Glaucoma Detection Based on Joint Optic Disc and Cup Segmentation Using Dense Prediction Transformer

Dindin Inas Candra Wiguna[1], Ema Rachmawati[2], Gamma Kosala[3]
[1,2,3]*School of Computing, Telkom University, Bandung, Indonesia*
[1]dindininas@student.telkomuniversity.ac.id, [2]emarachmawati@telkomuniversity.ac.id, [3]gammakosala@telkomuniversity.ac.id

***Abstract*—Glaucoma is an eye disease resulting from damage to the optic nerve. Glaucoma can cause vision loss if not treated quickly and appropriately. In this case, the observation of an ophthalmologist is needed to check it. However, an expert doctor's observation is subjective, so it takes a long time and is inconsistent. So a computer-aided diagnostic (CAD) system was built to detect glaucoma early on, fundus image analysis automation, time efficiency through optical disc and cup segmentation, and Cup to Disc Ratio (CDR) calculations. Several previous studies have proposed using models based on Vision Transformers (ViT), Convolutional Neural Networks (CNN), and a combination of the two. However, the ViT model has problems: the number of model computations increases when the image size also increases, and the CNN-based encoder-decoder model has a large size and is slow in its calculations. Therefore, a segmentation method that utilizes transformers as encoders and convolution as decoders was chosen, namely the Dense Prediction Transformer (DPT), which can process data in parallel. This research discusses the implementation of the DPT with a case study of the optical disc and cup segmentation on fundus images using the ORIGA dataset. The result shows that DPT is outperformed slightly by the Segmenter model with mIoU at 4.9%.**

Keywords—glaucoma, segmentation, disc, cup, Dense Prediction Transformer.

I. INTRODUCTION

The eye is one of the most essential organs in the human body. In addition to acting as an organ of the body, the eye also acts as a sense of sight. As one of the essential parts of the human organ, the eye is inseparable from various disease attacks, both from within and from outside the eye. Attacks from outside the eye are usually irritation caused by entering small objects or dust into the eye. Apart from irritation, there are also diseases caused from within the human body, such as cataracts, myopia, color blindness, short-sightedness, glaucoma, and many more. Glaucoma is a disease caused by damage to the optic nerve, which can lead to blindness [1]. As many as 76 million people worldwide suffer from glaucoma in 2020, and are expected to increase to 111.8 million people in 2040 [2]. In Indonesia, in 2015, as many as 65,774 people suffered from glaucoma; in 2017, it increased to 427,091 people [2]. Glaucoma, if it is late in treatment, can cause total blindness. Early detection of glaucoma is essential. Retinal image segmentation includes Optic Disc, Optic Cup, or Blood Vessel segmentation [2].

Several studies have been carried out regarding glaucoma before [3] was conducted by Artem Sevastopolosky using U-Net to segment disc and cup parts separately. They used the DRISON-DB, RIM-ONE-v.3, and DRASHTI-GS datasets. Their research yielded the best IoU of 0.89, Dice 0.94, and Prediction time of 0.1s. A similar study [4] applied A Large-scale Database and CNN to detect glaucoma. This research resulted in an accuracy of 85.2%, an AUC of 0.916, and an F2-score of 0.837 in the RIM-ONE dataset. Other studies [5] apply the implementation of a transformer model for segmentation, namely the Segmentation Transformer (SETR). They used the ADE20K and Pascal Context dataset, with the best Pixel Acc and mIoU results of 83.46 and 50.28, and the best mIoU results for the Pascal Context dataset were 55.83.

Vision Transformers (ViT) is part of Deep Learning. ViT is a model for image classification using attention and eliminating repetition and convolution operations over patching an image [6]. The ViT patching method involves breaking the image into many pieces, arranging them linearly, adding the embedding position, and then vectorizing the vector sequence for the transformer encoder [7]. ViT has the advantage that memory usage is manageable for the learning process because there is no repetition, and the parameters needed in this architecture are only a few [8]. Another advantage of ViT is in capturing image information globally from an image [9].

