U-Net with Hierarchical Bottleneck Attention for Landmark Detection in Fundus Images of the Degenerated Retina

Shuyun Tang, Ziming Qi, Jacob Granley, and Michael Beyeler(✉)

University of California, Santa Barbara, CA 93106, USA
{shuyun,zimingqi,jgranley,mbeyeler}@ucsb.edu

Abstract. Fundus photography has routinely been used to document the presence and severity of retinal degenerative diseases such as age-related macular degeneration (AMD), glaucoma, and diabetic retinopathy (DR) in clinical practice, for which the fovea and optic disc (OD) are important retinal landmarks. However, the occurrence of lesions, drusen, and other retinal abnormalities during retinal degeneration severely complicates automatic landmark detection and segmentation. Here we propose HBA-U-Net: a U-Net backbone enriched with hierarchical bottleneck attention. The network consists of a novel bottleneck attention block that combines and refines self-attention, channel attention, and relative-position attention to highlight retinal abnormalities that may be important for fovea and OD segmentation in the degenerated retina. HBA-U-Net achieved state-of-the-art results on fovea detection across datasets and eye conditions (ADAM: Euclidean distance (ED) of 25.4 pixels, REFUGE: 32.5 pixels, IDRiD: 32.1 pixels), on OD segmentation for AMD (ADAM: Dice coefficient (DC) of 0.947), and on OD detection for DR (IDRiD: ED of 20.5 pixels). We further validated the design of our network with an ablation study. Our results suggest that HBA-U-Net may be well suited for landmark detection in the presence of a variety of retinal degenerative diseases.

Keywords: Deep learning · Landmark detection · Segmentation · Self-attention · Fundus · Fovea · Optic disc · Retinal degeneration · Age-related macular degeneration · Diabetic retinopathy · Glaucoma

1 Introduction

Age-related macular degeneration (AMD), glaucoma, and diabetic retinopathy (DR) are three of the most common causes of blindness in the world [2]. Fundus photography has routinely been used to document the presence and severity of these retinal degenerative diseases in clinical practice. Among the landmarks of interest are the fovea, which is a small depression in the macula, and the optic disc (OD), which is where the optic nerve and blood vessels leave the retina.

H. Fu et al. (Eds.): OMIA 2021, LNCS 12970, pp. 62–71, 2021.
https://doi.org/10.1007/978-3-030-87000-3_7

However, detecting retinal abnormalities associated with these diseases (e.g., drusen in AMD, hemorrhage in DR) is a labor-intensive and time-consuming process, thus necessitating the need for automated fundus image analysis.

In recent years, numerous methods have been proposed for retinal structure detection. Jiang *et al.* [7] proposed an encoder-decoder network with deep residual structure and recursive learning mechanism for robust OD localization, followed by an end-to-end region-based convolutional neural network (R-CNN) for joint optic disc and cup segmentation [8]. Similarly, numerous studies have employed various convolutional neural network (CNN) models for fovea localization (e.g., [1,15]). Although fovea and OD are spatially correlated with each other, only a few studies (e.g., [10,22]) have focused on joint fovea and OD segmentation. Furthermore, models trained on healthy eyes tend not to generalize well to diseased eyes due to retinal abnormalities. A notable exception is Kamble *et al.* [9] who achieved state-of-the-art (SOTA) performance on landmark detection for AMD and glaucoma using a modified U-Net++ with an EfficientNet encoder. However, there is potential merit in combining convolutional backbone networks with attentional mechanisms [19] to highlight retinal abnormalities that may be important for landmark detection in the degenerated retina.

To develop a segmentation model that is well suited for retinal degeneration, we propose HBA-U-Net: a U-Net backbone enriched with hierarchical bottleneck attention. The main contributions of this work are:

1. We propose a hierarchical bottleneck attention (HBA) block: a novel attention mechanism that combines and refines self-attention [19], channel attention [21], and relative-position attention [13] to highlight retinal abnormalities important for landmark detection in the degenerated retina.
2. We integrate the HBA block into bottleneck skip connections across all layers of a U-Net backbone network to form HBA-U-Net, and test the network's performance on three benchmark datasets for retinal degeneration: ADAM [4] for AMD, REFUGE [11] for glaucoma, and IDRiD [12] for DR.
3. We validate the design of HBA-U-Net with an ablation study.
4. We demonstrate SOTA performance on fovea detection across datasets and eye conditions, on OD segmentation for AMD, and on OD detection for DR.

