding the CBAM module.Subsequently, it is inputted into a 12-layer Transformer module, followed by reshaping. The reshaped data is then passed to the decoder for upsampling four times. The feature maps in the decoder are fused with the feature maps from the encoder at the same level through skip connections. Finally, a 1×1 convolution is performed to complete the optic disc and optic cup segmentation.

A. Experimental Details

1) DATA PREPROCESSING

Image data preprocessing aims to improve the quality of images, reduce noise, enhance features, and prepare for subsequent model training. The image preprocessing methods used in this study are as follows:

Noise Removal Using Thresholding: Noise in the images is removed using a thresholding method. By appropriately selecting a threshold, bright noise points in the image are converted to black, thereby eliminating bright noise points in the image.

Data Augmentation: To address the challenge of insufficient glaucoma fundus image data and to improve the performance and robustness of the model, a series of data augmentation techniques are employed. These techniques involve randomly applying the following operations to the images:

Random Rotation: The dataset is randomly rotated to generate images with different angular inclinations.

Random Flipping: Horizontal and vertical flipping operations are applied to produce mirrored images.

Random Scaling: Images are randomly scaled, maintaining the content of the image while changing its size, thereby increasing the diversity and richness of the dataset.

Image Translation: Random translation of images is performed, changing the position of objects in the images and increasing the variability and diversity of the dataset.

2) EXPERIMENTAL ENVIRONMENT AND PARAMETER SETTINGS

Under the Windows 11 operating system and utilizing 16GB of memory with the PyCharm platform, this study employed PyTorch version 2.0.0 and conducted research based on Python 3.10. The computations were performed using an NVIDIA GeForce RTX3070Ti graphics card.

During training, the input images were uniformly resized to a size of 448×448, corresponding to real label image sizes of 224×224. The Adam optimizer was used with an initial learning rate and weight decay both set to 0.0001. The batch size was set to 16, and the number of epochs was set to 300.

III. RESULTS AND DISCUSSION

A. Evaluation Metrics

To evaluate the segmentation results of the improved model, the following evaluation metrics were employed in this study: The Dice coefficient and Intersection over Union (IOU) were used to assess the performance of the algorithm in optic disc and optic cup segmentation.

1) DICE COEFFICIENT

The Dice coefficient is calculated by taking twice the intersection area of the predicted region and the ground truth region, divided by their sum. It measures the similarity between two images and is one of the most commonly used evaluation metrics for image similarity in medical segmentation.

The Dice coefficient ranges from 0 to 1, with a higher value indicating better segmentation performance. A Dice coefficient of 1 means perfect overlap, indicating that the two sets are identical, while a value of 0 means no overlap, indicating that the two sets have no common elements.

Moreover, the Dice coefficient is insensitive to imbalanced data, making it widely applicable in fields such as medical image segmentation.

The formula for calculating the Dice coefficient is as follows:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}$$

In which, A represents the segmentation algorithm's generated result, and B stands for the dataset label. A∩B denotes the intersection of the two sets, while |A| and |B| represent the number of elements in each set, respectively.

2) IOU

IOU (Intersection over Union), is another commonly used metric for evaluating the performance of image segmentation models.

The formula for calculating IOU is as follows:

$$\text{IOU} = \frac{|A \cap B|}{|A \cup B|}$$

Where A represents the segmentation algorithm's generated result, and B denotes the dataset label. A∩B signifies the intersection of the two sets, while A∪B represents their union. The IOU value ranges from 0 to 1, with values closer to 1 indicating a higher degree of overlap between the segmentation result and the ground truth, thus indicating better segmentation performance. An IOU of 1 indicates a perfect overlap between the predicted and ground truth results; conversely, an IOU of 0 indicates no overlap between the predicted and ground truth results.

B. Experiment Results

1) ABLATION EXPERIMENT

The improved model proposed in this paper utilizes two modules: the Focus module and the CBAM module. Therefore, four sets of experiments were designed: TransUnet, TransUnet+Focus, TransUnet+CBAM, and TransUnet+Focus+CBAM. Each set was trained on the ORIGA-650 dataset for 300 epochs, and the final results were compared.

