Multi-level Dual-attention Based CNN for Macular Optical Coherence Tomography Classification

Sapna S Mishra, Bappaditya Mandal, and N. B. Puhan

***Abstract*—In this letter, we propose a multi-level dual-attention model to classify two common macular diseases, age-related macular degeneration (AMD) and diabetic macular edema (DME) from normal macular eye conditions using optical coherence tomography (OCT) imaging technique. Our approach unifies the dual-attention mechanism at multi-levels of the pre-trained deep convolutional neural network (CNN). It provides a focused learning mechanism by taking into account both multi-level features based attention focusing on the salient coarser features and self-attention mechanism attending higher entropy regions of the finer features. Our proposed method enables the network to automatically focus on the relevant parts of the input images at different levels of feature subspaces. This leads to a more locally deformation-aware feature generation and classification. The proposed approach does not require pre-processing steps such as extraction of region of interest, denoising and retinal flattening, making the network more robust and fully automatic. Experimental results on two macular OCT databases show the superior performance of our proposed approach as compared to the current state-of-the-art methodologies.**

***Index Terms*—Attention mechanism, Age-related Macular Degeneration (AMD), Diabetic Macular Edema (DME), Multi-level dual-attention, Optical Coherence Tomography (OCT)**

I. Introduction

Macular region present in the retina of the eye is mainly responsible for the central vision. Degraded macular health results in poor vision or loss of sight. Two such diseases that adversely affect the macula are age-related macular degeneration (AMD) and diabetic macular edema (DME) [1], [2]. These diseases, if left untreated, may lead to partial or complete vision loss. The progression of the diseases can be restricted or slowed down with proper care and necessary supplements, if detected in early stages [3], [4]. Due to the sight-threatening effects of these diseases, intensive research is being carried out to develop computer-assisted techniques for timely and accurate diagnosis of these diseases. The optical coherence tomography (OCT) is an imaging technique used to capture a three-dimensional view of the tissues in order to resolve the depth information [5]. Employing OCT imaging technique for acquiring macular images helps to perform the objective layer-wise analysis of the macula which makes the detection of AMD and DME easier. During the scanning process, captured images are corrupted by speckle noise and suffer random inclinations which make the analysis of OCT-scans a very difficult task.

S. S. Mishra, and N. B. Puhan are with the School of Electrical Sciences, Indian Institute of Technology of Bhubaneswar, Bhubaneswar 752050, India (e-mail: ssm14@iitbbs.ac.in; nbpuhan@iitbbs.ac.in)
B. Mandal is with the School of Computing and Mathematics, Keele University, Newcastle ST5 5BG, United Kingdom. (e-mail: b.mandal@keele.ac.uk).

Over the last decade, numerous handcrafted feature based methods have been proposed for macular OCT classification [6]–[10]. Authors in [6] classified the OCT volumes based on histogram of gradients of images but the method required retinal flattening and region of interest (RoI) extraction which affect the adaptability of the technique. Similarly, in [7], authors used linear binary pattern features for classification where motion blurred and shadowed scans are neglected. These conventional methods are database specific and semi-automatic in nature. These problems have been alleviated by the deep convolutional neural network (CNN) learning based methods which automatically extract the features and gives superior performance in many medical-image classification applications [11], [12]. In [13], a CNN is used to classify the surrogates generated using the statistical features of each OCT B-scan, however it involves a time-consuming denoising step. While in [14], a mixture of CNNs (experts) model is used for classification. It utilizes retinal flattening and volume of interest generation. Karri *et al.* [15] and Ji *et al.* [16] employed transfer learning for macular OCT classification where they fine-tuned GoogleNet and InceptionV3 network, respectively.

A common problem to these existing methodologies is that they involve pre-processing steps such as denoising, retinal flattening and RoI extraction which make the methods less automated, database dependent and time-consuming in nature. To develop a fully automated and robust technique, we need to eliminate these steps. Attention mechanism has been explored for image captioning [17], voice activity detection [18], speech emotion recognition [19] and question answering [20]. For biomedical imaging, attention has been used for report generation [21], disease classification [22], [23], organ segmentation [24] and localization [25]. In [26], authors have introduced attention mechanism for macular OCT classification where the proposed deep network requires a large number of model parameters, but their performance evaluation is limited.

II. Hybrid Attention Mechanism

In recent works, dual-attention has been examined for scene segmentation [27], visual question answering [28] and image classification [29]. Our multi-level dual-attention mechanism (DAM) consists of two attention blocks which utilize the information from different convolution layers of a deep CNN as shown in Fig. 1. Unlike the existing modules, where the inputs are taken from a single convolution layer and do not consider the crucial information in coarser scales, our attention modules impart focus on the salient features of the input image in two different feature subspaces, allowing the network to learn relevant features in coarser as well as finer subspaces.

