DeepGrading: Deep Learning Grading of Corneal Nerve Tortuosity

Lei Mou, Hong Qi, Yonghuai Liu, Yalin Zheng, Peter Matthew, Pan Su, Jiang Liu, Jiong Zhang, Yitian Zhao

***Abstract*— Accurate estimation and quantification of the corneal nerve fiber tortuosity in corneal confocal microscopy (CCM) is of great importance for disease understanding and clinical decision-making. However, the grading of corneal nerve tortuosity remains a great challenge due to the lack of agreements on the definition and quantification of tortuosity. In this paper, we propose a fully automated deep learning method that performs image-level tortuosity grading of corneal nerves, which is based on CCM images and segmented corneal nerves to further improve the grading accuracy with interpretability principles. The proposed method consists of two stages: 1) A pre-trained feature extraction backbone over ImageNet is fine-tuned with a proposed novel bilinear attention (BA) module for the prediction of the regions of interest (ROIs) and coarse grading of the image. The BA module enhances the ability of the network to model long-range dependencies and global contexts of nerve fibers by capturing second-order statistics of high-level features. 2) An auxiliary tortuosity grading network (AuxNet) is proposed to obtain an auxiliary grading over the identified ROIs, enabling the coarse and additional gradings to be finally fused together for more accurate final results. The experimental results show that our method surpasses existing methods in tortuosity grading, and achieves an overall accuracy of 85.64% in four-level classification. We also validate it over a clinical dataset, and the statistical analysis demonstrates a significant difference of tortuosity levels between healthy control and diabetes group. We have released a dataset with 1500 CCM images and their manual annotations of four tortuosity levels for public access. The code is available at: `https://github.com/iMED-Lab/TortuosityGrading`.**

***Index Terms*— Corneal confocal microscopy, corneal nerve, tortuosity grading, interpretability, deep learning.**

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China (LR22F020008, LZ19F010001), in part by the Youth Innovation Promotion Association CAS (2021298), in part by the National Science Foundation Program of China (61906181 and 62103398), in part by the Ningbo 2025 S&T Megaprojects (2019B10033 and Grant 2019B10061). (Lei Mou and Hong Qi contributed equally to this work.) (Corresponding author: Yitian Zhao, yitian.zhao@nimte.ac.cn)

L. Mou, J. Zhang and Y. Zhao are with the Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China; J. Zhang and Y. Zhao are also with The Affiliated Ningbo Eye Hospital of WenZhou Medical University, Ningbo, China; Y. Liu and P. Matthew are with Department of Computer Science, Edge Hill University, Ormskirk, UK; Y. Zheng is with Department of Eye and Vision Science, University of Liverpool, Liverpool, UK; J. Liu is with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China; P. Su is with the School of Control and Computer Engineering, North China Electric Power University, Baoding, China; H. Qi is with the Department of Ophthalmology, Peking University Third Hospital, Beijing, China; L. Mou is also with the University of Chinese Academy of Sciences, Beijing, China.

I. Introduction

Corneal confocal microscopy (CCM) is a non-invasive corneal imaging technique that is widely used clinically, especially for assessing the sub-basal nerve plexus [1]. Existing studies [2]–[4] suggest that morphological changes of the corneal nerves, such as nerve fiber branching, density, tortuosity, length and so on, are closely related to a variety of ocular and systemic diseases. Among these changes, tortuosity is the most widely used criterion to characterize diversifications of the corneal nerve fibers. Several clinical studies have shown that corneal nerve tortuosity is a potential and valuable biomarker for further pathological analysis of systemic or ocular diseases such as diabetic neuropathy (DN), dry eye disease, unilateral herpes zoster, herpes simplex keratitis, acute acanthamoeba and fungal keratitis [5]–[14]. To assess the relationship between corneal nerve tortuosities in normal and diseased images, tortuosity has been usually divided into 3 to 5 grades in different studies [1], [2], [5], [15], [16]. In this paper, we grade corneal nerves into four tortuosity levels, as in [5], [16], since the four-level tortuosity highlights significant variation in the corneal nerve fibers of patients with diabetic neuropathy [8], [17] and dry eye disease [9], [11]. For example, a study by Kallinikos *et al.* [8] demonstrate significant differences in corneal nerve tortuosity between four clinical groups in diabetic patients with neuropathy, i.e. nerve fiber tortuosity is significantly greater in the severe neuropathy group than in the control, mild and moderate neuropathy patient groups. Klisser *et al.* [17] also demonstrate a significant reduction in corneal nerve tortuosity in individuals with both T1DM (type 1) and T2DM (type 2) relative to controls, as well as a significant difference in tortuosity between diabetic participants with neuropathy compared to those without. In addition, Ma *et al.* [11] conclude from a detailed statistical analysis that increased corneal nerve tortuosity is linked to frequencies of ocular discomfort, visual function disturbance, and tear film instability in dry eye disease. Fig. 1 illustrates

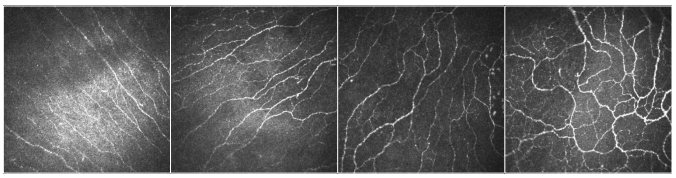

Fig. 1. Examples of CCM images with four different tortuosity levels, from left to right: levels 1 to 4.

four CCM images, in which the images are classified into four levels of corneal nerve tortuosity [2]: almost straight (*level 1*), mild tortuous (*level 2*), quite tortuous *(level 3*), and heavily tortuous (*level 4*).

Currently, tortuosity is mainly measured by analyzing the pixels of the corneal nerve structure extracted from a CCM image, which represents the anatomical structure of the corneal nerves. The measurement of tortuosity has been studied extensively [18]–[22]. As there is no generally accepted definition or standard description for the tortuosity of curvilinear structures [16], various curvilinear structure tortuosity measures have been proposed: length-based [23], [24], angle-based [19], [21], [25] and curvature-based [16], [22], [26]. Clinical parameters derived after tedious manual fiber tracing are subjective, and tedious to retrieve. Moreover, grading corneal nerve tortuosity by means of these subjective methods may lead to substantial inter-and intra-observer variation and human error [2], [25]. In addition, manual screening for corneal nerve tortuosity estimation is time consuming and thus a serious obstacle to the introduction of large-scale screening procedures.

Therefore, a number of automated curvilinear structure tortuosity assessment methods have been proposed to reduce the errors of manual tracing, and to increase speed of diagnosis. Conventionally, vessel tortuosity calculation requires several steps, including preprocessing, segmentation, vessel network splitting and curvature calculation. Considering that the errors introduced in each processing step in such pipelines may accumulate and cause information loss, Bekkers *et al.* [26] propose for the first time a tortuosity measure method based on the theory of best-fit exponential curves in the roto-translation group SE(2), which does not rely on vascular pre-segmentation but act directly on the retinal image, effectively advancing the workflow of the downstream tortuosity measurement. Grisan *et al.* [23] divide retinal vessels into segments with constant-sign curvature and then combine each evaluation of these segments to complete the automated evaluation of tortuosity. To address the limited repeatability and inter-observer agreement due to subjective and visual assessment, Roberto *et al.* [2] propose a fully automated framework for image-level tortuosity estimation. It consists of a hybrid segmentation method and a highly adaptable tortuosity estimation algorithm, with a feature selection strategy that attempts to identify the most discriminative nerve fibers from the CCM image for the representation of image-level tortuosity. Inspired by [26], Zhao *et al.* [16] propose to classify tortuosity directly from pixel-wise curvature measurements obtained through analysis of exponential curves in a 3D position-orientation space that is derived from enhanced CCM images, rather than measuring tortuosity based on a pre-tracked error-prone corneal nerve fiber map. These automated methods usually first track the corneal nerves, then compute tortuosity of individual nerve fibers, and finally perform a weighted fusion of the obtained measurements for the representation of image-level tortuosity. However, CCM images contain corneal nerve fibres with different numbers, lengths and distributions, and therefore it is problematic to use the tortuosity measurements of individual nerve fibres for the representation of the degree of tortuosity of the whole image.

Neural networks have greatly improved performance in the past decade for the tasks of image classification [27], [28], object detection [29], segmentation [30]–[33] and visual question answering (VQA) [34], etc. in the field of computer vision and image processing. Although neural network-based methods have achieved performance improvements in many tasks, how neural networks exactly work are often a mystery, i.e., it is difficult to explain their underlying mechanisms and behavior, leading to the resistance in their adoption for their applications. To address this limitation, visualization-based interpretability methods [35]–[38] have been developed to explain how neural networks arrive at decisions from a human cognition perspective. For example, [39] and [40] attempt to visualize the different convolutional neural network (CNN) layers to improve the interpretability of the model outputs and predictions. Moreover, Liao *et al.* [41] propose to integrate the object activation map with the location information into the CNN to enhance classification ability of their glaucoma detection model. In [42], CAMs [35] are further binarized and applied over original medical images to generate masks of regions of interest (ROIs): the generated temporary ROIs are then used synchronously with the original images as input for further training.

Most of the aforementioned tortuosity grading methods rely on handcrafted filters to extract tortuosity representations. These methods have the advantage of being highly interpretable, but are not well adapted to the assessment of large amounts of nerve tortuosity with variable morphology. With the development of deep learning techniques, learning-based methods can be well adapted to large-scale image analysis. However, in the corneal nerve image tortuosity grading the differences between adjacent levels is relatively subtle: the nerves are comprised of hard-to-discriminate, fine-grained features. Therefore, models directly transferred from existing deep learning-based classification methods cannot obtain the desired accuracy in corneal nerve tortuosity grading, being less capable of extracting fine-grained, more discriminative features at different levels.