TABLE 1: Ablation experiment

Model	Focus	CBAM	Dice (disc)	Dice (cup)	IOU (disc)	IOU (cup)
TransUnet			0.952	0.887	0.908	0.797
TransUnet+Focus	√		0.979	0.928	0.959	0.865
TransUnet+CBAM		√	0.961	0.893	0.924	0.807
Proposed	√	√	0.984	0.947	0.968	0.902

Based on the results of the ablation experiments shown in Table 1, it can be observed that introducing the model with the Focus enhancement leads to a significant improvement in both Dice coefficient and IOU. Similarly, the model with the CBAM enhancement also shows improved performance. However, when combining both the Focus and CBAM enhancements, the model achieves the highest performance across all metrics. Specifically, compared to the original model, the model with the combined Focus and CBAM enhancements demonstrates a 3.4% increase in the Dice coefficient for the optic disc, along with a 6.6% increase in IOU. Similarly, for the optic cup, there is an 6.8% increase in the Dice coefficient and a 13.2% increase in IOU. From the above findings, it can be concluded that the performance of the Focus module surpasses that of the CBAM module. Additionally, the improved TransUnet model used in this study exhibits better performance compared to the original model, enabling more accurate segmentation of the optic disc and optic cup.

2) COMPARISON EXPERIMENT

To evaluate the effectiveness of the segmentation algorithm, this paper compares the proposed optic disc segmentation model with segmentation models such as UNet, Res-UNet, FCN, and TransUnet. The results of glaucoma optic disc segmentation by different segmentation models are shown in Table 2.

TABLE 2: Ablation experiment

Model	Dice（disc）	Dice（cup）	IOU（disc）	IOU（cup）
UNet	0.937	0.809	0.881	0.679
Res-UNet	0.943	0.858	0.892	0.751
FCN	0.892	0.795	0.805	0.659
Proposed	0.984	0.947	0.968	0.902

According to the content of Table 2, it is clear that the improved TransUnet model used in this study performs excellently in terms of performance. It possesses higher Dice coefficients and IOU indicators for both optic cup and optic disc, demonstrating a more accurate performance in optic cup and optic disc segmentation tasks.

C. Segmentation Results Display

The segmentation results of optic cup and optic disc using different models on ORIGA-650 are shown in Fig. 5. In the figure, the white areas represent the optic cup, while the gray areas represent the optic disc. The first column shows the original images, the second column shows the segmentation results of FCN, the third column shows the segmentation results of U-Net, the fourth column shows the segmentation results of Res-UNet, the fifth column shows the results of the proposed method, and the last column shows the Ground Truth.

It can be seen that the proposed method outperforms other methods in segmentation accuracy. From the second row, it is evident that for images with low contrast, the proposed method shows a significant improvement in optic cup segmentation.

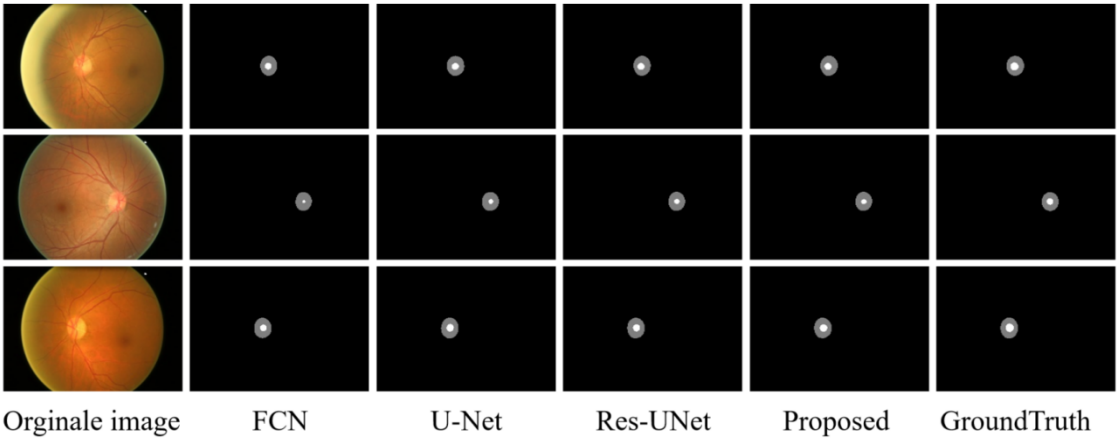

Figure 5: Segmentation results of different methods

IV. CONCLUSION

This paper proposes an improved method for optic disc (OD) and optic cup (OC) segmentation using the Focus module and CBAM module integrated into TransUnet. TransUnet effectively integrates local and global contextual information, combining the advantages of UNet and Transformer. Additionally, the introduction of the Focus module enhances the model by expanding the channel dimension through slicing the medical images, thus retaining more information at the same training image size (equivalent to downsampling without information loss). Furthermore, the adoption of the CBAM module provides the model with better attention mechanisms, effectively enhancing the accuracy of optic cup and optic disc segmentation results.