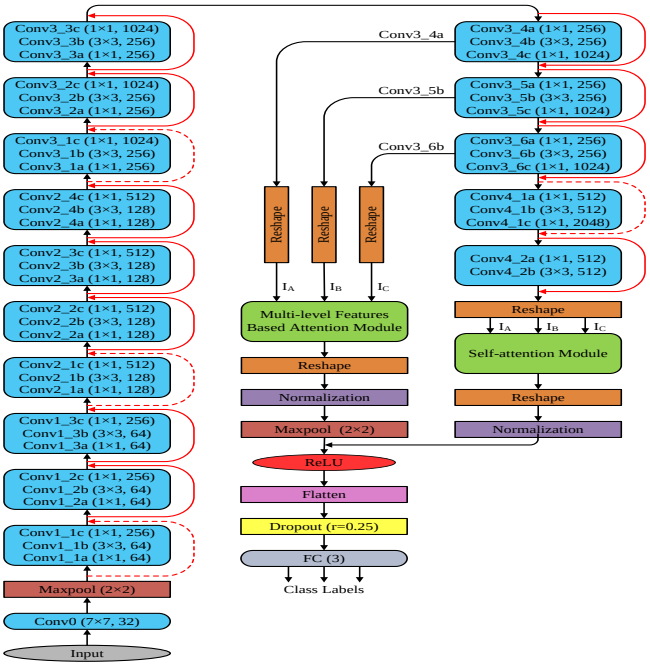

Fig. 1. Architecture of the proposed multi-level DAM network. Conv represents the convolutional layer, maxpool stands for max-pooling layer and $FC(3)$ denotes the fully connected layer with the number of neurons = 3. The first two numbers in the bracket of each Conv and Maxpool layer represents the filter dimensions while the last number in the Conv layers denotes the number of channels in that layer. The solid red lines show skip connections while the dotted red lines denote the skip connection with dimension increase.

A. Preparing the 'Deep Network'

For our proposed network, the ResNet50 architecture [30] has been chosen as the base model. Although replacing the base model with other deep pre-trained networks does not significantly affect the network performance as shown in subsection IV-B. The layers of ResNet learns the residual of the desired function with the help of the added skip connections and hence are trained to explicitly fit a residual mapping. The fully connected layer, average pooling layer and last convolutional layer of the ResNet have been removed. The initial framework is pre-trained on the ImageNet database [31]. The pre-trained weights are fine-tuned using the OCT database images. This approach of transfer learning guides the network to converge faster and overcomes the challenge of lack of large datasets. The features are extracted from different intermediate layers of the architecture and fed to the attention modules, as shown in Fig. 1. The complete architecture includes attention modules, a flatten, a dropout of 0.25 to prevent overfitting and a final dense layer activated by softmax function.

B. Multi-level Dual Attention Mechanism (DAM)

The proposed multi-level DAM aims to improve the performance of the existing network without using any external assistance. In [32], multi-head attention (MHA) has been used for machine translation working over 1-dimensional data. In our approach, we have extended its implications on 2-dimensional images that are classified based on spatial information involving deformations in the macular regions. Unlike MHA, the coefficients here are obtained using the feature maps of convolutional layers and a dual attention mechanism is developed. The proposed model includes attention modules, multi-level features based attention module and self-attention module which employ scaled scalar dot product for the computation of the alignment score. The multi-level DAM is formulated as:

$$\Phi_{att} = \psi(downsample(norm(\Phi_{R_{att_1}})) + norm(\Phi_{R_{att_2}})), \tag{1}$$

where $\Phi_{R_{att_1}}$ denotes reshaped output of multi-level featured based attention module and $\Phi_{R_{att_2}}$ is the reshaped output of self-attention module. ψ denotes the ReLU function, $norm(f_{att})$ represents normalization function to convert the attended feature, f_{att} to have mean = 0 and standard deviation = 1. $downsample(f)$ denotes downsampling of the input f by 2, f by performing maxpooling over it. The output of the multi-level DAM is the addition of normalised coarser and finer attended features. Hence. it utilizes information from various intermediate levels which leads to multi-level deformation-aware feature generation.

1) Multi-level Features Based Attention Module: The multi-level features based attention module requires three input tensors where the first two inputs are used to calculate the attention coefficients which then attend the weights of the third tensor, as shown in Fig. 1. The first attention module takes reshaped output matrices of three different layers of the network as the inputs, I_A, I_B and I_C.

