Image and Vision Computing 146 (2024) 105042

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

AES-Net: An adapter and enhanced self-attention guided network for multi-stage glaucoma classification using fundus images

Dipankar Das [a], Deepak Ranjan Nayak [a,*], Ram Bilas Pachori [b]

[a] Department of Computer Sceince and Engineering, Malaviya National Institute of Technology Jaipur, India
[b] Department of Electrical Engineering, Indian Institute of Technology Indore, India

ARTICLE INFO

Keywords:
Multi-stage glaucoma classification
Fundus image
Spatial-adapter module
Enhanced self-attention module (ESAM)
AES-Net

ABSTRACT

Glaucoma is a progressive eye condition that can lead to permanent vision loss. Therefore, on-time detection of glaucoma is critical for making an effective treatment plan. In recent years, enormous attempts have been made to develop automated glaucoma classification systems using CNNs through images. In contrast, limited methods have been proposed for diagnosing different glaucoma stages. It is mainly owing to the lack of large publicly available labeled datasets. Also, fundus images exhibit a high inter-stage resemblance, redundant features and minute size variations of lesions, making the conventional CNNs difficult to classify multiple stages of glaucoma accurately. To address these challenges, this paper proposes a novel adapter and enhanced self-attention based CNN framework named AES-Net for effective classification of glaucoma stages. In particular, we propose a spatial adapter module on top of the backbone network for learning better feature representations and an enhanced self-attention module (ESAM) to capture global feature correlations among the relevant channels and spatial positions. The ESAM assists in capturing stage-specific and detailed-lesion features from the fundus images. Extensive experiments on two multi-stage glaucoma datasets indicate that our AES-Net surpasses CNN-based existing approaches. The Grad-CAM++ visualization maps further confirm the effectiveness of our AES-Net.

1. Introduction

Glaucoma is one of the leading causes of irreversible blindness among individuals across the globe [1]. It is a chronic eye disorder that harms the optic nerve and is occurred due to the escalation of intraocular pressure [2]. Therefore, the clinical indications of this disorder are mostly evident in the optic nerve regions. The number of glaucoma patients was around 64.3 million in the year 2013 and is estimated to be 113 million by 2040 [3,4]. It has been observed that glaucoma affects individuals within the age range of 40 to 80. Early detection and on-time treatment can avoid blindness caused due to glaucoma. However, most glaucoma patients are oblivious about their disorders during the early stage due to the absence of noticeable signs [1]. Therefore, it is highly recommended for a regular eye checkup to detect glaucoma at an early stage and for better medical care, thereby averting the loss of eyesight [5]. Fig. 1 illustrates the impact of the vision loss at different glaucoma stages. Glaucoma, at first, results in peripheral vision loss and finally leads to blindness if left untreated.

The cup-to-disc ratio has been widely considered as a gold standard metric to diagnose glaucoma from fundus images. A high cup-to-disc ratio is regarded as an indicator of glaucoma. Therefore, early detection of glaucoma necessitates the use of qualified specialists [1]. These processes are manual, time-consuming, and laborious. Furthermore, due to a shortage of proficient ophthalmologists, it is difficult to carry out glaucoma testing for all suspicious persons. Hence, it is paramount to develop an automated diagnosis system that uses fundus images to detect distinct stages of glaucoma to reduce the burden on ophthalmologists and provide a fast and accurate decision.

The initial investigations on glaucoma identification involved the segmentation of optic cup and disc regions and subsequent evaluation of the CDR values [6–8]. The efficacy of these CDR evaluation-based methods is heavily reliant on segmentation accuracy, and they frequently encounter poor sensitivity issues. Thereafter, a plethora of ML-based automated methods have been proposed for glaucoma screening that combines conventional classifiers like support vector machine (SVM), naive Bayes (NB), etc., with various handcrafted features like wavelet energy, Gabor entropy, multiresolution features, wavelet correntropy features, etc. [5,9–12]. A few such multi-stage ML-based methods have recently been applied for diabetic retinopathy classification through fundus images [13,14]. Even though these

* Corresponding author.
E-mail addresses: 2022rcp9028@mnit.ac.in (D. Das), drnayak.cse@mnit.ac.in (D.R. Nayak), pachori@iiti.ac.in (R.B. Pachori).

https://doi.org/10.1016/j.imavis.2024.105042
Received 2 February 2024; Received in revised form 22 March 2024; Accepted 18 April 2024
Available online 24 April 2024

methods resulted in satisfactory performance, the major challenges lies in manually choosing the feature extraction methods and classifiers.

To address these problems, numerous efforts have been made by researchers toward developing automated methods using deep learning (DL) techniques, specifically CNN, due to its capability in learning effective feature representations [15–18]. Chen et al. [15] reported a CNN architecture, consisting of four convolutional followed by two linear layers. In [19], fusion of holistic and local features extracted from pre-trained CNN models was used for glaucoma screening. In [1], a multi-stream CNN was proposed that extracts features from the areas of the local disc region as well as the global fundus image for improved classification performance. Pal et al. [20] designed an architecture dubbed as G-EyeNet with a deep convolutional auto-encoder network for glaucoma screening. In [16], an end-to-end learnable customized CNN architecture was presented, whereas in [21], a multi-branch framework was developed using the features extracted from several DL frameworks and domain knowledge for enhanced detection performance. Phasuk et al. [22] developed an ensemble network to classify glaucoma, which essentially consists of two segmentation networks and three classification networks. In [17], the authors developed an InceptionV3-based framework for the classification of glaucoma. A compact evolutionary convolutional network using a real-codded GA was presented in [23] for glaucoma screening. However, it suffers from comprehensive end-to-end learning capability. Later, in [24], an end-to-end learnable model was proposed for optic disc and cup segmentation and classification. In [25] the efficacy of several state-of-the-art pre-trained CNN frameworks was verified for glaucoma classification. Juneja et al. [18] developed CoGNet, a deep CNN architecture based on Xception, to detect glaucoma in fundus images. These approaches have been primarily designed to classify fundus images as normal or glaucoma categories, and are unable to identify various stages of glaucoma.

Multi-stage glaucoma classification, on the other hand, is paramount and challenging due to large inter-stage resemblance, redundant information, and minute size variation of lesions in fundus images, as shown in Fig. 2. To this end, a scanty ML-based approaches have been presented for identifying glaucoma stages [2,26]. But, these approaches do not enjoy end-to-end learning and have been evaluated on balanced and small datasets. As a result, these methods may fail to adapt efficiently to unknown data, making their real-world application more difficult. The CNN-based methods can effectively deal with these challenges, however, have not yet been largely explored for multi-stage glaucoma classification. In [27], a customized CNN was developed and validated on a larger multi-stage dataset. Such CNNs might fail to capture the detailed lesion information from fundus images. To deal with these problems, attention techniques have been integrated with CNN frameworks, enabling the network to focus on vital features while suppressing unimportant information [28–35]. Tian et al. [36] developed GC-Net, an attention-based CNN framework, by combining a class attention block (CAB) with a global attention block (GAB) for classification of glaucoma stages. In [37], a global self-attention guided network named GS-Net was proposed to enable learning of prominent features from fundus images of glaucoma stages. Recently, Das et al. [38] proposed CA-Net based on cascaded attention and validated it on two multi-stage and a

Fig. 2. Illustration of fundus images of (a) normal (b) early-stage, and (c) advanced-stage glaucoma patients.

binary glaucoma classification datasets. Although these attention-based CNN models outperformed classical CNN models, they still lack of capturing prominent feature correlations, which are critical for improving classification performance. Further, they lead to high computational overhead.

To address these challenges, the primary objective of this study is to develop an advanced deep-learning framework that can accurately classify different stages of glaucoma using fundus images. We present a novel adapter and enhanced self-attention guided CNN (AES-Net) for the classification of glaucoma stages. The spatial adapter module helps to adjust the spatial representation for better learning, while the attention module facilitates in emphasizing prominent lesion features while ignoring extraneous information in the fundus images. The experimental results and comparisons indicate that the proposed AES-Net surpasses the recently proposed approaches in classification performance.