To address these limitations, we propose a novel two-stage corneal nerve tortuosity grading network, DeepGrading, based on a CNN embedded within a visually interpretable mechanism. It first predicts the coarse grading and the class activation regions of the image, using a pre-trained network refined through a proposed bilinear attention (BA) module. In contrast to existing attention modules, the proposed BA module attempts to model the interdependencies and global context of features in different channels by bilinear pooling and residual learning. Secondly, a novel auxiliary network (AuxNet) is proposed for the prediction of the additional grading over the identified activation regions of interest in the automatic corneal curve segmentation map [43], leading the global coarse and local independent gradings to be integrated for more accurate final grading results. More specifically, this paper makes the following four main contributions:

- We propose a novel framework to automatically learn the tortuosity features of corneal nerve fibers in CCM images. A data-driven neural network-based model has the advantage

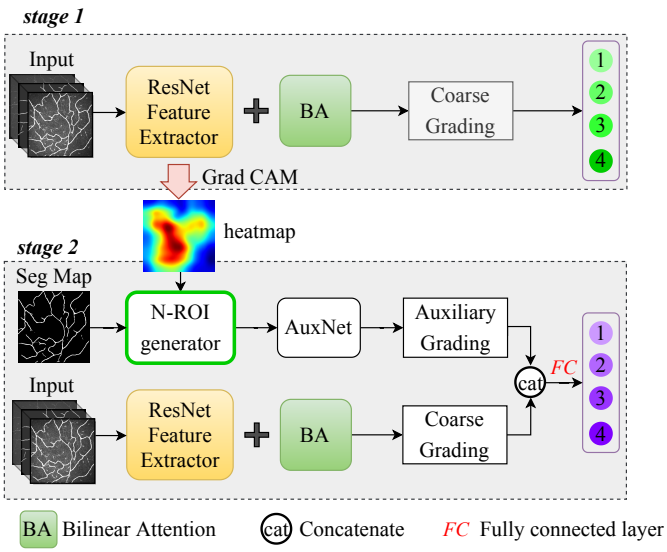

Fig. 2. Schematic diagram of DeepGrading. The green and purple circles indicate tortuosity levels graded by the two different stages, respectively. It is important to note that *stage 1* provides input for *stage 2*. Even though the same backbone network has been run in stage 1 and stage 2 respectively, they have different purposes: the former outputs N-ROIs, while the latter outputs the coarse grading.

of adaptively learning curvature-related features and mining more discriminative semantic features, rather than depending on manually-designed filters, as in conventional methods.

- To better extract fine-grained varied features between corneal nerves, we propose a novel BA module that can model the global context and *long-range* dependencies of nerve fibers by capturing second-order statistics of high-level tortuosity features. Moreover, this proposed BA module can be embedded as an independent unit in other classification models.
- To enhance the interpretability of the grading model and further improve the grading performance, we propose an AuxNet for feature re-extraction of nerve fiber segments within tortuosity activation regions. The proposed method not only obtains the decisive nerve fiber segments in the image, but also effectively narrows down the tortuosity feature mining range, and improves the grading performance from an interpretability perspective.
- The proposed method has been trained and evaluated on a newly-constructed corneal nerve tortuosity dataset, and the trained model has also been further tested over another publicly available database and a clinical dataset about healthy control and diabetic group. To encourage reproducible research on the topic, we have released this new dataset for public access[1].

II. Proposed Method

In this section, we describe the proposed DeepGrading framework in detail. Overall, the proposed framework consists of two stages, as shown in Fig. 2. The first stage (***stage 1***) is a bilinear attention network (BANet) that contains a fine-tuned backbone, followed by a bilinear attention module for the prediction of the global coarse grading and the activation regions of interest (ROIs) of a given CCM image. In the second stage (***stage 2***), we propose an auxiliary network (AuxNet) to grade the activated nerve fibers in the identified ROIs (N-ROI). We finally perform refined tortuosity grading by integrating the global and local auxiliary predictions of BANet and AuxNet, respectively.

A. Architecture of BANet

Existing classification methods based on learning strategies can achieve promising performance in many applications [27], [44], [45], but one of their major limitations is that they require large amounts of data for training. Direct training of tortuosity level grading models based on a relatively small number of CCM images may lead to a high risk of over-fitting. To speed up the training and enable faster convergence and thus improve the learning efficiency, we chose a pre-trained ResNet18 [27] as the backbone, and fine-tune it with CCM images and their segmentation maps. Grading nerve fibers of varying tortuosity in CCM images is a fine-grained recognition task. In view of this, inspired by Bilinear CNN [46], we propose a bilinear attention (BA) module to extract discriminative features of nerve fibers at different tortuosity levels. Fig. 3 shows its architecture.

In designing the fine-tuned BANet, we replace the original classification (fully-connected) layer of ResNet18 with the proposed BA module, shown as the green box in Fig. 3, where l_i (i=1, 2, 3, and 4) denotes the ith layer of ResNet18. To facilitate the description of BA, we denote the output feature of l_4 as F, where $F \in \mathbb{R}^{C\times W\times H}$, and C, W, and H represent the number of channels, the width and the height of F, respectively. In the proposed BA, we first reshape F into $\mathbb{R}^{C\times N}$, where $N = W \times H$. The number of channels in a neural network represents the high-level features extracted by the intermediate hidden layers, and the final classification is determined by these high-level features. Theoretically, the higher the number of the channels, the more feature dimensions the neural network contains and the more likely they include redundant information. As in [37], some channels are visualized to demonstrate that some of them attend to specific classes. However, there are also some channels that fail to provide relevant features for the classification but could not be sufficiently visualized. Therefore, it is necessary to design a channel attention model that applies the correlations between the extracted features to enhance the relevant ones and weaken the irrelevant ones, which leads more reliable and expressive features to be finally extracted for the fine-grained tortuosity classification. Then, we obtain the bilinear features of the corneal nerves by performing the matrix outer product of F and its transpose F^{T}, i.e.

$$\mathcal{B}_c = F \times F^{\mathrm{T}}. \tag{1}$$

where $\mathcal{B}_c \in \mathbb{R}^{C\times C}$ indicates the channel-related bilinear features of the corneal nerves. Further, a bilinear vector of dimension C^2 is obtained by reshaping $\mathbf{B}_c$ so that it can be used by the classification function (i.e., the fully connected layer). The obtained bilinear vector is then subjected to an element-wise sign operation sign$(\cdot)$ followed by the square root and

[1] https://doi.org/10.5281/zenodo.5880419

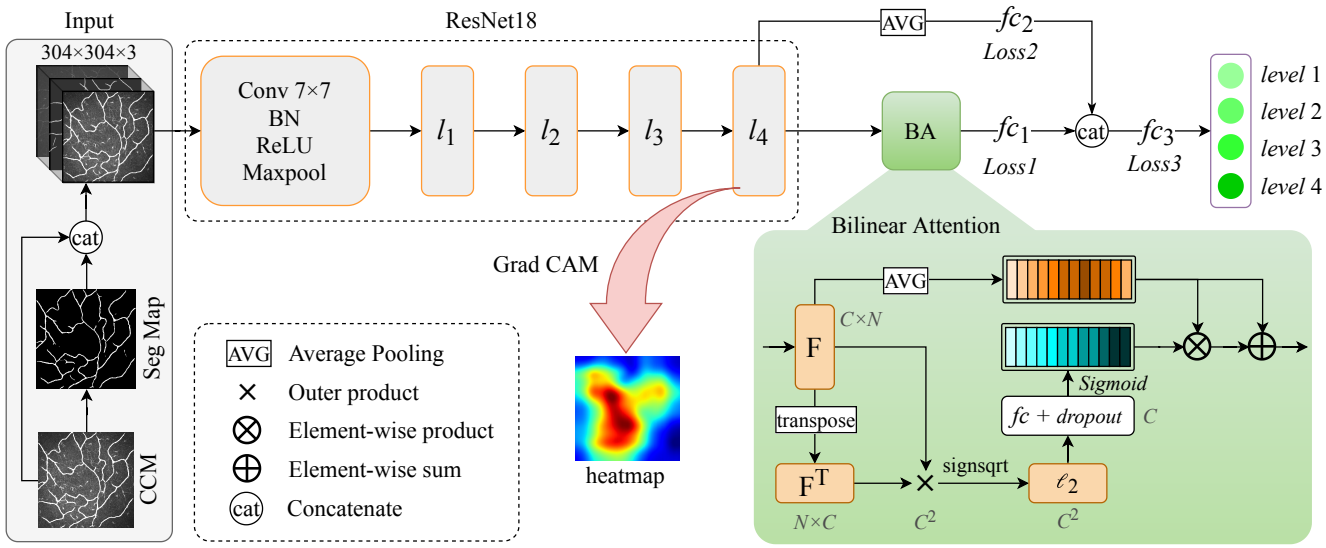

Fig. 3. Schematic diagram of the BANet (***stage 1***). The concatenation of the CCM, Seg Map and then the CCM again is used as the input to the network. l_i $(i = 1, 2, 3, 4)$ indicates the ith residual layer of ResNet18. fc, fc_1, fc_2 and fc_3 represent the fully connected layers. Note that we can obtain the tortuosity activation regions based on the trained BANet and Grad CAM [36] at this stage.

ℓ_2 normalization to enhance the feature expression [47]:

$$\hat{\mathcal{B}}_c = \frac{\text{sign}(\mathcal{B}_c)\sqrt{|\mathcal{B}_c|}}{\left\|\text{sign}(\mathcal{B}_c)\sqrt{|\mathcal{B}_c|}\right\|_2}, \tag{2}$$

where $\hat{\mathcal{B}}_c \in \mathbb{R}^{C^2}$ is the enhanced bilinear vector. We then feed $\hat{\mathcal{B}}_c$ into a fully connected ($\mathcal{FC}$) and a dropout (ψ) layers for dimensionality compression, where ψ is used to mitigate over-fitting, and the $\mathcal{FC}$ is used to compress the number of channels of $\hat{\mathcal{B}}_c$ to C. The compressed bilinear vector is finally normalized by a $Sigmoid$ function to generate the channel weights ω_{ba}:

$$\omega_{ba} = \frac{1}{1 + \mathrm{e}^{-\psi(\mathcal{FC}(\hat{\mathcal{B}}_c))}}, \tag{3}$$

where $\omega_{ba} \in \mathbb{R}^C$. In addition, we feed F into a global average pooling (GAP) layer to obtain a feature vector with the same dimension as that of ω_{ba}, denoted as ω_{avg}. We can then obtain the bilinear attention vector $\mathcal{B}_{att}$ of the corneal nerves by performing residual learning [27] through element-wise product ($\otimes$) and element-wise sum ($\oplus$) sequentially, i.e.:

$$\mathcal{B}_{att} = \omega_{avg} \oplus (\omega_{avg} \otimes \omega_{ba}). \tag{4}$$

Such residual connection allows us to insert our proposed BA module into any classification network of interest.

