

# Investigating Self-Attention in Swin-Unet Model for Disc and Cup Segmentation

Jehua Kusuma Dewa<sup>1</sup>, Ema Rachmawati<sup>2</sup>, Gamma Kosala<sup>3</sup>

<sup>1,2,3</sup> School of Computing, Telkom University, Bandung, Indonesia

<sup>1</sup>jehuakusumadewa@student.telkomuniversity.ac.id, <sup>2</sup>emarachmawati@telkomuniversity.ac.id, <sup>3</sup>gammakosala@telkomuniversity.ac.id

**Abstract**—Glaucoma is a condition of the eyes that results from damage to the optic nerve, which could potentially lead to loss of vision if not addressed promptly. Because glaucoma rarely shows early symptoms in the sufferer, it requires regular observation of the retinal fundus by an ophthalmologist to find out if this eye disorder appears. Doctors' observations are subjective, so they are inconsistent and take a long time. As a result, a computer-aided diagnostic (CAD) system was created that can automate the examination of retinal fundus images to detect glaucoma in its initial phase. In a consistent and time-saving manner by using optic disc and cup segmentation and cup-to-disc ratio (CDR) computation. CAD systems can also be used as decision support by doctors. Several previous image segmentation studies have proposed using convolution neural network (CNN) and Vision Transformer (ViT) based models and their combinations. However, the encoder-decoder model based on CNN is large and is slow in computation. The ViT model has the problem that the computational amount of the model increases when the image size also increases. Therefore, the segmentation method uses a Swin Transformer-based encoder-decoder model, Swin-Unet, which has the advantage of a self-attention mechanism performed on local windows and has linear computation. This paper presents a case study of optical disc and cup segmentation using the Swin-Unet method with the REFUGE dataset. The vCDR calculation with a threshold of 0.63 yields an accuracy of 94%. The IoU score results using the REFUGE dataset resulted in a score of 84% for the disc part and 80% for the cup part.

**Keywords**—glaucoma, segmentation, CDR, disc, cup, Swin-Unet

## I. INTRODUCTION

Visuals are crucial in the realm of computer vision for problem-solving. Image segmentation is a critical component of many visual understanding systems. Image segmentation plays a significant role in diverse contexts, including medical image analysis. The analysis of medical images proves beneficial in identifying and managing illnesses within healthcare systems. One of the medical image segmentation problems is optical disc and cup segmentation which is helpful for diagnosing glaucoma. Most of the global populace suffers from glaucoma, with more than 64 instances reported in 2013. Per forecasts, the number is estimated to escalate by achieving a total count of 80 million and 11.8 million cases by 2020 and 2040. Glaucoma ranks second among the main factors contributing to complete loss of vision across the globe. An approximate 4.5 million individuals face blindness due to this condition on a worldwide scale [1].

Currently, multiple research works address the topic of image segmentation, such as using U-Net for medical image segmentation [2]–[4], a U-Net network based on a convolution neural network (CNN) consisting of an encoder-decoder with skip connection. Recently, the

application of transformers in computer vision, often called Vision Transformer (ViT), especially image segmentation models, has been proposed [5]–[7]. There are also methods of combining CNN and Transformer for image segmentation [8]–[10]. However, the CNN-based encoder-decoder model is large and is slow to compute. The ViT model has the problem that the computational amount of the model increases as the image size increases [11].

Swin-Transformer is the latest algorithm developed from the vision transformer (ViT) proposed by Liu et al. [12]. The Swin Transformer method uses a hierarchical representation of the feature map, unlike ViT, which does not have a hierarchical feature map. With the hierarchical representation, starting by creating small patches and then merging neighboring patches as the layer increases, this model is appropriate for segmentation because pixel-level segmentation can be more accurate when the patches start small and thus capture information better [12].

This study proposes a new optic disc and cup segmentation system based on the retinal fundus image. The proposed system is based on a swin-transformer with the advantage of more efficient computation because the self-attention mechanism is performed in local windows. The main contributions of this research are as follows: (i) an optical disc and cup segmentation system is developed using a swin-transformer approach, which is different from previous studies that generally use CNN and Transformer; (ii) the developed swin-transformer-based segmentation system provides segmentation performance that is quite comparable to existing methods for the same dataset.