1.1. Contribution

The main contributions of this paper can be outlined as follows:

- We propose a spatial adapter and enhanced self attention-guided CNN dubbed as AES-Net for effective multi-stage glaucoma classification using fundus images.
- We introduce a spatial adapter module on top of the pretrained CNN that facilitates adapting the network to the glaucoma classification task by adjusting the spatial representation, thereby enhancing the generalization capability of the network.
- We introduce an enhanced self attention module (ESAM) just after the adapter module to selectively extract more detailed-lesion features from fundus images. It includes a channel self-attention module and spatial self-attention module that are connected in a parallel fashion to capture global feature correlations along channel and spatial dimensions, thereby improving the overall performance. The combination of the adapter and ESAM is termed as AESAM in our work.
- We validate our AES-Net on two multi-stage glaucoma datasets. To verify the model's effectiveness, we conduct ablative experiments and comparisons with cutting-edge attention mechanisms. Further, we derive a comparative analysis with a set of pre-trained CNN

Fig. 1. Illustration of impact on the vision of (a) normal (b) early-stage, and (c) advanced-stage glaucoma patients.

architectures and state-of-the-art (SOTA) glaucoma classification approaches.

By addressing the challenges above, the study intends to achieve state-of-the-art multi-stage glaucoma classification performance utilizing advanced deep learning techniques, with potential application in medical practice and public health. The target audience includes researchers and professionals working in medical imaging, machine learning, and artificial intelligence, as well as practitioners in the field of ophthalmology. Furthermore, those interested in healthcare technology development would benefit from knowing about the advances made by AES-Net in diagnosing glaucoma stages.

2. Proposed methodology

The overall framework of our proposed AES-Net is shown in Fig. 3. The AES-Net consists of a backbone network, a spatial-adapter, an ESAM, and a classifier. The backbone network is used to extract high-level feature representation. The spatial adapter module aids to adapt the network to the target task. The ESAM incorporates channel self-attention and a spatial self-attention that learns global feature correlations among channels and spatial positions, facilitating the extraction of more detailed and discriminative features from the fundus images. Subsequently, an average pooling and a FC layer are used to perform classification of glaucoma stages. In this section, we provide a detailed discussion on our proposed AES-Net and its basic building blocks.

2.1. Backbone

As illustrated in Fig. 3, we provide fundus images as input to a backbone network, a pretrained CNN which can be any ImageNet pre-trained CNN architecture. The backbone network generates hierarchical features. The feature maps from the last convolution layer of the backbone network are then passed to the spatial-adapter module.

2.2. Spatial adapter module

While utilizing ImageNet pre-trained CNN architecture for transfer learning, the inclusion of an adapter module aids in fine-tuning the model on downstream tasks. It allows the model to adjust its spatial representations as per the requirements of the target task while keeping the information gained during the pre-training phase, thereby increasing the efficacy of transfer learning. Motivated by the work reported in [39], we introduce a spatial adapter module in our network that provides additional flexibility by allowing the network to better generalize across the glaucoma datasets, enhancing the overall performance and robustness of the model. Additionally, it helps in capturing rich semantic information from fundus images. The spatial adapter module utilizes the output feature maps of the backbone network as input and its effectiveness is verified in Section 3.3.4. As shown in Fig. 4, it consists of a layer normalization followed by a depth-wise convolution layer placed between two 1×1 bottleneck layers. Further, the feature maps acquired

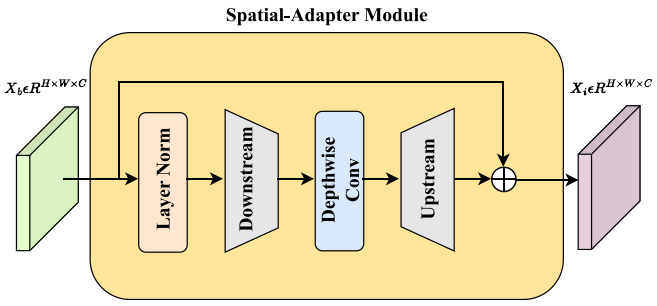

Fig. 4. Architecture of spatial adapter module.

after the subsequent operations are combined using a residual connection to obtain the output of spatial adapter module, $X_i \in \mathbb{R}^{H \times W \times C}$, which can numerically be expressed as follows:

$$adapter(X_b) = X_b \oplus Up(DWConv(Down(LN(X_b)))) \quad (1)$$

Where X_b represents the input feature map to the spatial adapter module, *LN* represents the layer normalization operation, *Down* denotes the downsample operation using 1×1 convolution layer, *DWConv* represents the depth-wise convolution operation, *Up* denotes the upsample operation using 1×1 convolution layer, $\oplus$ represents the addition operation. It is worth noting that all the convolution layers are followed by the Swish activation function.

2.3. Enhanced self-attention module (ESAM)

We introduce ESAM to our proposed framework to selectively learn rich discriminative and stage-specific features from lesion regions, leading to enhanced feature learning capabilities and improved performance. The design of ESAM is related to self-attention [40] and GSAM [37]. However, self-attention in [40] supports interactions between spatial positions only. In contrast, GSAM allows capturing interactions among all channels and possible spatial position pairs, leading to a high computational burden. Further, it involves interaction among unimportant positional elements. Hence to avoid these issues, we proposed ESAM which comprises of two self-attention modules, CSAM and SSAM as shown in Fig. 5.

2.3.1. Channel self-attention module (CSAM)

The channel self-attention is introduced to capture global channel interactions in an efficient manner within the feature map. The input tensor, $X_i \in R^{H \times W \times C}$, is supplied to each of the three paths in CSAM. In the first path, we apply global max pooling (GMP), a 1×1 Conv operation, and a reshape operation in sequence to generate a query tensor $q_c \in R^{1 \times C}$. While in the second path, we use GAP followed by a downsampling operation using 1×1 Conv operation to reduce the number of channels. The resultant tensor is then reshaped and transposed to acquire the key tensor $k_c \in R^{\dot{c} \times 1}$, where $\dot{c} = C/4$. Then we multiply the query and key tensors and then use the softmax function to generate the

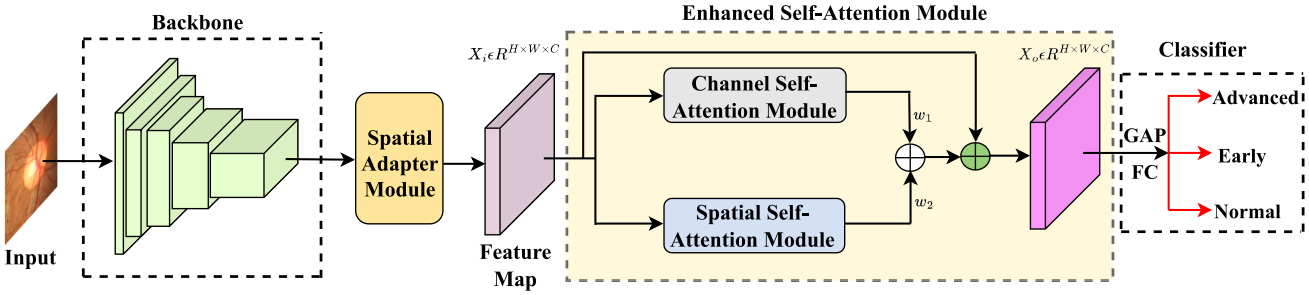

Fig. 3. Overall architecture of the proposed adapter and enhanced self-attention guided network (AES-Net) for automated classification of glaucoma stages.

