Attention-based U-net: Joint Segmentation of Layers and Fluids from Retinal OCT Images

M. Melinščak
Zagreb University of Applied Sciences, Zagreb, Croatia
martina.melinscak@tvz.hr

Abstract **– Since its introduction in 2015, U-net has become state-of-the-art neural network architecture for biomedical image segmentation. Although many modifications have been proposed, few novel concepts were introduced. Recently, some significant breakthroughs have been achieved by introducing attention or, more specifically, Transformers. Many attempts to incorporate self-attention mechanisms into solving computer vision tasks resulted in Vision Transformer (ViT). As ViT has some downsides compared to convolutional neural networks (CNNs), neural networks which merge advantages from both concepts prevail, especially in small data regimes we often face in medicine. U-net architecture still outperforms ViT models as their high accuracy relies on massive data. This paper investigates how attention added in U-net architecture affects results. We evaluate the outcomes on a publicly available dataset which consists of 1136 retinal optical coherence tomography (OCT) images from 24 patients suffering from neovascular age-related macular degeneration (nAMD). Also, we compare results to previously published results, and it could be noted that the Attention-based U-net model achieves higher Dice scores by a significant margin. The code is publicly available.**

Keywords – Vision Transformer; attention; convolutional neural networks; U-net; automatic segmentation; retinal optical coherence tomography images

I. INTRODUCTION

Optical coherence tomography (OCT) allows non-invasive 3D imaging where high-resolution images provide information about the retinal structure. Consequently, OCT imaging found an essential role in ophthalmology. Morphological features that can be seen and measured from OCT tomograms, such as the thickness of individual retinal layers, shapes, spatial distributions, and optical properties of various lesions and blood vessels, can serve as identifiers in the diagnosis of retinal diseases [1]. Some of the most often diseases diagnosed by OCT imaging are diabetic retinopathy (DR), age-related macular degeneration (AMD), retinal vein occlusion (RVO), central serous retinopathy (CSR), and glaucoma.

There is increased demand for automatic segmentation due to a lack of ophthalmologists and an increased prevalence of patients suffering from retinal diseases (primarily because of population ageing and the diabetes epidemic). However, detailed inspection of 3D volumes is time-consuming, prone to errors, and mostly impossible within a typical clinical setup. Also, there is no consensus between different ophthalmological opinions. On the other hand, tremendous advances in computer vision have been made, but we still lack a reliable model for automatic segmentation. Deep learning models are still too brittle and unable to generalize well on unseen data, which makes automatic segmentation an open and vital research field.

Since introducing AlexNet architecture, which achieves breakthrough results on an ImageNet benchmark [2], convolutional neural networks (CNNs) have become state-of-the-art architectures for computer vision tasks. Architectures that followed (e.g., VGG [3], GoogleNet [4], ResNet [5], Inception v3 [6], MobileNet [7], and EfficientNet [8]) have had more layers and parameters. However, there were minor enhancements – convolutions were still the predominant paradigm in computer vision. Mostly, concepts introduced in the following years were various regularization techniques and overcoming problems like overfitting due to an enormous number of parameters.

In a highly cited paper, "Attention is all you need," Vaswani et al. introduced the Transformers. Primarily Transformers were used for natural language processing (NLP), and mostly they gained popularity due to the latest large language models (LLM) applications. Next, Dosovitskiy et al. introduced a Vision transformer (ViT) [9]. Unlike convolutional networks, Transformers lack inductive bias like locality and translational equivariance. Furthermore, Transformers work with image patches, and similarly, like in multi-layer perceptron (MLP), all patches attend to all others, making the ViT model enormously computationally costly. Finally, the authors highlighted the remaining challenges, such as solving segmentation and detection tasks with Transformers. Although Transformers became highly popular, they were appropriate for tasks with accessible large amounts of data and immense computing power. Soon after, there were many proposals on how to merge the advantages of self-attention architectures and convolutional neural networks.