The material of this work is divided into several sections. Section II describes previous work on image segmentation and its application, as well as literature reviews. Section III goes into great length on Swin-Unet approaches. The experimental and analytical results are presented in Section IV. Section V contains conclusions.

## II. RELATED WORKS

Some diseases cannot be seen with the naked eye. It is necessary to do check-ups or examinations with specialists periodically, and the diagnosis is subjective, which makes one doctor and other different opinions usually take a very long time and is less effective. As a result, image segmentation is critical in medical image analysis to aid in developing an accurate CAD system [13]. A computer-aided diagnostic (CAD) system is built that is useful for automating disease diagnosis. CAD systems can also be used as decision support by doctors. Several pre-existing segmentation models exist, such as model segmentation based on CNN.

In CNN-based methods, spatial information such as texture and shape, as well as edge details, are given more attention because CNN focuses on obtaining a local

representation. Still, the global relationship is limited to the local representation. Classic CNN architectures, such as FCN [14], generate features using a pooling layer. This layer has the potential to interfere with the accuracy of spatial information. To improve the accuracy of spatial information, U-Net [15] is proposed by applying a skip connection between shallow and deep layers so that spatial information from shallow layers is not lost. However, the CNN-based model has a significant drawback because it cannot maintain the long-range relationship and localization of the convolutions in the CNN-based model [16].

The attention mechanism was first widely used in Natural Language Processing (NLP). The way this mechanism works is to mimic the attention system in the human brain. When humans look at an object, they pay attention to which parts are more important to focus on. This mechanism is applied to the network by giving weight to important parts higher than other parts in the network. One of the Attention mechanism implementations is the multi-head attention mechanism in Transformer [17]. By utilizing this mechanism, transformers have effectively accomplished many tasks within the natural language processing domain. After success in NLP, transformers also entered the field of computer vision, including Vision Transformers (ViT) [18]. Although Transformer does not have an inductive bias compared to CNN, this model can obtain state-of-the-art (SOTA) image recognition in its performance when pre-training with large datasets such as JFT-300M and ImageNet-22K. The downside of utilizing the ViT model is that as the image dimensions increase, so does the computation workload.

Inductive bias is introduced with the Swin-Transformer to improve the Transformer in image recognition [12]. Swin-Transformer also has more efficient computation and linear computational complexity due to the application of shifted windows multi-head attention mechanism. This mechanism also ensures information exchange between patches in different windows. The Swin-transformer model's arrangement is based on a hierarchical structure and has the flexibility of being a common backbone of the network rather than a Transformer-based model.

Considering the benefits offered by the Swin-Transformer, the Swin-Unet segmentation method was proposed, where this method obtained SOTA in its performance in segmentation using public medical image datasets [13]. Some applications of Swin-Unet in segmentation also exist [19]. This paper uses the Swin-Unet method to segment optical disc and cup.

### III. OPTIC CUP AND DISC SEGMENTATION BASED ON SWIN UNET

The segmentation system in this paper starts by preprocessing the input data and dividing it for training and testing. The Swin-Unet model is trained with training data. The model is tested with test data then the segmentation test results are calculated as vCDR value, after which glaucoma is detected with several predetermined thresholds. The illustration is shown in Fig. 1.