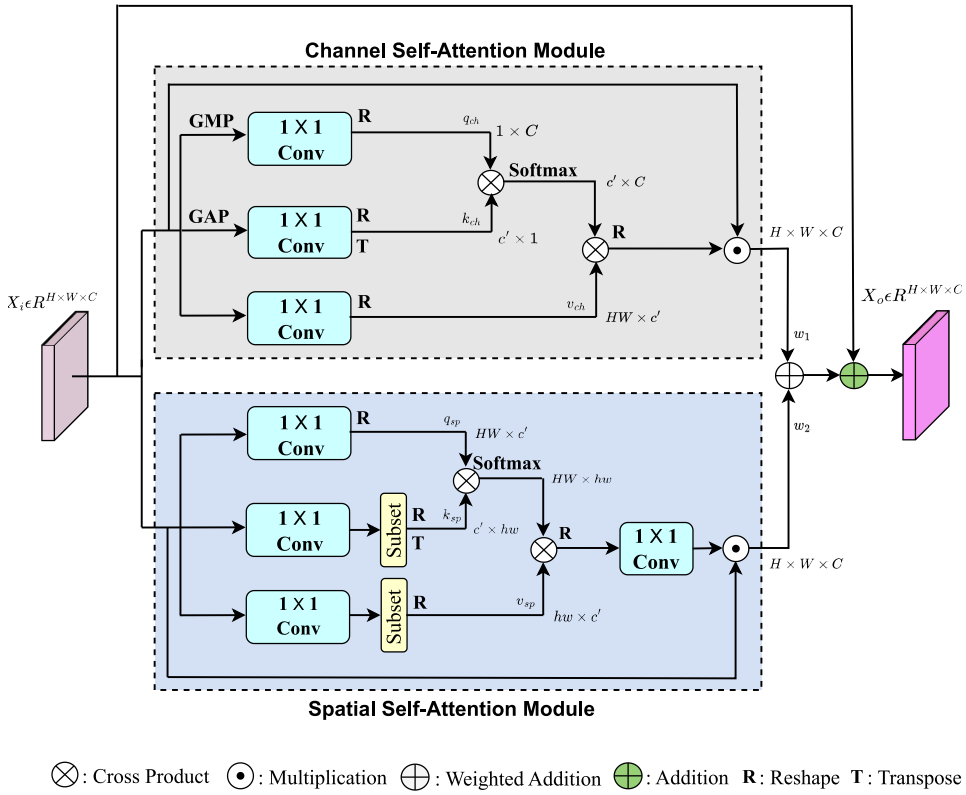

Fig. 5. Overview of the enhanced self-attention module (ESAM).

pair-wise channel attention weights $X_c^{'} \in R^{c^{'} \times C}$.

$$X_c^{'} = \sigma(k_c \otimes q_c) \tag{2}$$

where, $\otimes$ and σ represent the cross-product and softmax activation, respectively. Further, we harness 1×1 Conv operation over X_i and reshape the resultant tensor to acquire the value tensor $v_c \in R^{HW \times c^{'}}$.

$$X_c^{''} = v_c \otimes X_c^{'} \tag{3}$$

Next, v_c and $X_c^{'}$ are multiplied and reshaped back to a tensor $X_c^{''} \in R^{H \times W \times C}$. At last, the obtained tensor $X_c^{''}$ is again multiplied element-wise with the input feature map, X_i to attain the final channel attention map, X_c. Numerically, it can be expressed as follows:

$$X_c = X_c^{''} \odot X_i \tag{4}$$

where, $\odot$ represents element-wise multiplication. It is noteworthy that every 1×1 Conv block is followed by an LN and Swish activation.

2.3.2. Spatial self-attention module (SSAM)

The lesion size in fundus images of multi-stage glaucoma of patients varies significantly. The SSAM is introduced to capture the interaction between the relevant spatial positions, leading to less computational overhead. We give $X_i \in R^{H \times W \times C}$ as input to three parallel paths. The first path uses a 1×1 Conv operation of $c^{'}$ channels accompanied by a reshape operation to acquire the query tensor $q_s \in R^{HW \times c^{'}}$, where $c^{'} = C/4$. While in the second and third path, we utilize down-sampling operation to obtain subsets of key and value tensors of lower dimensions to reduce the number of operations. Hence, in the second path, the input is passed through a 1×1 Conv block of $c^{'}$ channels accompanied by a bilinear interpolation operation to obtain a reduced subset tensor of shape $h \times w \times c^{'}$. The obtained tensors are then reshaped and transposed to acquire the key tensor $k_s \in R^{c^{'} \times hw}$. After that, the obtained query and key tensors are multiplied accompanied by a softmax to generate the pair-wise spatial attention weight $X_s^{'} \in R^{HW \times hw}$.

$$X_s^{'} = \sigma(q_s \otimes k_s) \tag{5}$$

In the final path, a similar down-sampling is performed over X_i accompanied by a reshape operation to achieve the value tensor $v_s \in R^{hw \times c^{'}}$. Both v_s and $X_s^{'}$ are made to multiply and reshaped to $H \times W \times c^{'}$ which is followed by 1×1 Conv operation with C channels to get $X_s^{''} \in R^{H \times W \times C}$.

$$X_s^{''} = conv\left(X_s^{'} \otimes v_s\right) \tag{6}$$

At last, the obtained maps are element-wise multiplied with X_i to attain the global spatial attention feature map, X_s. Numerically it can be expressed as follows:

$$X_s = X_s^{''} \odot X_i \tag{7}$$

2.3.3. Fusion of attention feature maps

We combine the attention feature maps from two modules (CSAM and SSAM) using the weighted addition approach followed by a residual connection to attain the final attention feature map, X_o. The fusion of features enables the model to capture more category-specific features by focusing on prominent lesion areas and suppressing irrelevant parts, while the residual connection assists in a better flow of features within the network. Mathematically, it can be represented as follows:

$$X_o = X_i \oplus (w_1 X_c \oplus w_2 X_s) \tag{8}$$

Where w_1 and w_2 represent the trainable scalar weights.

2.4. Classifier

The classifier performs the classification of glaucoma stages by providing the attention feature maps through a GAP layer and an FC

layer with softmax activation.

3. Experiments and results

In this section, we present the datasets used, implementation details and results of proposed AES-Net along with state-of-the-art (SOTA) methods. We compare the efficiency of the ESAM with existing attention methods. Furthermore, we conduct ablation experiments and generate saliency maps using Grad-CAM++ to better understand the interpretability of our model.

3.1. Dataset used

Two datasets, Harvard Dataverse V1 (HDV1) and LMG have been used to evaluate AES-Net and other SOTA glaucoma classification approaches. The HDV1 dataset [27] comprises 1542 fundus images (786 normal, 289 early, and 467 advanced-stage), acquired at Kim's Eye Hospital, South Korea, using an AFC 330 non-mydriatic auto fundus camera. Later, a comparably larger dataset, LMG, was released in [38], consisting of 1582 fundus images (800 normal, 301 early, and 481 advanced-stage glaucoma). These images have been obtained from two different datasets, HDV1 and RIM-ONE r1 dataset [41]. The LMG dataset includes 40 fundus images (14 normal, 12 early, and 14 advanced-stage) from the RIM-ONE r1 dataset. The images of the LMG dataset are fixed at 240×240 pixels. Further, we utilized the RIM-ONE Extended [38] dataset to verify the efficacy of our AES-Net on a binary glaucoma classification setting, which consists of a total of 623 retinal fundus images (373 normal and 250 glaucoma).

3.2. Implementation details and performance metric

Training setting: To ensure a fair comparison, we adopt the data split approach for both datasets as followed in [38]. In particular, the train-test split constitutes 70–30% of the dataset, and the train-validation split includes 70–30% of the training set. The training set consists of 754 images, the validation set of 324 images, and the testing set of 464 images for the HDV1 dataset. Similarly, for the LMG dataset, the training set has 772 images, the validation set has 333 images, and the test set has 477 images. We resize the fundus images to 224×224 resolution. To reduce overfitting, various augmentation approaches, scaling, rotation, and vertical and horizontal flips, have been adopted. We initially set the hyperparameters such as the initial learning to 0.001 for the first two epochs of the pre-training phase and 0.0001 for the fine-tuning phase for 50 epochs with a decay factor of 0.8. Each model is trained with a categorical cross-entropy loss and Adam optimizer [42]. We set the batch size to 16 for all experimentation. The details of resource requirements are provided in Section 3.4. The proposed model has been developed on the Keras framework with TensorFlow backend. All of the experiments are performed on a workstation with an Intel Xeon processor, RAM of 128 GB, and an NVIDIA Tesla V100 graphic card with 32 GB of memory.

Several performance measures, namely accuracy (A_{cc}), precision (P_r), sensitivity (S_{en}), F1-Score, and area under the ROC curve (AUC), are taken into consideration to evaluate the performance of AES-Net.

3.3. Experimental results

This section provides the results obtained by the proposed AES-Net by adopting different backbones, comparison with glaucoma screening methods, and popular attention mechanisms. In addition, the visualization results of Grad-CAM++ and t-SNE are provided.

