NerveFormer: A Cross-Sample Aggregation Network for Corneal Nerve Segmentation

Jiayu Chen[1], Lei Mou[2], Shaodong Ma[2], Huazhu Fu, Lijun Guo, Yalin Zheng, Jiong Zhang[2], ✉, Yitian Zhao[2],✉

[1] School of Ningbo University
[2] Cixi Institute of BioMedical Engineering,
Chinese Academy of Sciences, Ningbo, China
zhangjiong@nimte.ac.cn

Abstract. The segmentation of corneal nerves in corneal confocal microscopy (CCM) is of great importance to the quantification of clinical parameters in the diagnosis of eye-related diseases and systematic diseases. Existing works mainly use convolutional neural networks to improve the segmentation accuracy, while further improvement is needed to mitigate the nerve discontinuity and noise interference. In this paper, we propose a novel corneal nerve segmentation network, named NerveFormer, to resolve the above-mentioned limitations. The proposed NerveFormer includes a Deformable and External Attention Module (DEAM), which exploits the Transformer-based Deformable Attention (TDA) and External Attention (TEA) mechanisms. TDA is introduced to explore the local internal nerve features in a single CCM, while TEA is proposed to model global external nerve features across different CCM images. Specifically, to efficiently fuse the internal and external nerve features, TDA obtains the *query* set required by TEA, thereby strengthening the characterization ability of TEA. Therefore, the proposed model aggregates the learned features from both single-sample and cross-sample, allowing for better extraction of corneal nerve features across the whole dataset. Experimental results on two public CCM datasets show that our proposed method achieves state-of-the-art performance, especially in terms of segmentation continuity and noise discrimination.
Keywords: Corneal nerve segmentation · Transformer · Cross-sample

1 Introduction

The morphological characteristics of the corneal subbasal nerves, such as length, density, tortuosity [20, 13, 17], are closely related to many ocular or systemic diseases. In clinical practice, corneal nerve images acquired by corneal confocal microscopy (CCM), are commonly used to assist ophthalmologists with rich pathological information, for studying disease-related alternations. Manual annotation of nerve fiber is able to support the quantitative analysis, however, it is time-consuming and subjective. Moreover, the inter- and intra-observer variations introduced by manual labeling greatly diminish the accuracy of quantitative assessment of corneal nerves [21]. Therefore, a fully automated and accurate corneal nerve segmentation method is essentially needed.

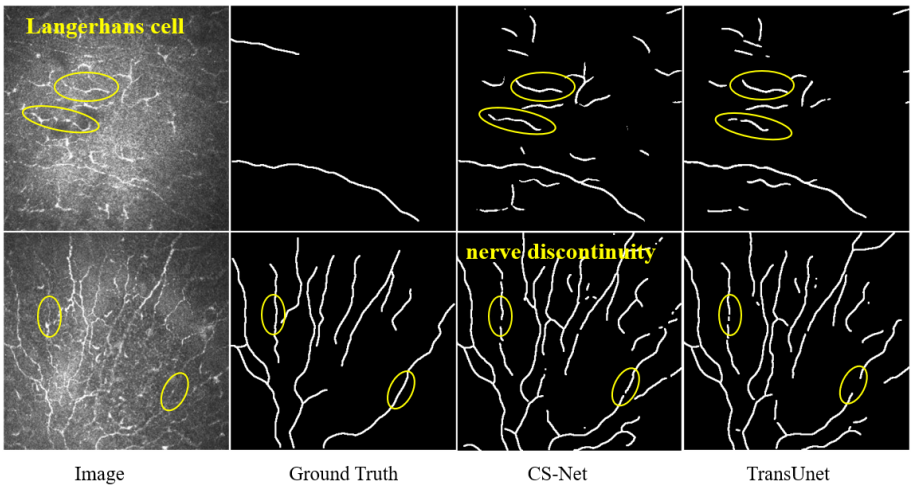

Fig. 1. Typical corneal nerves segmentation issues in CCM images. Top row: interference of Langerhans cells; Bottom row: nerve discontinuity. The automated results were obtained by two recent methods (CS-Net [10] and TransUnet [2]).