A fully-connected layer (fc_1 in Fig. 3) is then applied over the BA features $\mathcal{B}_{att}$ to perform a coarse tortuosity grading: i.e., $\Theta^{(1)} = \mathcal{FC}(\mathcal{B}_{att})$, $\Theta^{(1)} \in \mathbb{R}^4$ (4 tortuosity levels). As in the original ResNet18, we pass the output of l_4 through AVG and apply a fully connected layer to generate another coarse grading result $\Theta^{(2)} \in \mathbb{R}^4$, shown as fc_2 in Fig. 3. Two different branches generate two different weighted grading results. To optimize the grading performance, we integrate the two grading results ($\Theta^{(1)}$ and $\Theta^{(2)}$) by concatenation and feed them into another classifier (fc_3 in Fig. 3) for the final tortuosity prediction $\Theta^{(a)}$. Since the last classifier fc_3 is an integration of fc_1 and fc_2, we simultaneously minimize the cross-entropy of the three predicted probabilities and labels to obtain better grading performance.

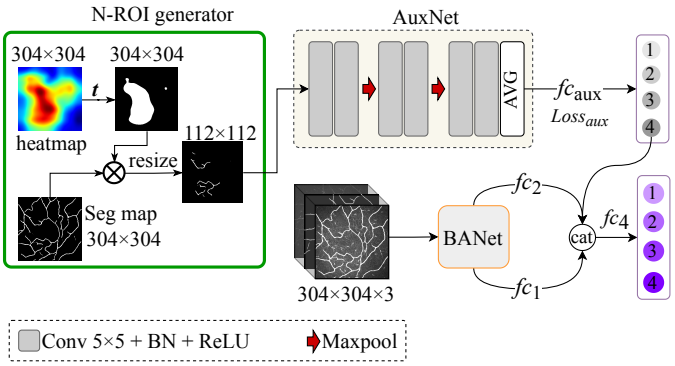

Fig. 4. Schematic diagram of the tortuosity grading framework (***stage 2***), which consists of a backbone classifier and an auxiliary classifier. The circles with numbers indicate predicted tortuosity levels. All the training samples of the auxiliary classifier are binary images.

B. Architecture of stage 2

The high similarity in tortuosity between two adjacent tortuosity levels of the corneal nerves poses a great challenge for their grading. Therefore, we need to find more fine-grained discriminative features. In addition, CCM images with higher tortuosity levels tend to contain fewer tortuous nerve fiber segments as well. Therefore, mining the decisive tortuosity features in an image is a key step to improve the grading performance. To this end, we follow [36] to obtain the gradient-based class activation mapping (Grad CAM) of the last layer (l_4) of BANet to visualize the tortuosity gradient and identify the most likely representative regions of each grade in each CCM image. A Grad CAM heatmap depicts the tortuosity level responses of the CCM image. However, any given activation region generated by Grad CAM will only roughly represent

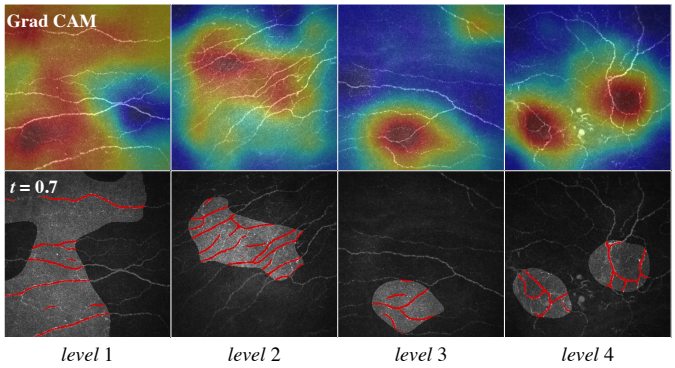

Fig. 5. Four example results of Grad CAM and N-ROI generator. Top row shows the activated gradient of tortuosity generated by Grad CAM heatmap, and the bottom row indicates the activated regions at $\boldsymbol{t} = \mathbf{0.7}$. The red pixels indicate the nerve fibers within the activated regions.

the tortuosity within that region, failing to accurately locate the specific nerve fiber segments at the corresponding tortuosity level. To address this limitation, we propose a generator to obtain nerve fiber segments within the regions of interest, namely N-ROI generator (as shown in the green box in Fig. 4), so that we can more accurately characterize the tortuosity of the image using the activated fiber segments. To achieve this, we obtain a binarized class activation mask by thresholding the normalized Grad CAM heatmap, where the threshold is defined as $\boldsymbol{t}$: we then chose $\boldsymbol{t} = 0.7$ in this paper, which will be justified and explained in Section III-E.3) below. Fig. 5 shows four examples of activated regions generated by Grad CAM, and their corresponding activated fibre segments generated by our proposed N-ROI generator.

Once the activated nerve fiber segments have been identified, we can use them to assist the BANet for further refinement of the grading. This is because the nerve fiber segments within the N-ROI have more fine-grained discriminative features when compared to the whole CCM image, as demonstrated by the statistical comparison in Fig. 8 in the next section. Therefore, we design an auxiliary network (AuxNet) independent of the BANet to grade the N-ROIs, as illustrated in Fig. 4. In the AuxNet, we resize the acquired activated nerve fiber region to 112×112 pixels and treat it as input, denoted as $\mathcal{I}_{aux}$. It is worth noting that we use 112×112 pixels instead of 304×304 pixels as input to reduce the GPU memory consumption and to minimize the damage to the nerve fiber topology caused by over downsampling. Next, in the first step, $\mathcal{I}_{aux}$ passes through two successive convolutional layers with a kernel size of 5, followed by a ReLU activation and a batch normalization layer. Then a max-pooling layer is applied to reduce the feature dimension, and to increase the receptive field of the kernel. The second step takes the output of the previous step as input and repeats the same operations as the previous step. The last step is identical to the previous step, except that the max-pooling layer is not included. The AuxNet then outputs features of dimension 28×28, which are fed into an AVG layer, followed by a fully connected layer (fc_{aux} in Fig. 4). For convenience, we denote the output of the AuxNet as $\Theta^{(3)}$. Thus, we integrate $\Theta^{(i)}$ (i=1, 2, and 3) by concatenation (cat$(\cdot)$) to generate a new feature vector and feed it into a classifier: i.e.,

$$\Theta = \mathcal{FC}\left(\text{cat}\left(\Theta^{(1)}, \Theta^{(2)}, \Theta^{(3)}\right)\right), \tag{5}$$

where Θ is the output of the proposed framework. Compared to the BANet, the final classifier (fc_4 in Fig. 4) of the proposed DeepGrading method integrates the predictions of the AuxNet, and we therefore additionally minimize the cross-entropy of AuxNet before the integration.

C. Loss function

There are two stages in the proposed method: ***stage 1*** and ***stage 2***, and each stage includes a training step and a testing step. We first apply the CS-Net [43] to automatically segment the raw CCM images (denoted as C) and obtain the segmentation maps (denoted as S) of the nerve fibers. Since the pre-trained ResNet18, which serves as the backbone of the proposed method, takes 3-channel images as input (denoted as $\mathbf{I}$), we then concatenate the raw CCM images with the corresponding segmentation masks to generate 3-channel images that can be fed into the pre-trained model: i.e.,

$$\mathbf{I} = \text{cat}\left(C, S, C\right). \tag{6}$$

An inadequate number of CCM images produces a network that is insufficiently complex to characterize all the samples, which drives the network to become over-confident. Inspired by [48], we introduce label smoothing to reduce the risk of over-fitting and to improve the generalization ability of the network. Usually, the final fully connected layer of the neural network outputs a vector z of dimension K, and outputs a probability q_i for i^{th} class of all the K classes after a *Softmax* operator: $q_i = \exp(z_i)/\sum_{j=1}^{K}\exp(z_j)$. It is important to note that the proposed BANet outputs N predictions, where $N = 3$ in the first stage: $\Theta^{(1)}$, $\Theta^{(2)}$ and $\Theta^{(a)}$, and $N = 4$ in the second stage: $\Theta^{(1)}$, $\Theta^{(2)}$, $\Theta^{(3)}$ and Θ. Therefore, we need to minimize the cross-entropy between the nth prediction and the hard label p_i by:

$$\mathcal{L}_{ce} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K} -p_i \log(q_i^n), \tag{7}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss of all the N predictions, q_i^n denotes the probability of the ith class in the nth predictions, and p_i is 1 for the correct class and 0 for the other classes. For a network trained by label smoothing with parameter α, we modify the hard label p_i to a soft label p_i^{ls}, where $p_i^{ls} = p_i(1-\alpha) + \alpha/K$, and we follow [48] to set $\alpha = 0.1$. Hence, the cross entropy between p_i^{ls} and q_i^n can finally be written as:

$$\mathcal{L}_{ls} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K} -\left(p_i(1-\alpha) + \alpha/K\right)\log(q_i^n). \tag{8}$$

In ***stage 2***, we obtain the N-ROIs of all the segmentation maps by performing the processes described in Section II-B. Then we load all the original CCM images, segmentation maps, and N-ROIs to train/validate/test DeepGrading. The same labels are used for computing losses in both ***stage 1*** and ***2***.

III. Experimental Setup and Results

A. Datasets

All the data described in this section have appropriate approvals from the institutional ethics committees, and written informed consent was obtained from each participant in accordance with the Declaration of Helsinki.