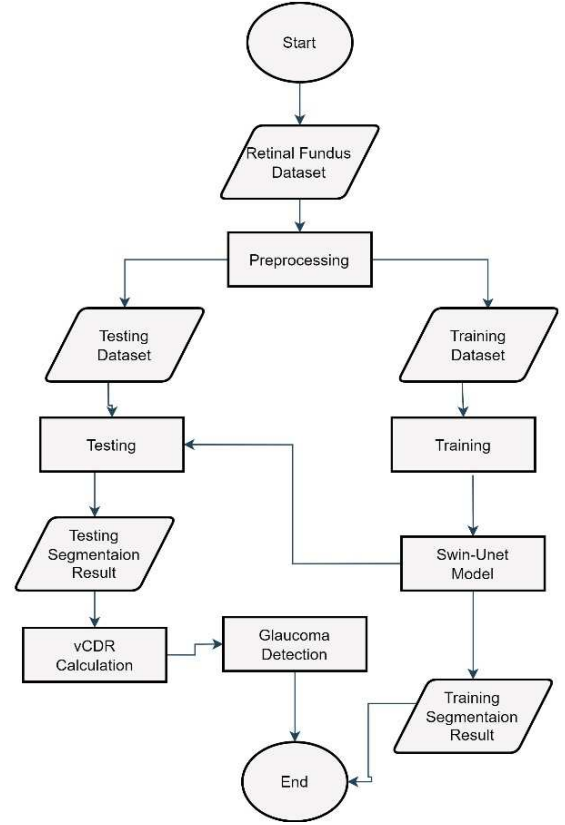


Fig. 1. Flowchart of retinal fundus image semantic segmentation system using Swin-Unet.

#### A. Preprocessing

The dataset obtained is very large (2124 x 2056 and 1634×1634 pixels). For that, resize is needed to speed up and facilitate the training of the model to be made. The resize size that is used is 224 x 224 pixels, as illustrated in Fig. 2. After that, normalization is carried out for the input pixel value of the RGB image into the interval [0, 1] and for the annotation or mask pixel value is converted to categorical [0, 1, 2], where 0, 1, 2 represent background, disc, and cup respectively.

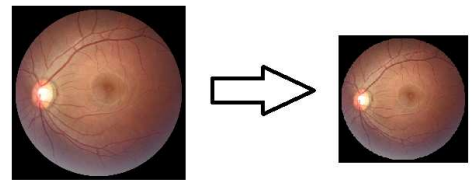


Fig. 2. Resize image.

#### B. Swin-Transformer

The Swin Transformer method uses a hierarchical feature map representation, unlike ViT, which does not have a hierarchical feature map, as can be seen in Fig. 3. With a hierarchical representation, starting by making a small patch and then combining neighboring patches as layers increase, this model is appropriate for segmentation because segmentation based on pixel-level can be more accurate when the patch starts small so it can capture better information [12].

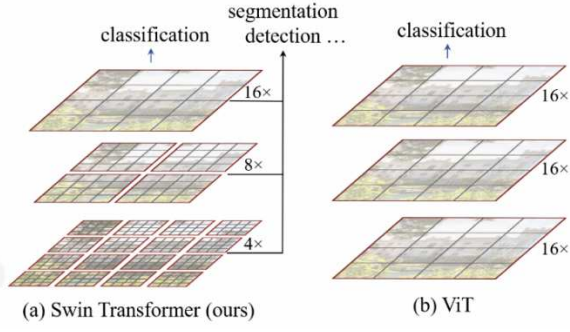


Fig. 3. (a) hierarchical representation of Swin Transformer map features (b) representation of ViT map features [12].

The structure of the Swin Transformer is composed of four phases, which can be seen in Fig. 4. There are four major components: patch partition, linear embedding, switch transformer block, and patch merging. Patch partitioning is dividing an image into several small non-overlapping patches. After that, linear embedding is applied to provide tokens on each patch. Then a swin transformer block is applied, which can be seen in Fig. 8. To get a hierarchical representation, patch merging is applied [12].

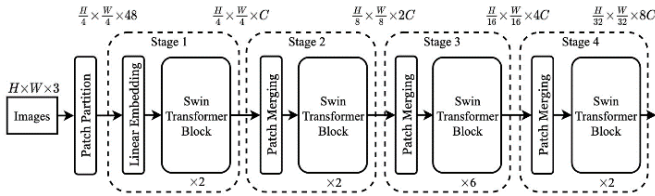


Fig. 4. Swin Transformer architecture [12].

Local computation in non-overlapping windows yields linear computation. The quantity of patches allocated to each window remains unchanged, which is why it is linear to the image resolution size. A shifted window allows relationships between neighbors to be established in windows that do not overlap on the previous layer, as seen in Fig. 5.