2 Methods

2.1 Model Architecture

HBA-U-Net. The proposed network architecture is illustrated in Fig. 1. First, ImageNet pretrained ResNet-50 blocks were used as encoders to obtain feature maps at different spatial resolutions. These feature maps, along with the original image, were then fed into a modified U-Net structure [10,14] with HBA blocks added to skip connections. The outputs of the HBA blocks were up-sampled and aggregated to produce the final fovea and OD segmentation mask.

Our goal was to incorporate HBA blocks into the U-Net without drastically increasing the computational complexity. Consistent with [16], we noticed that adding a self-attention mechanism to the bottleneck layers (a shrinking path, the

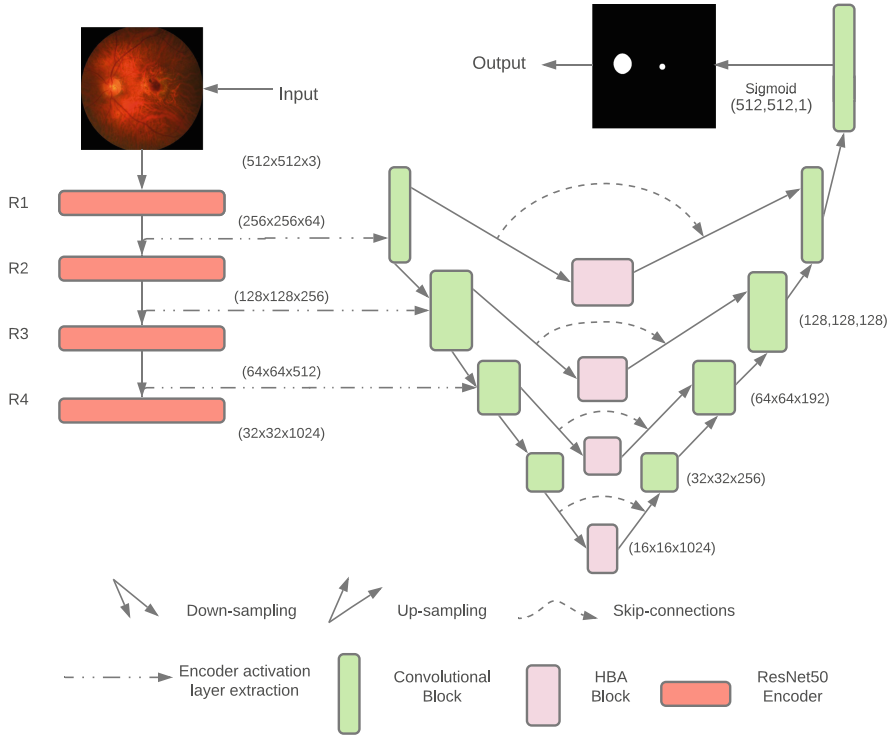

Fig. 1. HBA-U-Net architecture. A U-Net enriched with a novel attention block and re-designed skip-connection paths jointly locates the fovea and segments the optic disc. ResNet-50 was used as encoder. Note that the number of local bottlenecks and the down/up-sampling projection rate depends on the image dimensions.

attention module, and an expanding path) significantly boosted the network's performance. However, the original U-Net contains only a single bottleneck layer (between the last down-sampling block and the first up-sampling block). To incorporate multiple HBA blocks into the network, we therefore re-designed the U-Net by creating local bottleneck structures in each skip-connection pair (see Fig. 1). After each down-convolution block, the features were down-sampled by pooling and passed to the HBA block, followed by up-sampling to the original size. In this way, the pairs of down/up-sampling convolution blocks could be treated as local bottleneck structures operating at different spatial resolutions.

HBA Block. Recently, attention mechanisms have seen widespread adoption in various tasks [19]. Inspired by [16,21], our HBA block (Fig. 2) consisted of channel, content, and relative-position attention modules, each described in detail below. We denote the query, key, value, input feature map, relative height logit, and relative weight logit as q, k, v, F, R_h, R_w, respectively.