Dense Prediction Transformer (DPT) is a modification of the convolutional network. The DPT architecture utilizes a transformer as an encoder and a Convolutional as a decoder. In particular, DPT uses ViT as the backbone [10]. Ranftl et al. conducted research using DPT with the ADE20K dataset [10]. The study results say that the segmentation prediction for DPT is smoother and globally coherent than the full-convolutional method. If DPT is trained using a significant amount of data, its enormous potential can be realized.

In this study, we propose the application of DPT in glaucoma segmentation for retinal fundus images. The proposed system has the advantage of capturing image information globally and minor memory usage for learning since there is no looping. Furthermore, the model is compared with Segmenter to analyze the model.

This article is divided into five sections. The first section explains the importance of this research and describes the method in general related to glaucoma cases. The second section describes various previous methods and experiments in detail. The third section explains the proposed process thoroughly. The fourth section shows the experimental results and analysis. The fifth section is the conclusion.

II. RELATED WORKS

Artem Sevastopolosky conducted related research on the detection of glaucoma [3]. The study uses U-Net to segment the disc and cup sections separately. Uses DRISON-DB, RIM-ONE-v.3, and Drishti-GS datasets. This research yielded the best IoU of 0.89, Dice 0.94, and Prediction time of 0.1s.

Similar research was conducted by Fu et al. [11], which was conducted using the M-Net method to segment fundus images. The study used the ORIGA and SCES datasets. With his research results, the highest AUC score was 0.8508 with the M-Net + PT method, and the pure M-Net method only got

an AUC score of 0.8019 for the ORIGA dataset. Whereas for the SCES dataset, the M-Net + PT method produces an AUC score of 0.8998, and for pure M-Net, it has an AUC score of 0.8396.

Another study [4] conducted by Li et al. The study applied A Large-scale Database and CNN to detect glaucoma. This research resulted in an accuracy of 85.2%, an AUC of 0.916, and an F2-score of 0.837 in the RIM-ONE dataset.

One of the studies conducted by Sreng et al. [12] used DeepLabV3 to segment the disc and cup portions of fundus images. They used the RIM-One, ORIGA, DRISHTI-GS, and ACRIMA datasets, with the results of the accuracy of each dataset being 97.37%, 90.00%, 86.84%, and 99.53%, as well as the Area Under Curve results (AUC) respectively were 100%, 92.06%, 91.67%, and 99.98%. In the same year, Saxena et al. conducted research related to glaucoma detection using the Convolutional Neural Network (CNN) [13]. They used the ORIGA and SCES datasets and the AUC score as CNN performance measurement tools. And the result of the AUC score for the ORIGA dataset using the CNN method is 0.822, which is 0.60 smaller than the SCES dataset, which is 0.882.

The other research was conducted by Li et al. [14]. The research succeeded in developing and implementing a transformer model for case studies of medical image segmentation, and the case solved was segmenting the REFUGE dataset disc and cup. The result of the developed transformer model is called the SEGTRAN model. In the same year, Zheng [5] developed research conducting trials of implementing a transformer model for segmentation, namely the Segmentation Transformer (SETR), by trying a form of SETR decoder. They used the ADE20K and Pascal Context datasets, with the best mIoU and Pixel Acc results of 50.28 and 83.46 for the ADE20K dataset, and the best mIoU for the Pascal Context dataset was 55.83. The research was conducted by Gupta et al. [15] to detect glaucoma using the CLAHE and EfficientNet methods. This research uses the DRASHTI-GS dataset, and the results get the best mIoU of 0.910 and Dice Similarity of 0.960. convolutional multi-layer neural network classification.

III. Our Proposed System

The system is constructed using a dataset of fundus images and corresponding annotations. 80% of the dataset is utilized for training the model, with 60% allocated for training and 20% for validation. The remaining 20% of the dataset is reserved for the testing phase. Before training, both the images and annotations undergo a preprocessing step. The training images are then used to train the DPT, while the testing images are employed to assess the model's performance. The illustration is shown in Fig. 1.