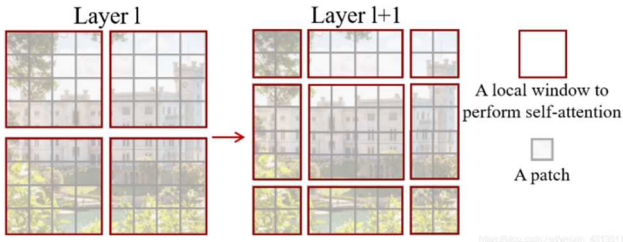


Fig. 5. shifted window approach [12].

More efficient batch computation, with the strategy used, involves shifting cyclically towards the upper left corner, which can be seen in Fig. 6. Through the new sub-windows formed from the shifted window, masking is carried out, which makes the computation only limited to each sub-window. Then perform a cyclic shift that maintains the identical number of partitions as the default window divisions. This is what makes it effective [12].

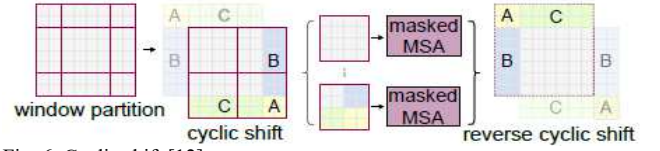


Fig. 6. Cyclic shift [12].

### C. Swin-Unet

Swin-Unet is a Swin Transformer-based encoder-decoder model. Fig. 7 depicts the model architecture. There is encoder, bottleneck, decoder, and skip connections. The encoder works as in subsection B. The bottleneck is built from 2 swin transformer blocks based on Fig. 8, while the dimensions and resolution of the features have not changed.

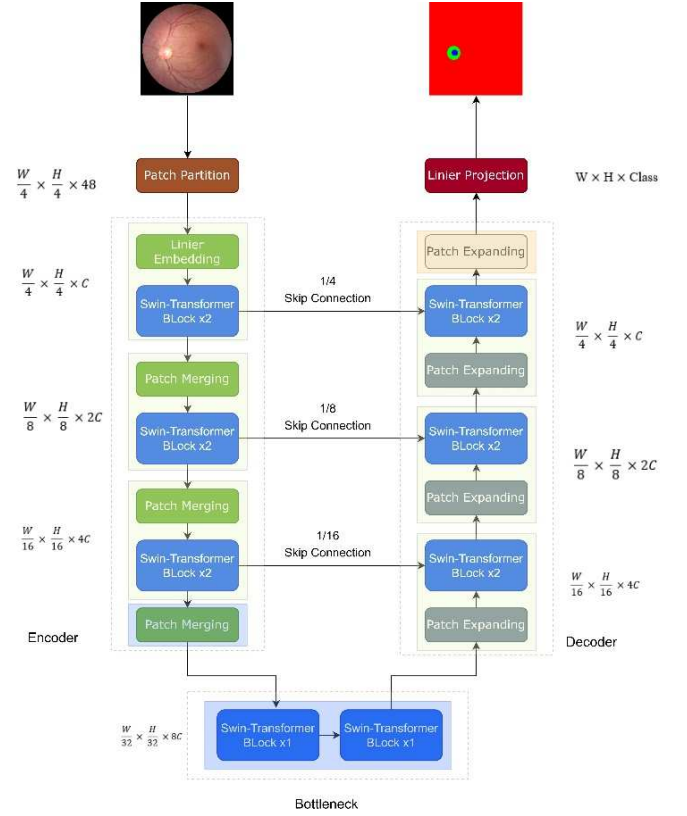


Fig. 7. Swin-Unet architecture.

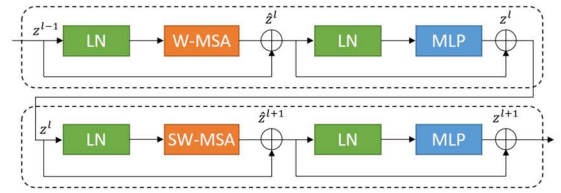


Fig. 8. Swin transformer block [13].

