Zhang *et al.*
Health Information Science and Systems (2022) 10:3
https://doi.org/10.1007/s13755-022-00170-2

Health Information Science and Systems

RESEARCH

Mixed pyramid attention network for nuclear cataract classification based on anterior segment OCT images

Xiaoqing Zhang[1,2*†], Zunjie Xiao[2†], Xiaoling Li[3], Xiao Wu[2], Hanxi Sun[2], Jin Yuan[4], Risa Higashita[2,5] and Jiang Liu[1,2,3,6*]

Abstract

Nuclear cataract (NC) is a leading ocular disease globally for blindness and vision impairment. NC patients can improve their vision through cataract surgery or slow the opacity development with early intervention. Anterior segment optical coherence tomography (AS-OCT) image is an emerging ophthalmic image type, which can clearly observe the whole lens structure. Recently, clinicians have been increasingly studying the correlation between NC severity levels and clinical features from the nucleus region on AS-OCT images, and the results suggested the correlation is strong. However, automatic NC classification research based on AS-OCT images has rarely been studied. This paper presents a novel mixed pyramid attention network (MPANet) to classify NC severity levels on AS-OCT images automatically. In the MPANet, we design a novel mixed pyramid attention (MPA) block, which first applies the group convolution method to enhance the feature representation difference of feature maps and then construct a mixed pyramid pooling structure to extract local-global feature representations and different feature representation types simultaneously. We conduct extensive experiments on a clinical AS-OCT image dataset and a public OCT dataset to evaluate the effectiveness of our method. The results demonstrate that our method achieves competitive classification performance through comparisons to state-of-the-art methods and previous works. Moreover, this paper also uses the class activation mapping (CAM) technique to improve our method's interpretability of classification results.

Keywords: Nuclear cataract, Classification, Mixed pyramid attention, CNN, AS-OCT images

Introduction

With an aging population globally, cataract will become the first cause for visual impairment and blindness in 2030 [1]. Cataract surgery and early intervention are two effective methods to improve cataract patients' vision and life quality, reducing blindness ratio and social burden. Nuclear cataract (NC) is a common age-related, yet reversible cataract type, associated with different factors, such as, increasing age, lifestyle factors, and genetic factors [2]. The clinical symptoms of NC are gradual clouding and progressive hardening of the nucleus region in the crystalline lens structure. Ophthalmologists have used several ophthalmic images (e.g., slit lamp images) over the past years to diagnose NC severity levels based on the clinical cataract classification systems. Lens Opacity Classification System III (LOCS III) [3] is a commonly accepted clinical cataract classification system for NC diagnosis, which is built on slit-lamp images. E.g., ophthalmologists usually use silt-lamp images to diagnose NC, but this manual diagnosis mode is subjective and highly relies on clinical experience and knowledge.

According to actual clinical diagnosis requirements and opacity development in the nucleus region, we can categorize the severity levels of NC into three levels based on the LOCS III. Level 1: Mild cataract (NC grade < 3),

*Correspondence: 11930927@mail.sustech.edu.cn; liuj@sustech.edu.cn
†Xiaoqing Zhang and Zunjie Xiao have contributed equally to this work.
[2] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[6] Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Southern University of Science and Technology, Shenzhen 518055, China
Full list of author information is available at the end of the article

is asymptomatic. Level 2: Moderate cataract (NC grade = 3), is symptomatic. Level 3: Severe cataract (NC grade > 3), is symptomatic severely. Clinical interventions, e.g., Kary Uni eye drops, can slow the opacity development of mild NC patients. For patients with moderate NC, clinical progress follow-up is necessary. Patients with severe NC should undergo cataract surgery. Figure 1 provides three representative severity levels of NC on AS-OCT images: mild (b), moderate (c), and severe (d).

Anterior segment optical coherence tomography (AS-OCT) imaging method is a quick, non-invasive, objective, user-friendly, and high-resolution, compared with other ophthalmic imaging modes like fundus imaging. Ophthalmologists and scholars have gradually used AS-OCT images for anterior segment ocular disease diagnosis and scientific research purposes. [4, 5] proposes deep convolutional neural network (CNN) models for automatic corneal structure segmentation, which can be used to assist ophthalmologists in locating corneal structure and diagnosing corneal diseases accurately. Fu et al. [6–9] used deep learning methods to detect angle-closure glaucoma on AS-OCT images for helping ophthalmologists diagnose glaucoma objectively and obtained promising performance. For clinical cataract diagnosis, AS-OCT image is able to capture the lens structure, including nucleus-, cortex-, and capsule- regions clearly compared with slit-lamp image and fundus image, which is vital for diagnosing different cataract types. Scholars have recently studied the opacity correlation between NC severity levels and clinical features from the nucleus region on AS-OCT images. E.g., Wong et al. [10] analyzed the opacity correlation between the severity levels of NC and average density with Spearman's correlation analysis method, and statistical results indicated that the opacity correlation relationship between them is strong. [11–14] also obtains similar opacity correlation coefficient values between them. Overall, existing clinical research provides the clinical support for AS-OCT image-based NC classification automatically and a potential contribution for cataract surgery planning, it is because clinical research [13, 10] has suggested that high intra-class and inter-class repeatability of AS-OCT image-based NC diagnosis.

Apart from clinical NC research on AS-OCT images, Zhang et al. [15] first proposed a CNN model named GraNet to predict NC severity levels automatically by using AS-OCT images and achieved poor performance without considering the relationship between NC and the lens structure. [16] uses intensity-based statistics method to extract clinical features from the nucleus region and then utilizes random forest (RF) to classify NC severity levels. Xiao et al. [17] presented a gated channel attention network to predict NC and got good classification results. Furthermore, we obtain two findings according to existing literature of NC research: 1) [11–18] obtains different correlation coefficients on different nucleus regions through the average density values of AS-OCT images, e.g., the bottom half region and the whole region; 2) clinical features play different roles in NC diagnosis like mean and maximum. We question whether clinical prior knowledge of NC can be converted into feature representation of CNNs to improve classification performance.