Many studies [1, 3, 4, 10, 16, 18, 19] have been conducted to automatically segment the corneal nerve fibers. Early works mainly focus on designing various enhancement filters, such as Gabor filters [4] and Gaussian filters [1] to obtain geometric features for better nerve fiber extraction. However, these methods rely heavily on low-level features and require parameter tuning by hand, and thus segmentation performance is limited. Deep learning-based models have recently been established to improve segmentation accuracy, by exploring high-level features rather than relying on low-level ones. For example, Colonna *et al.* [3] utilized U-Net [12] for end-to-end corneal nerve fibres tracking. Mou *et al.* [11] introduced a channel and spatial dual attention network (CS^2-Net) to improve the segmentation performance of nerve fibers.

Although the above-mentioned deep learning methods have achieved promising segmentation performances, they are still deficient in tackling the interference of background artifacts, e.g., Langerhans cells (LC) [14, 18], as shown in the top row of Fig. 1 - the LCs are often falsely identified as nerve fibers. In addition, many methods [10, 2] failed to preserve continuities of the nerve fibers especially in low-quality CCM images, as shown in the bottom row of Fig. 1. Recently, Transformer [15] has been widely applied in medical image, with its powerful global modeling capabilities. For example, the ability of the network to extract local and global information can be enhanced by embedding the transformer module in the CNN framework [2, 5]. MCTrans [9] proposed to use transformer-self-attention and transformer-cross-attention to enhance feature representation of the network.

Inspired by the above approaches, we consider to establish a transformer-based network by incorporating potentially valuable attention modules to learn more informative corneal nerve features for better segmentation performance. In this paper, a novel corneal nerve segmentation model named NerveFormer is introduced. It specifically targets at reducing the interferences caused by background artifacts in CCM images, and enhancing the continuity of nerve fiber segmentation. The main contributions are summarized as follows:

- A new corneal nerve segmentation method is proposed with a transformer-based deformable and external attention module (DEAM), which not only learns internal nerve features from single CCM image, but also learns shared nerve properties across multi-CCM images.
- A transformer-based external attention (TEA) is introduced to further fuse the internal features extracted by transformer-based deformable attention (TDA) with the shared properties of the external CCM images, allowing our method to obtain the most discriminative nerve features to alleviate the background artifacts interferences, and thus improves the ability to preserve the nerve continuity.

2 Proposed Method

The proposed method consists of a pre-trained feature extractor (i.e. encoder), a deformable and external attention module (DEAM), and a decoder. The architecture is shown in Fig. 2. We employed a ResNet34 model pre-trained on ImageNet as encoder. DEAM is a fusion of transformer-based deformable attention (TDA) and external attention (TEA), Where N represents the number of DEAM modules. The decoder is designed to recover the dimensions of the feature map layer by layer.

2.1 CNN encoder

Given an input CCM image with height H and width W. We can obtain feature maps $\mathbf{F}_l \in \mathbb{R}^{C \times \frac{H}{2^{l+1}} \times \frac{W}{2^{l+1}}}$ of the l^{th} layer by the encoder, where C represents the number of channels and $l \in \{1, 2, 3, 4\}$. To feed the extracted features into DEAM, we first flatten the features in the last three layers of the encoder into one-dimensional (1D) sequences and map them to the same channel dimension, respectively. Later, all these three sequences were concatenated into one sequence to extract the multi-scale information of the corneal nerves. To recover the lost spatial information caused by the flattening operation, we compute the position coordinates of each dimension using a sine and cosine functions of different frequencies [15], which is subsequently summed pixel-by-pixel with the multi-scale features to construct the input of the DEAM.

2.2 Deformable and external attention module (DEAM)

The proposed DEAM consists of two major components: TDA and TEA.