Based on the window partition mechanism, the swin transformer block can be formulated into 4 parts, as shown in (1), (2), (3), (4).

$$\hat{Z}^l = W - \text{MSA}(\text{LN}(Z^{l-1})) + Z^{l-1} \quad (1)$$

$$Z^l = \text{MLP}(\text{LN}(\hat{Z}^l)) + \hat{Z}^l \quad (2)$$

$$\hat{Z}^{l+1} = \text{SW-MSA}(\text{LN}(Z^l)) + Z^l \quad (3)$$

$$Z^{l+1} = \text{MLP}(\text{LN}(\hat{Z}^{l+1})) + \hat{Z}^{l+1} \quad (4)$$

The Swin transformer block comprises the LayerNorm (LN) layer, multi-head self-attention module, residual connection, and two MLP layers with non-linearity GELU. This block employs two types of multi-head self-attention modules: window-based multi-head self-attention module (W-MSA) and shifted window-based multi-head self-attention module (SW-MSA).  $\hat{Z}^l$  and  $Z^l$  is the results of the SW-MSA and MLP module of the  $l$  block. The swin-unet model uses a self-attention mechanism which can be formulated, as shown in (5).

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (5)$$

The matrices denoted by  $Q$ ,  $K$ , and  $V \in \mathbb{R}^{M^2 \times d}$  correspond to the query, key, and value respectively.  $M^2$  is the number of patches in the windows,  $d$  represents the number of dimensions of the query or key, and  $T$  represents the transpose, the  $K$  matrix is transposed to avoid dimensional conflicts when multiplying matrices. And the value of  $B$  is the bias value in the matrix  $B \in \mathbb{R}^{(2M-1) \times (2M+1)}$ . The decoder consists of an expanding patch that functions to reshape the extracted features into features that have a higher resolution ( $2 \times$  up-sampling) and reduce the feature dimensions to half of the original. Skip connections combine shallow and deep features to eliminate lost spatial information due to down-sampling. Then a linear layer is applied so that the combined features have the same dimensions as the up-sampled features [13].

#### D. IoU

This research uses the Intersection over Union (IoU) value to evaluate the segmentation performance. IoU is a standard measurement for segmentation by measuring how many ground truth pixels are correctly predicted by the prediction mask (area of overlap), with all pixels from the ground truth and prediction mask (area of union) [20]. IoU can be formulated, as shown in (6).

$$\text{IoU} = \frac{TP}{(TP + FP + FN)} \quad (6)$$

Ground truth is the region in the image that belongs to a class. True Positive (TP) is an area that is predicted to be in the ground truth. False Positive (FP) is an area predicted not to enter the ground truth. False Negative (FN) occurs when the model fails to indicate a specific area that is actually true based on the ground truth.

#### E. vCDR

Vertical cup-to-disc ratio (vCDR) is the vertical diameter ratio of the optic cup and optic disc. Glaucoma is often identified in three ways: intraocular pressure (IOP) measurement, visual field testing, and optic nerve head or optic disc (ONH) examination. Experts are more likely to take the ONH test. The cup-to-disc ratio (CDR) is a frequent test ophthalmologists use to assess ONH [21]. In this study, vCDR is used to determine whether a person has glaucoma based on the vCDR value obtained by calculating the mask/annotation of the image. vCDR can be formulated, as shown in (7).

$$\text{vCDR} = \sqrt{\frac{\text{Vertical area of optic cup}}{\text{Vertical area of optic disk}}} \quad (7)$$

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

Swin-Unet is a good segmentation model because it applies several mechanisms such as windows partition, shifted windows, and self-attention and has an architecture like U-Net. Subsections A describes the data used in the experiments, subsections B describes parameter setup, subsections C describes model comparison, and subsections D describes glaucoma detection.

### A. Datasets

The dataset used in this study is the REFUGE dataset comprising 1200 JPEG format retinal fundus images that were captured by ophthalmologists or technicians using two different instruments, namely Zeiss Visucam 500 (which produced 400 images with a resolution of 2124 x 2056 pixels) and Canon CR-2 (which produced 800 images having a resolution of 1634 x 1634 pixels) [22]. Example images and their related masks sourced from the REFUGE dataset are shown in Fig. 9. In this study, the data is divided into 3 parts with a ratio of 80% for training, 10% for validation, and 10% for testing.