In the proposed HBA block, content attention (blue box in Fig. 2) attended to individual pixels in each spatial feature map. For each attention head, dense

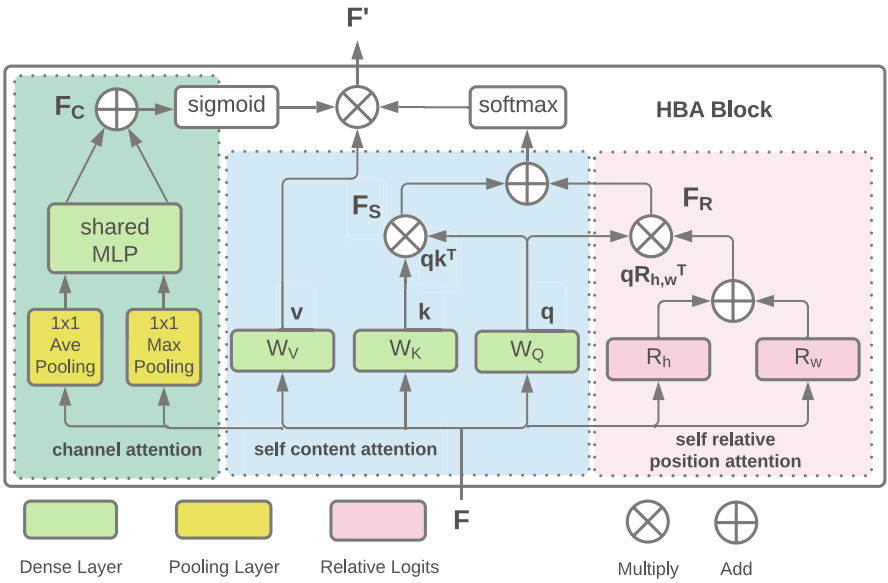

Fig. 2. HBA block architecture, consisting of channel attention (green box outputs F_C), content attention using multi-head self-attention (blue box outputs F_S), and relative position attention (pink box outputs F_R). (Color figure online)

layers (W_Q, W_K, W_V) were used to calculate the query ($q = W_Q(F)$), key ($k = W_K(F)$), and value ($v = W_V(F)$) for each pixel. The output of the content attention was an attention score (F_S) between key (k) and query vectors (q):

$$\mathrm{F_S} = qk^T. \tag{1}$$

Inspired by [13,16], we included relative-position attention (pink box in Fig. 2) to encode the relative position of different retinal landmarks (e.g., to relate the fovea to the OD location). Relative logits were used to store the x and y offsets (R_h and R_w) between each key and query. These were added and the relative positional attention score $\mathrm{F_R}$ was computed using the dot product:

$$\mathrm{F_R} = q(R_h + R_w)^T. \tag{2}$$

In a U-Net, spatial information is encoded to different channels through down/up-sampling. We believe channel-wise attention is well suited to utilize this information in the bottleneck layers, which usually have many channels. We therefore used the channel attention module proposed in [21] (green box in Fig. 2). The input feature map F was passed in parallel to average pooling and max pooling layers, compressing each channel to one value. These two feature maps were forwarded through a single, shared multi-layer perceptron (MLP) with one hidden layer and added to compute the final channel attention score (F_C):

$$\mathrm{F_C} = MLP(AvgPool(F)) + MLP(MaxPool(F)). \tag{3}$$

In contrast to a conventional transformer, the value vector was scaled not only by the content attention score (F_S), but also according to the relative-position attention score (F_R) and the channel attention score (F_C). The output of the HBA block (F′) is given as follows:

$$\mathrm{F}' = softmax(F_S + F_R)\sigma(F_C)v, \tag{4}$$

where the softmax was applied across attention heads and σ denotes the sigmoid function.

2.2 Datasets

We evaluated our model on three prominent datasets for retinal degeneration: ADAM [4] for AMD, REFUGE [11] for glaucoma, and IDRiD [12] for DR.

ADAM was released as part of a Grand Challenge at a satellite event of the ISBI 2020 conference. The dataset contains 400 fundus images at either 2124×2056 or 1444×1444 resolution, 87 of which depict eyes at various stages of AMD progression (typical signs include the presence of drusen, exudation, and hemorrhage), and the rest are from healthy controls. ADAM includes ground-truth OD segmentation masks and fovea image coordinates.

REFUGE was released as part of a Grand Challenge of the OMIA5 workshop at MICCAI 2018. The dataset contains 1200 fundus images at either 2124×2056 or 1634×1634 resolution, 120 of which depict eyes with glaucoma, and the rest are from healthy controls. REFUGE includes ground-truth OD segmentation masks and fovea image coordinates.