• **CORN-3** [16] is a subset of the publicly available CORneal Nerve (CORN) dataset[2], which was constructed by iMED Lab, Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, China. CORN has three subsets, and was designed for corneal nerve segmentation [43], enhancement [49], and tortuosity grading [16] purposes, respectively. CORN-3 contains 403 CCM images of corneal subbasal epithelium from 103 normal and pathological subjects. Each image has a resolution of 384×384 pixels, covering a field of view of $400 \times 400\mu m^2$. CORN-3 also provides reference fiber centerlines, annotated by an ophthalmologist. Two senior ophthalmologists (both with more than 15 years' clinical experience) were invited to independently perform intra-observer (twice) and inter-observer reproducible labeling of tortuosity (*level 1* through *level 4*) based on their clinical experience and previous studies [50], [51], and their consensus were finally selected as the ground truth, where *level 1* through *level 4* contain 54, 212, 10, and 29 images, respectively.

• **CORN**^{1500} is a new corneal nerve dataset which was particularly designed for the tortuosity grading task. We considered CORN-3 to be a relatively small dataset, and the image numbers at the different tortuosity levels are extremely imbalanced. These limitations might easily lead to model overfitting. To this end, the newly-constructed CORN^{1500} dataset aims to expand the datasets available for tortuosity grading. As with CORN-3, all 1500 CCM images in CORN^{1500} were acquired using a Heidelberg Retina Tomograph equipped with a Rostock Cornea Module (HRT-III) microscope. Each image has a resolution of 384×384 pixels covering a field of view of $400 \times 400\mu m^2$. All the images were manually annotated by the same two ophthalmologists as before into four tortuosity levels, with *levels 1* through *4* containing 214, 461, 364, and 461 images, respectively. We have made the CORN^{1500} available[3] online.

• A **clinical dataset** is constructed and used to demonstrate the effectiveness of the proposed method for grading the corneal nerve tortuosity in a clinical setting. This data was collected by the Department of Ophthalmology, Peking University Third Hospital, Beijing, China. A total of 354 CCM images were collected, including 160 and 194 images from healthy and diabetic subjects, respectively. All the CCM images in this dataset were acquired using the same device, and their imaging resolution and field of view are also in line with those of the CORN-3 and CORN^{1500} datasets above.

B. Implementation details

We implemented the proposed DeepGrading on a PyTorch platform with dual NVIDIA GPUs (GeForce GTX Titan Xp). We employed mini-batch stochastic gradient descent (SGD), with a weight decay of 0.0001, to optimize the model during training. To speed up the training and the convergence of the model, we used the pre-trained ResNet18 over ImageNet to initialize its parameters. To fine-tune its weights and biases, we set the initial learning rate before fc layer in the pre-trained ResNet18 to 0.001, while setting the initial learning rate to 0.01 in the proposed bilinear attention module, as well as in the fully connected layers. The cosine annealing learning rate strategy [52] was employed to adjust the learning rate in the training phase. We set the batch size to 64, and the maximum epochs to 200.

We evaluated the network using a 5-fold cross-validation method on the whole CORN^{1500} dataset, where each fold contains 1000, 200 and 300 CCM images for training, validation and testing, respectively. CCM images were divided according to the percentages of images at each *level* in the dataset. Early stopping strategy was used to avoid over-fitting: i.e., we stopped training if the validation accuracy didn't improve beyond 100 consecutive epochs. In ***stage 1***, we trained the BANet and saved the model with the best performance on the validation set of CORN^{1500}. We then used Grad CAM to obtain the activation regions of the corneal nerves. The Grad CAM maps and CCM images were then fed into the DeepGrading network in ***stage 2*** for training.

The preprocessing pipelines for training and testing were different. During training, we applied random image rotation, random brightness and contrast adjustment, random horizontal and vertical flipping, and random cropping to 304×304 pixels. Finally the gray maps were normalized, with a mean of 0.339 and a variance of 0.138, where the mean and variance were computed based on the training set. However, for the N-ROI images, we resized them to a size of 112×112 pixels in order to reduce the GPU resource usage.

During testing, we resized each image and segmentation map to 304×304 pixels, and each N-ROI image to 112×112 pixels. Please note that we need to generate the N-ROIs through the BANet and Grad CAM during testing, and then normalize the gray channel, as in training. We did not perform any random augmentations during testing.

C. Evaluation metrics

In order to quantitatively evaluate the proposed method, we follow [2] to compute the weighted accuracy ($wAcc$), sensitivity (wSe), and specificity (wSp) metrics:

$$wAcc = \sum_{l=1}^{K} r_l \frac{TP_l + TN_l}{TP_l + TN_l + FP_l + FN_l}, \tag{9}$$

$$wSe = \sum_{l=1}^{K} r_l \frac{TP_l}{TP_l + FN_l}, \tag{10}$$

$$wSp = \sum_{l=1}^{K} r_l \frac{TN_l}{TN_l + FP_l}, \tag{11}$$

where TP_l, TN_l, FP_l, and FN_l are the true positives, true negatives, false positives, and false negatives, respectively, for the lth level ($l = 1, 2, \cdots, K$). $K = 4$ denotes the total number of the tortuosity levels. r_l is the percentage of all the available images belonging to the lth level.

[2] https://imed.nimte.ac.cn/data-index.html
[3] https://doi.org/10.5281/zenodo.5880419

TABLE I

PERFORMANCE OF THE BASELINE METHODS AND THE PROPOSED METHOD ON **CORN^{1500}** AT INDIVIDUAL LAYER LEVEL AND OVERALL. ALL THE METRICS ARE EXPRESSED IN PERCENTAGE (%).

Methods	Metrics	*level1*	*level2*	*level3*	*level4*	*overall*	$p <$
Annunziata [2]	$wAcc$	80.27	72.31	69.81	81.30	75.92	.001
	wSe	63.19	51.27	53.75	68.33	59.14	
	wSp	84.37	85.68	70.33	90.57	82.74	
M4 [16]	$wAcc$	78.56	70.24	75.69	88.72	78.30	.001
	wSe	66.12	53.31	61.87	80.43	60.93	
	wSp	80.55	78.31	77.69	88.90	81.36	
VGG16	$wAcc$	87.33	76.67	74.00	87.33	80.82	.001
	wSe	72.09	48.91	61.64	72.83	62.67	
	wSp	89.88	88.94	77.97	93.75	87.88	
ResNet18	$wAcc$	89.33	75.33	79.00	90.33	82.83	.001
	wSe	65.12	55.43	64.38	81.52	67.00	
	wSp	93.39	84.13	83.70	**94.23**	88.45	
ResNet34	$wAcc$	92.33	78.33	73.33	86.67	81.68	.001
	wSe	67.44	61.96	61.64	70.65	65.33	
	wSp	96.50	85.58	77.09	93.75	87.58	
ResNet18-BP	$wAcc$	89.33	79.67	78.00	88.33	83.30	.001
	wSe	**81.40**	59.78	60.27	75.00	67.67	
	wSp	90.66	88.46	83.70	**94.23**	89.39	
ResNet34-BP	$wAcc$	89.33	79.33	77.00	88.33	82.96	.001
	wSe	79.07	52.17	63.01	79.35	67.00	
	wSp	91.05	**91.35**	81.50	92.31	89.20	
BANet18	$wAcc$	89.00	77.67	80.00	**90.67**	83.85	.001
	wSe	76.47	59.78	53.42	85.87	68.67	
	wSp	91.05	85.58	**88.55**	92.79	89.30	
BANet34	$wAcc$	90.67	77.67	76.00	87.67	82.19	.001
	wSe	65.12	55.43	54.79	85.87	66.00	
	wSp	94.94	87.50	82.82	88.46	87.72	
DeepGrading	$wAcc$	**92.33**	**82.00**	**81.67**	89.33	**85.64**	—
	wSe	59.52	**68.82**	**66.67**	**87.10**	**72.67**	
	wSp	**97.67**	87.92	86.40	90.34	**89.67**	

TABLE II

GRADING PERFORMANCE OF STATE-OF-THE-ART METHODS ON **CORN-3** IN TERMS OF ACCURACY, SENSITIVITY AND SPECIFICITY IN PERCENTAGE (%).

Methods	Metrics	*level1*	*level2*	*level3*	*level4*	*overall*	$p <$
Annunziata [2]	$wAcc$	85.80	78.00	75.90	84.30	79.00	.001
	wSe	71.80	64.40	66.00	70.70	66.30	
	wSp	86.70	78.00	75.90	84.30	79.00	
M4 [16]	$wAcc$	88.40	**80.30**	80.00	86.60	82.40	.001
	wSe	74.30	68.00	68.80	73.80	71.70	
	wSp	**90.10**	80.10	81.40	88.10	85.90	
BANet18	$wAcc$	**90.32**	76.18	80.89	96.53	80.80	.001
	wSe	96.30	69.81	58.33	**93.10**	71.96	
	wSp	89.40	83.25	89.15	96.79	86.63	
BANet34	$wAcc$	88.83	74.69	80.15	96.28	79.60	.001
	wSe	96.30	62.74	66.67	86.21	69.98	
	wSp	87.68	87.96	85.08	97.06	87.81	
DeepGrading	$wAcc$	86.85	79.90	**87.10**	**98.51**	**84.10**	—
	wSe	**98.15**	**70.75**	**72.22**	89.66	**76.18**	
	wSp	85.10	**90.05**	**92.54**	**99.20**	**90.71**	

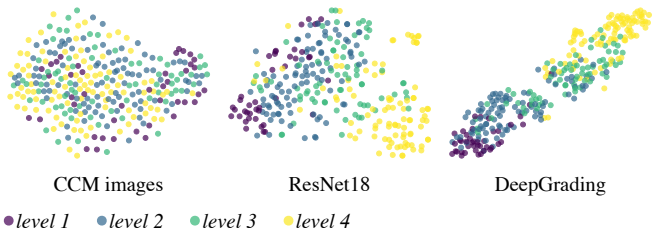

Fig. 6. The t-SNE visualization of the CCM images, ResNet18 and DeepGrading over the test images. The raw images were used as a comparison baseline for the t-SNE [55] visualization, while for ResNet18 and DeepGrading, the feature vectors of their penultimate layers were used instead.