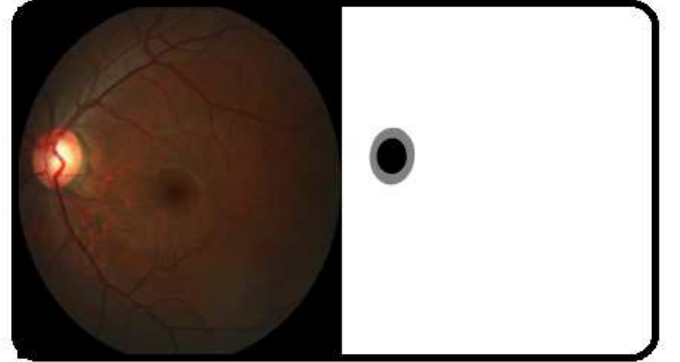


Fig. 9. Optic cup and disc segmentation.

### B. Parameter Setup

To acquire the optimal model during training, we examined different Swin-Unet hyperparameters, including the embedding dimension, MLP node, and head dimension.

#### Embedding dimension

By changing the number of embedding dimensions to 36, 96, 192, and 256 respectively, we explored the influence of different embedding dimension on the segmentation performance. As seen in Fig. 10, the accuracy of the model's ability to segment increases as there is an increase in the quantity of embedding dimensions. As a result, the number of embedding dimensions is adjusted to 192 in this study to make the model more resilient.



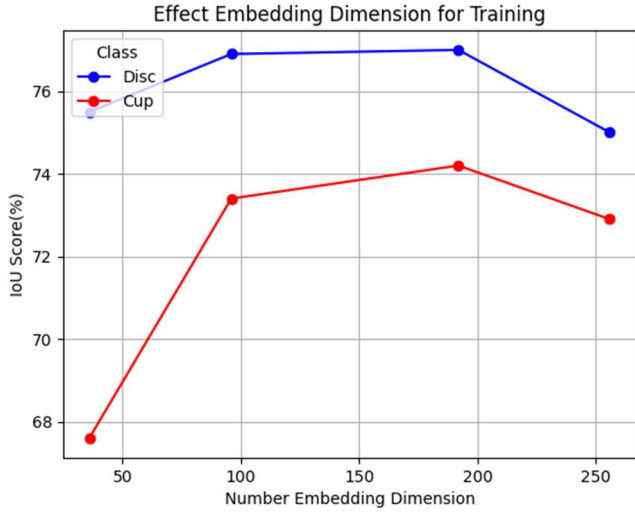


Fig. 10. The impact of the number of embedding dimensions on training.

### MLP node

We investigated the impact of different MLP nodes on the segmentation performance of the proposed model by adjusting the number of MLP nodes to 768, 1536, 3072, 6144, and 9216, respectively. As seen in Fig. 11, the accuracy of the model's ability to segment increases as there is an increase in the quantity of MLP nodes. As a result, the number of embedding dimensions is set to 6144 in this study to make the model more resilient.

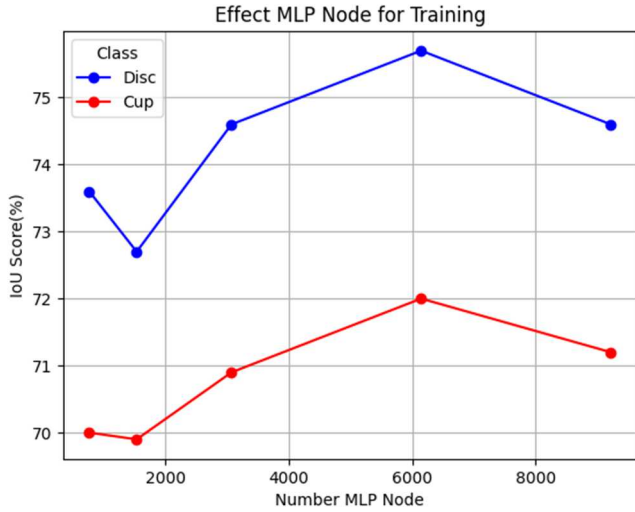


Fig. 11. The impact of the number of mlp nodes on training.

