VISION TRANSFORMER-BASED RETINA VESSEL SEGMENTATION WITH DEEP ADAPTIVE GAMMA CORRECTION

Hyunwoo Yu[⋆†] *Jae-hun Shim*[⋆†] *Jaeho Kwak*[⋆†] *Jou Won Song*[⋆†] *Suk-Ju Kang*[†]

[†] Department of Electronic Engineering, Sogang University, Seoul, Korea
[⋆] Equal contribution

ABSTRACT

Accurate segmentation of the retina vessel is essential for the early diagnosis of eye-related diseases. Recently, convolutional neural networks have shown remarkable performance in retina vessel segmentation. However, the complexity of edge structural information and the changeable intensity distribution depending on retina images reduce the performance of the segmentation tasks. This paper proposes two novel deep learning-based modules, channel attention vision transformer (CAViT) and deep adaptive gamma correction (DAGC), to tackle these issues. The CAViT jointly applies the efficient channel attention (ECA) and the vision transformer (ViT), in which the channel attention module considers the interdependency among feature channels and the ViT discriminates meaningful edge structures by considering the global context. The DAGC module provides the optimal gamma correction value for each input image by jointly training a CNN model with the segmentation network so that all the retina images are mapped to a unified intensity distribution. The experimental results show that our proposed method achieves superior performance compared to conventional methods on widely used datasets, DRIVE and CHASE DB1.

Index Terms— Vision Transformer, Channel Attention, Gamma Correction, Retina Vessel Segmentation

1. INTRODUCTION

Retina vessel segmentation is crucial in the early diagnosis of eye-related diseases. For example, diabetes and hypertension can be diagnosed from observing the retina vessel as they cause morphological changes in the blood vessels of the retina. However, manual annotations performed by ophthalmologists are very time-consuming and laborious. This brought demands for automatic retina vessel segmentation to alleviate the burdens of manual annotations.

Recently, deep learning algorithms have shown remarkable performance in the accurate retina vessel segmentation. U-Net [1] has shown considerable performance in the field of medical image segmentation by applying the skip connection to the traditional encoder-decoder architecture. Due to its performance, many studies attempted to augment the U-Net structure to enhance the performance of retina vessel segmentation. Zhang et al. [2] incorporated an edge-based mechanism to U-Net architecture, and Wu et al. [3] proposed a multiscale network followed network (MS-NFN) to enhance the performance of retina vessel segmentation.

However, these methods do not consider the interdependency among feature channels. To tackle this issue, many studies attempt to alter the structure of the skip connection in U-Net to focus on the details. AG-Net [4] utilized an attention mechanism named an attention guide filter to skip connection to better preserve the structural information. ECA-Net [5] proposed the efficient channel attention (ECA) module which applies channel attention to skip connection. CAR-UNet [6] further developed the ECA module to propose the modified efficient channel attention (MECA) which jointly applies max pooling and average pooling to the attention mechanism to aggregate more spatial information.

Consideration of global spatial relations is important in retrieving structural information from noisy images such as retina vessel images. SA-CNN [7] used the self-attention module to capture global relationships in the spatial domain in the CT denoising task. Recently, the success of vision transformer (ViT) [8] has attracted considerable attention to various deep learning tasks such as object detection and image segmentation. Its strength lies in the ability to model global contexts by considering long-range relations. Trans-UNet [9] fully exploited this advantage to apply ViT into the transformer of the U-Net structure to enhance the performance in medical image segmentation.

Another factor that makes retina vessel segmentation challenging is the varying intensity distribution of the retina images. The tonal quality of the retina image varies greatly depending on the system (i.e. camera type, lighting) on which the image is taken. Therefore, models may fail to generalize to fit unseen intensity distributions of the test images. To solve this, Sun et al. [10] proposed channel-wise random

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1004208), the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2018-0-01421) supervised by the IITP(Institute of Information communications Technology Planning Evaluation)