In the case of medical image segmentation, introducing U-net architecture [10] made a significant breakthrough. Furthermore, encoder-decoder architecture with skip connections enables concurrent capturing of localization and context. Many modifications to the U-net architecture were proposed [11]–[14]. Nevertheless, U-net-based architectures remain predominant in biomedical image segmentation, mainly due to the small data regime we often face in biomedical fields.

This paper is mainly inspired by Attention U-net [15]. We made all modifications to make it applicable to the

OCT dataset we use for evaluation. Also, we made a comparison to standard U-net architecture and state-of-the-art architectures like U-net++ architecture (a nested U-net architecture for medical image segmentation) [16] and U-net-like architecture [17] to see how adding attention affects results.

In summary, the main contributions of this work are as follows:

- Proposing an attention-based model for joint segmenting layers and fluids in OCT retinal images.
- Evaluating the model on the publicly available AROI dataset of 1136 B-scans from patients with neovascular AMD.
- Comparing obtained outcomes with previous results on the benchmark dataset and improving segmentation results by a significant margin.

The rest of this paper we organize as follows. First, section 2 mentions relevant related work using attention in ophthalmology. Then, in section 3, we describe the database we use for evaluation purposes and give a detailed description of the model we use. Next, in section 4, we provide results and discuss them in the following section; in the last section, we conclude by presenting the limitations of our work and possible future research directions.

II. Related work

Wang et al. [18] proposed a Multi-scale Transformer Global Attention Network (MsTGANet) for drusen segmentation on the Kermany dataset [19]. The Kermany dataset is the largest and most popular publicly available dataset for classifying OCT images into four classes: choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and images from healthy persons. Jiang et al. [20] used a Vision Transformer to classify images on the publicly available DUKE dataset [21]. As a result, images are classified into three classes: patients suffering from AMD, DME, and healthy persons. Cao et al. [22] proposed a regression method based on Transformer that performs segmentation of seven layers in OCT images from the publicly available DUKE DME dataset [23].

Playout et al. [24] presented Vision Transformer for image classification: the paper mainly focuses on classification and mechanisms for generating interpretable predictions via attribution maps. They also perform the segmentation of fluids in images from the AROI dataset. It is hard to compare results with other papers that used images from the AROI dataset as they only segment fluids and do not distinguish different types of fluids. Nevertheless, they show that segmentation with Vision Transformer outperforms classical CNN models.

A few papers [25]–[29] combine attention with CNNs. However, it is out of the scope of this paper to give an extensive survey of all published papers. Instead, we would like to refer the reader to the review study from Khan et al. [30].

III. Materials and methods

A. AROI dataset

To evaluate the proposed model and compare results with CNNs without attention, we use the publicly available AROI dataset [31]. Macular SD-OCT volumes were recorded with the Zeiss Cirrus HD OCT 4000 device: each OCT volume comprises 128 B-scans with a 1024 x 512 pixels resolution. There are annotated 1136 OCT B-scans from 24 patients suffering from late neovascular AMD. An ophthalmologist did all annotations. There are annotations for four boundaries between the layers: internal limiting membrane – ILM, retinal pigment epithelium (RPE), the boundary between the inner plexiform layer and inner nuclear layer (IPL/INL), and Bruch's membrane (BM). Also, fluids are annotated: pigment epithelial detachment (PED), subretinal fluid (SRF), and intraretinal fluid (IRF). Images are prepared for semantic segmentation with eight classes (Figure 1).

Data collection adhered to the tenets of the Declaration of Helsinki and the standards of Good Scientific Practice of Sestre milosrdnice University Hospital Center (Zagreb, Croatia). All patients signed informed consent, and the data were anonymized. The Ethics Committee of the Sestre milosrdnice University Hospital Center approved the presented study.

We opt for this dataset as it is publicly available, and annotations are provided for both layers and fluids. Also, results for human variability are given. Furthermore, images of patients suffering from neovascular AMD (in some cases simultaneously from geographic atrophy) are highly challenging for segmentation due to severe pathological changes. Commercial software for automatic segmentation associated with OCT devices usually works well for healthy persons but with significant errors in pathological cases.