Multi-level features based attention module is formulated as

$$\begin{aligned}\Phi_{att_1} \equiv &\big[I_A(196\times256), I_B(196\times256), I_C(196\times256)\big] \longrightarrow \\ &\big[FC_A(256),\ FC_B(256),\ FC_C(256)\big] \longrightarrow \big[R_A \\ &(\{196\times256\} \Rightarrow \{196\times32\times8\}),\ R_B(\{196\times256\} \\ &\Rightarrow \{196\times32\times8\}),\ R_C(\{196\times256\} \Rightarrow \{196\times32 \\ &\times8\})\big] \longrightarrow \odot\left(\sigma\left(\frac{\odot(a,b)}{\sqrt{d_k}}\right), c\right) \longrightarrow R(\{196\times32\times \\ &8\} \Rightarrow \{196\times256\}) \longrightarrow Add(out, I_C) \longrightarrow FC(32),\end{aligned} \tag{2}$$

where $R(\{x\} \Rightarrow \{y\})$ denotes reshape layer, reshaping input of shape $\{x\}$ to $\{y\}$. Fully connected layers are implied by $FC(n)$ with n being the number of neurons. The subscripts in reshape and fully connected layers depict that it is in either A^{th} or B^{th} or C^{th} input path. $\odot(g,h)$ signifies batch dot product of the tensors g and h, d_k denotes the number of channels of the input feature and σ indicates the softmax function. a, b and c are the outputs of layers R_A, R_B, and R_C, respectively. $Add(i,j)$ performs addition of matrices i, j. Here,

$$I_A = T_1(\psi(bn(O(Conv3\_4a)))) \tag{3}$$

$$and,\ T_1 = R(\{14\times14\times256\} \Rightarrow \{196\times256\}), \tag{4}$$

where bn denotes batch normalization and $O(l)$ represents output feature of layer, l. Similarly,

$$I_B = T_1(\psi(bn(O(Conv3\_5b)))), \tag{5}$$

$$I_C = T_1(\psi(bn(O(Conv3\_6b)))). \tag{6}$$

Here, names of the layers are in accordance with the Fig. 1. In this module, the matrix I_C is attended by the reshaped outputs of the two prior layers, I_A and I_B. These inputs undergo dot product function followed by scaling by a factor of $\sqrt{d_k}$. The coefficients are generated after passing this resultant through softmax function which optimizes the weights to impart higher probabilities to higher entropy regions that contribute more in the task of classification. These attention coefficients are then multiplied with linearly transformed I_C and the product is added to I_C itself. Finally, the module is appended by a dense layer which is activated by ReLU function which bestows an additional non-linearity to the attained weights. This approach assists the model to establish a complex relationship between the three layers and generates focused features. The coefficients of the module depend on the target labels and accordingly dictate the weights of the base CNN. This technique enables the network to utilize the information of coarser features preventing loss of any useful information. Thus, the network is trained to yield more focussed features as input to the classifier leading to better convergence.

2) Self-attention Module: The self-attention module is employed after the final feature extraction layer of the base model. In contrary to the first module, here the output of same convolutional layer acts as the attended feature as well as is employed for computing the attention coefficients, and hence named as self-attention. This technique aims for focussing on finer features of CNN, previous to the classification layer and to improve the network's performance. It is represented as,

$$\begin{aligned}\Phi_{att_2} \equiv & \big[I_A(49\times 512), I_B(49\times 512), I_C(49\times 512)\big] \longrightarrow \\ & \big[FC_A(512),\ FC_B(512),\ FC_C(512)\big] \longrightarrow \big[R_A(\{49 \\ & \times 512\} \Rightarrow \{49\times 64\times 8\}),\ R_B(\{49\times 512\} \Rightarrow \\ & \{49\times 64\times 8\}),\ R_C(\{49\times 512\} \Rightarrow \{49\times 64\times 8\})\big] \\ & \longrightarrow \odot\left(\sigma\left(\frac{\odot(a,b)}{\sqrt{d_k}}\right), c\right) \longrightarrow R(\{49\times 64\times 8\} \Rightarrow \\ & \{49\times 512\}) \longrightarrow Add(out, I_C) \longrightarrow FC(32)\end{aligned} \tag{7}$$

where $I_A = I_B = I_C = I$, $I = T_2(\psi(bn(O(Conv4\_2b))))$,

$$T_2 = \ R(\{7\times 7\times 512\} \Rightarrow \{49\times 512\}). \tag{8}$$

The multi-level attention layer is designed with a larger input size, shown in (2) whereas the self-attention module is developed for finer features so, has a smaller input, shown in (7). The obtained output features of both the attention modules are reshaped into 4-dimensional tensors, $\Phi_{R_{att_1}}$ and $\Phi_{R_{att_2}}$ and their addition is given in (1).