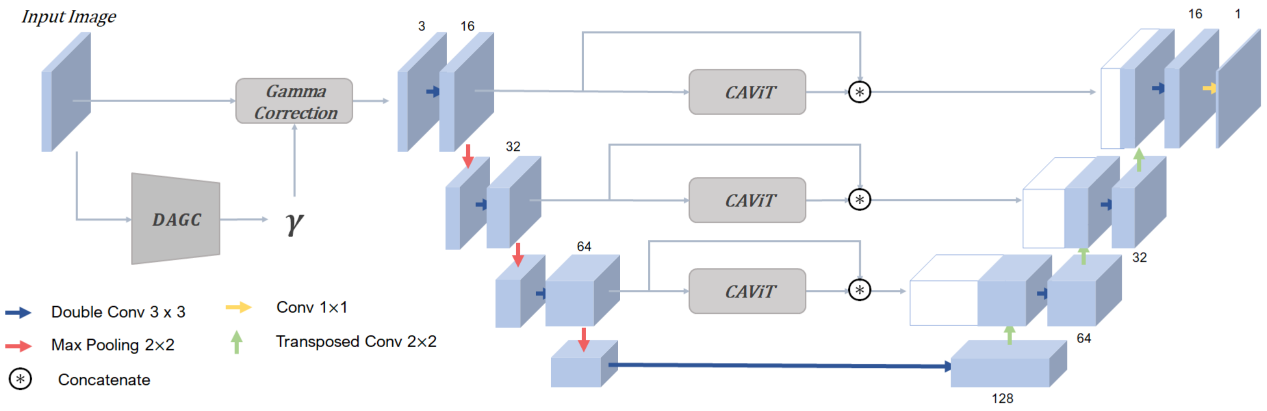


Fig. 1. Diagram of the proposed model. The DAGC module based on ResNet18 architecture is applied before the U-Net architecture. Predicted gamma value is used to perform gamma correction on the input image, which is the input to the U-Net architecture. The CAViT block is applied to the skip connection.

gamma correction (CWRGC) for retina vessel augmentation to introduce data under various illumination settings into the training set. Huang et al. [11] proposed the novel adaptive gamma correction (NAGC), which utilized histograms to perform the adaptive gamma correction to generalize the intensity of the brain image. However, conventional methods such as adaptive gamma correction cannot guarantee that the gamma-corrected intensity distribution is very suitable for the neural network.

This paper proposes two novel deep learning-based modules, channel attention vision transformer (CAViT) and deep adaptive gamma correction (DAGC) to enhance the performance of retina vessel segmentation. The CAViT block jointly applies the ECA module and ViT to the skip layer of U-Net. The ECA module is jointly used with a convolution layer to leverage the interdependency among channel features. From the output of the ECA module, the ViT captures the global relation between image patches to sort out meaningful edge structures. We integrate the CAViT block to the skip connection of the U-Net backbone architecture to enhance the ability to extract important edge structures. The DAGC utilizes CNN to generate gamma values that generalize input retina images to similar intensity distribution. By jointly training this DAGC module with the segmentation network, gamma values can be learned to map the intensity distribution, from which the segmentation network can learn best. The proposed method is evaluated on two widely used datasets for retina vessel segmentation: DRIVE and CHASE DB1. The results show that our proposed method achieves superior performance to conventional methods on both datasets.

2. PROPOSED METHOD

The overall architecture of the proposed method is shown in Fig. 1. In this section, we describe the building blocks of the proposed method in detail: the CAViT and the DAGC.

2.1. Channel Attention Vision Transformer

As presented in Fig. 2., CAViT is composed of an ECA module, a convolution layer, and a ViT. Features transferred to the skip connection of U-Net are very low-level features. Therefore, they contain features with useless information along with the useful edge structures that need to be preserved.

The ECA module, first proposed in ECA-Net [5], measures the interdependency among channels to provide weights to the channel features according to their importance with little computational complexity. In the CAViT block, we used an ECA module so that the module can focus on important channel features.

ViT is a self-attention-based network that learns the relationship between image patches, thereby strongly focusing on the global context of the image. The ViT decomposes a given image into patches and performs linear projection to transform each patch into a 1D patch embedding vector. These embedding vectors are linearly projected into key, query, and value. Then the matrix multiplication is performed between the query and the transposed key, to output the attention score, which is multiplied by the value to produce the final output feature. This process enables the ViT to globally reference the relationship among image patches.