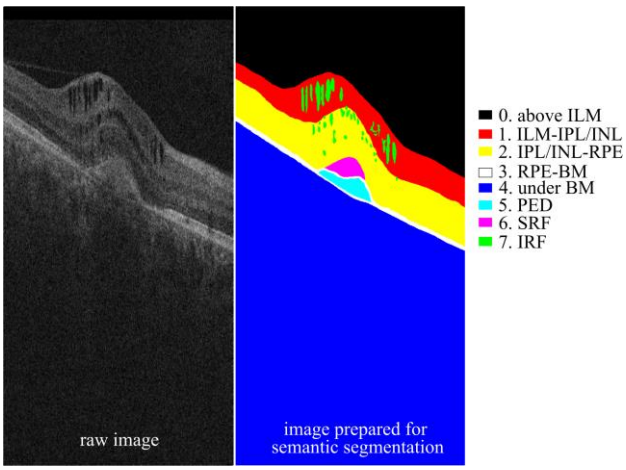

Figure 1. Example of an image from the AROI dataset [31].

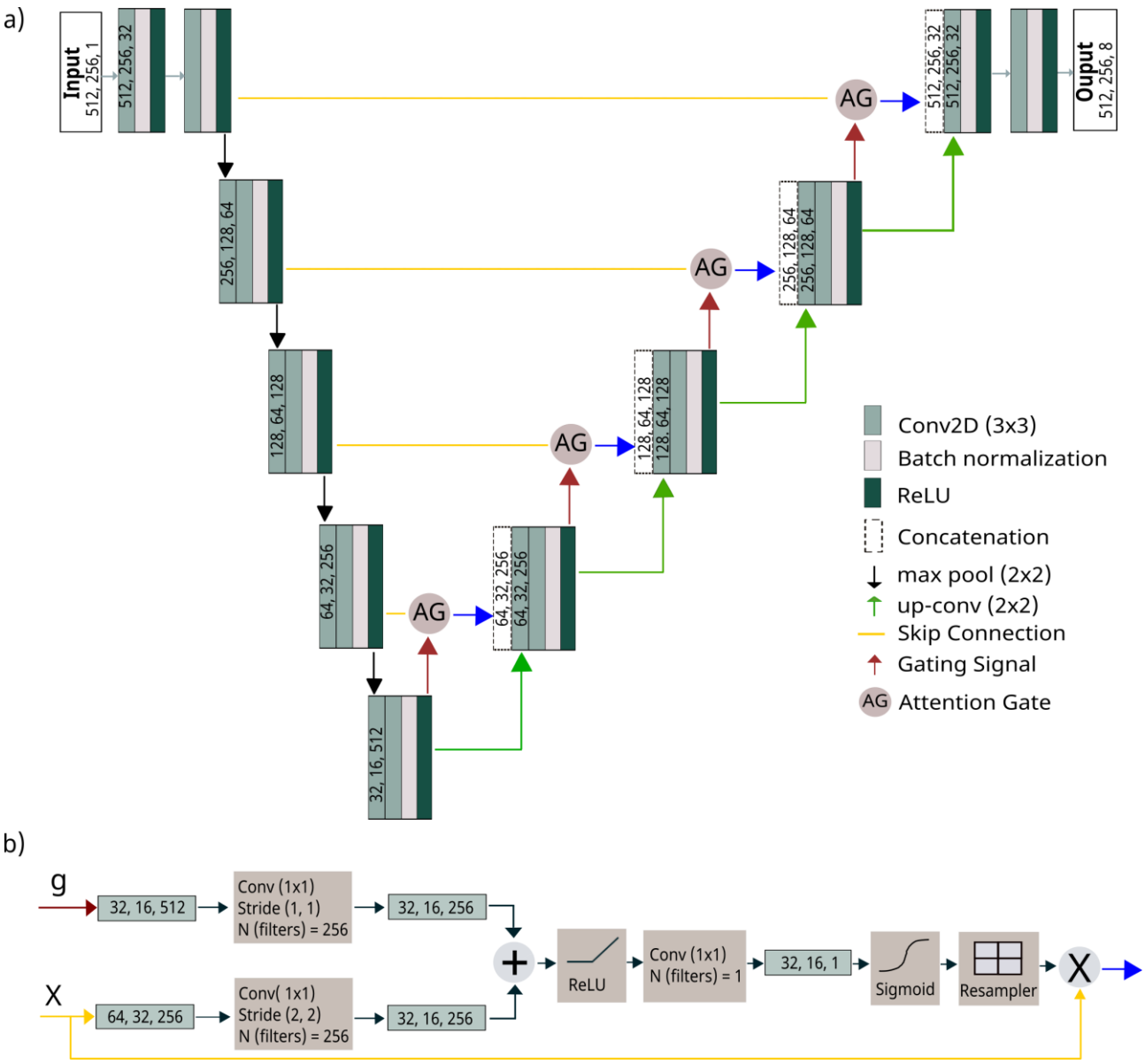

Figure 2. a) Proposed network architecture: unlike in standard U-net architecture, there is an Attention Gate instead of a direct skip connection. b) Attention Gate in which signal X from skip connection and a gating signal G from the next lowest layer of the network is merged.

B. *Proposed network architecture*

A paper from Oktay et al. [15] primarily inspires the architecture we use, in which they introduce the attention mechanism for pancreas segmentation. Standard U-net consists of an encoder, decoder, and skip connections. Each block in the encoder consists of convolutions, activation function (ReLU), and Batch normalization [32]. In addition, there is max-pooling between two blocks, which reduces the image size by half. Similarly, in the decoder, there is upsampling, which doubles the size of the image. Finally, skip connections combine information from the encoder with more spatial information and the decoder with more locally relevant features. The downside of standard U-net architecture is that there are plenty of filters from initial layers without much information.

Attention added to U-net architecture is a way to highlight relevant activations during the training, and in that way, it should lead to better generalization. There are two types of attention: hard and soft attention. Hard attention usually highlights relevant features by cropping. Therefore, it is not differentiable and cannot be used with backpropagation. Contrary, with soft attention, the relevant parts of images get larger weights and vice versa. It is differentiable, and therefore it can be used with backpropagation. With training, attention is increasingly focused on regions of interest (ROI).