D. Comparison with state-of-the-art methods

To demonstrate the superior performance of the proposed DeepGrading method, we selected several baseline methods for performance comparison on the CORN^{1500} dataset, including Annunziata [2], M4 [16], VGG16 [53], ResNet18, and ResNet34 [27]. Moreover, we embedded the proposed BA module into ResNet18 (denoted BANet18) and ResNet34 (denoted BANet34), in order to verify the performance of the BA module, independently of DeepGrading. In addition, bilinear pooling (BP) [28] was also embedded into ResNet18 (denoted ResNet18-BP) and ResNet34 (denoted ResNet34-BP), respectively, to highlight by contrast the performance of the proposed BA module. Similarly, for all the learning-based comparison methods, we initialized their parameters pre-trained on ImageNet and fine-tuned them to achieve the best performance by following the strategies used in training the proposed method. We fine-tuned ResNet18 and ResNet34 directly based on the official code provided by PyTorch. We implemented in Python ResNet18/34-BP, BANet18/34, and the methods in [2] and [9], and carefully fine-tuned them for the optimal performance according to the implementation details provided in the original papers. To verify the superiority of the proposed method, we computed the performance measures both at an individual level and overall. In addition, we obtained the predicted probabilities of all the methods and expanded the labels and predicted probabilities into the *one-vs-all* manner. Then AUC (area under the ROC curve) statistical significance test was performed using Delong's test [54][4]. All the statistics hereinafter were performed on AUC similarly unless stated otherwise. The test results are presented in Table I. They show that the proposed DeepGrading produces superior grading of the corneal nerve tortuosity with an overall accuracy of 85.64%, as evidenced by all p-values < 0.001.

To fairly compare the proposed method, we reported classification results of Annunziata [2], M4 [16], and the proposed method on Dataset CORN-3. The tortuosity of the corneal nerve in the CORN-3 images was manually divided into four levels, which coincided with the number of levels that can be graded by the proposed method. Accordingly, we used the 1200 images from CORN^{1500} as the training set and the remaining 300 images as the validation set to select the model parameters. Finally, CORN-3 was served as the test set to verify the grading performance of DeepGrading: the results are shown in Table II. These results show that the proposed DeepGrading achieves the best performance on the CORN-3 dataset, although the overall grading accuracy decreases by 1.54%. These results demonstrate that the proposed method offers better generalization capability than the existing methods.

To demonstrate in-depth that the proposed DeepGrading can better extract the tortuosity features of nerve fibers, we use t-SNE [55] to visualize and interpret the high-dimensional

[4] https://www.medcalc.org/

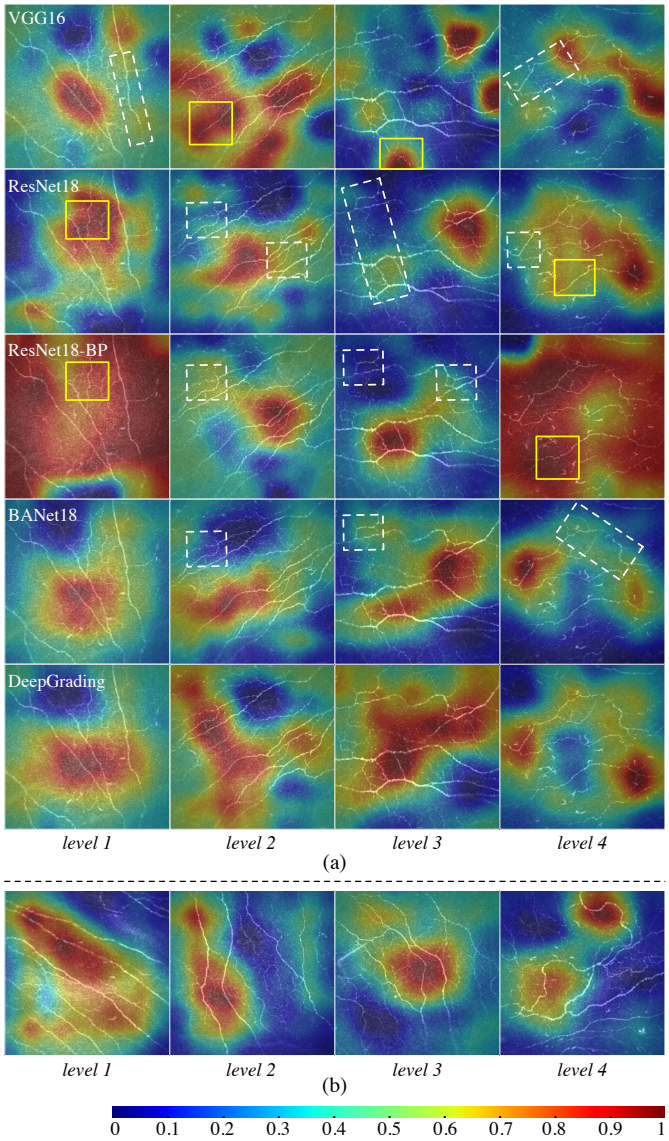

Fig. 7. Grad CAM heatmaps. (a): Heatmaps obtained using different methods over the images randomly selected at each tortuosity level. From the top to the bottom: the class activation maps of VGG16, ResNet18, ResNet18-BP, and the proposed BANet18 and DeepGrading, respectively. The nerve fibers within the dashed boxes basically determine the tortuosity of the image, but they were not activated. For the images at levels 1 and 4 in tortuosity, ResNet18-BP activated regions without nerve fibers as well as lesion, as shown in the yellow box. (b): Heatmaps obtained using the proposed method over the other randomly selected images with fewer nerve fiber bifurcations at each tortuosity level.

features acquired by the network. Fig. 6 illustrates the t-SNE visualization of the high-dimensional features of the original CCM images, ResNet18 and DeepGrading. Intuitively, the features of the original CCM images at different tortuosity levels are mixed, hard to decide the clustering centers. The clusters become clearer after the CCM images have been processed by ResNet18. However, it is still difficult to delineate the clustering boundaries between *levels 1*, *2*, and *3*. The integration of AuxNet enables the proposed DeepGrading to focus on capturing the fine-grained variations among *levels 1*, *2* and *3*, as evidenced by the t-SNE visualization with more compact clusters in Fig. 6.

The class activation maps for VGG16, ResNet18, ResNet18-BP, BANet18, and DeepGrading are illustrated in Fig. 7(a). The heatmaps reflect visual representations of the feature weights learned by the grading network. A darker red color in the heatmap indicates that the region contributes more to the tortuosity grading, and conversely a darker blue color indicates less or even no contribution of a region to the tortuosity grading. Although all the methods were able to locate the regions in the CCM images that determine fiber tortuosity, there are significant differences in local details. For *level 1*, where the nerve fibers are almost straight, VGG16 in the first column of Fig. 7(a) performed worst, with narrowly mapped activation regions. In contrast, DeepGrading located the activation regions that contained most of the straight nerve fibers. As the local variation in the CCM images became more significant with higher tortuosity levels, all the networks tended to activate the most representative regions. However, in the class activation visualizations of *levels 2* and *3*, VGG16 contains mis-activated regions, whereas ResNet18-BP activated almost all nerve fibers in *levels 1* and *4*. In sharp contrast, the proposed BANet improved the tortuosity activation accuracy, and DeepGrading further refined the identification of the most representative fibers. An interesting observation is that the heatmaps with higher tortuosity of nerve fibers in Fig. 7(a) highlight the regions with higher densities of bifurcations. To demonstrate that the proposed method can also localize the regions with a lower density of bifurcations, we randomly chose four other images at each tortuosity level and the experimental results are illustrated in Fig. 7(b). We can observe from Fig. 7(b) that the proposed method can learn tortuosity features even in the CCM images with a low density of bifurcations. These visualizations clearly explain why the proposed DeepGrading achieved better tortuosity grading performance. It is also worth noting that we model the concerns of the clinicians when grading tortuosity by generating the Grad CAM heatmaps, and thus allow the grading model to possess better interpretability.

E. Ablation study

To better demonstrate the curvature grading performance of the proposed DeepGrading, we conduct a series of ablation experiments of the procedure on the **CORN**^{1500} dataset.

1) Role of BA: In this section, we document a series of comparisons in terms of $wAcc$, wSe, and wSp, intended to verify the performance of the BA module. We validated the proposed BA module on the original CCM images (denoted CCMs), the corneal nerve segmentation maps (denoted SegMaps) and the concatenation of the two (denoted CCMs + SegMaps). The ablation results are shown in Table III, where ✓and × indicate that we trained the model with or without a BA module, respectively. We can see that the grading performance of the model was improved by adding a BA module. Concatenating the CCM image with the segmentation map is equivalent to adding hard attention to the CCM image, guiding the model to focus more on the anatomical structure of the corneal nerves when extracting features, as evidenced by the results in Table III. In summary, concatenating the segmentation map with the

TABLE III

ABLATION STUDY OF BA MODULE. ALL THE METRICS ARE EXPRESSED IN PERCENTAGE (%). **CCMS**: CORNEAL CONFOCAL MICROSCOPY IMAGES, **SEGMAPS**: CORNEAL NERVE SEGMENTATION MAPS, **CCMS + SEGMAPS**: CONCATENATION OF CCMS AND SEGMAPS.

Ablation Sets	w/ BA	$wAcc$	wSe	wSp	p
CCMs	×	81.46	66.42	86.33	0.026
	✓	81.81	64.67	87.50	
SegMaps	×	80.93	65.16	85.84	$< .01$
	✓	81.18	66.17	85.90	
CCMs + SegMaps	×	82.91	68.42	88.76	$< .001$
	✓	**83.85**	**68.67**	**89.30**	

TABLE IV

PERFORMANCE OF THE PROPOSED BANET AND THE PROPOSED AUXNET UNDER DIFFERENT THRESHOLDS. ALL THE METRICS ARE EXPRESSED IN PERCENTAGE (%).

Methods		$wAcc$	wSe	wSp	p
BANet		83.85	68.67	89.30	$< .001$
BANet + AuxNet	$t = 0.5$	85.34	72.13	90.03	0.024
	$t = 0.6$	85.29	71.00	**90.21**	< 0.01
	$t = 0.7$	**85.64**	**72.67**	89.67	—
	$t = 0.8$	85.01	70.33	89.84	< 0.01

CCM image and the proposed BA module can significantly improve tortuosity grading performance ($p < 0.05$).