The proposed CAViT block fully exploits this advantage of the ViT. The output feature from the ECA module, which is weighted with the attention map, is put into the ViT, where the ViT processes the discriminated feature maps globally across the spatial dimension. By doing this, the ViT can weaken meaningless edge structures, such as noise, and extract only the meaningful features. The joint structure of the ECA module and the ViT in the CAViT block enables the segmentation network to recover important edge structures while mitigating the interference of meaningless edge structures from the low-level features passed on to the skip layer. The effects of the CAViT block are presented in Fig. 4.

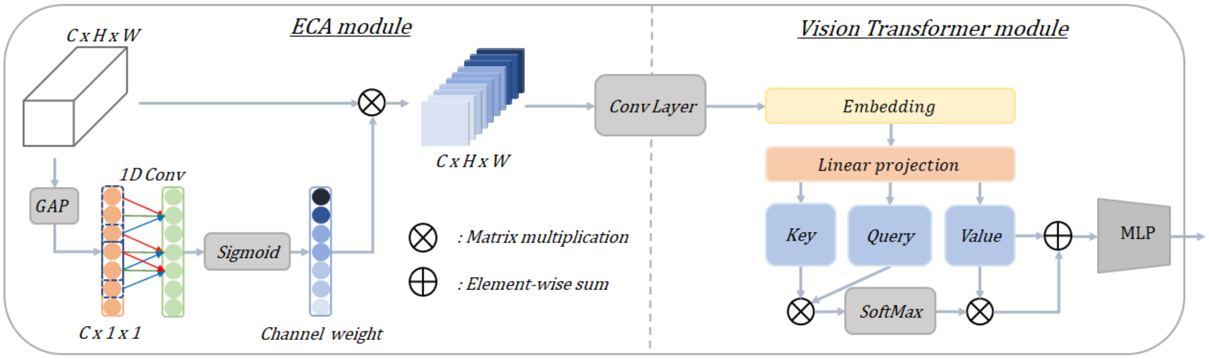


Fig. 2. Overall architecture of the proposed CAViT block. The ECA module, a convolution layer and a ViT module is sequentially connected to form the CAViT block.

2.2. Deep Adaptive Gamma Correction

Although there have been attempts to adaptively adjust the gamma correction value to unify the intensity distribution of medical image data [11], these methods do not guarantee the intensity distribution best suited for the segmentation network. Therefore, we introduce a novel deep learning-based adaptive gamma correction method called DAGC.

We utilized the ResNet18 [12] classification model as the backbone architecture for the novel CNN-based gamma predictor. To modify the classification model into a gamma predictor, we modified the last fully connected layer of ResNet18 to predict a single gamma correction value. The DAGC module is trained simultaneously with the segmentation network, so the gamma prediction and the segmentation map are jointly trained to supplement each other. The DAGC module takes an input retina image to predict a gamma value for the input image, and the predicted gamma value is used to perform gamma correction on the input image, which maps the intensity of the image to a common intensity distribution. The gamma-corrected image is then fed into the segmentation network to produce the final result. The effects of the DAGC module are displayed in Fig. 3.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Environments

We evaluated our proposed method using two public retinal fundus image datasets: DRIVE and CHASE DB1. The DRIVE dataset consists of 40 fundus images of 565×584 pixel resolution, divided into 20 training images and 20 testing images, respectively. The CHASE DB1 dataset consists of 28 fundus images with 999×960 pixel resolution divided into 20 training images and 8 testing images each.

The resolutions of the input images in each skip connection were set to the multiples of 16 as the image patch sizes of the ViT were each set to 16×16, 8×8, and 4×4. Therefore, the resolution of the DRIVE and CHASE DB1 dataset were padded to 592×592 and 1008×1008. To enhance the robustness of the network, we applied standard rotation and flipping data augmentation. In addition, inspired by [10], we applied gamma correction with random gamma values in the range of [0.7, 1.3], to acquire robustness on lighting variation of the retina images. We used field of view (FoV) masks to only apply random gamma correction on the oculi region. Because CHASE DB1 does not provide the official masks, as does DRIVE, we used simple thresholding to acquire masks on the CHASE DB1 dataset, similar to the method used in IterNet [13]. To obtain objective results, we evaluated on the fundus images cropped to their initial size.