We use soft attention to keep all advantages of backpropagation, or more precisely, soft attention is added to skip connections to restrain activations from irrelevant regions. Inputs to the Attention Gate are signals G and X, as shown in Figure 2. G is the gating signal. It comes from the deeper part of the network and has better feature representations. On the other hand, X is a signal from a skip connection. As it comes from the previous layer, it has better spatial information than G. Attention Gate combines X and G. We use stride equals two for X to reduce the image size to be able to sum both signals. By

summing them, aligned weights get larger and unaligned weights are suppressed. After the ReLU activation function, weights range from zero to infinity, so they should be constrained with the sigmoid function, which brings them back between zero and one. Then, upsampling is done to get the original size of X. Finally, X and weights are multiplied element-wise; in that way, vector X is rescaled based on relevance.

C. Training the model

We have used the same parameters for training as the previously published paper [33] to enable better comparison: the categorical cross-entropy loss was used to train the model, the batch size was set to 4, and the AdaBound optimizer [34] was used. Images were resized to 512x256 pixels (i.e., half of the original size). Early stopping was used to prevent overfitting. Also, data augmentation was applied (horizontal flipping and rotating in a small range of angles [-8º, 8º]). As the previously published paper recommended, we used K-fold cross-validation with K equals 6.

The model was trained on Google Colab [35] with a GPU. The model was implemented in Python, using the Keras library with the TensorFlow backend.