### Head dimension

We investigated the effect of different head dimensions on the segmentation performance of the proposed model by adjusting the number of head dimensions to 12, 32, 64, and 74, respectively. As seen in Fig. 12, the accuracy of the model's ability to segment increases as there is an increase in the quantity of head dimensions. As a result, the number of head dimensions is adjusted to 64 in this study to make the model more resilient.

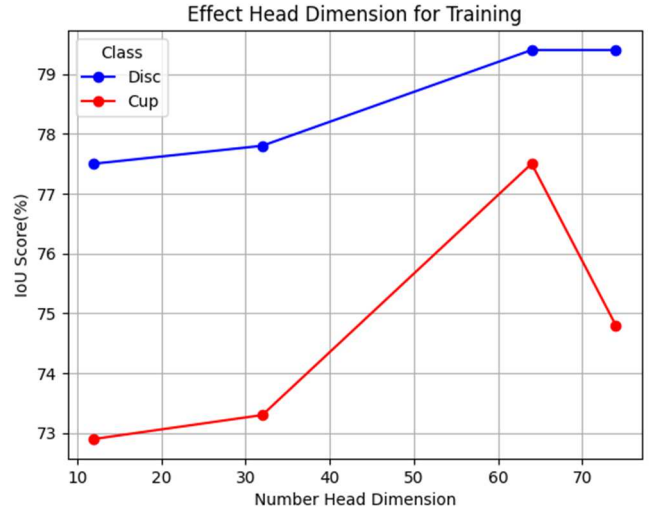


Fig. 12. The impact of the number of head dimensions on training.


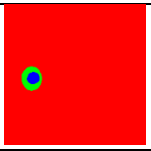
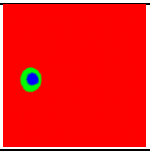
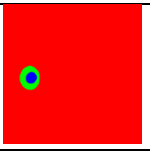
### C. Results

Swin-Unet and U-net have the same architecture. The only difference is that Swin-Unet is based on Swin-transformer, and Unet is based on CNN or convolution. In this part, we show the comparison of the IoU score of Swin-Unet compared to U-Net. As shown in Table I, Swin-Unet performs best in the REFUGE dataset compared to U-Net with intersection over union. The sample of prediction result comparison can be seen in Table II.

TABLE I. ACCURACY OF SEGMENTATION USING TWO DIFFERENT METHODS ON THE REFUGE DATASET.

Model	Training IoU Score		Testing IoU Score	
	Disc	Cup	Disc	Cup
U-Net	84.0	81.1	83.0	80.0
Swin-Unet	<b>84.3</b>	<b>80.5</b>	<b>84.0</b>	<b>80.5</b>

TABLE II. SEGMENTATION RESULT OF 2 DIFFERENT METHODS ON THE REFUGE DATASET

Image	Ground Truth	Predicted Result	
		Swin-Unet	U-Net
			
(IoU Disc; IoU Cup; vCDR)		(90.3; 90.7; 0.75)	(87.9; 83.9; 0.56)


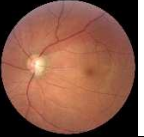
### D. Glaucoma detection

To identify glaucoma, we calculate the vCDR value from the cup and disc mask, and we use the vCDR value as a threshold to determine whether there's glaucoma in the fundus image. If the vCDR value is the same as the threshold or more, we label it glaucoma, and lower than the threshold, we label it non-glaucoma. We use various vCDR thresholds, namely 0.80, 0.70, and 0.63. Table III shows the result of glaucoma accuracy prediction based on the vCDR threshold. We give a sample for glaucoma detection based on the vCDR value in Table IV.