2) Role of AuxNet: In this section, we verify the superiority of the proposed AuxNet. First of all, we compare the proposed BANet and BANet + AuxNet (DeepGrading) on the test set to demonstrate the grading performance of DeepGrading. All the experimental results are listed in Table IV. Statistical analysis shows that the proposed DeepGrading performs significantly better than the BANet ($p < 0.05$), as evidenced by increases of 1.79% and 4.00% in terms of $wAcc$ and wSe respectively.

*3) Determination of parameter **t**:* As described in Section II-B, the acquisition of N-ROIs requires a threshold $\boldsymbol{t}$. Therefore, to find the optimal threshold $\boldsymbol{t}$ that enables AuxNet to better assist DeepGrading for grading tortuosity, we constructed four experiments with different threshold values ($\boldsymbol{t} = \{0.5, 0.6, 0.7, 0.8\}$). The tortuosity grading results are shown in Table IV. They show that the proposed DeepGrading (BANet + AuxNet) achieved the best tortuosity grading performance at $\boldsymbol{t} = 0.7$ ($p < 0.05$, Delong's test), which justifies the fine-tuning of the threshold parameter.

4) Role of N-ROI: Furthermore, to verify that the thresholded Grad CAM mask and the N-ROI generator filter out the representative tortuosity level region, we trained AuxNet with the CCM images (denoted as AuxNet-CCM) and the CCM images multiplied with the thresholded Grad CAM (denoted as AuxNet-tCCM) as input, respectively. The grading performance is illustrated in Table V. It can be observed that the grading performance improves significantly ($p < 0.0001$) when the thresholded Grad CAM was directly multiplied with the CCM, which confirms the effectiveness of the N-ROI generator.

5) Impact of segmentation on grading: The proposed method, especially AuxNet, relies on nerve fiber segmentation for tortuosity grading. Therefore, to verify the impact of neural fiber segmentation on tortuosity grading, we trained the backbone ResNet18 and DeepGrading with neural fibers

TABLE V

GRADING PERFORMANCE OF AUXNET-CCM AND AUXNET-tCCM IN TERMS OF ACCURACY, SENSITIVITY AND SPECIFICITY. ALL THE METRICS ARE EXPRESSED IN PERCENTAGE (%)

Methods	Metrics	*level1*	*level2*	*level3*	*level4*	*overall*
AuxNet-CCM	$wAcc$	87.67	75.00	76.00	74.00	76.70
	wSe	35.71	44.09	34.72	**94.62**	56.33
	wSp	**96.12**	**88.89**	**89.04**	64.73	82.45
AuxNet-tCCM	$wAcc$	**89.00**	**87.33**	**80.33**	**78.67**	**83.20**
	wSe	**52.38**	**86.02**	**69.44**	54.84	**67.67**
	wSp	94.96	87.92	83.77	**89.37**	**88.36**

TABLE VI

CORNEAL NERVE SEGMENTATION PERFORMANCE OF DIFFERENT METHODS AND THEIR IMPACT ON THE PROPOSED BANET ON TORTUOSITY GRADING. **UPPER TABLE**: SEGMENTATION PERFORMANCE; **LOWER TABLE**: GRADING PERFORMANCE. ALL THE METRICS ARE EXPRESSED IN PERCENTAGE (%)

Methods	SEN	FDR
U-Net	77.57 ± 1.45	33.11 ± 2.08
CS-Net	83.80 ± 0.98	25.56 ± 0.28

Methods	ACC	SEN	SPE	p
ResNet18 (U-Net)	80.35	62.33	86.82	< 0.001
DeepGrading (U-Net)	83.33	68.33	87.81	
ResNet18 (CS-Net)	82.83	67.00	88.45	< 0.001
DeepGrading (CS-Net)	85.64	72.67	89.67	

segmented by U-Net [31] and CS-Net [43], respectively, where U-Net and CS-Net were trained on the CORN-1 [43] dataset. The segmentation and the tortuosity grading performance are shown in the upper and lower tables of Table VI, respectively, where the segmentation performance is measured in sensitivity (SEN) and false discovery rate (FDR) [43]. They show that the grading performance of our proposed method on non-ideal segmentation (lower table of Table VI, second row) remains on the whole superior to the competing method on ideal segmentation (lower table of Table VI, third row), demonstrating that the proposed method can significantly ($p < 0.001$) improve the tortuosity grading performance.

IV. DISCUSSION

The proposed DeepGrading, in which the AuxNet module was designed for challenging grading level prediction and developed from the perspective of interpretability [36], achieved on the whole the best results among the tested methods. Inspired by Bilinear CNN [28], we proposed a BA model using the attention mechanism to exploit features with high-level semantic information that might contribute to nerve fiber tortuosity grading, thereby improving grading performance. The experimental results demonstrate that the proposed method can better classify corneal nerves into different tortuosity levels. The tortuosity of the nerves is related to disease status: the more critical the grading accuracy, the more reliable the method.

A. Interpretable exploration of grading performance

By observing the Grad CAM heatmap in Fig. 7, it is clear that the network pays more attention to local complex variations in the CCM images, especially at higher tortuosity

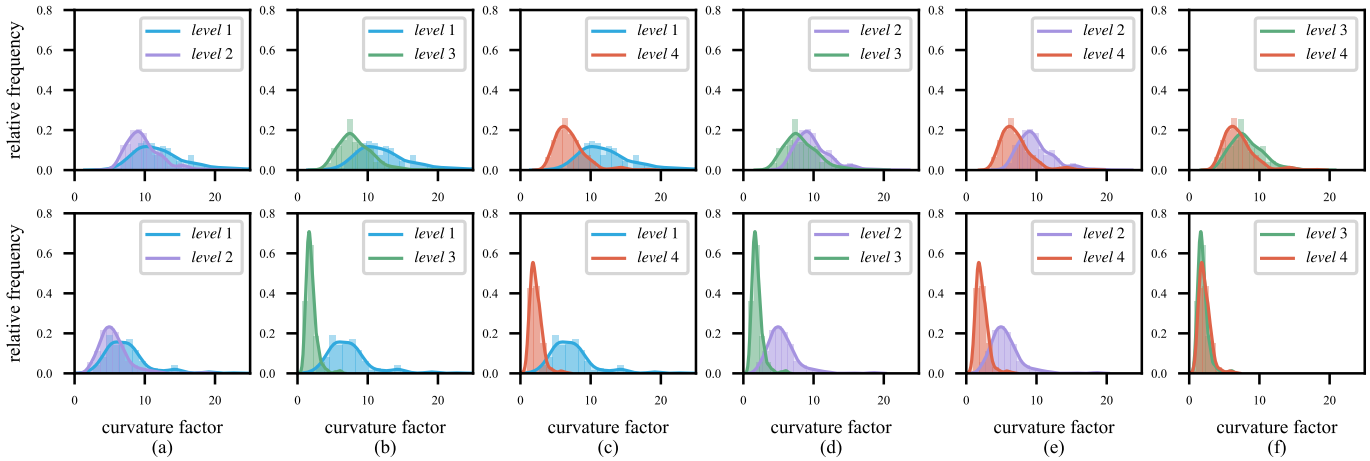

Fig. 8. Pairwise contrast histograms with Gaussian kernel density estimation (KDE) in 4 levels of nerve fibers in global CCM images and N-ROIs. The first row shows the tortuosity density of global fiber segments, and the second row shows the tortuosity density of N-ROIs. The horizontal and vertical axis indicate the curvature factors and the relative frequency of the training images of CORN^{1500}, respectively.

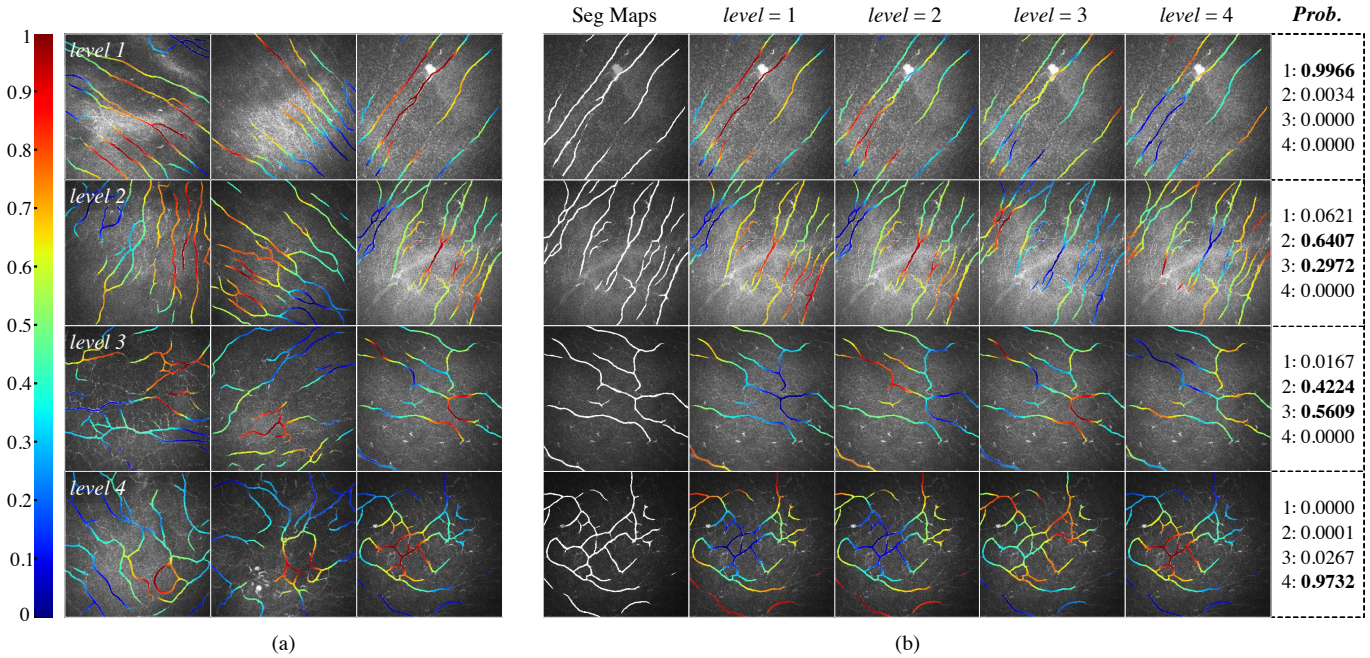

Fig. 9. Visual examples of activated nerve fiber segments. (a) indicates the nerve fiber segments that were activated in different CCM images in each of the 4 levels; (b) indicates the nerve fiber segments and their probabilities that were activated at $level = n$ $(n = 1, 2, 3, 4)$ in the same CCM image in each of the 4 levels. Best viewed in color. ***Prob.*** indicates the credibility when the tortuosity of the CCM image is predicted to be $level\ n$.