IV. Results

We use a Dice score as an evaluation metric as it is the most common evaluation metric for semantic segmentation. In Table 1 are reported Dice scores (mean and standard deviation) for each class and the inter- and intra-observer errors. Also, the prediction errors from published results [33] are reported for the standard U-net model, U-net-like model, and U-net++ model. U-net-like model [17] has encoder-decoder architecture and is enhanced so that each block in the encoder and decoder has residual blocks inspired by ResNet [5], but it lacks skip connections. U-net++ architecture (a nested U-net architecture for medical image segmentation) [16] is inspired by DenseNet architecture [36]. Therefore, dense blocks and convolution layers exist between the encoder and decoder instead of direct skip connections.

It could be noted that the Dice scores for Attention-based U-net are higher in the case of all three fluids. Also, the Dice scores are higher for the area between the IPL/INL (inner plexiform layer and inner nuclear layer) and RPE (retinal pigment epithelium) and the area between RPE and BM (Bruch's membrane). Significantly, segmentations of IRF and SRF are diagnostically most important, and Dice scores are much improved for those classes. The Dice scores for segmenting IRF achieved by the standard U-net model and the Attention-based U-net model are 0.480 (±0.252) and 0.563 (±0.146), respectively. Similarly, the Dice scores for segmenting SRF achieved by the standard U-net model and the Attention-based U-net model are 0.513 (±0.294) and 0.600 (±0.276), respectively.

Figure 3 shows images in case of no severe pathological changes, in case of moderate pathological changes, and a case of extreme pathological changes. In the first case, it could be observed that an Attention-based U-net is the only model which detects SRF. Other models cannot determine SRF when it occupies a small region. In the second case, it could be noticed that Attention-based U-net segmentation prediction is better than predictions from all other models. Better segmentation prediction is even more visible in the third case of extreme pathological changes (all three fluids are present).

V. Discussion

From the overall results, it could be observed that Dice scores are still lower for any model predictions than for inter-observer variability. However, it is questionable how appropriate this comparison is as ophthalmologists never perform manual segmentation as part of clinical practice. Also, there is no consensus among the ophthalmologist on which segmentation accuracy we need. Probably it depends on the segmentation purpose, whether it is just detecting some pathological changes or a case where fluid volume is guidance for therapy.

Also, it could be seen that higher Dice scores are accomplished with standard U-net for highly represented classes and are background from a diagnostical point of view (e.g., the area under BM and the area above ILM). Conversely, lower Dice scores achieved with Attention-based U-net for these classes result from paying more attention to regions of higher interest (e.g., IRF, SRF, PED, RPE-BM, IPL/INL-RPE). As mentioned in previous work [33], a considerable class imbalance makes segmentation extremely challenging, and adding an attention mechanism has improved model performance.

TABLE I. The Dice score (mean and standard deviation) in the inter-observer case, the intra-observer case, for the standard U-net model, the U-net-like model, the U-net++ model, and the Attention-based U-net model.

	Above ILM	**ILM - IPL/INL**	**IPL/INL - RPE**	**RPE - BM**	**Under BM**	**PED**	**SRF**	**IRF**
Inter-observer [33]	0.982 (0.072)	0.950 (0.111)	0.948 (0.112)	0.699 (0.129)	0.989 (0.114)	0.860 (0.301)	0.876 (0.366)	0.735 (0.280)
Intra-observer [33]	0.998 (0.003)	0.973 (0.008)	0.970 (0.117)	0.778 (0.092)	0.998 (0.001)	0.912 (0.242)	0.924 (0.331)	0.844 (0.140)
Standard U-net [33]	**0.995** (0.011)	**0.950** (0.028)	0.923 (0.083)	0.669 (0.129)	**0.988** (0.016)	0.638 (0.173)	0.513 (0.294)	0.480 (0.252)
U-net-like [33]	0.995 (0.004)	0.899 (0.040)	0.890 (0.066)	0.476 (0.132)	0.988 (0.014)	0.533 (0.139)	0.365 (0.291)	0.040 (0.061)
U-net++ [33]	0.992 (0.011)	0.944 (0.032)	0.924 (0.064)	0.641 (0.133)	0.986 (0.017)	0.622 (0.159)	0.465 (0.297)	0.432 (0.265)
Attention-based U-net	0.991 (0.021)	0.945 (0.032)	**0.932** (0.051)	**0.682** (0.126)	0.985 (0.021)	**0.674** (0.147)	**0.600** (0.276)	**0.563** (0.146)