A. Dense Prediction Transformer

Dense Prediction Transformer (DPT) follows the proven encoder-decoder structure that has achieved significant results in dense prediction tasks. DPT utilizes vision transformers as the core component, demonstrating how the encoder's output representation can be efficiently converted into dense predictions. Additionally, insights into the effectiveness of this approach can be seen in. Fig. 2 (left) provides an overview of the complete architecture.

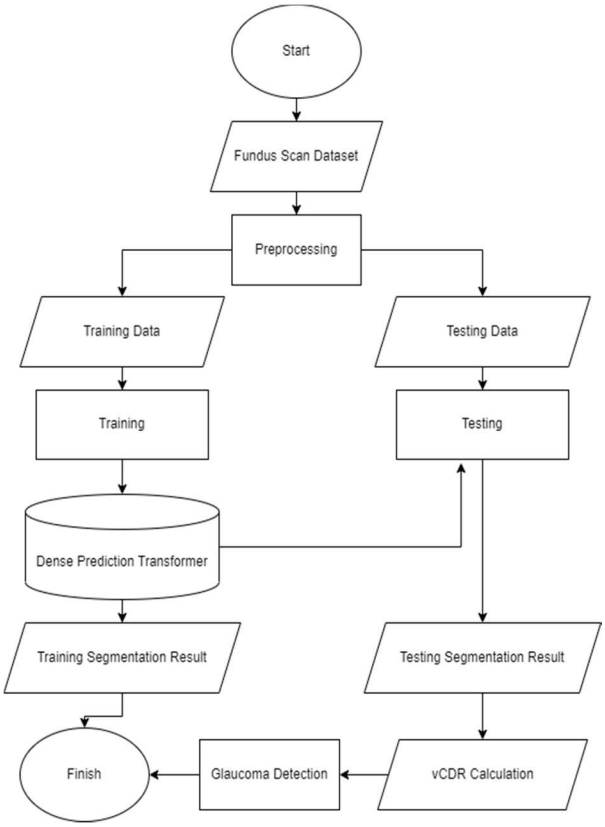


Fig. 1. Proposed system overview: Glaucoma detection.

Transformer encoder. The transformer encoder, specifically ViT, processes an image's bag-of-words representation [16]. Each image patch or deep feature is embedded as a single "word" into a feature space. We'll use the term "tokens" to refer to these embedded "words" throughout this discussion. Transformers use successive blocks of multi-headed self-attention (MHSA) to change the representation. The interaction between the tokens is made possible by MHSA, changing the representation.

During calculations, the transformer consistently preserves the token count. This ensures that the initial embedding's spatial resolution is preserved during the transformer process because each token corresponds to an image patch. Additionally, multi-headed self-attention (MHSA) operations work naturally globally, enabling each token to be aware of and affect one another. As a result, at every point after initial implantation, the transformer possesses a global receptive field. Contrast this with convolutional networks, which continuously widen their receptive fields as features move through progressively more convolutional and downsampling layers.

Specifically, ViT employs a process to extract a patch embedding from the image. This entails processing non-overlapping square patches of size pixels taken from the image. Linear projection is used to flatten these patches into vectors and insert them individually. Since transformers operate as set-to-set functions, they do not inherently retain spatial position information for each token. Learnable position embedding, which incorporates spatial information into an overall representation, is paired with image embedding to alleviate this restriction. Inspired by prior work in natural