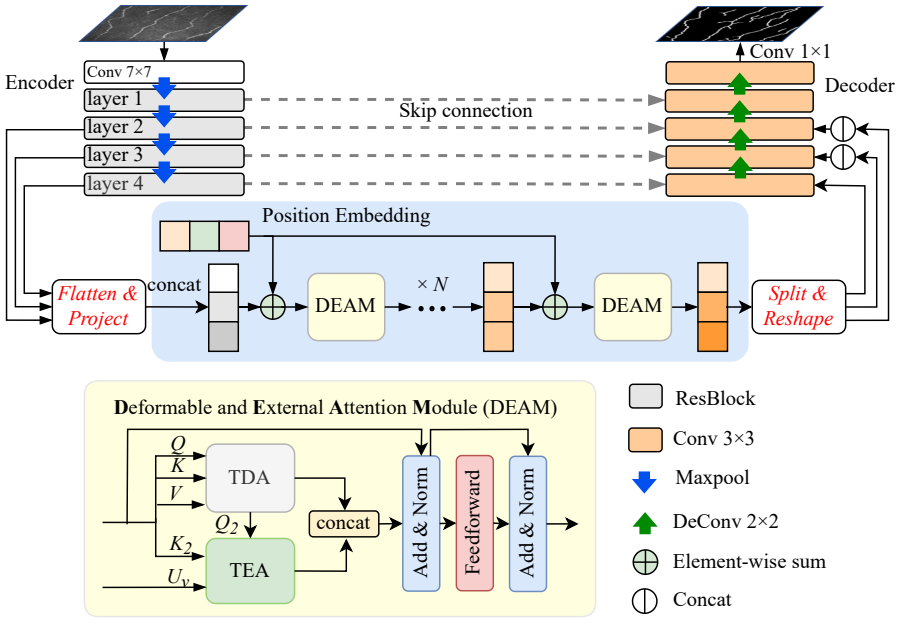

Fig. 2. Diagram of the proposed NerveFormer.It contains three main components: a pre-trained encoder, a deformable and external attention module (DEAM), and a CNN decoder, where the fused attention of transformer-based deformable attention (TDA) and external attention (TEA) constitute the key components of DEAM.

TDA: Inspired by deformable attention [22], we motivate our model to focus only on a small fraction of key sampling points around the reference point, without considering the spatial size of the feature map, thus alleviating the challenge of large feature resolution.

The TDA can be formulated as:

$$D\left(z_q, p_q, x\right) = \sum_{h=1}^{H} W_h \left[\sum_{k=1}^{K} A_{hqk} \cdot W_h' x\left(p_q + \Delta p_{hqk}\right)\right], \tag{1}$$

where $x \in \mathbb{R}^{C \times H \times W}$ is the input feature map, z_q and p_q are the content feature and reference point of the q^{th} query element, respectively. W_h is the encoding of the key element and A_{hqk} is the weight of the kth key element. K is the number of sampled key elements, and Δp_{hqk} is the sampling offset of the k^{th} sampling point in the h^{th} head. The TDA enables the proposed DEAM to learn more crucial information in a single CCM image.

TEA: Self-attention in a typical transformer is a linear combination of self-values to refine the input features. However, self-attention in this style only considers the relationship between elements (i.e., nerve fiber pixels) in a single

CCM image, failing to explore the shared characteristics of the same elements across different CCM images. For a small number of CCM images containing Langerhans cells, we can further explore the shared characteristics of nerve fibers in different CCM images to better discriminate corneal nerves from Langerhans cells and background artifacts. Therefore, we introduce an external attention (EA) mechanism [8] to address the limitation of self-attention. The EA is defined by $\mathrm{EA} = (\alpha)_{i,j} = \mathrm{Norm}(QU_k^{\mathbf{T}}))U_v$, where U_k and U_v are two learnable external memory units. $(\alpha)_{i,j}$ is the similarity between i^{th} pixel and j^{th} row of U_k, U_v. $Q \in \mathbb{R}^{S \times d}$ indicates the set of $query$, where S is the number of elements, d is the number of feature dimensions. The EA is specifically designed to learn shared characteristics across images, while in our case, we aim to establish a more complete representation module by embedding both of the internal and external feature information of corneal nerves. Thus, we propose a transformer-based external attention (TEA) module by taking TDA as its query input. Comparing with original external attention, our TEA can learn the fused features from a single CCM image itself and external CCM images, which are beneficial to our model for learning the shared characteristics of multiple corneal nerves and also the discriminative features against background artifacts (e.g., Langerhans cells). In addition, multi-head attention [15] is employed to improve the capacity of single head attention by capturing different relations between tokens. Finally, we build up a multi-head transformer-based external attention (TEA) in our model, which can be written as:

$$\mathrm{TEA}_{multi\_head} = \sum_{h}^{H} \mathrm{Norm}\left(Q_{2h} K_{2h}^{\mathbf{T}}\right) U_v, \tag{2}$$

where H is the number of heads, Q_2 and K_2 are the output features of TDA and output features of encoder, respectively.

2.3 CNN decoder

The 1D feature map output from DEAM is split and reshaped into three two-dimensional (2D) feature maps, based on the dimensions of the l^{th} $(l = 2, 3, 4)$ layer in the encoder. In the decoder, we gradually upsample the feature maps to the input resolution using a five-layer CNN module, where each layer consists of a deconvolutional layer and a 3×3 convolutional layer. In addition, a skip connection is added to the corresponding layer between the encoder and the decoder, to retain more low-level and spatial information. Finally, a 1×1 convolutional layer is applied to the features to generate the segmentation map. Mean square error (MSE) loss and Dice coefficient (DC) loss are employed to compute the error between segmentation map and ground truth in a 4:6 ratio.

Table 1. Comparisons of results between different methods.

Methods	CCM-1				CCM-2			
	SEN	FDR	DICE	AUC	SEN	FDR	DICE	AUC
U-Net [12]	0.8425	0.2444	0.7943	0.9165	0.8100	0.2071	0.7965	0.9012
CE-Net [6]	0.8584	0.2114	0.8174	0.9225	0.8390	0.1993	0.8171	0.9159
CS-Net [10]	0.8532	0.1869	0.8294	0.9210	0.8363	0.1940	0.8183	0.9147
MDACN [18]	0.8486	0.1847	0.8282	0.9188	0.8144	0.2061	0.7952	0.9033
TransUnet [2]	0.8578	0.1878	0.8317	0.9232	0.8278	0.1929	0.8148	0.9103
MCTrans [9]	0.8600	0.1860	0.8325	0.9242	0.8395	0.1891	0.8230	0.9164
UTNet [5]	0.8559	0.1827	0.8325	0.9224	0.8263	0.1889	0.8162	0.9098
NerveFormer	**0.8738**	**0.1813**	**0.8432**	**0.9314**	**0.8541**	**0.1864**	**0.8317**	**0.9236**

3 Experiments

3.1 Datasets and implementation details

CCM-1 and CCM-2 are two subsets of a public dataset CORN-1 [10]. CORN-1 includes a total of 1698 CCM images of the corneal basal nerves, of which 1578 are available for CCM-1 and 120 for CCM-2. These images were acquired by using Heidelberg Retina Tomography equipment with a Rostock Cornea Module (HRT-III) microscope. Each image has a resolution of 384×384 pixels, and with 1×1 pixel centerline annotation, which was traced by an ophthalmologist using the open source software ImageJ[1]. In the CCM-1 and CCM-2 datasets, the ratio of training set, validation set, and test set is set as 3:1:1. Our method was implemented in PyTorch framework and all the experiments are run with two NVIDIA GPUs (Tesla V100). The adaptive moment estimation (Adam) with an initial learning rate of 0.0003 served as the optimizer, and the cosine annealing strategy is introduced to update the learning rate. The batch sizes were set as 32 and 8 for training models on CCM-1 and CCM-2, respectively. Random rotations in the range of $[-30°, 30°]$ and random horizontal and vertical flips were set with a probability of 0.5 for data augmentation.