IDRiD was released as part of a Grand Challenge at ISBI 2018. The dataset contains 516 images at 4288×2848 resolution divided into 413 train images and 103 test images, all of which contain pathological conditions such as DR and diabetic macular edema. IDRiD includes ground-truth image coordinates for the fovea and OD center, but not segmentation masks.

2.3 Implementation Details

Data Preprocessing and Augmentation. First, we resized every image in the dataset to 512×512 pixels. Second, we followed [9] to generate circular segmentation masks from the ground-truth fovea coordinates and combined them with the ground-truth OD segmentation masks. Third, we applied random image rotations (uniformly sampled from $[-0.2, 0.2]$ rad), and horizontal/vertical flips to augment the original dataset on-the-fly. Fourth, we split the data 85-15 into train and test sets and held out 20% of the training images for validation.

Training Procedure. The model was trained using the adam optimizer, the Dice loss [17], and early stopping, with a custom learning rate scheduler (start rate 0.0025, decay rate 0.985 after 150 epochs), and batch size 8 for 500 epochs. Initial weights were pre-trained on ImageNet. The model was implemented using Keras 2.4.3 (Python 3.7) and run on an NVIDIA Tesla K80 (12 GB of RAM). The code is available at github.com/bionicvisionlab/2021-HBA-U-Net.

Evaluation Metrics. We evaluated the performance of the model using Euclidean distance (ED) [10], where only image coordinates were given, and Dice coefficient (DC), where segmentation masks were given. Since none of the three datasets came with fovea segmentation masks, we followed [10] to create a circular disc centered over the ground-truth fovea coordinates, which was then used to train our network. After training, we recovered predicted coordinates by extracting the centroid of the predicted segmentation mask using scikit-image.

3 Experiments and Results

3.1 Joint Fovea and OD Detection in the Degenerated Retina

Table 1 summarizes our results on three prominent datasets for retinal degeneration: ADAM for AMD, REFUGE for glaucoma, and IDRiD for DR.

HBA-U-Net achieved SOTA performance on fovea detection across all datasets (ADAM: ED 25.4 px; REFUGE: ED 32.5 px; IDRiD: 32.1 px) and thus across eye conditions, despite the fact that these datasets were previously used in Grand Challenges that featured convolutional [9], attentional [23], and adversarial [20] approaches, some of which had a considerably larger number of trainable parameters. Because all three datasets are relatively new, the number of published results is still relatively small.

HBA-U-Net also achieved SOTA performance on OD segmentation for AMD (DC of 0.947, on par with [9]) and on OD detection for DR (ED of 20.5). Our OD segmentation was slightly worse than competing models, with the SOTA belonging to [20], a patch-based morphology-aware segmentation network.

However, please note that the test data of these challenges is not made available to the public. To offer a fair comparison across models, we therefore re-implemented a number of commonly used alternative network architectures and compared their performance using our own train/test split. These alternative

Table 1. Landmark detection on ADAM, REFUGE, and IDRiD. Note that Challenge test data is not publicly available. ED: Euclidean Distance. DC: Dice Coefficient.

		Fovea	Optic Disc	
	Model	ED	ED	DC
ADAM	Aira matrix [9] (ISBI 2020 Challenge Winner)	26.2	–	**0.947**
	HBA-U-Net (this paper)	**25.4**	–	**0.947**
REFUGE	Fu *et al.* [5]	–	–	0.936
	Zhang *et al.* [23]	–	–	0.953
	Kamble *et al.* [9]	35.2	–	0.957
	Wang *et al.* [20]	–	–	**0.960**
	HBA-U-Net (this paper)	**32.5**	–	0.947
IDRiD	DeepDR (IDRiD Subchallenge-3 Winner, on-site)	64.5	21.1	–
	ZJU-BII-SGEX (IDRiD Subchallenge-3 Winner, online)	45.9	25.6	–
	HBA-U-Net (this paper)	**32.1**	**20.5**	–

Table 2. Landmark detection for different reimplemented models tested on ADAM, REFUGE, and IDRiD. ED: Euclidean Distance, DC: Dice Coefficient, F: Fovea, OD: Optic Disc.