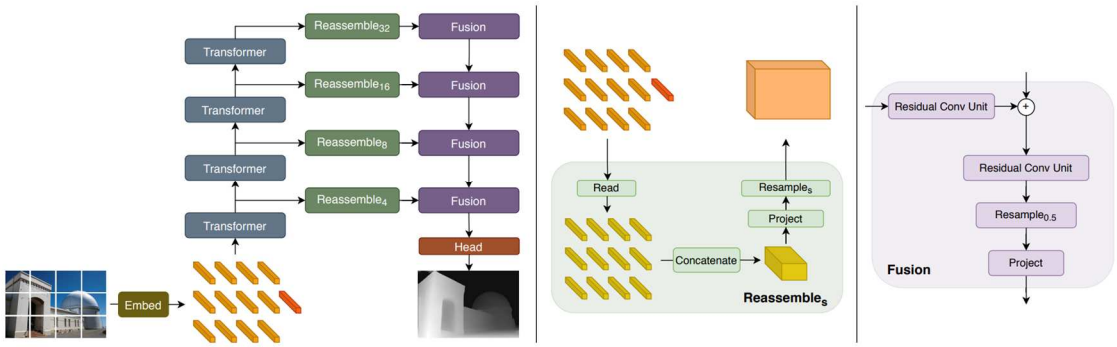


Fig. 2. Left: Architecture overview. Center: Reassemble. Right: Fusion [10].

language processing (NLP), ViT introduces a special learned token not directly associated with specific image content. This token is known as the readout token. The readout token aggregates data into a global visual representation, which is then used for classification tasks. The result of applying the embedding procedure to an image of size $H \times W$ pixels is a set of $t^0 = \{t_0^0, \dots, t_{N_p}^0\}$, $t_n^0 \in \mathbb{R}^D$ tokens, where $N_p = \frac{HW}{p^2}$, t_0 refers to the readout token, and D is the feature dimension of each token.

The input tokens are changed into new representations t^l using L transformer layers, where l is the output of the l-th transformer layers. Vaswani et al. [6] defined several variations of this fundamental design and chose the Vit-Base variation, which has 12 layers of transformers and a patch-based embedding process. Projecting the flattened patches to dimensions of D = 768 and D = 1024, respectively, is a step in the embedding process for ViT-Base. Notably, both feature dimensions exceed the number of pixels in the input patch. As a result, the embedding procedure can learn and store information if it proves advantageous for the task at hand. As a result, the features of the input patch can be resolved with pixel-level accuracy.

Convolutional decoder. The decoder combines the set of tokens to create image-like feature representations at different resolutions. These feature representations are then incrementally merged to form the ultimate dense prediction. To accomplish this, introduce a straightforward three-stage *Reassemble* operation. This operation enables the retrieval of image-like representations from the output tokens of any layer in the transformer encoder.

$$Reassemble_s^{\widehat{D}}(\mathrm{t}) = (Resample_s \circ Concate \circ Read)(t),$$

$\widehat{D}$ is the output feature dimension and s is the recovered representation's output size ratio relative to the input picture.

First, map the $N_p + 1$ tokens to a set of N_p tokens amenable to spatial concatenation into an image-like representation, as shown in (1).

$$\text{Read: } \mathbb{R}^{N_p+1\times D} \rightarrow \mathbb{R}^{N_p\times D}. \tag{1}$$

This operation plays a crucial role in handling the readout token effectively. Although the readout token may not have a specific role in the dense prediction task, it can still be valuable for capturing and distributing global information. Therefore, three variants of this mapping model were studied to explore its potential usefulness, as shown in (2), (3), and (4).

$$\text{Read}_{ignore}\,(t) = \{t_1, \dots, t_{N_p}\} \tag{2}$$

ignores the readout token completely,

$$\text{Read}_{add}\,(t) = \{t_1 + t_0, \dots, t_{N_p} + t_0\} \tag{3}$$

adding the representations transfers the data from the readout token to the other tokens, and

$$\text{Read}_{proj}\,(t) = \{mlp(cat(t_1, t_0)), \dots, mlp(cat\left(t_{N_p}, t_0\right))\} \tag{4}$$

The method concatenates the readout token with all other tokens to propagate information from the readout token to the other tokens. Then, using a linear layer followed by a GELU non-linearity, the concatenated representation is projected back to the initial feature dimension D.