Over the years, attention mechanism [19, 20] has become a vital component of CNNs, enabling to augmenting feature representations of feature maps for obtaining expected classification performance. Squeeze-and-excitation (SE) is a widely used attention method, which reconstructs the inter-dependent relationship among channels and recalibrates the feature maps. The spatial pyramid attention (SPA) [21] method extracts global-local feature representations with the pyramid pooling method for boosting the representational power of a CNN. In [22], Residual Attention

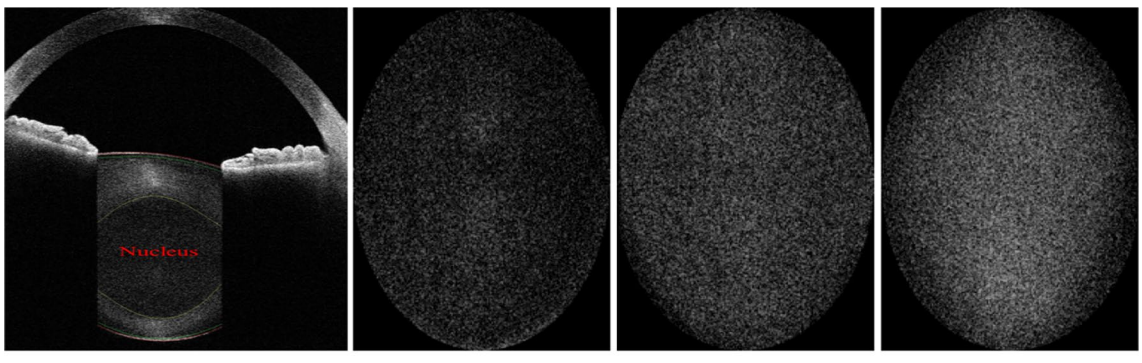


Fig. 1 Three nuclear cataract severity levels based on AS-OCT images (**a**). Mild nuclear cataract (**b**) with slight opacity but is asymptomatic. Moderate nuclear cataract (**c**) with moderate opacity and is symptomatic. Severe nuclear cataract (**d**) with severe opacity and is symptomatic obviously

Network is proposed to enhance further the classification performance, which provides a learning paradigm to combine the residual connection mechanism with the attention mechanism. Convolutional block attention block (CBAM) [23] and bottleneck attention module (BAM) [24] extend the idea of SE by combining channel attention method with spatial attention method sequentially and concurrently. Specially, these attention methods adopt global pooling methods, e.g., global average pooling method (GAP), to extract local and global feature representations from feature maps, which can be taken as other forms of clinical features (mean and maximum) of NC on AS-OST images. Motivated by the relationship between global feature representation of CNNs and clinical features of NC, this paper develops a novel attention-based network named Mixed Pyramid Attention Network (MPANet) by infusing the clinical prior knowledge of NC, to predict NC severity levels automatically on AS-OCT images, as shown in Fig. 2a.

In the MPANet, we design an effective mixed pyramid attention (MPA) block (Fig. 2c), consisting of a group convolution layer, mixed pyramid pooling (MPP) structure, and multi-layer perception (MLP). The group convolution layer enhances the feature representation difference of feature maps with two individual convolution partitions. It is followed by the MPP, which extracts different feature representation types and local-global feature representations of the feature map from each channel with the MPP method. Finally, it uses a learnable MLP to construct the interaction between channels for adjusting the relative importance of feature maps. A clinical AS-OCT image dataset with 7,919 images from 335 participants (average age is 69.40±9.97) is used to demonstrate the effectiveness of MPANet. The results show that our proposed MPANet achieves better performance compared with state-of-the-art attention-based CNNs and previous methods. A public OCT dataset is used to verify the generation ability of the MPANet, and results demonstrate the superiority of our method

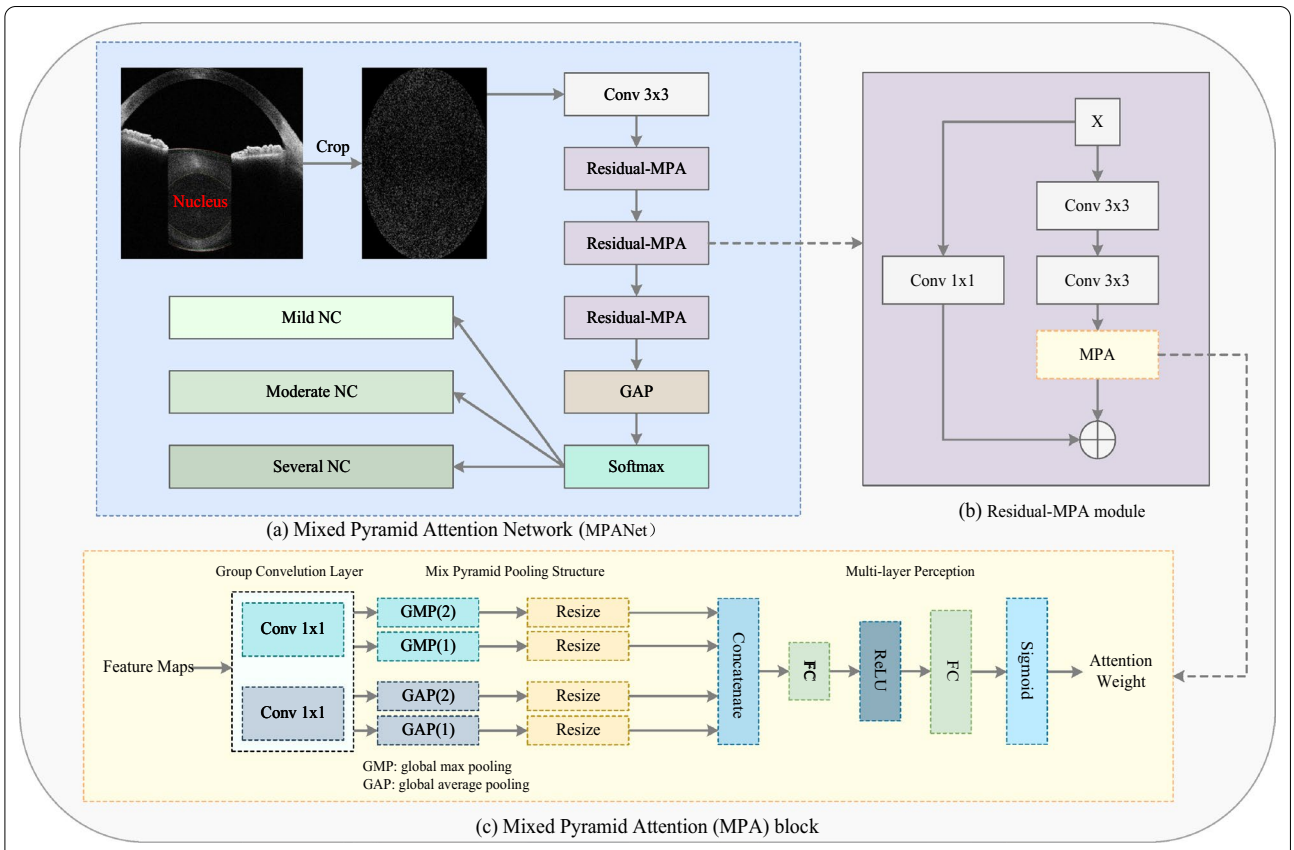


Fig. 2 The architecture of the Mixed Pyramid Attention Network (MPANet). We devise the MPA block by integrating clinical prior knowledge, and then utilize the MPA block to construct the Residual-MPA module by plugging it into the Residual module. MPANet (a) is used for NC classification by using the nucleus region from AS-OCT images, comprised of multiple Residual-MPA modules. We use a deep CNN model to acquire the nucleus region automatically. MPA block comprises a group convolution layer, a mixed pyramid pooling structure, and multi-layer perception. Green and blue pointwise convolutions (Conv. 1 × 1) denote two learned feature representation types

through comparisons to existing literature and attention-based CNNs. Furthermore, we utilize the class activation mapping (CAM) technique to localize what and where our MPANet focuses on, enhancing predicted outputs' interpretability.