3.3.1. Comparison between SOTA CNN architectures

In the suggested framework, the spatial adapter and ESAM are integrated immediately after the backbone network. We select a variety of contemporary pre-trained CNN models, namely, VGG [43], InceptionV3 [44], ResNet-50 [45], MobileNet [46], Xception [47], EfficientNetB1 [48], and DenseNet-169 [49], as backbone networks and verify their effectiveness along with spatial adapter and ESAM. These models are trained using the transfer learning approach, where we replace the final layer with a customized FC layer of three nodes. Table 1 and Table 2 compare the classification results of all aforementioned backbone networks with and without ESAM over HDV1 and LMG dataset, respectively. The results indicate that the proposed model incorporating our proposed ESAM consistently enhances the performance across both datasets. Furthermore, it is evident that DenseNet-169, coupled with the adapter and ESAM, surpasses all other state-of-the-art CNN models. Notably, it achieves a higher accuracy of 86.20% and 84.48% on the HDV1 and LMG datasets, respectively. The combination of adapter and ESAM is named as AESAM, while DenseNet-169 with AESAM is named as AES-Net in this paper. It is noteworthy that the evaluation of all backbone architectures are performed subject to similar setup to ensure fair comparison. The ROC curves of DenseNet-169 in the presence of AESAM on both the datasets are shown in Figs. 6 and 7. It can be observed that the inclusion of AESAM with DenseNet-169 results in a higher AUC score on both the datasets compared to the baseline. Figs. 8 and 9 depict the confusion matrix of the DenseNet-169 and AES-Net model for HDV1 and LMG datasets, respectively. It can be seen that the proposed AES-Net accurately classifies more samples of all three classes than the DenseNet-169 on both datasets. Also, it can be observed that the misclassification of early-stage samples is relatively high compared to other stages. This is mainly due to slight lesion size variations and similarities of early-stage samples with samples of other classes.

3.3.2. Visualization results

To better understand the impact of AESAM in our framework and analyze its explainability, we use Grad-CAM++ [50] visualizations. The generated saliency maps help determine whether or not the model effectively focuses on critical areas within the fundus images. Fig. 10 shows the heatmap visualization results and the corresponding fundus images selected from the overall LMG dataset. Remarkably, our model without AESAM concentrates on unimportant areas, whereas the model with AESAM successfully localizes and accentuates glaucomatous regions in fundus images. The capacity of the model to detect discriminative areas improves the interpretability of AES-Net, making it ideal for accurately diagnosing various stages of glaucoma. To further test the explainability and interpretability power of AES-Net, t-SNE [51] plots are drawn as depicted in Figs. 11 and 12, which help in visualizing the features learned by the baseline model (DenseNet-169) and AES-Net, on HDV1 dataset and LMG dataset, respectively, in the 2D embedding space. The results demonstrate that features learned by AES-Net have higher separability than the baseline model on both datasets, indicating

Table 1
Performance comparison of different pre-trained networks on HDV1 dataset.

Method	A_{cc} (%)	P_r (%)	S_{en} (%)	F1-score (%)	AUC
VGG-19	79.09	78.10	79.09	78.48	0.9156
VGG-19 + AESAM	80.17	79.43	80.17	79.43	0.9213
Xception	81.46	81.47	81.46	81.37	0.9195
Xception + AESAM	81.55	82.36	81.55	81.95	0.9217
ResNet-50	82.75	82.43	82.75	82.46	0.9360
ResNet50 + AESAM	84.48	84.03	84.48	84.11	0.9383
InceptionV3	80.60	80.25	80.60	80.42	0.9244
InceptionV3 + AESAM	82.80	83.40	82.80	82.45	0.9360
EfficientNetB1	79.74	78.95	79.74	79.21	0.9056
EfficientNetB1 + AESAM	81.13	80.37	81.13	80.46	0.9089
MobileNet	80.17	79.43	80.17	79.43	0.9238
MobileNet + AESAM	81.55	81.07	81.55	80.60	0.9237
DenseNet-169	83.83	83.35	83.87	83.41	0.9375
DenseNet-169 + AESAM	**86.20**	**85.32**	**85.77**	**85.46**	**0.9493**

Table 2
Performance comparison of different pre-trained networks on LMG dataset.

Method	A_{cc} (%)	P_r (%)	S_{en} (%)	F1-score (%)	AUC
VGG-19	79.03	77.95	79.03	78.10	0.9070
VGG-19 + AESAM	80.17	79.88	80.17	79.46	0.9158
Xception	79.24	78.41	79.24	78.44	0.9071
Xception + AESAM	81.34	81.02	81.34	80.96	0.9190
ResNet-50	81.55	82.36	81.55	81.34	0.9244
ResNet-50 + AESAM	82.97	82.33	82.97	82.48	0.9339
InceptionV3	79.87	79.17	79.87	79.07	0.9018
InceptionV3 + AESAM	81.46	81.16	81.46	81.27	0.9159
EfficientNetB1	77.14	77.25	77.14	77.03	0.9034
EfficientNetB1 + AESAM	78.86	78.54	78.86	78.65	0.9038
MobileNet	80.29	79.88	80.29	79.63	0.9055
MobileNet + AESAM	81.46	81.35	81.46	81.37	0.9195
DenseNet-169	82.80	82.58	82.80	82.61	0.9216
DenseNet-169 + AESAM	**84.48**	**84.27**	**84.48**	**84.34**	**0.9414**

its better learning capabilities and thereby leading to higher classification performance. It is worth noting that the t-SNE is used to visualize the feature representation of the GAP layer over a set of images randomly chosen from the whole dataset.

3.3.3. Comparison with previous attention mechanisms

To quantify the efficacy of our proposed AESAM, we compare it with recent attention mechanisms, namely squeeze and excitation (SE) [28], convolutional block attention module (CBAM) [29], bottleneck attention module (BAM) [52], global context (GC) [53], self-attention (SA) [40], global attention block (GAB) [30], triplet attention (TA) [31], and global self attention module (GSAM) [37], and the comparative analysis on HDV1 and LMG dataset are depicted in Table 3 and 4, respectively. It is important to mention that the top-performing backbone network (i.e DenseNet-169) is only considered for this comparison. Further like AESAM, all of the existing attention modules are stacked on top of the backbone network. Our proposed AESAM surpasses popular attention modules on the two multi-stage glaucoma datasets. Notably, in comparison to the baseline, it enhances the accuracy by 1.68% and 2.37% on the LMG and HDV1 datasets, respectively. Also, it can be observed that a comparable performance is achieved by GSAM. Further, the number of

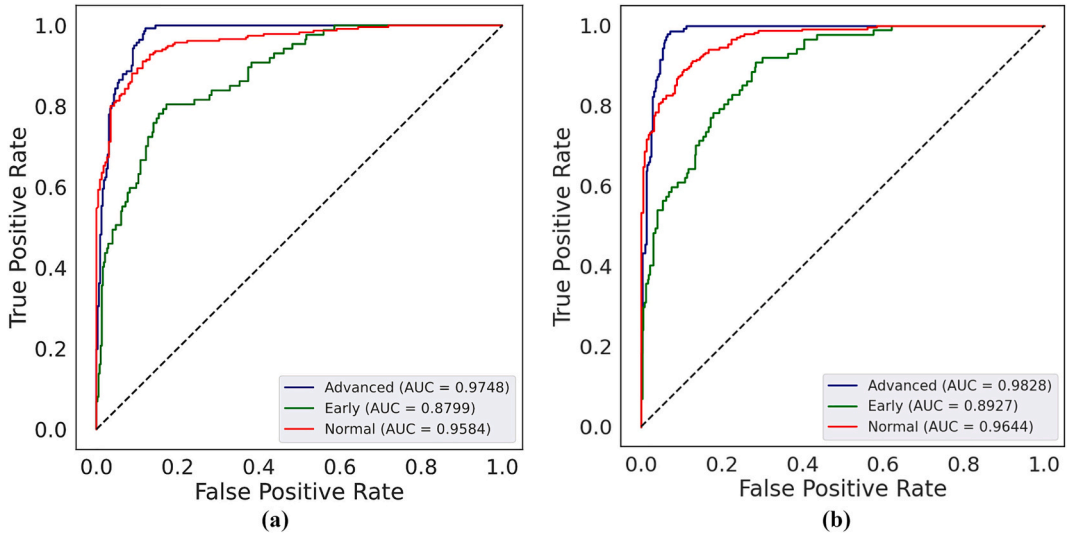

Fig. 6. The ROC curves of (a) DenseNet-169, and (b) DenseNet-169 with AESAM, for HDV1 dataset.