The Np tokens acquired after a Read block can be rearranged into an image-like representation by placing each token according to the location of the initial patch in the image. A spatial concatenation operation is applied, resulting in a feature map with a size of size $\frac{H}{p} \times \frac{W}{p}$ with D channels, as shown in (5).

$$\text{Concatenate: } \mathbb{R}^{N_p\times D} \rightarrow \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times D}. \tag{5}$$

finally, it sends this representation to a spatial resampling layer, which enlarges the representation to size $\frac{H}{s} \times \frac{W}{s}$ with D characteristics per pixel, as shown in (6).

$$\text{Resample}_s\text{: } \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times D} \rightarrow \mathbb{R}^{\frac{H}{s}\times\frac{W}{s}\times D}. \tag{6}$$

this process is carried out by first projecting the input representation to $\widehat{D}$ using 1×1 convolutions, then performing either a (stride) 3×3 convolution when $s \geq p$ or a strided 3×3 transpose convolution when $s < p$ to implement spatial downsampling and upsampling procedures, respectively.

To complete this process, the feature maps extracted from the successive stages are combined using a feature fusion block based on RefineNet [17] (refer to Fig. 2, right). They progressively increased the representation resolution in each

fusion stage by a factor of two. The representation resolution was eventually reduced to half of the input image. An output head with task-specific functions is added to provide the final forecast. A schematic representation of the overall architecture is shown in Fig. 1.

Handling varying image sizes. The DPT can accommodate images of various sizes like fully convolutional networks can. The embedding process can produce a variable number of picture tokens, indicated as Np, if the image size is divisible by p (the size of the patches). The transformer encoder is a set-to-set design and can easily handle different token counts. To encrypt patch sites, the embedding position, however, relies on the size of the image. To ensure that the position embedding is the right size, linear interpolation is performed. It is important to note that the interpolation of each image can be done dynamically. If the input image lines up with the convolutional decoder's stride (32 pixels), both the reassemble and fusion modules can easily change token counts after the embedding method and transformer phases.

B. *Evaluation Metric*

The performance of this system for fundus scan image segmentation is evaluated using intersection over union (IoU). IoU compares the number of pixels in the ground truth predicted by the prediction mask with a whole pixel of ground truth and prediction masks [18], as shown in (7).

$$IoU = \frac{True\ Positive}{True\ Positive+False\ Positive+False\ Negative}. \quad (7)$$

A model's overall performance can be measured by averaging the IoU of each class (mIoU).

C. *Vertical Cup to Disc Ratio*

Vertical cup-to-disc ratio (vCDR) is a measurement that represents the ratio of the optical cup's vertical diameter to the optical disc's vertical diameter. CDR is a well-known feature used in the assessment of glaucoma. Three methods are frequently used to diagnose glaucoma: measurement of intraocular pressure (IOP), testing of the visual field, and inspection of the optic nerve head or disc (ONH). Professionals more frequently take the ONH test. Ophthalmologists often administer the cup-to-disc ratio (CDR) test to patients to determine ONH [19]. In this study, the vCDR is employed to determine the presence of glaucoma by calculating the vCDR value by analyzing the image's mask or annotation. The formulation for vCDR can be expressed as shown in (8).

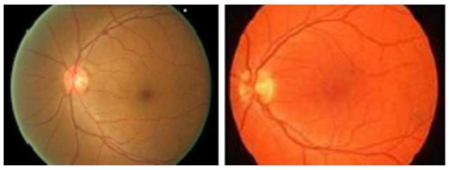
$$\text{vCDR} = \sqrt{\frac{Vertical\ area\ of\ optic\ cup}{Vertical\ area\ of\ optic\ disc}} \quad (8)$$

IV. Experimental Results and Analysis

This section explains the dataset, model training phase, experiment result, and analysis.

A. *Datasets*

The ORIGA dataset [20] was collected through the Singapore Malay Eye Study (SiMES). The ORIGA dataset contains 650 retinal images annotated by trained professionals from Singapore and is an open dataset with a resolution of 2518 x 2048. an example of the ORIGA dataset can be seen in Fig. 3.



a b

Fig. 3. a) Eye's healthy fundus; b) Glaucoma in the eye's fundus [20].