The main contributions of this paper are summarized as follows:

- We propose a novel attention-based CNN architecture named Mixed Pyramid Attention Network (MPANet) for automatic NC classification on AS-OCT images by incorporating the clinical prior knowledge of NC: relative importance of clinical features and correlation between different nucleus regions and NC severity levels.
- In the MPANet, we construct a novel mixed pyramid attention (MPA) block for learning different feature representation types and local-global feature representation information adaptively. Moreover, this paper exploits three MPA variants: MPA-A, MPA-B, and MPA-C, **testing which factors affect the performance of our MPA** .
- We conduct experiments on a clinical AS-OCT image dataset and a public OCT dataset, and the results demonstrate that the MPANet surpasses strong baselines and previous works. This paper also uses the CAM method to visualize the classification results to improve interpretability.

The rest of this paper is organized as follows: Section 2 introduces our MPANet framework in detail. In Section 3, dataset introduction and experiment setting are presented. We discuss results and validate the general performance of our method in Section 4 and Section 5. Finally, we conclude and present future work in Section 6.

Methodology

A mixed pyramid attention (MPA) block can be taken as a computational unit which aims at incorporating the clinical prior knowledge into attention-based CNNs for boosting their representational power. Given the feature tensor $X = [x_1, x_2, ..., x_C] \in R^{C\times H\times W}$ as the input for MPA, and it generates the augmented representations $X^{'} = [x_1^{'}, x_2^{'}, ..., x_C^{'}] \in R^{C\times H\times W}$.

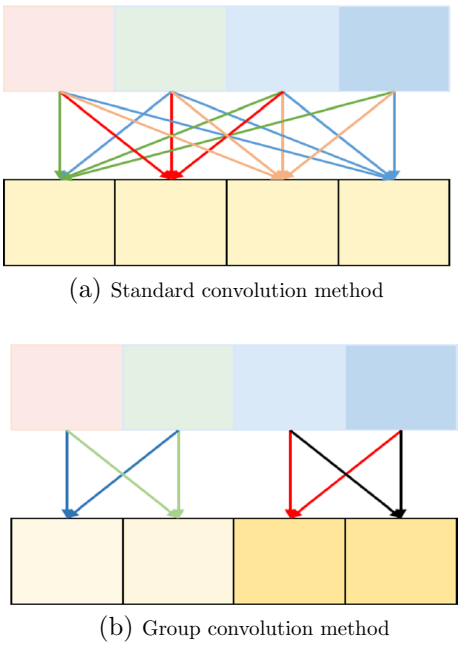
Mixed pyramid attention block

Figure 2c presents the overall framework of our mixed pyramid attention block architecture, which is comprised of a group convolution layer, a mixed pyramid pooling structure, and multi-layer perception (MLP). We will illustrate these tree components and their effects step by step in the following.

Group convolution layer

The group convolution (GC) method has been widely used to design efficient CNN architectures [25–27], since it can reduce the convolution redundancy as well as improve the general performance. Figure 3 provides a comparison example of standard convolution method and group convolution method.

Considering the advantages of the group convolution method, this paper first uses it to learn different feature representations for enhancing their difference in the MPA block, where we split convolution kernels into two convolution partitions. Thus, two convolution partitions can independently learn different feature representations from the previous layer, as shown in Fig. 2 (blue and green colors represent two convolution partitions, respectively). Specifically, two convolution partitions correspond to two individual pooling operations in the mixed pooling pyramid structure correspondingly. Furthermore, we set the convolution kernel size to 1×1 (named pointwise convolution, Conv. 1x1) for two convolution partitions, and the number of convolution kernels of each convolution partition is equal. This is because the pointwise convolution method is capable of clustering feature representations from previous feature maps according to literature [28, 29].



(a) Standard convolution method

(b) Group convolution method

Fig. 3 Toy example comparisons between standard convolution method and group convolution method

Mixed pyramid pooling structure

Following the group convolution method, we design a mixed pyramid pooling structure (MPP) to extract local-global feature representation and different feature representation types simultaneously. It first uses multi-scale global average pooling (GAP) and multi-scale global max pooling (GMP) operations to capture local-global feature representation types from feature maps correspondingly, which are generated by two independent convolution partitions in the group convolution layer. In the MPP, multi-scale GAP extracts local-global channel-wise average feature representations and multi-scale GMP extracts local-global channel-wise salient feature representations.

The motivation to adopt these two pooling operations as follows: *1) clinical findings have suggested that max density and mean density have varying levels of significance with NC severity levels* [10, 14] *(Noted that max and mean are two significant indicators for clinical NC diagnosis). Furthermore, these two clinical features can be viewed as channel-wise average feature representation and channel-wise salient feature representation of CNNs. 2)* [11–18] *indicates that top half- and bottom half-nucleus regions have different correlations with NC severity levels.* Thus, we set two pyramid pooling scales for GAP and GMP: 2×2 and 1×1. Two pooling methods can adaptively learn local-global feature representations and two feature representation types, e.g., 2×2 scale pooling operation can capture four local feature representations. In contrast, 1×1 scale pooling operation can capture the global feature representation of the feature map from each channel. GAP and GMP operations capture both channel-wise average and salient feature representations. Then we convert the extracted feature representations generated by two pooling operations into 1D vectors and concatenate them together along with channel axis and can be formulated as follows:

$$\begin{aligned} Z = Concate([\mu_1, \mu_2, \mu_3, \mu 4, \mu, Max_1, \\ Max_2, Max_3, Max_4, Max], axis = 1), \end{aligned} \tag{1}$$

where $Z = [z_1, z_2, ..., z_C]$, $z_C \in R^{10\times 1}$, $\mu_1, \mu_2, \mu_3, \mu 4, Max_1, Max_2, Max_3, and Max_4$ denote local channel-wise average and salient feature representations from each channel of feature maps; μ and Max denote global channel-wise average and salient feature representations from each channel of feature maps;.