3.2 Comparison with the state-of-the-art methods

To demonstrate the superiority of the proposed NerveFormer, we employed several state-of-the-art methods for the comparison. CNN-based methods: U-Net [12], CE-Net [6], CS-Net [10], MDACN [18]; Transformer-based methods: TransUnet [2], MCTrans [9] and UTNet [5]. For fair comparison, all these methods are trained and validated on CCM-1 and CCM-2 with the same data split as ours. Fig. 3 qualitatively illustrates corneal nerve segmentation results from challenging cases. By observing the results in the first two rows, we can conclude that the proposed method obtains more complete segmentations with better nerve fiber continuity than CS-Net [10] and TransUnet [2], as indicated by the red arrows. Moreover, we can see that our method presents better immunization

[1] https://imagej.nih.gov/ij/

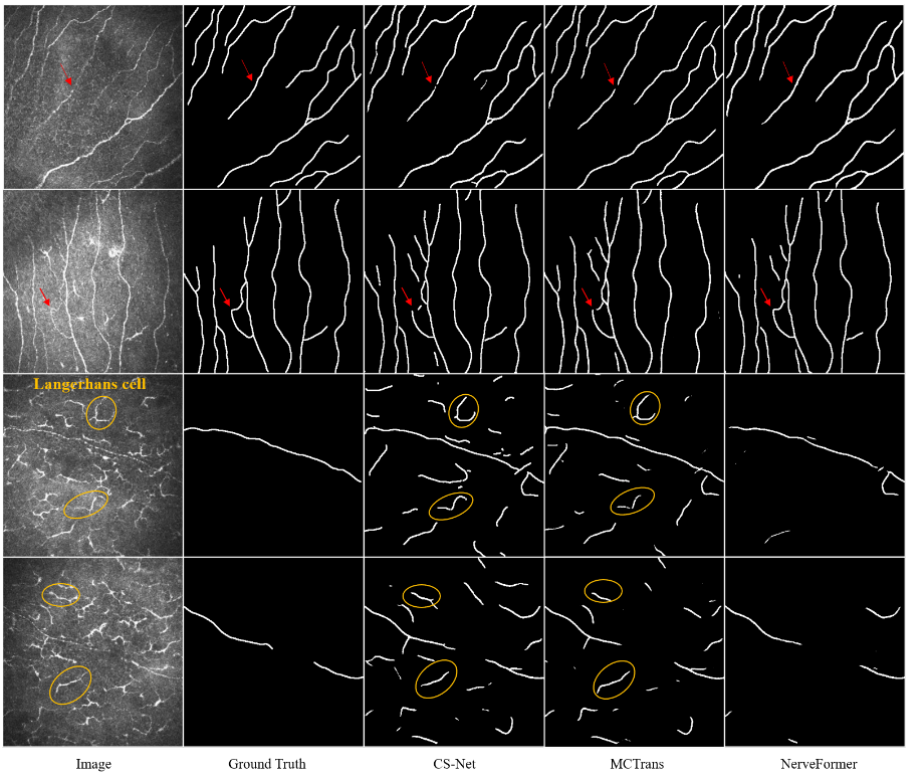

Fig. 3. Automated nerve segmentation results in CCM images with low-contrast, and heavy background artifacts (e.g. Langerhans cells).

against background artifacts, as demonstrated in the last two rows in Fig. 3. Specifically, as illustrated in the third row, we can observe that both CS-Net and TransUNet incorrectly identify the curve-like Langerhans cells as corneal nerves, while the proposed NerveFormer effectively distinguishes between them. The proposed method also enables better extraction of corneal nerves in regions covered by background noise, especially in the case of tiny corneal nerves. The above analyses indicate that the proposed network effectively explores the internal and external features in different CCM images, and thus can improve the performance of the corneal nerve segmentation in challenging CCM images.

To facilitate better objective performance evaluation of our NerveFormer, we calculated the following metrics: *sensitivity* (SEN) = TP / (TP + FN), *false discovery rate* (FDR) = FP / (FP + TP), *dice* coefficent (DICE) = 2 × TP / (2 × TP + FP + FN), , where TP denotes true positive, FN denotes false negative, and FP denotes false positive and area under the ROC curve (AUC). We follow [7] to compute the evaluation metrics in terms of a three-pixel tolerance region around the ground truth centerline. Based on quantitative

Table 2. Ablation study results of the proposed model on the CCM-1 and CCM-2 datasets.