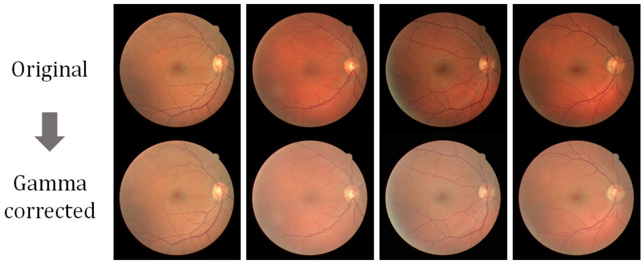
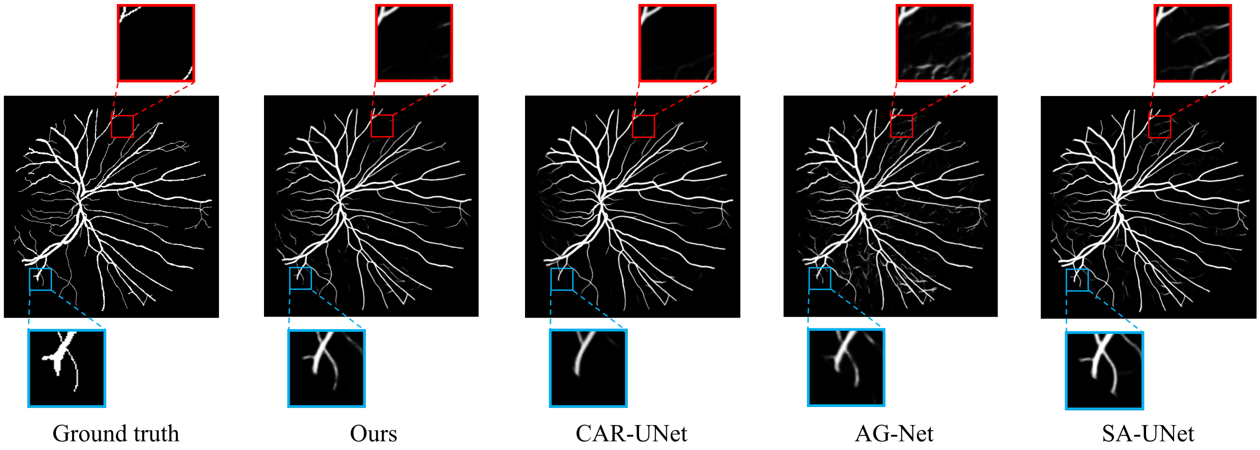


Fig. 3. Results of gamma correction on the predicted gamma values on selected samples in DRIVE test set. Note that although the intensity mapping of the original retina image varies greatly, gamma corrected images show similar intensity distributions.

3.2. Training Settings

We generally followed the training settings in CAR-UNet [6]. The feature channel after the first convolution was set to 16. For the optimizer, we used Adam optimizer with binary cross-entropy as the loss function. For both DRIVE and CHASE DB1, we trained the network for 100 epochs using the step learning rate scheduler with the initial learning rate of 1×10^{3} with gamma 0.5 applied at epoch 30, 50, 70, and 90. The batch size was set to 2.



Ground truth Ours CAR-UNet AG-Net SA-UNet

Fig. 4. Comparison of the results of a selected image from DRIVE test set. The red boxes denote regions with meaningless noisy edge structures and blue boxes denote regions with meaningful edge structures that need to be preserved.

Table 1. Comparison of the proposed methods to other methods on DIVE and CHASE DB1 datasets. The best results in each category are printed in bold.

Metric	Spe	Sen	Acc	AUC
DRIVE				
U-Net[1]	0.9820	0.7537	0.9531	0.9755
AG-Net[4]	0.9848	0.8100	0.9692	0.9856
Pyramid U-Net[14]	0.9807	**0.8213**	0.9615	0.9815
CAR-UNet[6]	0.9849	0.8135	0.9699	0.9852
SA-UNet[15]	0.9840	0.8212	0.9698	**0.9864**
Ours	**0.9872**	0.7924	**0.9700**	0.9854
CHASE DB1				
U-Net[1]	0.9701	0.8288	0.9578	0.9772
AG-Net[4]	0.9848	0.8186	0.9743	0.9863
Pyramid U-Net[14]	0.9787	0.8035	0.9639	0.9832
CAR-UNet[6]	0.9839	0.8439	0.9751	0.9898
SA-UNet[15]	0.9835	**0.8573**	0.9755	**0.9905**
Ours	**0.9881**	0.7968	**0.9761**	0.9888

3.3. Ablation Studies

To show that the DAGC module performs as intended, we applied gamma correction on selected samples using the predicted gamma value from the DAGC module. Fig. 3. shows 4 raw retina images from the DRIVE test set with the corresponding gamma-corrected images. As shown in the sample images, despite the great variance of the tonal quality of the 4 sample retina images, gamma-corrected images show very similar intensity distribution for all four images, thereby showing that the DAGC successfully maps all retina images to a uniform intensity distribution as intended.