Multi-layer perception network

The MPP generates mixed feature representations Z, which cannot represent the inter-dependencies of channels directly. Thus, we convert mixed feature representations into channel-wise weights by a simple multi-layer perceptron (MLP) network. Like SE, the MLP adopts two fully connected (FC) layers to construct the inter-dependencies between channels. The First FC layer is used to squeeze different global-local feature representation types with the dimensionality reduction d for better efficiency, like an encoder operation. The second FC layer reconstructs the dependency relationship of intra-channels, which is like a decoder operation. The operations of two FC layers can be formulated as follows:

$$G = \sigma(W_2\delta(W_1 Z)), \tag{2}$$

where $W_2 \in R^{C\times d}$, $W_1 \in R^{d\times C}$, $\delta, G \in R^{C\times 1}$, and σ denote the learnable weights of two fully-connected layers, Relu function, attention weights, and sigmoid function. Finally, the input X is reweighed by the attention weights, thus, the output can be obtained by:

$$X^{'} = G \cdot X, \tag{3}$$

To study the effects of d on the performance of the network, we adopt a reduction ratio r to control the value of d through the following equation:

$$d = max(C/r, M), \tag{4}$$

where M represents the minimal value of d by manual setting, and we set M and r to 32 and 16 in the experiments.

Discussion To exploit which factors affect the performance of our MPA block, this paper develops three MPA variants: MPA-A, MPA-B, and MPA-C.

MPA-A: The number of convolution kernel sizes of each convolution partition in the group convolution layer is equal to the previous convolutional layer has. GMP and GAP only use 1×1 pooling scale size.

MPA-B: Two convolution partitions in the group convolution layer using half the convolution kernel sizes as the previous layer adopts. GMP and GAP also only use the 1×1 pooling scale.

MPA-C: Each convolution partition in the group convolution layer has the same number of convolution kernel sizes as the previous layer contains. GMP and GAP also use 1×1 and 2×2 pooling scales.

Network architecture

This paper uses ResNets [30] as backbone networks to verify the effectiveness of our method according to two reasons. First, Resets are widely used CNN architectures and have achieved surpassing performance; many state-of-the-art attention-based CNN models like SENet built on ResNet, which provide strong baselines to evaluate our MPANet's performance. Second, the residual connection method (residual block) can alleviate the gradient vanishing problem in deep CNN models. We incorporate Residual block into MPA block titled Residual-MPA module,

and construct our MPANet through stacking Residual-MPA modules, a GAP layer, and a classifier, as shown in Fig. 2. We adopt ResNet18 and ResNet34 as baselines in this paper because these two models are commonly used and can achieve competitive results on limited datasets. Following modern attention-based CNNs, this paper adopts softmax function and cross-entropy loss as the classifier and loss function, respectively.

Dataset and experiment setting

Datasets

AS-OCT image dataset

This paper collected a clinical AS-OCT image dataset through the CASIA2 ophthalmology device (Tomey Corporation, Japan). AS-OCT images capture the whole anterior chamber structure of each eye, as shown in Fig. 1a. Considering NC severity levels is only related to the nucleus region according to clinical research, as shown in Fig. 1b–d.We use a deep segmentation network [31] to crop the nuclear region automatically, as shown in Fig. 2 (top left).

The AS-OCT image dataset contains 335 participants, and the total number of eyes is 437 (the number of right eyes is 213, and the number of left eyes is 224). The average age of participants is 69.40±9.97. We collect the number of AS-OCT images from each eye is 20, and the total number of AS-OCT images is 7,919. Since we manually remove poor-quality images with the help of experienced clinicians. The dataset collection of this paper is conducted according to the tenets of the Helsinki Declaration. Because of the retrospective nature and fully anonymized usage of the dataset, we are exempted by the medical ethics committee to inform the patients. Given lacking clinical cataract classification systems built on AS-OCT images, NC labels of AS-OCT images are mapped from slit-lamp images, which three experienced ophthalmologists labeled based on LOCS III. Moreover, clinical NC research [10, 13] have proved that high intra-class and inter-class repeatability for NC diagnosis on AS-OCT images, which provides strong support for automatic AS-OCT image-based NC classification.

We divide the dataset into two disjoint subsets based on eye level: training dataset (307 eyes) and testing dataset (130 eyes). The training and testing datasets do not contain AS-OCT images from the same eye. The AS-OCT images in the training and testing datasets are 5,551 and 2,368, respectively. 10% training dataset is used as validation dataset. Table 1 shows the NC severity level distribution on AS-OCT images. For data augmentation, we use random flipping (horizontal and vertical directions) and rotating (-20-20 degrees) method for the training dataset. We then normalize AS-OCT images with channels' means and standard deviations in the training by following the standard practice. We only normalize AS-OCT images with channels' means and standard deviations for the testing dataset. We resize AS-OCT images into 224x224 for both training and testing datasets.

Table 1 Severity level distribution of nuclear cataract on AS-OCT image dataset

Dataset	Severity Levels		
	Mild	Moderate	Severe
Training	1969	2185	1397
Testing	949	978	441
Total	2918	3163	1838

USUD dataset

It is an OCT image dataset of diabetic macular edema and age-related macular degeneration (AMD), which is collected and released by University of California, San Diego. The dataset comprises two sub-datasets: training and testing datasets. training dataset has 108,312 images: 37,206with choroidal neovascularization (CNV); 11,349 with diabetic macular edema (DME); 8,617 with drusen, and 51,140 normal. The testing dataset has 1000 images, and each class has the same number of images (250 images)-the more detailed introduction of the USUD dataset in [32]. In the experiments, we adopt the same dataset split method in [32].

Evaluation measures

Four commonly used evaluation measures are considered to evaluate the performance of methods [33, 34]: Accuracy (ACC), precision (PR), sensitivity (Sen), and F1 score, which are formulated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

$$Sen = \frac{TP}{TP + FN}, \tag{6}$$

$$PR = \frac{TP}{TP + FP}, \tag{7}$$

$$F1 = \frac{2 \times PR \times Sen}{PR + Sen}, \tag{8}$$

where TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively. **ACC:** the total number of AS-OCT images include mild, moderate, and severe are classified correctly. **F1:** F1 score is the harmonic mean of PR and

Sen, which is a significant indicator of assessing overall performance.

Sen assesses how many TP's AS-OCT images are classified correctly, which is a vital clinical diagnosis indicator.

Baselines

To evaluate the overall performance of our MPANet on two datasets thoroughly, we conduct the following baselines:

- Advanced attention methods: CBAM, efficient channel attention (ECA) [20], GCA, SPA, and SE are used to demonstrate the effectiveness of our MPA.
- State-of-the-art CNN models: GoogleNet, EfficientNet [35], SKNet [36], VGGNet, ResNeXt, GraNet, and BAM.
- Classical machine learning methods. We extracted eight clinical features from the lens nucleus region of AS-OCT images according to previous works [37, 16]: mean density, maximum density, entropy, intensity range, variance, skewness, absolute mean deviation, and median. Then, we use classical machine learning methods to classify NC's severity levels based on extracted features, like Naive Bayes (NB), decision tree (DT), random forest (RF), support vector machine (SVM), linear regression (LR), Adaboost, and XGboost.