Methods	CCM-1				CCM-2			
	SEN	FDR	DICE	AUC	SEN	FDR	DICE	AUC
Backbone	0.8446	0.2127	0.8128	0.9158	0.8359	0.2008	0.8151	0.9142
Backbone+TDA	0.8598	0.1895	0.8325	0.9243	0.8509	0.1960	0.8252	0.9219
Backbone+TEA	0.8534	0.1844	0.8315	0.9211	0.8487	0.1928	0.8256	0.9209
Backbone+DEAM	**0.8738**	**0.1813**	**0.8432**	**0.9314**	**0.8541**	**0.1864**	**0.8317**	**0.9236**

segmentation comparisons in Table 1, we can clearly observe that the proposed method outperforms the other state-of-the-art methods on both CCM-1 and CCM-2 datasets. Specifically, the proposed NerveFormer achieves the highest sensitivity and lowest FDR compared with the other methods. In detail, for CCM-1, our method is 1.6% higher and 2.59% lower than MCTrans [9] in terms of SEN and FDR, respectively, and 1.73% higher and 1.44% lower than MCTrans [9] for CCM-2, respectively. These objective quantitative analyses suggest that the proposed NerveFormer can effectively detect more corneal nerve pixels and at the same time reducing the falsely detected background artifacts, which is consistent with the observation in Fig. 3 that our method is more capable of discriminating between corneal nerves and Langerhans cells.

3.3 Ablation Study

In order to investigate the contributions of TDA and TEA in the proposed NerveFormer, we conduct the following ablation studies. The encoder-decoder architecture with the pre-trained ResNet34 are taken as the Backbone. The TDA and TEA modules are gradually added into the Backbone, namely Backbone+TDA, Backbone+TEA and Backbone+DEAM, to assess their contributions to corneal nerve segmentation. Table 2 illustrates the contributions of different modules to the segmentation performance on CCM-1 and CCM-2. Compared to the Backbone, the network with only TDA achieves better performance, with an increase of approximate 1.80%, 10.91%, 2.42% and 0.93% in SEN, FDR, DICE and AUC on CCM-1, respectively, and an increase of approximate 2.40%, 2.4%, 1.24%, 0.84% on CCM-2, respectively. By integrating TEA into the Backbone, we can observe an improvement of approximate 1.04%, 13.31%, 2.30% and 0.58% in SEN, FDR, DICE and AUC on CCM-1, respectively, and an improvement of approximate 1.53%, 4.00%, 1.29%, 0.73% on CCM-2, respectively. Furthermore, we found that the Backbone+TEA obtains a lower FDR than Backbone+TDA, indicating that TEA with constraints of external samples can be more robust to the interference of background cells. Finally, the proposed NerveFormer that incorporates both TDA and TEA (Backbone+DEAM) achieves the best performance on all the metrics.

3.4 Conclusion

In this paper, we have proposed a cross-sample aggregation network, i.e., NerveFormer, for corneal nerve segmentation. The proposed NerveFormer can effectively learn the internal and external corneal nerve features by integrating a proposed deformable and external attention module. Experimental results on the two publicly available corneal nerve datasets demonstrate that our method achieves state-of-the-art performance by enhancing segmentation continuity and suppressing background artifacts interference.

Acknowledgement. This work was supported in part by the National Science Foundation Program of China (62103398 and 61906181),Zhejiang Provincial Natural Science Foundation of China (LR22F020008), in part by the Youth Innovation Promotion Association CAS (2021298), in part by the Ningbo major science and technology task project (2021Z054) and in part by the AME Programmatic Fund (A20H4b0141).