All raw images have been converted into .PNG format and resized to 384 x 384 pixels. Images have also been annotated into two classes (disc and cup). Samples of the raw image are shown in Fig. 4, and the annotation is shown in Fig. 7.

B. *Result and Analysis*

We divide the dataset into 520 images for training, 65 for validation, and 65 for testing. We used the tool MMSegmentation [21] for the model's training, testing, and inference. We set the parameters for the configuration: batch size to 3, images resized to 384 x 384 pixels randomly cropped to 512 x 512 pixels, and iterative training to 5000. But some experiments change the scale to 512 x 512. Every 500 iterations, evaluation is done with training to track the procedure. This evaluation reports the model's performance in the IoU of each class and the mean IoU of all classes.

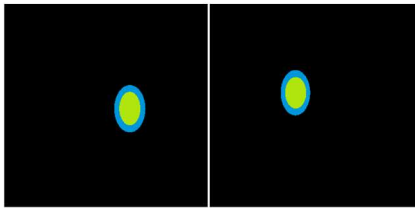


Fig. 4. Samples of Fundus annotation images.

After training, the model is tested using 65 images and evaluated. The evaluation also reports the IoU of each class and the mean IoU. This process is carried out on each model to be compared and analyzed. In this experiment, DPT is divided into four categories. What distinguishes one category is the difference in scale size, learning rate, and weight in each DPT category. More detailed information about the difference in scale size, learning rate, and weight in each category, as well as the mIoU results for val and test, can be seen in Table I. For the ORIGA dataset, as shown in Table I, model DPT performs at mIoU in the range of 70% to 77%; increasing the scale's size significantly affects DPT performance. Reducing the learning rate and weight does not perform in this case, as shown in Fig. 5.

TABLE I. Dense Prediction Transformer on Origa Dataset

Method	Scale	Learning Rate	Weight	mIoU	
				Val	Tes
DPT ViT-B	383x384	0.01	0.0005	70.31%	73.12%
DPT ViT-B	512x512			75.61%	74.61%
DPT ViT-B	383x384	0.00001	0.0001	70.26%	73.10 %
DPT ViT-B	512x512			76.74%	75.94%

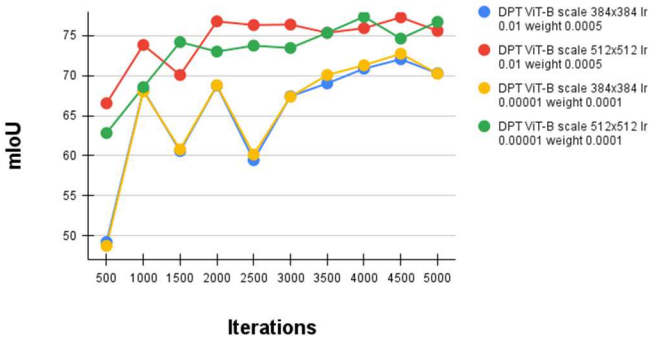


Fig. 5. mIoU graph of Dense Prediction Transformer in ORIGA dataset training phase.

C. *Comparison of Dense Prediction Transformer and Segmenter*

For comparison, the Dense Prediction Transformer and Segmenter models both use the Vision Transformer backbone; the only difference is that Dense Prediction uses multi-head self-attentions as their basic operation, and the pure Segmenter only uses self-attentions.

For ORIGA datasets, DPT performs slightly better than Segmenter, as shown in Table II. The Segmenter model has the best performance with mIoU of 71.84%, and the DPT model still has the best performance with mIoU of 76.74% with a margin of around 4.9%. IoU results from every 500 iterations of all models can be seen in Fig. 5.