Experiment setting

We implement our MPANet, its variants, comparable deep networks with the Pytorch platform, OpenCV, and Python. This paper uses the stochastic gradient descent (SGD) optimizer to optimize all deep networks and sets the SGD optimizer with a weight decay of 0.0005 and a momentum of 0.9. The batch size is 16, and training epochs are 100. We set the initial learning rate to 0.025 and decreased it by a factor of 5 every 20 epochs. All deep networks train from scratch. We conduct all experiments on a server with one TITAN V GPU (11 GB).

Results and analysis

Performance comparison with state-of-the-art attention blocks

Table 2 presents the NC classification results of the proposed MPA, its three variants, and state-of-the-art attention methods by using the same backbones (ResNet18 and ResNet34). Note that the base models denote ResNet18 and ResNet34. The results show that our MPA consistently improves the performance through comparisons to other attention methods. Remarkably, MPA outperforms SPA, GCA, and SE with **1.85%**, **3.66%** and **1.61%**, respectively. The results demonstrate that the effectiveness of our proposed MPA by infusing the clinical prior knowledge.

For comparison between MPA and its three variants, MPA and MPA-C get better NC classification results than MPA-A and MPA-B, verifying that the multi-scale pyramid pooling structure can extract different feature representation types and global-local feature representations efficiently. MPA outperforms MPA-C, which confirms that the group convolution method can improve classification performance by enhancing feature representation difference from feature maps.

Overall, the results demonstrate that MPA is more able to get better performance than advanced attention methods and strong backbones by considering the global-local feature representation and feature representation types with group convolution method and mixed pyramid pooling structure. Interestingly, we observe that not all attention methods achieve better performance by taking ResNet34 as the backbone than taking ResNet18 as the backbone. One possible reason to account for the results is that the number of parameters in ResNet34 is much more than ResNet18; thus, it needs massive data to train a good CNN model. However, available AS-OCT images of NC classification are limited, since it is challenging to collect massive medical data.

Performance comparison with strong baselines

We compare our MPANets with state-of-the-art deep networks and classical machine learning methods based on four evaluation measures, as shown in Table 3. The results

Table 2 Performance comparison of our MPA and state-of-the-art attention methods on the AS-OCT image dataset (The best results are marked in bold)

	Base	SE	CBAM	GCA	SPA	ECA	MPA	MPA-B	MPA-C	MPA
ResNet18	82.94	85.09	84.54	83.66	85.22	85.09	85.77	85.90	86.40	**86.70**
ResNet34	83.78	85.47	84.25	83.36	85.14	84.88	86.19	86.23	86.61	**86.99**

show that our MPANet get better classification results than advanced deep networks and machine learning methods. Specifically, MPA-D-Net-34 gets the best accuracy, the best sensivity, and the best F1 score with **86.99%** and **89.09%**, and **88.70%** respectively. It outperforms vanilla CNN models: ResNet, VGGNet, ResNeXt, and GoogleNet, above absolute **3.21%**, **2.61%**, **2.23%**, and **4.51%** of accuracy. Compared with attention-based CNNs, our MPA-D-Net-34 obtains 1.91% and **2.72%** absolute improvements of sensitivity than SENet34 and SPANet34 correspondingly. The results prove that combining the group convolution method with a mixed pyramid pooling structure is an efficient method for devising the attention module. This is because the MPA module can capture two feature representation types and local-global feature representations by introducing clinical prior knowledge.

Table 3 also presents NC classification results of seven machine learning methods on the AS-OCT image dataset. RF obtains the best performance through comparison to other machine learning methods. Our MPANet surpasses RF by noticeable gains of **4.39%** in the accuracy, **3.93%** in precision, **4.45%** in the sensitivity, and **4.18%** in F1, showing the superiority of our proposed method. Deep networks achieve better performance than machine learning methods. The GraNet achieves 84.56% accuracy and outperforms GraNet [15] (57.86%) by a significant improvement of 26.7%, demonstrating that NC severity levels are only associated with the nucleus region rather than the crystalline region lens region, which is also consistent with clinical research.

Table 3 NC classification results of machine learning methods and deep learning methods on AS-OCT image dataset (The best results are marked in bold)

Method	ACC	PR	Sen	F1
SVM [16]	82.19	83.64	85.05	84.29
NB	80.62	81.28	84.17	82.22
LR	81.80	85.36	82.65	83.84
DT	80.62	83.03	83.01	82.98
RF [16]	82.60	84.49	84.64	84.52
Adaboost	75.55	83.53	79.52	76.84
XGboost	82.18	84.45	84.18	84.31
GraNet [15]	85.05	85.66	87.25	86.37
VGG19	84.38	86.41	85.70	85.94
ResNet34	83.78	85.57	86.02	85.71
ResNeXt29	84.76	87.6	85.92	86.63
GoogleNet	82.48	86.28	84.21	84.03
EfficientNet	84.42	86.12	86.30	86.11
SKNet	85.68	88.22	86.78	87.32
SENet34	85.47	87.44	87.18	86.83
BAM	83.19	86.66	84.79	85.07
GCA-Net-18 [17]	83.66	85.97	84.68	85.26
ECANet-18	85.09	86.45	86.32	86.38
SPANet-18	85.22	88.13	86.37	86.92
CBAM-ResNet18	84.54	87.20	85.88	86.39
MPANet-18-C	86.40	**89.02**	88.78	88.02
MPANet-34-C	86.61	88.07	89.02	88.31
MPANet-18	86.70	88.26	89.06	88.59
MPANet-34	**86.99**	88.42	**89.09**	**88.70**

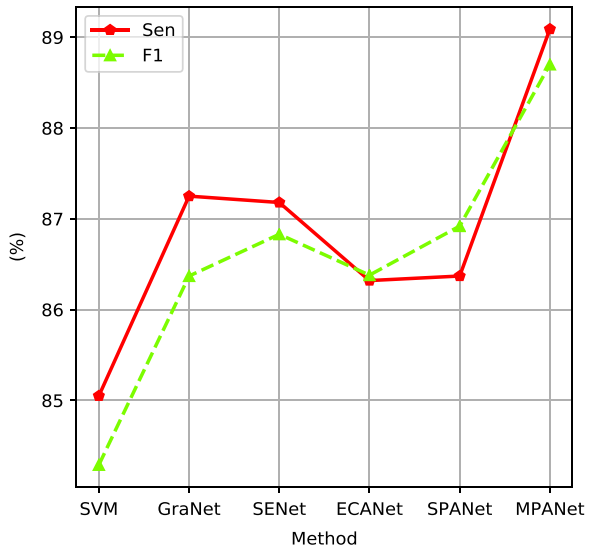


Fig. 4 Performance comparison of our MPANet and strong baselines in sensitivity and F1