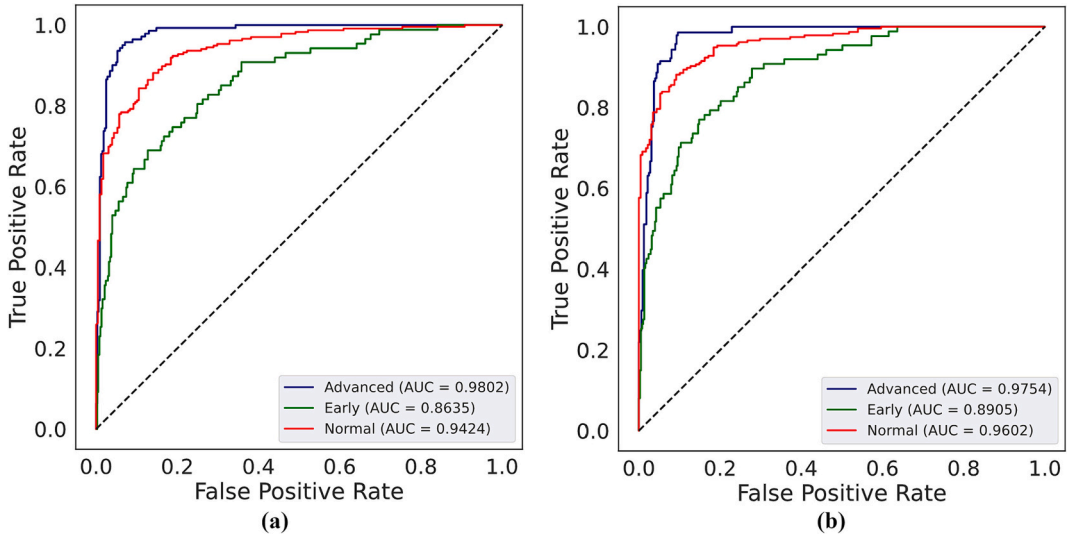

Fig. 7. The ROC curves of (a) DenseNet-169, and (b) DenseNet-169 with AESAM, for LMG dataset.

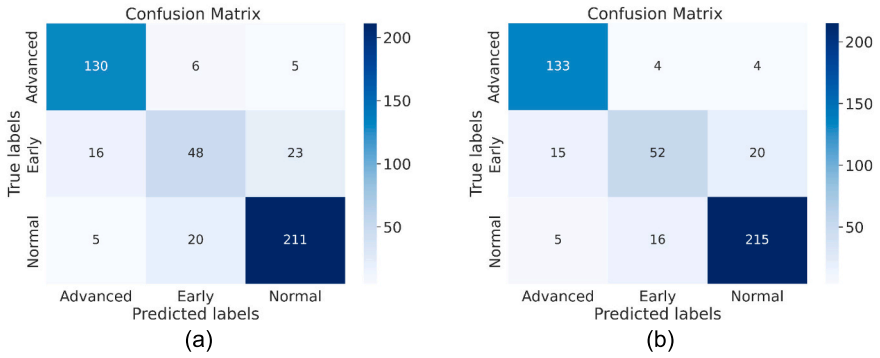

Fig. 8. Confusion matrix of (a) DenseNet-169 and (b) AES-Net on HDV1 dataset.

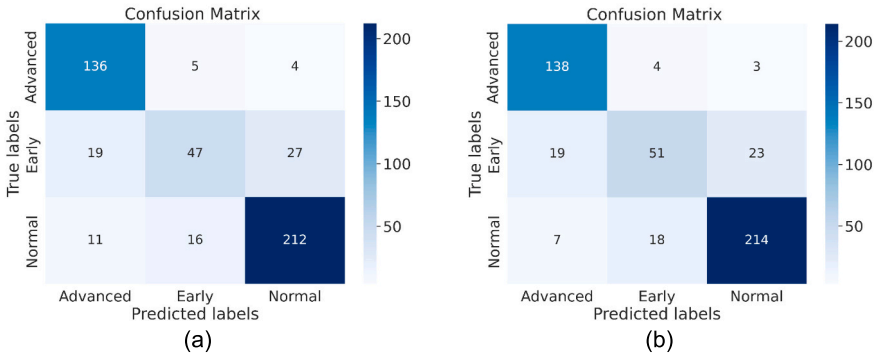

Fig. 9. Confusion matrix of (a) DenseNet-169 and (b) AES-Net on LMG dataset.

trainable parameters required by the model with AESAM is higher than that of many attention modules. However, compared to the best-performing attention mechanism, 'GSAM', our AESAM demands fewer parameters. The proposed attention module is a plug-and-play attention that could potentially be introduced in CNN models for performance improvement.

3.3.4. Ablation experiment

The effects of each part of the AESAM are confirmed through a set of ablation experiments. Table 5 and Table 6 report the experimental results on HDV1 and LMG datasets, respectively. In this experiment, we provide the effectiveness of adapter, CSAM and SSAM individually. A decrease in classification performance has been observed when adapter, CSAM or SSAM is combined individually with a backbone. However, these in together improves the accuracy by 2.37% and 1.68% on the HDV1 and LMG dataset, compared to the baseline, thereby demonstrating its efficiency and learning capabilities.

3.3.5. Comparison with SOTA glaucoma classification approach

We compare our proposed AES-Net against CNN-based SOTA glaucoma detection methods using two considered datasets as depicted in Tables 7 and 8. For fair comparison purpose, we implement the existing ML based methods [5,10] and CNN models [15,16,27,36] under similar implementation setup. From the table, it can be observed that our AES-Net demonstrates superior performance on both LMG and HDV1 datasets compared to previous ML based and CNN-based methods. In contrast to conventional CNN-based models [15,16,27], the AES-Net extracts rich discriminant features from fundus images, resulting in higher performance. In addition, our model outperforms the recently proposed attention-based models such as ResNet50 + GAB + CAB [36], GS-Net [37], and CA-Net [38]. This performance improvement can be attributed to the integration of the adapter and ESAM together with a pre-trained CNN, facilitating the extraction of stage-specific and discriminative features while suppressing redundant information. The proposed framework can hence serve as an efficient multi-stage glaucoma classification system for ophthalmologists to validate their screening results.

3.3.6. Performance comparison on other datasets

We further evaluate the performance of AES-Net on a binary glaucoma dataset, RIM-ONE Extended, to verify robustness of the model. The results of ML and DL-based glaucoma screening approaches are compared in Table 9. Our proposed AES-Net obtains higher performance than state-of-the-art ML and DL-based approaches. Furthermore, our model outperformed the recent attention-based techniques. It is mainly due to the capability of the model to learn better feature representation due to the inclusion of AESAM.

3.4. Discussion

Automated glaucoma screening through fundus images has become a popular topic in medical image analysis, utilizing ML and DL techniques. It is essential to detect glaucoma in its early stages to preserve vision. To this end, limited attempts have been made to classify glaucoma stages using ML and DL in the past years, and the results achieved by these methods still need potential improvements. This is mainly due to the scarcity of labeled data on various glaucoma stages, and the challenges lie in the subtle lesion size and shape variations among different classes. To cope with these issues, we propose a novel AES-Net model to classify various stages of glaucoma effectively. The proposed model supports end-to-end learning and can learn class-specific features with the help of the AESAM. The AES-Net has been evaluated on two benchmark datasets, and the results demonstrate its superiority over other attention mechanisms. Furthermore, ablation experiments have been performed to test the effect of each component of the proposed AESAM. Additionally, the effectiveness of our model has been verified by providing