The experimental results demonstrate that the proposed method achieves excellent performance on the ORIGA-650 dataset, confirming its effectiveness in retinal image segmentation. It provides reliable results for calculating the cup-to-disc ratio, assisting ophthalmologists in glaucoma screening. Future work should focus on cross-dataset validation to assess the model's generalization and robustness. Additionally, further investigations are needed to provide diagnostic suggestions based on the segmentation results.

FUNDING

This work was supported by the fund of Beijing Municipal Education Commission (Project No. 22019821001), Zhiyuan Science Foundation from Beijing Institute of Petrochemical Technology (Project No. 2023015), and Climbing Program Foundation from Beijing Institute of Petrochemical Technology (Project No. BIPTAA1-2021-004).

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

REFERENCES

[1] Tham Y C, Li X, Wong T Y, et al. “Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis”. Ophthalmology, 2014, 121(11):2081-2090.DOI:10.1016/j.ophtha.2014.05.013.

[2] Bian W L X. “Optic disc and optic cup segmentation based on anatomy guided cascade networkComputer Methods and Programs in Biomedicine: An International Journal Devoted to the Development,” Implementation and Exchange of Computing Methodology and Software Systems in Biomedical Research and Medical Practice, 2020, 197(1).

[3] Peng L, Lin L, Cheng P, et al. “FARGO: A Joint Framework for FAZ andRV Segmentation from OCTA Images” in International Workshop on Ophthalmic Medical Image Analysis. Springer, Cham, 2021. DOI:10.1007/978-3-030-87000-3_5.

[4] Maninis, K.-K., Pont-Tuset, J., Arbeláez, P., & Van Gool, L. “Deep Retinal Image Understanding,” Lecture Notes in Computer Science, 2016, 140–148. doi: 10.1007/978-3-319-46723-8_17

[5] Edupuganti V G, Chawla A, Kale A. “Automatic Optic Disk and Cup Segmentation of Fundus Images Using Deep Learning,” IEEE, 2018.DOI:10.1109/ICIP.2018.8451753.

[6] Sevastopolsky, A. “Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network,” Pattern Recognition and Image Analysis, 2017, 27(3), 618–624. doi:10.1134/s1054661817030269

[7] Ronneberger, O., Fischer, P., & Brox, T. “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 2015, 234–241. doi:10.1007/978-3-319-24574-4_28

[8] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu and X. Cao “Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation,” in IEEE Transactions on Medical Imaging, vol. 37, no. 7, pp. 1597-1605, July 2018, doi: 10.1109/TMI.2018.2791488.

[9] Shankaranarayana S M, Ram K, Mitra K,et al. “Joint Optic Disc and Cup Segmentation Using Fully Convolutional and Adversarial Networks,” in International Workshop on Fetal and Infant Image Analysis International Workshop on Ophthalmic Medical Image Analysis.2017.DOI:10.1007/978-3-319-67561-9_19.

[10] Y. Qin and A. Hawbani, “A Novel Segmentation Method for Optic Disc and Optic Cup Based on Deformable U-net,” in 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2019, pp. 394-399, doi: 10.1109/ICAIBD.2019.8837025.

[11] J. Dai et al., "Deformable Convolutional Networks," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764-773, doi: 10.1109/ICCV.2017.89.

[12] Shelhamer, E., Long, J., & Darrell, T. “Fully Convolutional Networks for Semantic Segmentation”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4), 640–651. doi:10.1109/tpami.2016.2572683.

[13] iek, zgün, Abdulkadir A, Lienkamp S S, et al. “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation.” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2016. DOI:10.1007/978-3-319-46723-8_49.

[14] Huang H, Lin L, Tong R, et al. “UNet 3+: A Full-Scale Connected UNet for Medical Image SegmentationarXiv, 2020.DOI:10.1109/ICASSP40776.2020.9053405.

[15] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. “UNet++: A Nested U-Net Architecture for Medical Image Segmentation,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. 2018. DOI:10.1007/978-3-030-00889-5_1.

[16] Dosovitskiy, A., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”, arXiv e-prints, 2020. doi:10.48550/arXiv.2010.11929.

[17] Chen, J., “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation”, arXiv e-prints, 2021. doi:10.48550/arXiv.2102.04306.

[18] Woo S, Park J, Lee J Y, et al. “CBAM: Convolutional Block Attention Module,” in European Conference on Computer Vision 2018, Springer, Cham, 2018. DOI:10.1007/978-3-030-01234-2_1.

[19] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi and J. Zhong, “Attention Is All You Need In Speech Separation,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 21-25, doi: 10.1109/ICASSP39728.2021.9413901.

[20] Deng J, Dong W, Socher R, et al. “ImageNet: a Large-Scale Hierarchical Image Database,” in 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA.IEEE, 2009.DOI:10.1109/CVPR.2009.5206848.