levels. This means that we can further explore the fine-grained features of local regions. In view of this, we propose AuxNet to guide the backbone network to focus on capturing local corneal nerve variations. The proposed AuxNet has contributed significantly in further improving grading performance, especially the procedure of generating N-ROIs using Grad CAM [36]. To demonstrate that the generated N-ROIs can contain more discriminative fine-grained features and make an interpretable comparison to explain why AuxNet can improve performance, we constructed a statistical comparison analysis of the tortuosity of local N-ROIs and global nerve fiber segments, where the nerve fiber curvature factor is obtained using the tortuosity estimation algorithm proposed in [56]. Fig. 8 illustrates the tortuosity density distribution of nerve fibers in the training images of CORN^{1500} based on the Gaussian kernel density estimation (KDE [57]) algorithm: the first and second rows represent the tortuosity density of global nerve fiber segments and N-ROIs, respectively. The horizontal axis represents the curvature factors, the vertical axis represents the relative frequency of the images with a particular curvature factor. By observing columns (a) to (f) in Fig. 8, we can see an increase in the discriminability of the tortuosity of *level 1* versus *levels 3* and *4*, and *level 2* versus *levels 3* and *4* in the N-ROIs, confirming that N-ROIs contain more discriminative fine-grained information than the CCM images as a whole. Especially, we proposed the AuxNet that takes the nerves in

TABLE VII
NUMBERS OF CCM IMAGES AT DIFFERENT TORTUOSITY LEVELS PREDICTED USING THE PROPOSED DEEPGRADING IN AN INDEPENDENT DATASET ABOUT THE HEALTHY CONTROL (HC) AND DIABETIC MELLITUS (DM) GROUPS.

Group	*level 1*	*level 2*	*level 3*	*level 4*	Total	p
HC	83	53	16	8	160	< 0.001
DM	14	32	95	53	194	

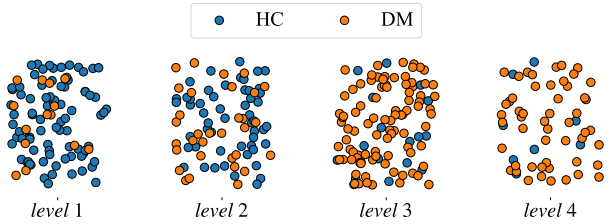

Fig. 10. Distribution of corneal nerve tortuosity levels predicted using the proposed DeepGrading in the healthy control (HC) and diabetic mellitus (DM) groups.

the ROIs as input. The overall grading accuracy is determined by BANet and AuxNet. Although *levels 3* and *4* are more challenging to differentiate, it is the nerves in the ROIs that become less indicative, not the nerves of the whole image. AuxNet was designed to improve the deficiencies in accuracy when classifying *levels 2* and *3* based on the whole images.

One of the purposes of the proposed method is to mimic the ophthalmologists performing tortuosity grading so as to predict the areas of interest to them, thus allowing us to further improve the grading performance based on the nerve fibers within the obtained highlighted areas. The Grad CAM heatmap in Fig. 7 initially explains that the proposed method can more accurately locate the tortuosity-determining regions. However, the Grad CAM heatmap is a coarse representation that cannot accurately show the tortuosity distribution of each fiber segment. We generated a probability map of the entire corneal nerve anatomy by overlaying the Grad CAM heatmap on the segmented corneal nerve fibers, as illustrated in Fig. 9. Compared with the Grad CAM heat map, the probability map accurately indicates the activation probability of the nerve fibers at the pixel level. As shown in Fig. 9(a), we can obtain the fiber segments that determine the tortuosity of the image at various tortuosity levels: i.e., the fiber segments with higher heating values can be used to determine the tortuosity. The probability map can be used not only to explain network performance, but also to aid clinical disease diagnosis, i.e., it can help physicians to identify the regions where tortuosity occurs as well as the severity of the tortuosity.

However, images with higher tortuosity levels usually contain nerve fiber segments with different tortuosities. To further explore and characterize the nerve fibers with different tortuosities in a single CCM image, we assigned a specific $level$ to Grad CAM when generating the tortuosity probability map: i.e., we assigned $level = n$ $(n = 1, 2, 3, 4)$, respectively. In addition, the grading probabilities of nerve fibers under different levels in the same image are output simultaneously, as illustrated in Fig. 9(b), where the nerve fibers that were activated under the assigned $level$ are highlighted in red. As in the first row of Fig. 9(b), most of the straight nerve fibers under *level 1* were activated as expected at $level = 1$ with a probability of 99.66%. In contrast, a small number of nerve fibers were still activated at $level = 2$, but the activation probability is as low as 0.34%. Similarly, the activation probabilities at $level = 3$ and $level = 4$ are 0, i.e., it is not credible that the nerve fibers are activated at either *level 3* or *4*. Therefore, we believe that the tortuosity of the nerve fibers reported in the first row is *level 1*. In the second row, the activation probabilities of the nerve fibers at $level = 2$ and $level = 3$ are 64.07% and 29.72%, respectively. Therefore, we believe that there exist two levels of nerve fiber segments with available locations in the CCM image. This is more evident in the third row of Fig. 9(b), where the nerve fibers are activated as *level 2* and *level 3* with 42.24% and 56.09% probabilities, respectively. As in the first row, the activation probability for $level = 1/2/3$ is too low to support that the CCM image in the fourth row contains both *levels 1*, *2* and *3* nerve fibers. Therefore, we conclude that this image contains only corneal nerves of tortuosity *level 4*, with an activation probability of 97.32%.

B. Clinic evaluation

Our tortuosity grading was further validated by the clinical practice. The clinical dataset was used to study the correlation between the tortuosity levels and pathology state. This dataset contains 160 and 194 images from healthy control (HC) and Diabetic Mellitus (DM) groups. We performed the automated tortuosity grading for each image and the results are shown in Table VII. Since the predicted tortuosity levels were discrete, Chi-square test was used to verify the statistically significant difference between the two gradings. We can conclude from Table VII that the tortuosity levels predicted by DeepGrading are significantly different in the HC and DM groups ($p < 0.001$). Fig. 10 depicts the distribution of the tortuosity of the CCM images in these two groups. It can be observed clearly that the tortuosity of the HC group is mainly distributed in *levels 1* and *2*, while that of the DM group falls mainly into *levels 3* and *4*. It may be concluded that DM is highly related to *levels 3* and *4*, but less related to *levels 1* and *2*, which is consistent with the findings of previous studies [6], [8], [58] that diabetic patients exhibit higher corneal nerve tortuosity. These findings indicate that our automatic tortuosity grading method has a potential to assist in differentiating the healthy subjects from those with diabetes.

Although the proposed corneal nerve tortuosity grading algorithm has yet to be applied for clinical diagnosis on a large scale, it has great potentials to be introduced into a real clinical pathway for early diagnosis and monitoring of many eye and systemic diseases. For example, due to the high prevalence of diabetic neuropathy (DN) in patients with type 1 diabetes, early detection of DN is very important for risk stratification of patients so that the endpoints of foot ulceration and premature death can be prevented [59]. CCM examination and analysis could represent a novel non-invasive method to accurately quantify nerve morphology, and thus aid in the diagnosis and determination of the severity of DN [60]. In order to make real

benefits to clinicians and patients, the proposed method can be further developed as a plug-in of existing software such as ImageJ[5], deployed as a web service, or distributed as a standalone software tool. These research tools can be freely used by clinicians to support their decision-making process at their own risks. If the algorithm will be commercially used as software as a medical device (SaMD) to make the diagnosis alone, regulatory approvals such as FDA or CE mark will be required in additional to rigorous clinical evaluations. On the other hand, continuous refinement and improvement will be always required to make the algorithm robust and accurate.

V. Conclusions and Future Work

Accurate assessment of corneal nerve fiber tortuosity is very important to facilitate the examination and diagnosis of many ophthalmic diseases. Due to the subjectivity and slow speed of diagnostic procedures in manual examination, it is particularly important to develop a fully automated corneal nerve tortuosity estimation method. In this paper, we proposed a novel corneal nerve tortuosity grading method based on the latest deep learning and interpretability mechanism, namely DeepGrading. The proposed DeepGrading achieves overall state-of-the-art grading performance using the proposed bi-linear attention module and an auxiliary grading network. It not only grades the overall tortuosity of CCM images but also can locate and grade the tortuosity of nerve fibers at different spatial locations in each CCM image. This method has huge potential to be introduced into clinics after large population studies in future. However, the CCM images involved in this work are of limited field of view relative to the entire cornea, and a small error in the estimation of clinical parameters may lead to misdiagnosis. Therefore, we will introduce CCM images with a larger field of view in the future work and include clinical reliability validation of the obtained tortuosity. On the other hand, we intend to apply this powerful method to other imaging modalities, such as fundus images and optical coherence tomography angiography (OCT-A) images, for improved diagnosis of eye-related diseases in the future.

References

[1] F. Scarpa, X. Zheng, Yuichi O., and Alfredo R., "Automatic evaluation of corneal nerve tortuosity in images from in vivo confocal microscopy," *Investigative Ophthalmology Visual Science*, vol. 52, no. 9, pp. 6404–6408, 2011.

[2] R. Annunziata, A. Kheirkhah, S. Aggarwal, P. Hamrah, and E. Trucco, "A fully automated tortuosity quantification system with application to corneal nerve fibres in confocal microscopy images," *Medical Image Analysis*, vol. 32, pp. 216–232, 2016.

[3] K. Edwards, N. Pritchard, D. Vagenas, A. Russell, R. Malik, and N. Efron, "Standardizing corneal nerve fibre length for nerve tortuosity increases its association with measures of diabetic neuropathy," *Diabetic Medicine*, vol. 31, no. 10, pp. 1205–1209, 2014.