III. Experimental Setup and Results

A. Evaluation Protocol

We have performed experiments on two databases: Duke [6] and NEH [14] database, using two evaluation protocols. The first protocol is leave patient(s) out (LPO), followed from [13], where one and two volumes of each case are taken out randomly as the test set for Duke and NEH database, respectively. The remaining portion is divided into training and validation set as 80% and 20%, respectively. The experimental process is repeated 10 times with randomly selected test cases and average of the experiments are reported. The second protocol is 5-fold cross-validation (CV) [14]. The models are trained using 8 GB NVIDIA GeForce GTX 1080 GPU. The parameters used for training and testing of the models are as follows: *batch size* = 16, *number of epochs* = 100 for LPO protocol and 50 for each fold of 5-fold CV, *decay* = $1e-6$ and *momentum* = 0.9. The SGD optimizer is used with the categorical cross-entropy loss function. The adaptive learning rate technique is adopted, initialized with 0.001.

TABLE I
Performance comparison of different methods on Duke database. Here, † denotes motion blurred and shadowed scans not considered and * denotes RoI extracted. Top part of table presents the results for 5-fold CV protocol whereas bottom part is for LPO protocol.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score	AUC
Wang *et al.* † [7] (5-fold CV)	98.00	-	98.00	-	0.984
MCME * [14] (5-fold CV)	-	98.33	97.78	0.9771	0.999
Multi-level DAM (5-fold CV)	**99.97 (+/- 0.06)**	**99.97 (+/- 0.06)**	**99.97 (+/- 0.06)**	**0.9996**	**1.0**
SurrogateAssisted [13] (LPO)	88.45	-	-	-	-
Multi-level DAM (LPO)	**95.57**	**95.29**	**96.04**	**0.956**	**0.9974**

B. Results on Duke Database

Duke database contains 45 volumes of OCT B-scans obtained from 15 subjects of each class and has a total of 3241 scans. Each scan (image) is resized to 224×224 and self replicated three times to obtain a three channel input for the network. In the learning phase, some of the resized OCT scans are horizontally flipped and translated by ± 40 pixels, generating an augmented training set. This helps in tackling the translation problem and reduces the inconsistency of a different number of right and left eyes in the database. Some random samples are rotated and added to counter the effect of inclination in the scans. The network is trained using the OCT scans and scan-level results are presented in this paper.

Table I shows the performance of the proposed method on the Duke database in terms of accuracy, precision, recall, F1-score and area under region of operating curve (AUC) using both the protocols as compared to other existing methods for 3-class macular OCT classification where its superior performance can be observed. Moreover, DAM does not require the tedious RoI selection process, computationally expensive denoising and retinal flattening pre-processing steps. The deformation-aware feature generation leads to improved classification performance of the network. Fine-tuned ResNet50 (without attention) architecture yields an average accuracy of 95% using 5-fold CV protocol and 71% using LPO protocol and has around 23.88 millions of model parameters, whereas our DAM network has 23.54 million parameters. Hence, the proposed multi-level DAM has lesser number of model parameters and better performance than the baseline ResNet50 and outperforms the existing state-of-the-art using both the protocols, as shown in Table I. Removal of pre-processing steps such as denoising and retinal flattening makes our method faster than the methods involving these steps whereas the elimination of RoI extraction leads to a more robust network.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON NEH DATABASE USING 5-FOLD CV PROTOCOL. * DENOTES ROI EXTRACTED.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score	AUC
WCME* [10]	-	95.21 (+/- 3.2)	94.6 (+/- 3.4)	0.9458	0.986
MCME* [14]	-	99.39 (+/- 1.21)	99.36 (+/- 1.33)	0.9934	0.998
LACNN [26]	-	99.33(+/- 1.49)	99.39 (+/- 1.49)	0.993	0.994
Multi-level DAM	**99.62 (+/- 0.42)**	**99.60 (+/- 0.39)**	**99.62 (+/- 0.42)**	**0.996**	**0.9997**

C. Results on NEH Database

The NEH database contains OCT volumes of 48 AMD, 50 DME, and 50 Normal cases. It consists up of a total of 4230 B-scans acquired from Noor eye hospital, Tehran. Similar to the Duke database, the data augmentation technique has been adopted here. Table II shows the performance comparison of the proposed network on the NEH database with the existing works using the 5-fold CV protocol. It can be deduced that the proposed method outperforms the current state-of-the-art methodologies in terms of given metrics on this database. For the LPO protocol, accuracy, precision and recall of 93.03%, 93.86% and 92.26% respectively, are obtained. AUC of 0.991 and F1-score of 0.928 have been achieved. There is no present work in the literature using the LPO protocol on NEH database to compare the obtained results. Fine-tuned ResNet50 architecture [30] yields an average accuracy of 64.45% for 5-fold CV and 59% for LPO protocol.