Furthermore, Fig. 4 visually presents sensitivity and F1 values of our MPANet and other five strong baselines: SVM, GraNet, SENet, ECANet, and SPANet. The horizontal axis represents our MPANet and comparable methods, and the vertical axis represents the values of sensitivity and F1. As previously introduced, sensitivity (red color) is a vital evaluation indicator clinically, and F1 (green color) is a commonly used evaluation indicator to evaluate the general performance of a method. According to Fig. 4, it can be seen that our MPANet significantly surpasses other strong baselines, proving the efficacy of method by incorporating clinical prior knowledge. To better understand the NC classification results of our MPANet, Fig. 5 shows the confusion matrix of it. The horizontal and vertical axes represent predicted results and ground truth, respectively. According to Fig. 5, sensitivity values of mild NC, moderate NC, and severe NC are 89.57%, 79.45%, and **98.19%** based on MPANet accordingly, showing it is challenging to predict moderate NC accurately as well as for clinical diagnosis. Our MPANet obtains 86.03%, 83.46%, and **96.44%** for mild NC, moderate NC, and severe NC in F1 score, respectively,

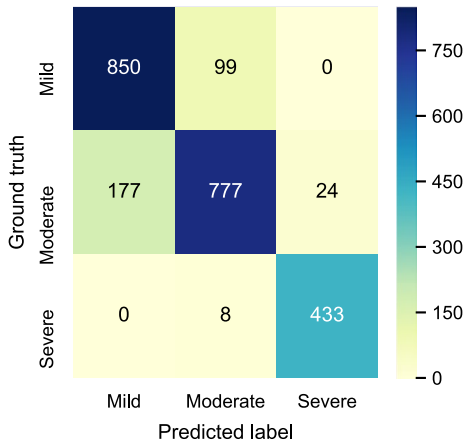


Fig. 5 The confusion matrix of MPANet

Table 4 Comparison of with different combinations based on the MPA when taking ResNet34 (The best results are marked in bold)

Method	GC	GMP	GAP	ACC
MPA-1	✗	✗	✓	86.28
MPA-2	✓	✗	✓	86.36
MPA-3	✗	✓	✗	85.98
MPA-4	✓	✓	✗	86.11
MPA-5	✗	✓	✓	86.74
MPA	✓	✓	✓	**86.99**

demonstrating the general performance of MPANet is good. Furthermore, we can get the kappa coefficient value of MPANet based on the confusion matrix, and the kappa coefficient is a vital indicator to assess diagnostic reliability. The kappa coefficient value of our MPANet is 0.7955, proving that it exhibits high reliability of NC diagnosis.

In this paper, AS-OCT images used for automatic classification, collected from NC participants with varying severity levels, and there are no AS-OCT images from normal participants without opacity. Hence, our proposed MPANet cannot be used for NC screening directly and only can be used for clinical diagnosis. In the future, we will plan to collect AS-OCT images from normal participants further to test the robustness and generation of our method.

Ablation study

Effects of different combinations

To further test which factors affect the performance of the MPA block, we conduct a number of ablation experiments, as shown in table 4. GC, GMP, and GAP represent group convolution method, global max pooling method, and global average pooling method. ✗ denotes that we do not use GC, GMP, or GAP in the MPA block, while ✓ represents we use GC, GMP, or GAP in the MPA block.

MPA-5 outperforms MPA-2 and MPA-4, indicating that the mixed pyramid pooling structure has a more significant effect on the NC classification results than the group convolution method. MPA-1 achieves better performance than MPA-3, showing GAP is more capable of learning important feature representation information than GMP, agreeing with clinical research. MPA achieves

Table 5 Comparisons of different gating operators based on the MPA block when taking ResNet34 (The best results are marked in bold)

Operator	ACC	F1
Tanh	85.73	87.48
Softmax	84.92	86.77
Sigmoid	**86.99**	88.70

Table 6 Comparisons of different r and M (The best results are marked in bold)

Ratio r	Dimensionality reduction d		
	8	**16**	**32**
8	85.73	86.19	84.76
16	85.64	85.30	**86.99**
32	86.15	85.81	85.56

Table 7 Performance comparison of the MPANet and state-of-the-art methods on USUD dataset (The best results are marked in bold)

Method	ACC	Sen	F1
LBP-SVM [38]	71.33	48.27	64.04
HOG-SVM [39]	78.90	66.20	–
MDFF [38]	93.93	91.76	91.46
VGG16 [38]	91.50	91.50	91.50
ResNet34 [32]	80.50	78.30	–
Inception [39]	90.30	90.00	–
LACNN [39]	90.20	88.10	–
LACNN-Inception [39]	93.00	91.60	–
SENet	94.16	90.00	91.49
ECA	94.40	91.83	92.08
SPA-Net	94.11	89.83	91.32
GCA-Net	94.94	92.12	92.79
BAM	94.89	91.95	92.69
CBAM	94.20	89.74	91.30
MPANet	**96.74**	**95.12**	**95.39**