TABLE II. ORIGA DATASET MODEL COMPARISON

Method	IoU	
	val	test
DPT ViT-B scale 384x384 lr 0.01 weight 0.0005	70.31%	73.12%
DPT ViT-B scale 512x512 lr 0.01 weight 0.0005	75.61%	74.61%
DPT ViT-B scale 384x384 lr 0.00001 weight 0.0001	70.26%	73.1 %
DPT ViT-B scale 512x512 lr 0.00001 weight 0.0001	76.74%	75.94%
Seg-T scale 384x384 lr 0.01 weight 0.0005	71.7%	71.22%
Seg-T scale 512x512 lr 0.01 weight 0.0005	71.84%	71.9%

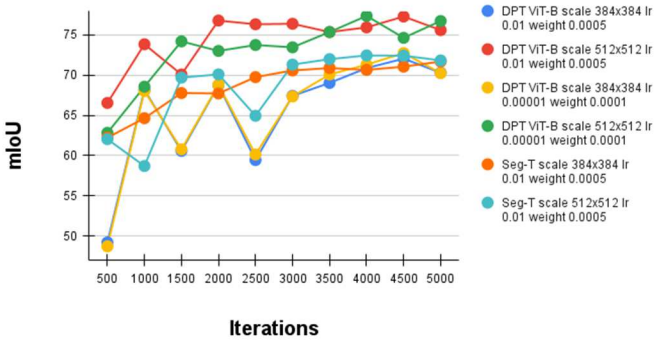


Fig. 6. mIoU graph of Dense Prediction Transformer and Segmenter.

TABLE III. STRONGEST PER-VARIANT PERFORMANCE

Class	IoU	
	DPT ViT-B	Seg-T
Disc	**77.11%**	71.85%
Cup	**78.38%**	74.97%

To compare the best variant of each model, a visual comparison of raw image results and segmentation results can be seen in Fig. 6. To compare the best variant of each model, a visual comparison of raw image and segmentation results can be seen in Table III. The DPT model outperformed the Segmenter model in 2 of 2 classes. In disc and cup classes, the DPT method works better.

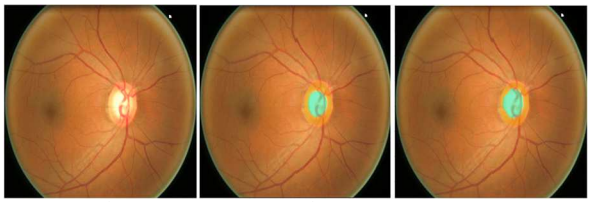
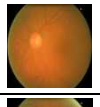
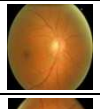


Fig. 7. Visual comparison between the raw image and model segmentation in the ORIGA dataset.

D. *Glaucoma Detection*

From Dense Prediction Transformer (DPT) result, we calculate the vCDR value from the cup and disc mask. We tried several threshold variations starting from 60, 63, 64, and 70 to assess whether the fundus image contains glaucoma. If the vCDR calculation result is equal to or more than the threshold, it is labeled as having glaucoma, and if it is lower than the threshold, then it is labeled as not glaucoma—detection results, as shown in Table IV. For the threshold, we set 63 because it gets 76% accuracy for detecting glaucoma labels.

TABLE IV. GLAUKOMA DETECTION RESULT

Image Name	Label Ground Truth	Label Prediction	IoU Score Disc	IoU Score Cup
	Glaucoma	Non-Glaucoma	74.84%	68.80%
	Glaucoma	Glaucoma	74.76%	70.41%
	Glaucoma	Glaucoma	75.53%	70,00%
	Non-Glaucoma	Non-Glaucoma	76.73%	75.20%
	Glaucoma	Glaucoma	76.52%	73.33%

V. Conclusion

Fundus image segmentation plays a significant role in the early detection of glaucoma. The capabilities of the Dense Prediction Transformer (DPT) method, which uses a Vision Transformer instead of a convolutional network with multi-head self-attention as the basic operation, can be demonstrated using the available dataset. We also compare it with Segmenter using the Vision Transformer as its backbone. In this study, our proposed system outperformed other models with the same batch size and iterations during the training phase. Similar or alternative methodologies can be explored in future research to enhance the model's performance in this scenario.