[4] J. Kim and M. Markoulli, "Automatic analysis of corneal nerves imaged using in vivo confocal microscopy," *Clinical and Experimental Optometry*, vol. 101, no. 2, pp. 147–161, 2018.

[5] P. Su, T. Chen, J. Xie, Y. Zheng, H. Qi, D. Borroni, Y. Zhao, and J. Liu, "Corneal nerve tortuosity grading via ordered weighted averaging-based feature extraction," *Medical Physics*, vol. 47, no. 10, pp. 4983–4996, 2020.

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

[6] A. Manfre, E. Brugin, A. Ghirlando, E. Moretto, and E. Midena, "Corneal diabetic neuropathy: A confocal microscopy study," *Investigative Ophthalmology & Visual Science*, vol. 47, no. 13, pp. 5569–5569, 2006.

[7] Albert A., Bernardo C., and Pedram H., "In vivo confocal microscopy in dry eye disease and related conditions," *Seminars in Ophthalmology*, vol. 27, no. 5-6, pp. 138–148, 2012.

[8] P. Kallinikos, M. Berhanu, C. O'Donnell, A. Boulton, N. Efron, and R. Malik, "Corneal nerve tortuosity in diabetic patients with neuropathy," *Investigative Ophthalmology & Visual Science*, vol. 45, no. 2, pp. 418–422, 2004.

[9] Y. Liu, Y. Chou, X. Dong, Z. Liu, X. Jiang, R. Hao, and X. Li, "Corneal subbasal nerve analysis using in vivo confocal microscopy in patients with dry eye: analysis and clinical correlations," *Cornea*, vol. 38, no. 10, pp. 1253–1258, 2019.

[10] B. Williams, D. Borroni, R. Liu, Y. Zhao, J. Zhang, J. Lim, B. Ma, V. Romano, H. Qi, M. Ferdousi, 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*, vol. 63, no. 2, pp. 419–430, 2020.

[11] B. Ma, J. Xie, T. Yang, P. Su, R. Liu, T. Sun, Y. Zhou, H. Wang, X. Feng, S. Ma, et al., "Quantification of increased corneal subbasal nerve tortuosity in dry eye disease and its correlation with clinical parameters," *Translational Vision Science & Technology*, vol. 10, no. 6, pp. 26–26, 2021.

[12] P. Hamrah, A. Cruzat, M. Dastjerdi, H. Prüss, L. Zheng, B. Shahatit, H. Bayhan, R. Dana, and D. Pavan-Langston, "Unilateral herpes zoster ophthalmicus results in bilateral corneal nerve alteration: an in vivo confocal microscopy study," *Ophthalmology*, vol. 120, no. 1, pp. 40–47, 2013.

[13] P. Hamrah, A. Cruzat, M. Dastjerdi, L. Zheng, B. Shahatit, H. Bayhan, R. Dana, and D. Pavan-Langston, "Corneal sensation and subbasal nerve alterations in patients with herpes simplex keratitis: an in vivo confocal microscopy study," *Ophthalmology*, vol. 117, no. 10, pp. 1930–1936, 2010.

[14] K. Kurbanyan, L. Hoesl, W. Schrems, and P. Hamrah, "Corneal nerve alterations in acute Acanthamoeba and fungal keratitis: an in vivo confocal microscopy study," *Eye*, vol. 26, no. 1, pp. 126–132, 2012.

[15] R. Annunziata, A. Kheirkhah, S. Aggarwal, B. Cavalcanti, P. Hamrah, and E. Trucco, "Tortuosity classification of corneal nerves images using a multiple-scale-multiple-window approach," in *Proceedings of the Ophthalmic Medical Image Analysis First International Workshop*. 2014, University of Iowa.

[16] Y. Zhao, J. Zhang, E. Pereira, Y. Zheng, P. Su, J. Xie, Y. Zhao, Y. Shi, H. Qi, J. Liu, et al., "Automated tortuosity analysis of nerve fibers in corneal confocal microscopy," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2725–2737, 2020.

[17] J. Klisser, S. S. Tummanapalli, J. Kim, J. Chiang, V. Khou, T. Issar, T. Naduvilath, A. Poynten, M. Markoulli, and A. Krishnan, "Automated analysis of corneal nerve tortuosity in diabetes: implications for neuropathy detection," *Clinical and Experimental Optometry*, pp. 1–7, 2021.

[18] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P.W. Pluim, R. Duits, and B. M. ter Haar Romeny, "Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores," *IEEE transactions on medical imaging*, vol. 35, no. 12, pp. 2631–2644, 2016.

[19] E. Bribiesca, "A measure of tortuosity based on chain coding," *Pattern Recognition*, vol. 46, no. 3, pp. 716–724, 2013.

[20] M. Alvarado-Gonzalez, W. Aguilar, E. Garduño, C. Velarde, E. Bribiesca, and V. Medina-Bañuelos, "Mirror symmetry detection in curves represented by means of the slope chain code," *Pattern Recognition*, vol. 87, pp. 67–79, 2019.

[21] Ö. Smedby, N. Högman, S. Nilsson, U. Erikson, A. Olsson, and G. Walldius, "Two-dimensional tortuosity of the superficial femoral artery in early atherosclerosis," *Journal of Vascular Research*, vol. 30, no. 4, pp. 181–191, 1993.

[22] W. Hart, M. Goldbaum, B. Côté, P. Kube, and M. Nelson, "Measurement and classification of retinal vascular tortuosity," *International Journal of Medical Informatics*, vol. 53, no. 2-3, pp. 239–252, 1999.

[23] E. Grisan, M. Foracchia, and A. Ruggeri, "A novel method for the automatic grading of retinal vessel tortuosity," *IEEE Transactions on Medical Imaging*, vol. 27, no. 3, pp. 310–319, 2008.

[24] E. Bullitt, G. Gerig, S. Pizer, W. Lin, and S. Aylward, "Measuring tortuosity of the intracerebral vasculature from mra images," *IEEE Transactions on Medical Imaging*, vol. 22, no. 9, pp. 1163–1171, 2003.

[25] P. Mehrgardt, S. Zandavi, S. Poon, J. Kim, M. Markoulli, and Mat. Khushi, "U-net segmented adjacent angle detection (usaad) for auto-

matic analysis of corneal nerve structures," *Data*, vol. 5, no. 2, pp. 37, 2020.

[26] E. J. Bekkers, J. Zhang, R. Duits, and B. M. ter Haar Romeny, "Curvature based biomarkers for diabetic retinopathy via exponential curve fits in se(2)," in *Ophthalmic Medical Image Analysis International Workshop*, 2015, pp. 113–120.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[31] O. Ronneberger, P. Fischer, and T. 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.

[32] L. Mou, Y. Zhao, H. Fu, Y. Liu, J. Cheng, Y. Zheng, P. Su, J. Yang, L. Chen, A. F. Frangi, M. Akiba, and J. Liu, "CS^2-Net: Deep learning segmentation of curvilinear structures in medical imaging," *Medical Image Analysis*, vol. 67, pp. 101874, 2021.

[33] Y. Ma, H. Hao, J. Xie, H. Fu, J. Zhang, J. Yang, Z. Wang, J. Liu, Y. Zheng, and Y. Zhao, "ROSE: A retinal oct-angiography vessel segmentation dataset and new model," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 928–939, 2021.

[34] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, vol. 31, Curran Associates, Inc.

[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[36] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[37] A. Chattopadhay, A. Sarkar, P. Howlader, and V. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

[38] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, and W. Qian, "An interpretable ensemble deep learning model for diabetic retinopathy disease classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2045–2048.

[39] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Transactions on Medical Imaging*, 2020.

[40] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[41] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, "Clinical interpretable deep learning model for glaucoma diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1405–1412, 2019.

[42] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927*, 2018.

[43] L. Mou, Y. Zhao, L. Chen, J. Cheng, Z. Gu, H. Hao, H. Qi, Y. Zheng, A. Frangi, and J. Liu, "Cs-net: Channel and spatial attention network for curvilinear structure segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 721–730.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[46] T. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[47] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*. Springer, 2010, pp. 143–156.

[48] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[49] Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, and Y. Zhao, "Structure and illumination constrained gan for medical image enhancement," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2021.

[50] P. Hertz, V. Bril, A. Orszag, A. Ahmed, E. Ng, P. Nwe, M. Ngo, and B. Perkins, "Reproducibility of in vivo corneal confocal microscopy as a novel screening test for early diabetic sensorimotor polyneuropathy," *Diabetic Medicine*, vol. 28, no. 10, pp. 1253–1260, 2011.

[51] N. Lagali, E. Poletti, D. Patel, C. McGhee, P. Hamrah, A. Kheirkhah, M. Tavakoli, I. Petropoulos, R. Malik, T. Utheim, et al., "Focused tortuosity definitions based on expert clinical assessment of corneal subbasal nerves," *Investigative Ophthalmology & Visual Science*, vol. 56, no. 9, pp. 5102–5109, 2015.

[52] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[54] E. DeLong, D. DeLong, and D. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[55] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[56] P. Su, Y. Zhao, T. Chen, J. Xie, Y. Zhao, H. Qi, Y. Zheng, and J. Liu, "Exploiting reliability-guided aggregation for the assessment of curvilinear structure tortuosity," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 12–20.

[57] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[58] S. De Cillà, S. Ranno, E. Carini, P. Fogagnolo, G. Ceresara, N. Orzalesi, and L. Rossetti, "Corneal subbasal nerves changes in patients with diabetic retinopathy: an in vivo confocal study," *Investigative Ophthalmology & Visual Science*, vol. 50, no. 11, pp. 5155–5158, 2009.

[59] S. Tesfaye, N. Chaturvedi, S. Eaton, J. Ward, C. Manes, C. Ionescu-Tirgoviste, D. Witte, and J. Fuller, "Vascular risk factors and diabetic neuropathy," *New England Journal of Medicine*, vol. 352, no. 4, pp. 341–350, 2005.

[60] P. Hossain, A. Sachdev, and R. A. Malik, "Early detection of diabetic peripheral neuropathy with corneal confocal microscopy," *The Lancet*, vol. 366, no. 9494, pp. 1340–1343, 2005.