TABLE III. THRESHOLD AND ACCURACY ON GLAUCOMA DETECTION

Threshold	Accuracy
0.80	70%
0.70	80%
<b>0.63</b>	<b>94%</b>

TABLE IV. GLAUCOMA DETECTION ON THE SEGMENTATION RESULT

Image	Label Ground Truth	Label Prediction	IoU Score Disc	IoU Score Cup
	Glaucoma	Non-Glaucoma	83.1	73.3
	Glaucoma	Glaucoma	85.9	78.6

## V. CONCLUSION

Segmentation of optic disc and cup images is very important for glaucoma detection because we can calculate the vCDR value to take early preventive measures against glaucoma. By utilizing the REFUGE dataset, which has been segregated into three segments for training, validation, and testing purposes, we can demonstrate the capability of two different models, Swin-Unet and U-Net. For better model performance for Swin-Unet, we set 3 hyperparameters to embed the dimensions: 192, MLP nodes: 6144, and head dimension: 64. Based on the experiments conducted, the cup and disk segmentation system based on Swin-Unet achieves good performance compared to U-Net.

## REFERENCES

- [1] 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-Disk Ratio computation in Retinal Fundus Images," 2019.
- [2] H. Xiong, S. Liu, R. V. Sharan, E. Coiera, and S. Berkovsky, "Weak label based Bayesian U-Net for optic disc segmentation in fundus images," *Artif Intell Med*, vol. 126, Apr. 2022, doi: 10.1016/j.artmed.2022.102261.
- [3] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 3–11. doi: 10.1007/978-3-030-00889-5\_1.
- [4] H. J. Sandoval-Cuellar, M. A. Vázquez Membrillo, G. Alfonso-Francia, J. C. Ortega Pedraza, and S. Tovar-Arriaga, "Optic Disc and Optic Cup Segmentation Using Polar Coordinate and Encoder-Decoder Architecture," in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 117–126. doi: 10.1007/978-3-030-89586-0\_9.
- [5] J. Gu *et al.*, "Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01178.
- [6] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00717.
- [7] S. Zuo, Y. Xiao, X. Chang, and X. Wang, "Vision transformers for dense prediction: A survey," *Knowl Based Syst*, vol. 253, 2022, doi: 10.1016/j.knosys.2022.109552.
- [8] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021. doi: 10.1007/978-3-030-87199-4\_16.
- [9] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021. doi: 10.1007/978-3-030-87193-2\_2.
- [10] H. Wang, X. Chen, T. Zhang, Z. Xu, and J. Li, "CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images," *Remote Sens (Basel)*, vol. 14, no. 9, May 2022, doi: 10.3390/rs14091956.
- [11] Y. Gu, Z. Piao, and S. J. Yoo, "STHardNet: Swin Transformer with HardNet for MRI Segmentation," *Applied Sciences (Switzerland)*, vol. 12, no. 1, Jan. 2022, doi: 10.3390/app12010468.
- [12] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00986.
- [13] H. Cao *et al.*, "Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2023. doi: 10.1007/978-3-031-25066-8\_9.
- [14] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 4, 2017, doi: 10.1109/TPAMI.2016.2572683.
- [15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. doi: 10.1007/978-3-319-24574-4\_28.
- [16] W. Gao *et al.*, "TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.00288.
- [17] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.
- [18] K. Han *et al.*, "A Survey on Vision Transformer," Dec. 2020, doi: 10.1109/TPAMI.2022.3152247.
- [19] J. Yao and S. Jin, "Multi-Category Segmentation of Sentinel-2 Images Based on the Swin UNet Method," *Remote Sens (Basel)*, vol. 14, no. 14, Jul. 2022, doi: 10.3390/rs14143382.
- [20] W. Zhao, H. Zhang, Y. Yan, Y. Fu, and H. Wang, "A semantic segmentation algorithm using FCN with combination of BSLIC," *Applied Sciences (Switzerland)*, vol. 8, no. 4, Mar. 2018, doi: 10.3390/app8040500.
- [21] 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.
- [22] J. I. Orlando *et al.*, "REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs," Oct. 2019, doi: 10.1016/j.media.2019.101570.