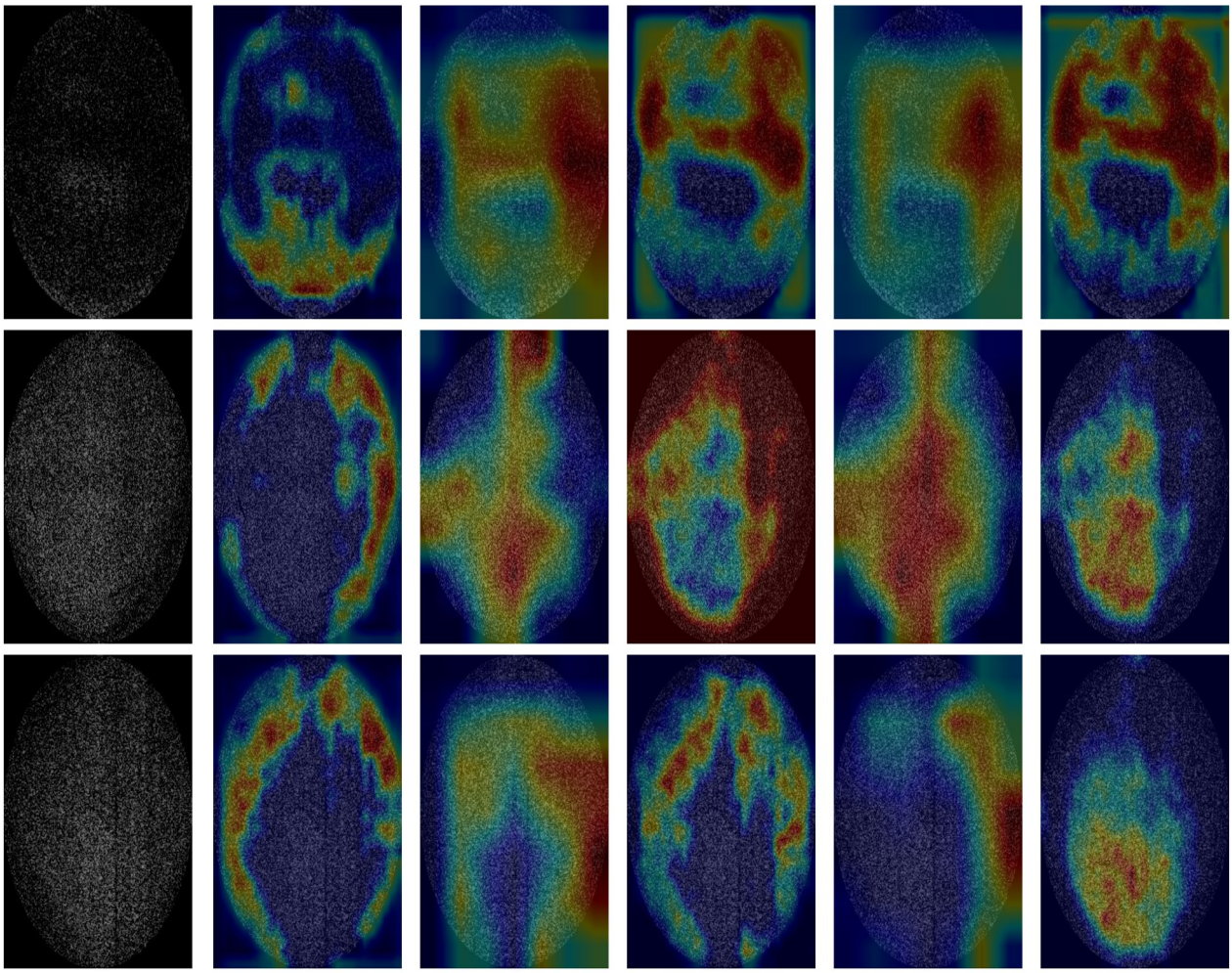


Fig. 6 The CAM visualization results of our proposed MPANet and other state-of-the-art attention-based CNNs. Row 1 to row 3 denotes the mild NC, moderate NC, and severe NC. The heat maps highlight the informative regions that CNNs learned for specific NC severity levels

better NC classification performance than MPA-5, demonstrating that GC, GMP, and GAP can boost the classification performance as previously discussed in Table 2.

Effects of different gating operators

Table 5 shows the classification results of three gating operators based on the MPA block. It can be seen that replacing sigmoid with tanh and softmax slightly worsens the performance of MPANet. The comparable results suggest that it is significant to design the gating operator, which is capable of highlighting useful channels efficiently.

Effects of dimensionality reduction d

Dimensionality reduction d is a vital factor to affect the performance and the computational cost of our MPANet, which are determined by two significant hyper-parameters: r and M. We conduct a series of experiments by setting different combinations of M and r for investigating the trade-off between the performance and the computational cost mediated by these two hyper-parameters, as shown in Table 6. It can be observed that increased/decreased complexity does not improve/worsen the performance of the MPANet. We set M and r to 32 and 16, respectively, and our method keeps a good trade-off between accuracy and complexity. In fact, using the identical M and r for different layers of a network may not be an optimal method

considering the varying roles of different layers played. Thus, further improvements can be obtained by tuning the M and r to meet the needs of a CNN architecture.

Validation on USUD dataset

We also compare our MPANet with advanced attention-based CNNs and previous works, as shown in Table 7. It can be seen that our MPANet gets **96.74%** accuracy, **95.12%** sensitivity, and **95.39%** F1, respectively, and significantly outperforms other comparable methods above absolute 1.8% at least on three evaluation measures. The results prove the generation ability of MPANet.

Visualization of improved interpretability

Figure 6 presents the CAM visualization results of our MPANet and other four state-of-the-art attention methods on the AS-OCT image dataset. It offers the three representative AS-OCT images of three NC severity levels and their CAM visualization results. We can see that our method is more capable of localizing opacity information of NC on AS-OCT images through comparisons to other attention-based CNNs. For example, our proposed MPANet pays more attention to the center- and bottom- nucleus regions for moderate and severe NC levels, agreeing with the conclusion of WHO Cataract Grading System [40] which suggests that clinicians should focus on the center- and bottom- nucleus regions in diagnosing NC. Overall, visualization results also explain why our method performs better than other attention-based CNNs, e.g., SENet.

Conclusion and future work

This paper presents an effective mixed pyramid attention network (MPANet) to predict severity levels of NC by using AS-OCT images automatically. In the MPANet, we design a mixed pyramid attention block for learning different feature representation types and local-global feature representations with the group convolution method and the mixed pyramid pooling structure. We conduct experiments on a clinical AS-OCT dataset, and the results show that our MPANet achieves 86.99% in accuracy and 89.09% in sensitivity accordingly, which outperforms previous methods and strong baselines. Moreover, we also conduct extensive experiments on a public OCT dataset, and MPANet also gets better performance than state-of-the-art methods, demonstrating its generation ability. Overall, our MPANet has great potential for clinical nuclear cataract diagnosis and cataract surgery planning on AS-OCT images.

In the future, we plan to collect more AS-OCT images to evaluate the performance of the MPANet from both NC and normal participants; it is because we only use the AS-OCT images from NC participants in this paper. Moreover, we will design lightweight and advanced attention mechanisms to enhance the deep network's interpretability and boost the classification results.

Acknowledgements

This work was supported in part by Guangdong Provincial Department of Education (2020ZDZX3043, SJJG202002), Guangdong Provincial Key Laboratory (2020B121201001), and Shenzhen Natural Science Fund (JCYJ20200109140820699) and the Stable Support Plan Program (20200925174052004).

Author details

[1]Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen 518055, China. [2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China. [3]School of Ophthalmology and Optometry, Wenzhou Medical University, Wenzhou 325035, China. [4]State Key Laboratory of Ophthalmology, Sun Yat-sen University, Guangzhou 510060, China. [5]Present Address: Tomey Corporation, Nagoya, Japan. [6]Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Southern University of Science and Technology, Shenzhen 518055, China.