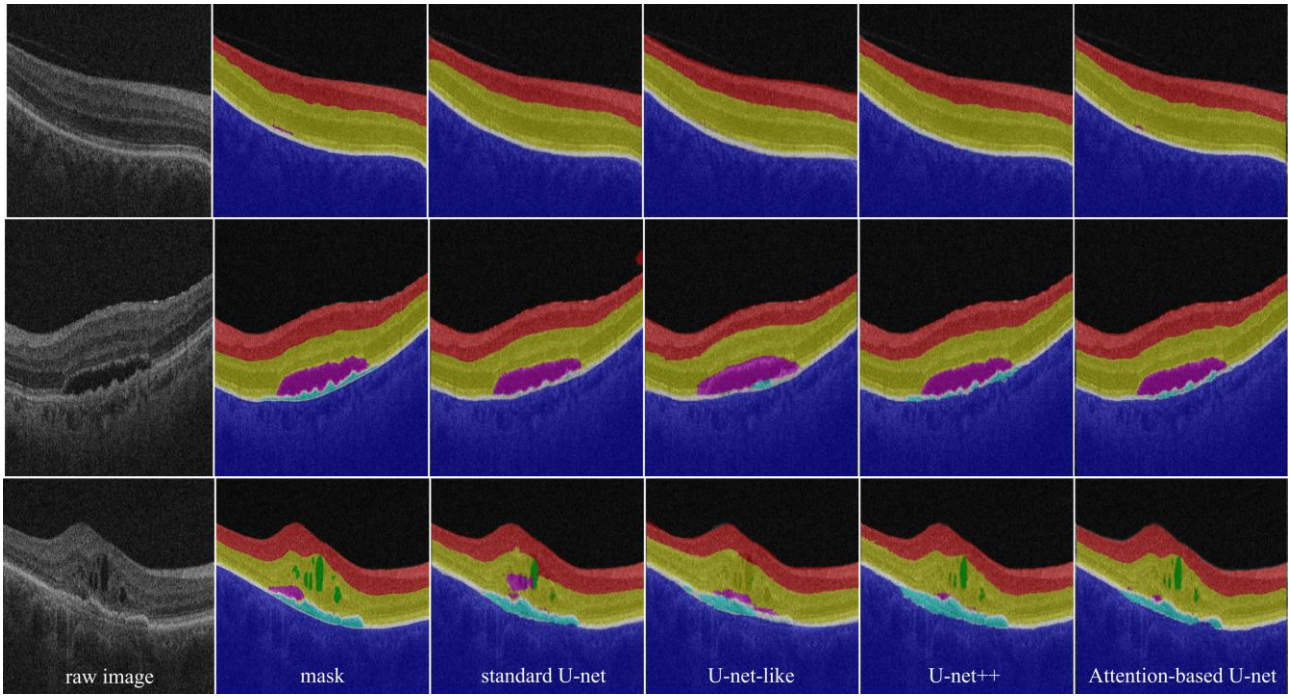

Figure 3. From left to right: raw image, mask (ground truth), predictions from the standard U-net model, U-net-like model, U-net++ model, and Attention-based U-net model. First raw: a case of minor pathological changes (only SRF is present). Second raw: a case of moderate pathological changes (SRF and PED are present). Third raw: a case of severe pathological changes (all fluids are present). All images are cropped to ROI.

By careful inspection of segmentation predictions, it could be spotted that some of the errors present in models without attention remain and probably could not be solved with attention mechanisms. It could be surmised that dominant errors are consequences of a lack of understanding of the retinal structure and pathological traits and hence could not be solved only based on attention mechanisms.