References

[1] R. Fan *et al.*, "Detecting Glaucoma from Fundus Photographs Using Deep Detecting Glaucoma from Fundus Photographs Using Deep Learning without Convolutions: Transformer for Improved Learning without Convolutions: Transformer for Improved Generalization Generalization," 2022, doi: 10.36227/techrxiv.19727314.v1.

[2] A. Mvoulana, R. Kachouri, and M. Akil, "Fully automated method for glaucoma screening using robust optic nerve head detection and unsupervised segmentation based cup-to-disc ratio computation in retinal fundus images," Computerized Medical Imaging and Graphics, vol. 77, 2019, doi: 10.1016/j.compmedimag.2019.101643.

[3] 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, Jul. 2017, doi: 10.1134/S1054661817030269.

[4] Li, L., Xu, M., Wang, X., Jiang, L., & Liu, H. (2019). Attention Based Glaucoma Detection: A Large-Scale Database and CNN Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10571-10580).

[5] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021, June). Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6877-6886). IEEE.

[6] A. Vaswani et al., "An image is worth 16*16 words: transformers for image recognition at scale," in *Advances in Neural Information Processing Systems*, 2017.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, *30*.

[8] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021, July). Perceiver: General Perception With Iterative Attention. In *International Conference on Machine Learning* (pp. 4651-4664). PMLR.

[9] Wassel, M., Hamdi, A. M., Adly, N., & Torki, M. (2022, August). Vision Transformers Based Classification for Glaucomatous Eye Condition. In *2022 26th International Conference on Pattern Recognition (ICPR)* (pp. 5082-5088). IEEE.

[10] Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12179-12188).

[11] 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," *IEEE Trans Med Imaging*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018, doi: 10.1109/TMI.2018.2791488.

[12] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win, "Deep Learning for Optic Disc Segmentation and Glaucoma Diagnosis on Retinal Images," *Applied Sciences (Switzerland)*, vol. 10, no. 14, Jul. 2020, doi: 10.3390/app10144916.

[13] Saxena, A., Vyas, A., Parashar, L., & Singh, U. (2020, July). A Glaucoma Detection Using Convolutional Neural Network. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 815-820). IEEE.

[14] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical Image Segmentation Using Squeeze-and-Expansion Transformers," in *IJCAI International Joint Conference on Artificial Intelligence, 2021*. doi: 10.24963/ijcai.2021/112.

[15] N. Gupta, H. Garg, and R. Agarwal, "A Robust Framework for Glaucoma Detection Using CLAHE and EfficientNet," *Visual Computer*, vol. 38, no. 7, pp. 2315–2328, Jul. 2022, doi: 10.1007/s00371-021-02114-5.

[16] Sivic, J., & Zisserman, A. (2008). Efficient Visual Search of Videos Cast as Text Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 591-606.

[17] Lin, G., Milan, A., Shen, C., & Reid, I. (2017). Refinenet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1925-1934).

[18] Zhao, W., Zhang, H., Yan, Y., Fu, Y., & Wang, H. (2018). A Semantic Segmentation Algorithm Using FCN with Combination of BSLIC. *Applied Sciences*, *8*(4), 500.

[19] R. Ali *et al.*, "Optic Disk and Cup Segmentation through Fuzzy Broad Learning System for Glaucoma Screening," *IEEE Trans Industr Inform*, vol. 17, no. 4, pp. 2476–2487, Apr. 2021, doi: 10.1109/TII.2020.3000204.

[20] Zhang, Z., Yin, F. S., Liu, J., Wong, W. K., Tan, N. M., Lee, B. H., ... & Wong, T. Y. (2010, August). Origa-light: An Online Retinal Fundus Image Database for Glaucoma Analysis and Research. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology* (pp. 3065-3068). IEEE.

[21] MMsegmentation Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," 2020.