References

1. Annunziata, R., Kheirkhah, A., Hamrah, P., Trucco, E.: Scale and curvature invariant ridge detector for tortuous and fragmented structures. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 588–595. Springer (2015)
2. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Colonna, A., Scarpa, F., Ruggeri, A.: Segmentation of corneal nerves using a u-net-based convolutional neural network. In: Computational Pathology and Ophthalmic Medical Image Analysis, pp. 185–192. Springer (2018)
4. Dabbah, M.A., Graham, J., Petropoulos, I., Tavakoli, M., Malik, R.A.: Dual-model automatic detection of nerve-fibres in corneal confocal microscopy images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 300–307. Springer (2010)
5. Gao, Y., Zhou, M., Metaxas, D.N.: Utnet: a hybrid transformer architecture for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 61–71. Springer (2021)
6. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging **38**(10), 2281–2292 (2019)
7. Guimarães, P., Wigdahl, J., Poletti, E., Ruggeri, A.: A fully-automatic fast segmentation of the sub-basal layer nerves in corneal images. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5422–5425. IEEE (2014)
8. Guo, M.H., Liu, Z.N., Mu, T.J., Hu, S.M.: Beyond self-attention: External attention using two linear layers for visual tasks. arXiv preprint arXiv:2105.02358 (2021)
9. Ji, Y., Zhang, R., Wang, H., Li, Z., Wu, L., Zhang, S., Luo, P.: Multi-compound transformer for accurate biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 326–336. Springer (2021)

10. Mou, L., Zhao, Y., Chen, L., Cheng, J., Gu, Z., Hao, H., Qi, H., Zheng, Y., Frangi, A., Liu, J.: Cs-net: channel and spatial attention network for curvilinear structure segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 721–730. Springer (2019)
11. Mou, L., Zhao, Y., Fu, H., Liu, Y., Cheng, J., Zheng, Y., Su, P., Yang, J., Chen, L., Frangi, A.F., et al.: Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging. Medical image analysis **67**, 101874 (2021)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
13. Su, P., Chen, T., Xie, J., Zheng, Y., Qi, H., Borroni, D., Zhao, Y., Liu, J.: Corneal nerve tortuosity grading via ordered weighted averaging-based feature extraction. Medical Physics **47**(10), 4983–4996 (2020)
14. Su, P.Y., Hu, F.R., Chen, Y.M., Han, J.H., Chen, W.L.: Dendritiform cells found in central cornea by in-vivo confocal microscopy in a patient with mixed bacterial keratitis. Ocular immunology and inflammation **14**(4), 241–244 (2006)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
16. Wei, S., Shi, F., Wang, Y., Chou, Y., Li, X.: A deep learning model for automated sub-basal corneal nerve segmentation and evaluation using in vivo confocal microscopy. Translational Vision Science & Technology **9**(2), 32–32 (2020)
17. Williams, B.M., Borroni, D., Liu, R., Zhao, Y., Zhang, J., Lim, J., Ma, B., Romano, V., Qi, H., Ferdousi, M., et al.: An artificial intelligence-based deep learning algorithm for the diagnosis of diabetic neuropathy using corneal confocal microscopy: a development and validation study. Diabetologia **63**(2), 419–430 (2020)
18. Yang, C., Zhou, X., Zhu, W., Xiang, D., Chen, Z., Yuan, J., Chen, X., Shi, F.: Multi-discriminator adversarial convolutional network for nerve fiber segmentation in confocal corneal microscopy images. IEEE Journal of Biomedical and Health Informatics (2021)
19. Zhang, D., Huang, F., Khansari, M., Berendschot, T.T., Xu, X., Dashtbozorg, B., Sun, Y., Zhang, J., Tan, T.: Automatic corneal nerve fiber segmentation and geometric biomarker quantification. The European Physical Journal Plus **135**(2), 266 (2020)
20. Zhao, Y., Zhang, J., Pereira, E., Zheng, Y., Su, P., Xie, J., Zhao, Y., Shi, Y., Qi, H., Liu, J., et al.: Automated tortuosity analysis of nerve fibers in corneal confocal microscopy. IEEE transactions on medical imaging **39**(9), 2725–2737 (2020)
21. Zhao, Y., Zheng, Y., Zhao, Y., Liu, Y., Chen, Z., Liu, P., Liu, J.: Uniqueness-driven saliency analysis for automated lesion detection with applications to retinal diseases. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–118. Springer (2018)
22. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)