3.4. Comparison to Conventional Methods

To verify the performance of our proposed method, we compared the performance of our proposed method with other retina vessel segmentation methods on DRIVE and CHASE DB1 datasets. The evaluation metrics used are specificity (Spe), sensitivity (Sen), accuracy (Acc), and area under the ROC Curve (AUC). The results are summarized in Table 1. The results show that our model scored the highest in specificity and accuracy in both DRIVE and CHASE DB1 datasets. For DRIVE dataset, our model achieved 0.23% higher specificity and 0.01% higher accuracy compared to the second-best models in each category. For CHASE DB1 dataset, our model achieved 0.33% higher specificity and 0.06% higher accuracy compared to the second-best models in each category.

We chose a sample image from DRIVE test set to visualize the strength of our proposed model when compared to other methods in Fig. 4. From the result segmented image, our model successfully excluded noisy meaningless edge structures not present in the ground truth image, while preserving detailed edge on meaningful edge structures. The result images show that no other models were able to achieve both simultaneously. The above results display the superiority of our method over conventional methods.

4. CONCLUSION

This paper presents a novel retina segmentation network with channel attention vision transformer (CAViT) and deep adaptive gamma correction (DAGC) module. The CAViT block consists of an ECA module and a ViT, where the ECA module considers the interdependency among feature channels and the ViT extracts meaningful edge structures from the weighted feature map from the ECA module by focusing on the global context. The DAGC utilizes CNN to train for optimal gamma value that maps any retina images to similar intensity distribution. Experimental results on DRIVE and CHASE DB1 datasets show the superiority of our method over conventional retina vessel segmentation methods.

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[2] Yishuo Zhang and Albert CS Chung, "Deep supervision with additional labels for retinal vessel segmentation task," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 83–91.

[3] Yicheng Wu, Yong Xia, Yang Song, Yanning Zhang, and Weidong Cai, "Multiscale network followed network model for retinal vessel segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 119–126.

[4] Shihao Zhang, Huazhu Fu, Yuguang Yan, Yubing Zhang, Qingyao Wu, Ming Yang, Mingkui Tan, and Yanwu Xu, "Attention guided network for retinal image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 797–805.

[5] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," .

[6] Changlu Guo, Márton Szemenyei, Yangtao Hu, Wenle Wang, Wei Zhou, and Yugen Yi, "Channel attention residual u-net for retinal vessel segmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1185–1189.

[7] Meng Li, William Hsu, Xiaodong Xie, Jason Cong, and Wen Gao, "Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2289–2301, 2020.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[10] Xu Sun, Huihui Fang, Yehui Yang, Dongwei Zhu, Lei Wang, Junwei Liu, and Yanwu Xu, "Robust retinal vessel segmentation from a data augmentation perspective," in *International Workshop on Ophthalmic Medical Image Analysis*. Springer, 2021, pp. 189–198.

[11] Zheng Huang, Guoli Song, and Yiwen Zhao, "Brain tumor screening using adaptive gamma correction and deep learning," in *Proceedings of the 2019 8th International Conference on Bioinformatics and Biomedical Science*, 2019, pp. 47–53.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Liangzhi Li, Manisha Verma, Yuta Nakashima, Hajime Nagahara, and Ryo Kawasaki, "Iternet: Retinal image segmentation utilizing structural redundancy in vessel networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3656–3665.

[14] Jiawei Zhang, Yanchun Zhang, and Xiaowei Xu, "Pyramid u-net for retinal vessel segmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1125–1129.

[15] Changlu Guo, Márton Szemenyei, Yugen Yi, Wenle Wang, Buer Chen, and Changqi Fan, "Sa-unet: Spatial attention u-net for retinal vessel segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1236–1242.