Model	ADAM		REFUGE		IDRiD	
	ED_F	DC_{OD}	ED_F	DC_{OD}	ED_F	ED_{OD}
U-Net [14]	70.7	0.741	65.2	0.806	87.1	53.7
EfficientNet encoded U-Net++ [9]	26.9	0.867	37.6	0.935	50.4	28.1
HBA-U-Net (this paper)	**25.4**	**0.947**	**32.5**	**0.947**	**32.1**	**20.5**

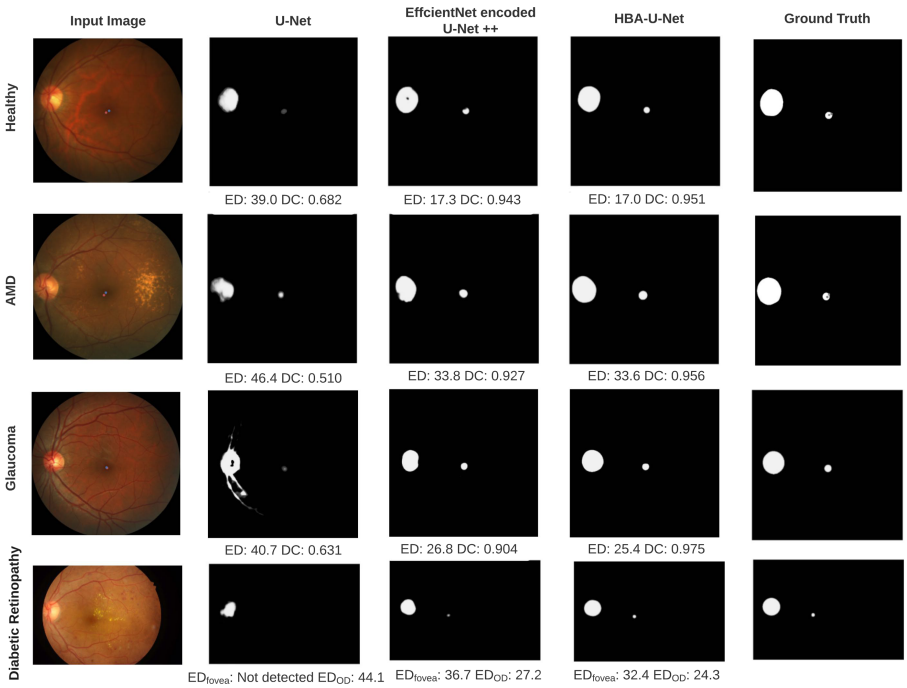

Fig. 3. Representative example predictions for a healthy eye (*top row*), AMD (*second row*), glaucoma (*third row*), and DR (*bottom row*). Predictions are shown for a re-implemented U-Net (*second column*), EfficientNet encoded U-Net++ with scSE blocks (*third column*), and HBA-U-Net (*fourth column*), and compared against ground truth (*rightmost column*). Error rates are given below each prediction panel.

networks included the classical U-Net [14] and an EfficientNet [18] encoded U-Net++ with scSE blocks (similar to [9]). Results are given in Table 2 and example predictions are shown in Fig. 3. HBA-U-Net outperformed the baseline models on all three datasets.

Table 3. Ablation studies on each network component. Starting from a U-Net backbone [10,14], we gradually added a ResNet-50 encoder [6], a standard self-attention block [13] ('Self-Att'), a single HBA block in the bottleneck ('HBA-1'), and HBA blocks across all levels of the hierarchy ('HBA-all').

U-Net	ResNet	Self-Att	HBA-1	HBA-all	Params	Fovea ED	OD DC
✓					8.7M	70.7	0.741
✓	✓				20.8M	34.8	0.902
✓	✓	✓			21.1M	29.8	0.925
✓	✓	✓	✓		21.3M	25.8	0.920
✓	✓	✓	✓	✓	22.2M	**25.4**	**0.947**

3.2 Ablation Study

To measure the impact of the HBA block on different versions of our proposed model architecture, we performed an ablation study on ADAM (see Table 3).

Starting with the original U-Net [10,14] as a baseline, we were able to reduce fovea ED by a factor of two by adding a ResNet-50 encoder [6]. Adding the original self-attention block [13] (without relative position and channel-wise attention; labeled 'Self-Att' in Table 3) at the bottleneck part of the U-Net improved fovea ED by ∼5%, but led to a ∼2% decrease in DC for OD segmentation. Upgrading the self-attention block to our proposed HBA block at the bottleneck part of the U-Net (labeled 'HBA-1') resulted in both the ED and DC improving by ∼4%. Finally, creating local bottlenecks with HBA blocks at each skip connection in the hierarchy (labeled 'HBA-all') led to SOTA performance.