Received: 21 January 2022 Accepted: 4 March 2022
Published online: 25 March 2022

References

1. Burton MJ, Ramke J, Marques AP, Bourne RRA, Faal HB. The lancet global health commission on global eye health: vision beyond 2020. Lancet Glob Health. 2021;9(4):e489–551.
2. Liu YC, Wilkins M, Kim T, Malyugin B, Mehta JS. Cataracts. Lancet. 2017;390(10094):600–12.
3. Chylack LT, Wolfe JK, Singer DM, Leske MC, Bullimore MA, Bailey IL. The lens opacities classification system iii. Arch Ophthalmol. 1993;111(6):831–6.
4. Dos Santos VA, Schmetterer L, Stegmann H, Pfister M, Messner A, Schmidinger G, Garhofer G, Werkmeister RM. Corneanet: fast segmentation of cornea oct scans of healthy and keratoconic eyes using deep learning. Biomed Opt Express. 2019;10(2):622–41.
5. Keller B, Draelos M, Tang G, Farsiu S, Kuo AN, Hauser K, Izatt JA. Real-time corneal segmentation and 3d needle tracking in intrasurgical oct. Biomed Opt Express. 2018;9(6):2716–32.
6. Fu H, Baskaran M, Xu Y, Lin S, Wong DWK, Liu J, Tun TA, Mahesh M, Perera SA, Aung T. A deep learning system for automated angle-closure detection in anterior segment optical coherence tomography images. Am J Ophthalmol. 2019;203:37–45.
7. Fu H, Li F, Sun X, Cao X, Liao J, Orlando JI, Tao X, Li Y, Zhang S, Tan M, et al. Age challenge: angle closure glaucoma evaluation in anterior segment optical coherence tomography. Med Image Anal. 2020;66:101798.
8. Fu, H, Xu, Y, Lin, S, Wong, D.W.K, Mani, B, Mahesh, M, Aung, T, Liu, J. Multi-context deep network for angle-closure glaucoma screening in anterior segment oct. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 356–363. Springer (2018)
9. Xu, C., Zhu, X, He, W, Lu, Y, Li, X. Fully deep learning for slit-lamp photo based nuclear cataract grading. In: MICCAI (2019)
10. Wong AL, Leung CKS, Weinreb RN, Cheng AKC, Cheung CYL, Lam PTH, Pang CP, Lam DSC. Quantitative assessment of lens opacities with anterior segment optical coherence tomography. Br J Ophthalmol. 2009;93(1):61–5. https://doi.org/10.1136/bjo.2008.137653.
11. de Castro A, Benito A, Manzanera S, Mompeán J, Canizares B, Martínez D, Marín JM, Grulkowski I, Artal P. Three-dimensional cataract crystalline lens imaging with swept-source optical coherence tomography. Invest Ophthalmol Vis Sci. 2018;59(2):897–903.
12. Grulkowski I, Manzanera S, Cwiklinski L, Mompeán J, De Castro A, Marin JM, Artal P. Volumetric macro-and micro-scale assessment of crystalline lens opacities in cataract patients using long-depth-range swept source optical coherence tomography. Biomed Opt Express. 2018;9(8):3821–33.

13. Makhotkina NY, Berendschot TT, van den Biggelaar FJ, Weik AR, Nuijts RM. Comparability of subjective and objective measurements of nuclear density in cataract patients. Acta Ophthalmol. 2018;96(4):356–63.
14. Wang, W, Zhang, J, Gu, X, Ruan, X, Liu,Y. Objective quantification of lens nuclear opacities using swept-source anterior segment optical coherence tomography. Br J Ophthalmol: bjophthalmol–2020–318334 (2021)
15. Zhang, X, Xiao, Z, Risa, H, Chen, W, Yuan, J, Fang, J, Hu, Y, Liu, J. A novel deep learning method for nuclear cataract classification based on anterior segment optical coherence tomography images. In: IEEE SMC (2020).
16. Zhang, X, Fang, J, Xiao, Z, Risa, H, Chen, W, Yuan, J, Liu, J. Research on classification algorithms of nuclear cataract based on anterior segment coherence tomography image. Comput Sci. https://doi.org/10.11896/jsjkx.201100085 (2022).
17. Xiao Z, Zhang X, Higashita R, Hu Y, Yuan J, Chen W, Liu J. Gated Channel Attention Network for Cataract Classification on AS-OCT Image. InInternational Conference on Neural Information Processing 2021 (pp. 357–368). Springer, Cham
18. Chen D, Li Z, Huang J, Yu L, Liu S, Zhao YE. Lens nuclear opacity quantitation with long-range swept-source optical coherence tomography: correlation to locs iii and a scheimpflug imaging-based grading system. Br J Ophthalmol. 2019;103(8):1048–53.
19. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: TPAMI (2018)
20. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. Eca-net: efficient channel attention for deep convolutional neural networks, 2020 IEEE. IEEE: In CVPR (2020)
21. Guo, J, Ma, X, Sansom, A, McGuire, M, Kalaani, A, Chen, Q, Tang, S, Yang, Q, Fu, S. Spanet: Spatial pyramid attention network for enhanced image recognition. In: ICME, pp. 1–6. IEEE (2020)
22. Wang, F, Jiang, M, Qian, C, Yang, S, Li, C, Zhang, H, Wang, X, Tang, X. Residual attention network for image classification. In: CVPR, pp. 3156–3164 (2017)
23. Woo, S, Park, J, Lee, JY, Kweon IS. Cbam: Convolutional block attention module. In: ECCV, pp. 3–19 (2018)
24. Park J, Woo S, Lee JY, Kweon IS. A simple and light-weight attention module for convolutional neural networks. IJCV. 2020;128(4):783–98.
25. Xie, S, Girshick, R, Dollár, P, Tu, Z, He, K. Aggregated residual transformations for deep neural networks. In: CVPR, pp. 1492–1500 (2017)
26. Zhang T, Qi GJ, Xiao B, Wang J. Interleaved group convolutions. In: CVPR, pp. 4373–4382 (2017)
27. Zhang, X, Zhou, X, Lin, M, Sun, J. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: CVPR, pp. 6848–6856 (2018)
28. Lin, M, Chen, Q, Yan, S. Network in network. arXiv preprint arXiv:1312.4400 (2013)
29. Zhang X, Zhao H, Zhang S, Li R. A novel deep neural network model for multi-label chronic disease prediction. Front Genet. 2019;10:351.
30. He, K, Zhang, X, Ren, S, Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
31. Cao G, Zhao, W, Higashita, R,Liu, J, Yang, M. An efficient lens structures segmentation method on as-oct images. In: EMBC; 2020
32. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122–31.
33. Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. Comput Methods Programs Biomed. 2020;196: 105608.
34. Zhang H, Niu K, Xiong Y, Yang W, He Z, Song H. Automatic cataract grading methods based on deep learning. Comput Methods Programs Biomed. 2019;182: 104978. https://doi.org/10.1016/j.cmpb.2019.07.006.
35. Tan, M, Le,Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
36. Li, X, Wang, W, Hu, X, Yang, J. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519 (2019)
37. Li H, Lim JH, Liu J, Mitchell P, Tan AG, Wang JJ, Wong TY. A computer-aided diagnosis system of nuclear cataract. IEEE Trans Biomed Eng. 2010;57(7):1690–8.
38. Das V, Dandapat S, Bora PK. Multi-scale deep feature fusion for automated classification of macular pathologies from oct images. Biomed Signal Process Control. 2019;54: 101605. https://doi.org/10.1016/j.bspc.2019.101605.
39. Fang L, Wang C, Li S, Rabbani H, Chen X, Liu Z. Attention to lesion: lesion-aware convolutional neural network for retinal optical coherence tomography image classification. IEEE Trans Med Imaging. 2019;38(8):1959–70.
40. Thylefors B, Chylack Jr LT, Konyama K, Sasaki K, Sperduto R, Taylor HR, West4 S. A simplified cataract grading system The WHO Cataract Grading Group. Ophthalmic Epidemiol. 2002;9(2):83–95