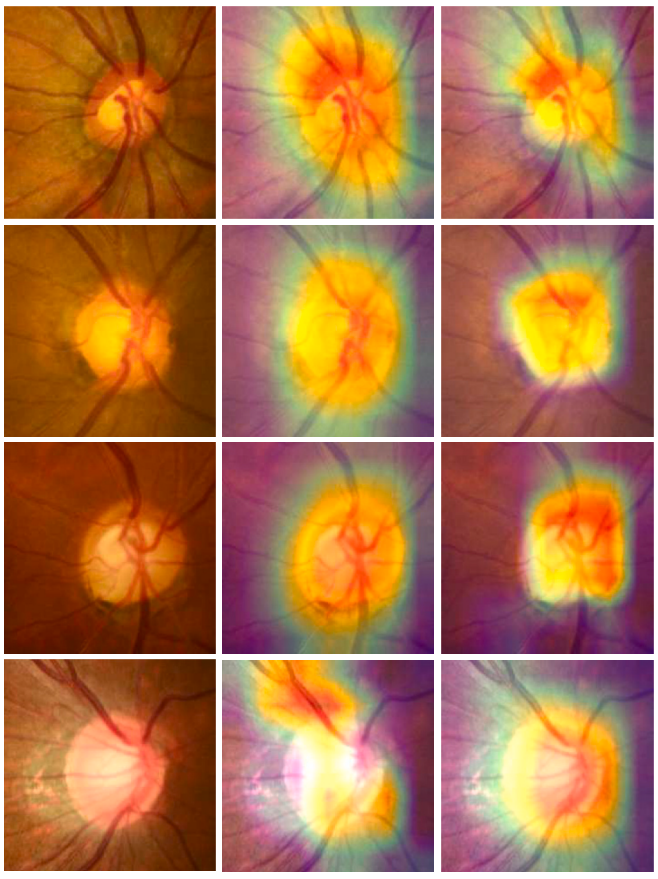

Fig. 10. Illustration of heatmaps generated using Grad-CAM++.

the visualization results in terms of heatmaps generated using Grad-CAM++ and t-SNE plots. The results indicated that our method achieved state-of-the-art glaucoma classification results. This is mainly attributed to the inclusion of AESAM, which consists of the spatial adapter module and enhanced self-attention module (ESAM). It helps the network to efficiently learn stage-specific detailed features from lesion areas in fundus images.

We report the number of trainable parameters and FLOPs as a reference for model complexity. The AES-Net has 18.96 million trainable parameters, and the FLOPs required are 7.07×10^9. The model takes around 2.28×10^3 sec to complete all the epochs during training, and it requires an inference time of 0.019 ms to process each image. Our AES-Net supports end-to-end learning and provides rapid glaucoma diagnosis, demonstrating its practical utility in clinical settings. Hence, ophthalmologists can utilize the proposed as a supplementary tool to cross-check their screening. Specifically, AES-Net offers faster clinical decisions and can be used to test glaucoma suspects on a large scale, which, on the other hand, avoids the scarcity of eye specialists in remote regions, thereby improving treatment plans and patient outcomes.

Although our model achieves higher performance than other SOTA methods, it still has some limitations. Our method performs better in the presence of limited data; however, its effectiveness needs to be explored using large and diverse datasets with images collected from different demographic regions. Our model may fail to detect very early and moderate stages of glaucoma which are crucial in clinical settings. Further, the proposed AES-Net demands a comparatively higher number of learnable parameters, leading to high computational complexity. Hence, a lightweight variant of the proposed model could be explored in future to reduce computational overhead.

4. Conclusion

This paper presented an attention-based CNN termed AES-Net for automatic multi-stage glaucoma screening. The AES-Net introduced a spatial adapter and an ESAM to learn more detailed and discriminative feature representations from fundus images. Extensive experiments on two multi-stage glaucoma classification datasets demonstrated the superiority of AES-Net over SOTA methods. The ablation experiments further verified the effectiveness of the proposed ESAM and spatial adapter module. Furthermore, the proposed model is end-to-end learnable and can be used as a supplementary tool by ophthalmologists in accurately identifying glaucoma stages.

Though AES-Net achieved higher performance compared to previous methods, there is still room for performance improvements. The potential reason for this could be the usage of limited and imbalanced

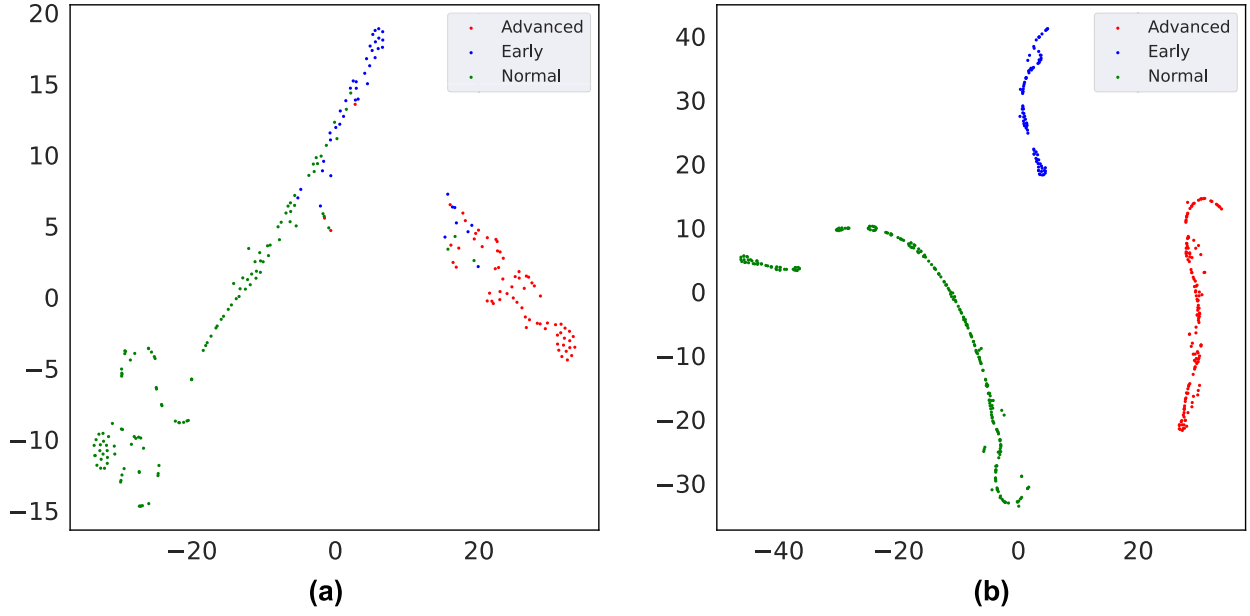

Fig. 11. t-SNE visualization of embedding space using (a) baseline (DenseNet-169), and (b) AES-Net, for the HDV1 dataset. Each scatter plot contains three colored clusters, in which each color denotes a novel class and each point represents an image sample.

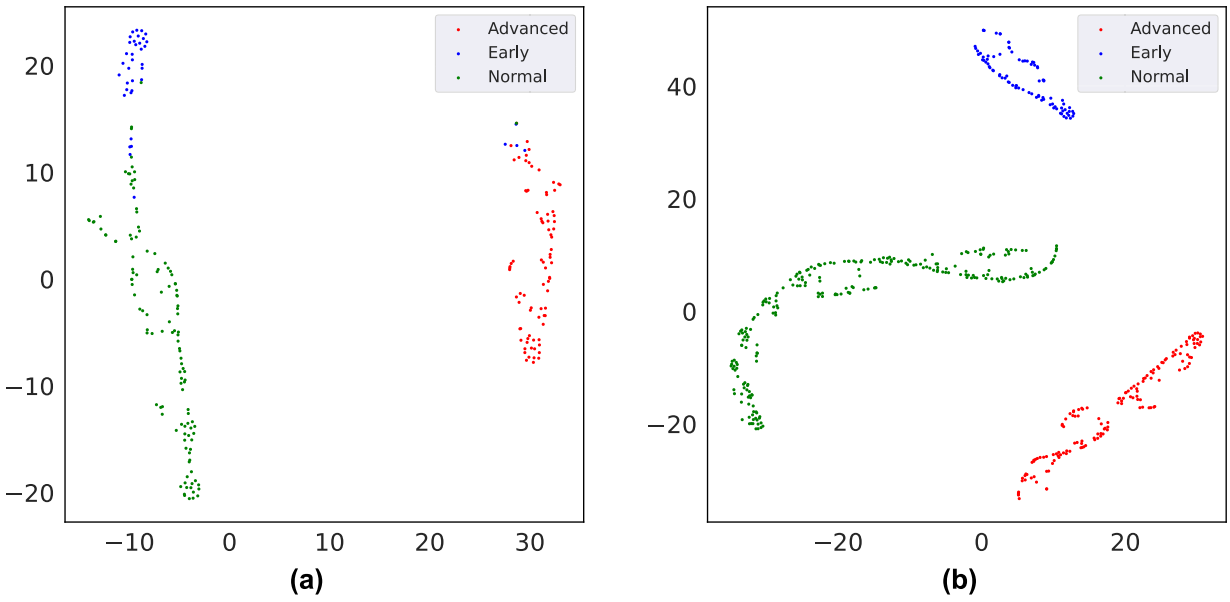

Fig. 12. t-SNE visualization of embedding space using (a) baseline (DenseNet-169), and (b) AES-Net, for the LMG dataset.