4 Conclusions

We have proposed a re-designed U-Net architecture with hierarchical bottleneck attention and demonstrated its utility for fundus analysis. The proposed network achieved SOTA performance on fovea detection across datasets and eye conditions, on OD segmentation for AMD, and on OD detection for DR.

Although self-attention, channel attention, and relative-position have been deployed separately in other computer vision tasks, here we refined, simplified, and combined their potential in segmenting retinal abnormalities. Furthermore, our ablation study demonstrates the benefit of the local bottleneck structures and HBA blocks for retinal landmark segmentation. Compared to content self-attention alone, HBA does not add much overhead: relative position attention does not have any learnable parameters and channel attention consists of a shared MLP with one hidden layer. Compared to other pure attention networks such as ViT [3], HBA blocks are more resourceful and better suited to work in combination with convolutional modules commonly used in segmentation tasks.

Overall our results suggest that HBA-U-Net may be well suited for landmark detection in the presence of a variety of retinal degenerative diseases.

References

1. Alais, R., Dokládal, P., Erginay, A., Figliuzzi, B., Decencière, E.: Fast macula detection and application to retinal image quality assessment. Biomed. Signal Process. Control **55**, 101567 (2020)
2. Blindness, G., Collaborators, V.I.: Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Global Health **9**(2), e144–e160 (2021)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2021)
4. Fu, H.: ADAM: Automatic detection challenge on age-related macular degeneration, January 2020. Publisher: IEEE type: dataset
5. Fu, H., et al.: Disc-aware ensemble network for glaucoma screening from fundus image. IEEE Trans. Med. Imaging **37**(11), 2493–2501 (2018). Conference Name: IEEE Transactions on Medical Imaging
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). iSSN: 1063–6919
7. Jiang, S., Chen, Z., Li, A., Wang, Y.: Robust optic disc localization by large scale learning. In: Fu, H., Garvin, M.K., MacGillivray, T., Xu, Y., Zheng, Y. (eds.) OMIA 2019. LNCS, vol. 11855, pp. 95–103. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32956-3_12
8. Jiang, Y., et al.: JointRCNN: a region-based convolutional neural network for optic disc and cup segmentation. IEEE Trans. Biomed. Eng. **67**(2), 335–343 (2020). Conference Name: IEEE Transactions on Biomedical Engineering
9. Kamble, R., Samanta, P., Singhal, N.: Optic disc, cup and fovea detection from retinal images using U-Net++ with EfficientNet encoder. In: Fu, H., Garvin, M.K., MacGillivray, T., Xu, Y., Zheng, Y. (eds.) OMIA 2020. LNCS, vol. 12069, pp. 93–103. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63419-3_10
10. Meyer, M.I., Galdran, A., Mendonça, A.M., Campilho, A.: A pixel-wise distance regression approach for joint retinal optical disc and fovea detection. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 39–47. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_5
11. Orlando, J.I., et al.: Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)
12. Porwal, P., , et al.: IDRiD: diabetic retinopathy - segmentation and grading challenge. Med. Image Anal. **59**, 101561 (2020)
13. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. arXiv:1906.05909 [cs], June 2019. arXiv: 1906.05909
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Sedai, S., Tennakoon, R., Roy, P., Cao, K., Garnavi, R.: Multi-stage segmentation of the fovea in retinal fundus images using fully Convolutional Neural Networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 1083–1086, April 2017. iSSN 1945–8452

16. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. arXiv:2101.11605 [cs], January 2021. arXiv: 2101.11605
17. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
18. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. arXiv:1905.11946 [cs, stat], September 2020. arXiv: 1905.11946
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
20. Wang, S., Yu, L., Yang, X., Fu, C.W., Heng, P.A.: Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Trans. Med. Imaging **38**(11), 2485–2495 (2019). Conference Name: IEEE Transactions on Medical Imaging
21. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. CoRR (2018)
22. Yu, H., et al.: Fast localization of optic disc and fovea in retinal images for eye disease screening. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2011, vol. 7963, p. 796317, March 2011
23. Zhang, Z., Fu, H., Dai, H., Shen, J., Pang, Y., Shao, L.: ET-Net: a generic edge-aTtention guidance network for medical image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 442–450. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_49