VI. Conclusion

This paper proposes a novel model for the automatic segmentation of biomarkers in retinal OCT images, which combines the attention mechanism with a convolutional neural network. By evaluating results on the publicly available dataset, it could be inferred that the attention mechanisms lead to improved segmentation predictions. Notably, errors due to class imbalance are decreased and lead to better fluid segmentation, which is of prime importance.

Although images in the AROI dataset are highly challenging for segmenting due to extreme pathological deviations in retinal structure, it remains to examine how attention mechanisms improve generalization on unseen images. The lack of extrapolation to unseen data is a significant deficiency of deep learning. The medical field is even more critical due to the small data regime.

This research shows that despite significant progress in the field, we still lack some fundamental concepts to overcome shortcomings. Moreover, even combining the most advanced methods, segmentation predictions are still deficient in some cases of severe pathological alterations. Nevertheless, in our opinion, much progress can be made by active learning and by introducing OCT devices with accompanying automatic segmentation models in clinical practice to help ophthalmologists partially and simultaneously enhance model performance via active learning.

Data and code availability

The AROI dataset is publicly available [31].

Code is available on GitHub:
https://github.com/mmelinscak/OCT-images-segmentation

References

[1] A. Mishra, A. Wong, K. Bizheva, and D. A. Clausi, "Intra-retinal layer segmentation in optical coherence tomography images," *Opt. Express*, vol. 17, no. 26, pp. 23719–23728, Dec. 2009, doi: 10.1364/OE.17.023719.

[2] "ImageNet." http://www.image-net.org/ (accessed Sep. 14, 2020).

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, Accessed: Jul. 07, 2021. [Online]. Available: http://arxiv.org/abs/1409.1556

[4] C. Szegedy *et al.*, "Going Deeper With Convolutions," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9. Accessed: Aug. 23, 2022. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, Accessed: Sep. 14, 2020. [Online]. Available: https://arxiv.org/abs/1512.03385v1

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[7] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,"

arXiv:1704.04861 [cs], Apr. 2017, Accessed: Jul. 07, 2021. [Online]. Available: http://arxiv.org/abs/1704.04861

[8] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, May 2019, pp. 6105–6114. Accessed: Aug. 23, 2022. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[9] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv:1505.04597 [cs]*, May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597

[11] J. Wang *et al.*, "S-D Net: Joint Segmentation and Diagnosis Revealing the Diagnostic Significance of Using Entire RNFL Thickness in Glaucoma," p. 10.

[12] Z. Gu *et al.*, "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019, doi: 10.1109/TMI.2019.2903562.

[13] J. I. Orlando *et al.*, "U2-Net: A Bayesian U-Net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological OCT scans," *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1441–1445, Apr. 2019, doi: 10.1109/ISBI.2019.8759581.

[14] W. Liu, Y. Sun, and Q. Ji, "MDAN-UNet: Multi-Scale and Dual Attention Enhanced Nested U-Net Architecture for Segmentation of Optical Coherence Tomography Images," *Algorithms*, vol. 13, no. 3, p. 60, Mar. 2020, doi: 10.3390/a13030060.

[15] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas." arXiv, May 20, 2018. doi: 10.48550/arXiv.1804.03999.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," *arXiv:1807.10165 [cs, eess, stat]*, Jul. 2018, Accessed: Mar. 30, 2020. [Online]. Available: http://arxiv.org/abs/1807.10165

[17] K. Team, "Keras documentation: Image segmentation with a U-Net-like architecture." https://keras.io/examples/vision/oxford_pets_image_segmentation/ (accessed Sep. 14, 2020).

[18] M. Wang *et al.*, "MsTGANet: Automatic Drusen Segmentation From Retinal OCT Images," *IEEE Trans. Med. Imaging*, vol. 41, no. 2, pp. 394–406, Feb. 2022, doi: 10.1109/TMI.2021.3112716.