Table 3
Comparison of our AESAM with existing attention mechanisms on HDV1 dataset.

Attention	A_{cc} (%)	F1-score (%)	AUC	#Para (M)
None	83.83	83.41	0.9375	12.64
SE [28]	83.87	83.41	0.9389	14.67
BAM [52]	84.04	83.72	0.9409	15.42
CBAM [29]	84.04	83.72	0.9409	14.95
GC [53]	83.87	82.34	0.9389	14.67
SA [40]	84.04	83.72	0.9403	15.42
GAB [30]	84.48	84.34	0.9414	18.19
TA [31]	84.69	84.52	0.9416	12.85
GSAM [37]	84.91	84.55	0.9454	23.74
AESAM (Ours)	86.20	85.46	0.9493	18.96

Table 4
Comparison of our AESAM with existing attention mechanisms on LMG dataset.

Attention	A_{cc} (%)	F1-score (%)	AUC	#Para (M)
None	82.80	82.58	0.9216	12.64
SE [28]	82.80	82.58	0.9248	14.67
BAM [52]	82.97	82.46	0.9256	15.42
CBAM [29]	83.43	83.22	0.9244	14.95
GC [53]	82.96	82.45	0.9256	14.67
SA [40]	83.40	82.98	0.9379	15.42
GAB [30]	83.40	83.27	0.9336	18.19
TA [31]	83.85	83.48	0.9357	12.85
GSAM [37]	83.43	83.22	0.9247	23.74
AESAM (Ours)	84.48	84.34	0.9414	18.96

datasets to validate our AES-Net. To cope with these issues, in the future, we investigate the GAN models to synthesize fundus images and pre-train AES-Net on generated images. The synthetic data helps by increasing the diversity of the dataset, providing more samples for the model to learn while training. This expanded diversity will help the model to generalize better to unseen data by learning from different variations that might not be present in original dataset. Furthermore, we plan to exploit the use of contrastive learning to avoid the class imbalance issue and meta-learning techniques [55] to enable the model to quickly adapt to new situations with a limited number of glaucoma samples. In addition, we will explore the use of transformer architecture [56] to classify multiple stages of glaucoma. The future work also includes the possible pilot studies or collaborations with ophthalmologists to test the practical utility of AES-Net in actual medical settings. Additionally, we intend to verify the effectiveness of AES-Net on edge cases or very early/late stages of glaucoma.

CRediT authorship contribution statement

Dipankar Das: Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Deepak Ranjan Nayak:** Conceptualization, Methodology, Software, Writing-Reviewing and Editing, Validation, Supervision. **Ram Bilas Pachori:** Writing – review & editing, Validation, Supervision, Formal analysis, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Deepak Ranjan Nayak reports financial support was provided by

Table 5
Results of ablation studies on HDV1 dataset.

Model	Adapter	CSAM	SSAM	A_{cc} (%)	F1-score (%)	AUC	Δ_{Acc}(%)
Densenet169	✘	✘	✘	83.83	83.41	0.9375	–
	✓	✘	✘	84.04	83.72	0.9409	0.21
	✘	✓	✘	84.69	84.52	0.9416	0.86
	✘	✘	✓	84.48	84.34	0.9414	0.65
	✘	✓	✓	85.34	84.93	0.9475	1.51
	✓	✓	✓	86.20	85.46	0.9493	2.37

Table 6
Results of ablation studies on LMG dataset.

Model	Adapter	CSAM	SSAM	A_{cc} (%)	F1-score (%)	AUC	Δ_{Acc}(%)
Densenet169	✘	✘	✘	82.80	82.58	0.9216	–
	✓	✘	✘	82.97	82.58	0.9248	0.17
	✘	✓	✘	83.85	83.48	0.9357	1.05
	✘	✘	✓	83.40	83.27	0.9336	0.60
	✘	✓	✓	84.04	83.72	0.9409	1.24
	✓	✓	✓	84.48	84.34	0.9414	1.68

Table 7
Comparison with previous glaucoma classification methods on HDV1 dataset.

Approach	A_{cc} (%)	F1-score (%)	AUC
Gabor entropy with NB [10]	62.18	58.77	0.7501
Gabor entropy with SVM [10]	68.27	65.21	0.7678
Wavelet energy with SVM [5]	71.21	68.05	0.7971
4-Layer CNN [15]	75.64	75.84	0.8900
InceptionV3 [27]	80.60	80.42	0.9244
Customized CNN [27]	78.45	78.42	0.9049
Customized CNN [16]	78.01	79.01	0.9097
Inception-ResNet-V2 [54]	81.03	80.91	0.9112
GC-Net [36]	82.75	82.75	0.9259
GC-Net [36]	82.75	82.75	0.9259
CA-Net [38]	85.34	84.92	0.9477
GS-Net [37]	84.91	84.55	0.9454
AES-Net (Ours)	**86.20**	**85.46**	**0.9493**

Table 8
Comparison with previous glaucoma screening methods on LMG dataset.

Approach	A_{cc} (%)	F1-score (%)	AUC
Gabor entropy with NB [10]	62.18	58.77	0.7501
Gabor entropy with SVM [10]	68.27	65.21	0.7678
Wavelet energy with SVM [5]	71.21	68.05	0.7971
4-Layer CNN [15]	76.93	76.02	0.9012
InceptionV3 [27]	79.87	79.07	0.9018
Customized CNN [27]	77.35	75.89	0.8969
Customized CNN [16]	78.19	76.61	0.9159
Inception-ResNet-V2 [54]	79.66	79.03	0.9046
GC-Net [36]	81.97	81.32	0.9244
GC-Net [36]	82.59	82.42	0.9247
CA-Net [38]	83.85	83.48	0.9257
GS-Net [37]	83.43	83.22	0.9247
AES-Net (Ours)	**84.48**	**84.34**	**0.9414**

Table 9
Comparison with previous glaucoma screening methods on RIM-ONE Extended dataset.

Method	Acc (%)	F1-score (%)	AUC
Gabor entropy with NB [10]	62.18	58.77	0.7501
Gabor entropy with SVM [10]	68.44	74.45	0.6632
Wavelet energy with SVM [5]	77.01	76.42	0.7561
4-Layer CNN [15]	80.21	83.70	0.7907
Customized CNN [27]	79.14	82.66	0.7818
InceptionV3 [27]	79.14	82.66	0.7818
Customized CNN [16]	77.54	80.00	0.7818
GC-Net [36]	88.23	90.35	0.8731
GC-Net [36]	89.30	91.22	0.8842
CA-Net [38]	92.51	93.85	0.9176
GS-Net [37]	92.08	93.54	0.9086
AES-Net (Ours)	**93.12**	**94.22**	**0.9234**

Science and Engineering Research Board. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source of the datasets used to validate our proposed model has been provided in the paper.

Acknowledgments

This work is supported by the Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India under project No. SRG/2020/001460.