[19] D. Kermany, K. Zhang, and M. Goldbaum, "Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images," vol. 3, Jun. 2018, doi: 10.17632/rscbjbr9sj.3.

[20] Z. Jiang *et al.*, "Computer-aided diagnosis of retinopathy based on vision transformer," *J. Innov. Opt. Health Sci.*, vol. 15, no. 02, p. 2250009, Mar. 2022, doi: 10.1142/S1793545822500092.

[21] P. P. Srinivasan *et al.*, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomed Opt Express*, vol. 5, no. 10, pp. 3568–3577, Sep. 2014, doi: 10.1364/BOE.5.003568.

[22] G. Cao, S. Zhang, H. Mao, Y. Wu, D. Wang, and C. Dai, "A single-step regression method based on transformer for retinal layer segmentation," *Phys. Med. Biol.*, vol. 67, no. 14, p. 145008, Jul. 2022, doi: 10.1088/1361-6560/ac799a.

[23] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, and S. Farsiu, "Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema," *Biomed Opt Express*, vol. 6, no. 4, pp. 1172–1194, Mar. 2015, doi: 10.1364/BOE.6.001172.

[24] C. Playout, R. Duval, M. C. Boucher, and F. Cheriet, "Focused Attention in Transformers for interpretable classification of retinal images," *Medical Image Analysis*, vol. 82, p. 102608, Nov. 2022, doi: 10.1016/j.media.2022.102608.

[25] X. Liu, S. Wang, Y. Zhang, D. Liu, and W. Hu, "Automatic fluid segmentation in retinal optical coherence tomography images using attention based deep learning," *Neurocomputing*, vol. 452, pp. 576–591, Sep. 2021, doi: 10.1016/j.neucom.2020.07.143.

[26] G. Lazaridis, M. Xu, S. S. Afgeh, and D. Garway-Heath, "Bio-Inspired Attentive Segmentation of Retinal OCT imaging," p. 10.

[27] X. Mao *et al.*, "Deep Learning with Skip Connection Attention for Choroid Layer Segmentation in OCT Images," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Jul. 2020, pp. 1641–1645. doi: 10.1109/EMBC44109.2020.9175631.

[28] X. Xi, X. Meng, Z. Qin, X. Nie, Y. Yin, and X. Chen, "IA-net: informative attention convolutional neural network for choroidal neovascularization segmentation in OCT images," *Biomed. Opt. Express*, vol. 11, no. 11, p. 6122, Nov. 2020, doi: 10.1364/BOE.400816.

[29] Z. Fu *et al.*, "MPG-Net: Multi-Prediction Guided Network for Segmentation of Retinal Layers in OCT Images," *arXiv:2009.13634 [cs, eess]*, Sep. 2020, Accessed: Apr. 06, 2022. [Online]. Available: http://arxiv.org/abs/2009.13634

[30] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10s, p. 200:1-200:41, Rujan 2022, doi: 10.1145/3505244.

[31] M. Melinščak, M. Radmilović, Z. Vatavuk, and S. Lončarić, "AROI: Annotated Retinal OCT Images database," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Sep. 2021, p. accepted for publication.

[32] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," p. 11.

[33] M. Melinščak, M. Radmilović, Z. Vatavuk, and S. Lončarić, "Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation," *Automatika*, vol. 62, no. 3–4, pp. 375–385, Oct. 2021, doi: 10.1080/00051144.2021.1973298.

[34] L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive Gradient Methods with Dynamic Bound of Learning Rate," *arXiv:1902.09843 [cs, stat]*, Feb. 2019, Accessed: Feb. 04, 2021. [Online]. Available: http://arxiv.org/abs/1902.09843

[35] "Google Colaboratory." https://colab.research.google.com/notebooks/intro.ipynb (accessed May 17, 2020).

[36] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *arXiv:1608.06993 [cs]*, Jan. 2018, Accessed: Jul. 19, 2020. [Online]. Available: http://arxiv.org/abs/1608.06993