References

[1] H. Fu, J. Cheng, Y. Xu, C. Zhang, D.W.K. Wong, J. Liu, X. Cao, Disc-aware ensemble network for glaucoma screening from fundus image, IEEE Trans. Med. Imaging 37 (11) (2018) 2493–2501.
[2] D. Parashar, D. Agrawal, 2-d compact variational mode decomposition-based automatic classification of glaucoma stages from fundus images, IEEE Trans. Instrum. Meas. 70 (2021) 1–10.
[3] H.A. Quigley, A.T. Broman, The number of people with glaucoma worldwide in 2010 and 2020, Br. J. Ophthalmol. 90 (3) (2006) 262–267.
[4] Y.-C. Tham, X. Li, T.Y. Wong, H.A. Quigley, T. Aung, C.-Y. Cheng, Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis, Ophthalmology 121 (11) (2014) 2081–2090.
[5] S. Dua, U.R. Acharya, P. Chowriappa, S.V. Sree, Wavelet-based energy features for glaucomatous image classification, IEEE Trans. Inf. Technol. Biomed. 16 (1) (2011) 80–87.
[6] G.D. Joshi, J. Sivaswamy, S. Krishnadas, Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment, IEEE Trans. Med. Imaging 30 (6) (2011) 1192–1205.
[7] J. Cheng, J. Liu, Y. Xu, F. Yin, D.W.K. Wong, N.-M. Tan, D. Tao, C.-Y. Cheng, T. Aung, T.Y. Wong, Superpixel classification based optic disc and optic cup segmentation for glaucoma screening, IEEE Trans. Med. Imaging 32 (6) (2013) 1019–1032.
[8] J. Cheng, F. Yin, D.W.K. Wong, D. Tao, J. Liu, Sparse dissimilarity-constrained coding for glaucoma screening, IEEE Trans. Biomed. Eng. 62 (5) (2015) 1395–1403.
[9] M.R.K. Mookiah, U.R. Acharya, C.M. Lim, A. Petznick, J.S. Suri, Data mining technique for automated diagnosis of glaucoma using higher order spectra and wavelet energy features, Knowl.-Based Syst. 33 (2012) 73–82.
[10] U.R. Acharya, E. Ng, L.W.J. Eugene, K.P. Noronha, L.C. Min, K.P. Nayak, S. V. Bhandary, Decision support system for the glaucoma using gabor transformation, Biomed. Signal Proc. Control 15 (2015) 18–26.
[11] S. Maheshwari, R.B. Pachori, U.R. Acharya, Automated diagnosis of glaucoma using empirical wavelet transform and correntropy features extracted from fundus images, IEEE J. Biomed. Health Inform. 21 (3) (2016) 803–813.
[12] T. Kausu, V.P. Gopi, K.A. Wahid, W. Doma, S.I. Niwas, Combination of clinical and multiresolution features for glaucoma detection and its classification using fundus images, Biocybernet. Biomed. Eng. 38 (2) (2018) 329–341.
[13] Y.B. Özçelik, A. Altan, Overcoming nonlinear dynamics in diabetic retinopathy classification: a robust ai-based model with chaotic swarm intelligence optimization and recurrent long short-term memory, Fract. Fraction. 7 (8) (2023) 598.
[14] Y.B. Özçelik, A. Altan, Classification of diabetic retinopathy by machine learning algorithm using entorpy-based features, in: Proceedings of the ÇAnkaya International Congress on Scientific Research, IKSAD Golbasi, Adiyaman Province, Turkey, 2023, pp. 10–12.
[15] X. Chen, Y. Xu, D.W.K. Wong, T.Y. Wong, J. Liu, Glaucoma detection based on deep convolutional neural network, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 715–718.
[16] U. Raghavendra, H. Fujita, S.V. Bhandary, A. Gudigar, J.H. Tan, U.R. Acharya, Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images, Inf. Sci. 441 (2018) 41–49.
[17] M.N. Bajwa, G.A.P. Singh, W. Neumeier, M.I. Malik, A. Dengel, S. Ahmed, G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection, in: In: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[18] M. Juneja, S. Thakur, A. Uniyal, A. Wani, N. Thakur, P. Jindal, Deep learning-based classification network for glaucoma in retinal images, Comput. Electr. Eng. 101 (2022) 108009.
[19] A. Li, J. Cheng, D.W.K. Wong, J. Liu, Integrating holistic and local deep features for glaucoma classification, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2016, pp. 1328–1331.
[20] A. Pal, M.R. Moorthy, A. Shahina, G-eyenet: A convolutional autoencoding classifier framework for the detection of glaucoma from retinal fundus images, in: In: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 2775–2779.
[21] Y. Chai, H. Liu, J. Xu, Glaucoma diagnosis based on both hidden features and domain knowledge through deep learning models, Knowl.-Based Syst. 161 (2018) 147–156.
[22] S. Phasuk, P. Poopresert, A. Yaemsuk, P. Suvannachart, R. Itthipanichpong, S. Chansangpetch, A. Manassakorn, V. Tantisevi, P. Rojanapongpun, C. Tantibundhit, Automated glaucoma screening from retinal fundus image using deep learning, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 904–907.
[23] D.R. Nayak, D. Das, B. Majhi, S.V. Bhandary, U.R. Acharya, Ecnet: an evolutionary convolutional network for automated glaucoma detection using fundus images, Biomed. Signal Proc. Control 67 (2021) 102559.
[24] Á.S. Hervella, J. Rouco, J. Novo, M. Ortega, End-to-end multi-task learning for simultaneous optic disc and cup segmentation and glaucoma classification in eye fundus images, Appl. Soft Comput. 116 (2022) 108347.
[25] V. Sunanthini, J. Deny, E. Govinda Kumar, S. Vairaprakash, P. Govindan, S. Sudha, V. Muneeswaran, M. Thilagaraj, Comparison of cnn algorithms for feature extraction on fundus images to detect glaucoma, J. Healthc. Eng. 2022 (2022).
[26] D. Parashar, D.K. Agrawal, Automated classification of glaucoma stages using flexible analytic wavelet transform from retinal fundus images, IEEE Sensors J. 20 (21) (2020) 12885–12894.
[27] J.M. Ahn, S. Kim, K.-S. Ahn, S.-H. Cho, K.B. Lee, U.S. Kim, A deep learning model for the detection of both advanced and early glaucoma using fundus photography, PLoS One 13 (11) (2018) e0207982.
[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[29] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[30] A. He, T. Li, N. Li, K. Wang, H. Fu, Cabnet: category attention block for imbalanced diabetic retinopathy grading, IEEE Trans. Med. Imaging 40 (1) (2020) 143–153.
[31] D. Misra, T. Nalamada, A.U. Arasanipalai, Q. Hou, Rotate to attend: Convolutional triplet attention module, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3139–3148.
[32] J. Zhu, Z. Li, J. Wei, Y. Zeng, H. Ma, Fine-grained bidirectional attentional generation and knowledge-assisted networks for cross-modal retrieval, Image Vis. Comput. 124 (2022) 104507.
[33] H. Tang, C. Yuan, Z. Li, J. Tang, Learning attention-guided pyramidal features for few-shot fine-grained recognition, Pattern Recogn. 130 (2022) 108792.
[34] Q. Zhu, Z. Li, W. Kuang, H. Ma, A multichannel location-aware interaction network for visual classification, Appl. Intell. 53 (20) (2023) 23049–23066.
[35] S. Yan, H. Tang, L. Zhang, J. Tang, Image-specific information suppression and implicit local alignment for text-based person search, IEEE Trans. Neural Networks Learn. Syst. (2023) 1–14, https://doi.org/10.1109/TNNLS.2023.3310118.
[36] H. Tian, S. Lu, Y. Sun, H. Li, Gc-net: Global and class attention blocks for automated glaucoma classification, in: In: 2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2022, pp. 498–503.
[37] D. Das, D.R. Nayak, Gs-net: Global self-attention guided cnn for multi-stage glaucoma classification, in: In: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 3454–3458.
[38] D. Das, D.R. Nayak, R.B. Pachori, Ca-net: a novel cascaded attention-based network for multi-stage glaucoma classification using fundus images, IEEE Trans. Instrum. Meas. 72 (2023) 1–10.
[39] J. Pan, Z. Lin, X. Zhu, J. Shao, H. Li, St-adapter: parameter-efficient image-to-video transfer learning, Adv. Neural Inf. Proces. Syst. 35 (2022) 26462–26477.
[40] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: ICML, 2019, pp. 7354–7363.
[41] F. Fumero, S. Alayón, J.L. Sanchez, J. Sigut, M. Gonzalez-Hernandez, Rim-one: An open retinal image database for optic nerve evaluation, in: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2011, pp. 1–6.
[42] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv preprint. arXiv:1412.6980, 2014.
[43] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv preprint. arXiv:1409.1556, 2014.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[46] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient Convolutional Neural Networks for mobile vision applications, arXiv preprint. arXiv:1704.04861, 2017.
[47] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[48] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[49] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[50] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 839–847.
[51] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (11) (2008).
[52] J. Park, S. Woo, J.-Y. Lee, I.S. Kweon, Bam: Bottleneck Attention Module, arXiv preprint. arXiv:1807.06514, 2018.
[53] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
[54] V.K. Velpula, L.D. Sharma, Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion, Front. Physiol. 14 (2023) 1175881.
[55] H. Tang, Z. Li, Z. Peng, J. Tang, Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 610–618.
[56] W. Li, Z. Li, X. Yang, H. Ma, Causal-vit: robust vision transformer by causal intervention, Eng. Appl. Artif. Intell. 126 (2023) 107123.