It is evident from Tables I and II that our proposed multi-level DAM consistently outperforms all other existing methods including the fine-tuned original ResNet50 architecture in both the databases. It yields deformation-aware predictions which produce superior results. The removal of pre-processing steps gives added superiority to our network by making it fully automatic, more generic and faster. Besides, incorporated attention mechanism eliminates the need for RoI extraction which reduces the chances of missing the pathology symptoms outside the peripheral region of the scan.

IV. ANALYSIS AND DISCUSSIONS

A. Analysis using Attention maps

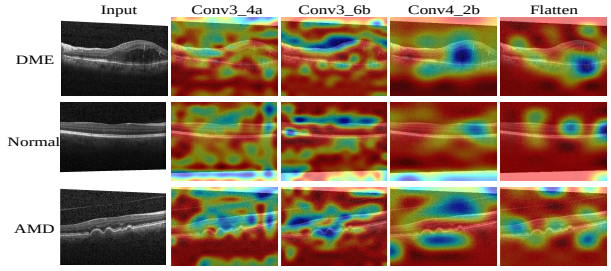

Fig. 2. Attention maps for samples of each case from Duke database obtained from various layers of the proposed Multi-level DAM network. Here, blue color denotes the highest attention while red denotes the lowest attention.

The attention maps of various intermediate layers used for attention modules along with the flatten layer of the proposed DAM network have been illustrated in Fig. 2 as well as in Supp A of the supplementary material attached with this paper where attention maps for samples of NEH database are also shown. Fig. 2 and the analysis in Supp A show that the relevant morphological deformations are automatically highlighted by the network and the focus converges as we move towards the classification layer. In Fig. 2, the attention maps show both coarser and finer focus over relevant regions of the input scans. Further, comparison of the attention maps of fine-tuned ResNet50 model and the proposed DAM network has been carried out in Supp A where better convergence of our network has been observed. Similarly, experimental results and analysis of advantages of dual attention over single self-attention module are presented in Supp B of supplementary.

B. Selection of Deep Networks, Input Image Size and Cross-database Analysis

TABLE III
PERFORMANCE OF MULTI-LEVEL DAM WITH DIFFERENT BASE PRE-TRAINED DEEP LEARNED NETWORKS.

Pre-trained Base Network	Model Parameters (in millions)	Duke Database			NEH Database		
		Acuuracy (%)	Precision (%)	Recall (%)	Accuarcy (%)	Precision (%)	Recall (%)
VGG-Face	16.35	100	100	100	99.81	99.81	99.81
ResNet50	23.54	99.97	99.97	99.97	99.62	99.60	99.62
VGG16	16.35	100	100	100	99.25	99.26	99.25

We have conducted several experiments to show that our proposed multi-level DAM can work with pre-trained deep networks, such as VGG-Face [33], ResNet50 and VGG16 [34]. The results including the number of model parameters are shown in Table III for the 5-fold CV protocol. It is evident from the experimental results that the selection of deep pre-trained framework for the proposed architecture does not affect much its diagnostic accuracy when evaluated on two macular OCT datasets. Supp C in the supplementary shows the experimental results with varying input image size, from where it is evident that reducing the input image size to one quarter leads to a slight reduction in performance on both the databases. Table IV in the supplementary, Supp D, shows the experimental results on cross-database analysis using our DAM approach on Duke and NEH databases.

V. CONCLUSIONS AND FUTURE WORK

This work proposes a multi-level DAM that helps in the robust classification of macular diseases: AMD and DME from normal macular imaging using retinal OCT scans. Our proposed multi-level DAM approach takes into account both multi-level features based attention arising from locally deformation-aware feature generation and self-attention mechanism focusing on higher entropy regions of the finer features. The newly developed DAM trains the network to focus on the relevant regions of the OCT B-scans and reduces the number of parameters of the base CNN. The proposed method does not require any pre-processing steps, such as RoI extraction, denoising and retinal flattening and hence a fully automatic and end-to-end trainable model is developed. Experimental results and detail analysis of our proposed approach show superiority over state-of-the-art methodologies. The effectiveness of the model can be explored for diagnosis of other macular pathologies with symptoms in peripheral region [3], [4].

REFERENCES

[1] S. Mehta, "Age-related macular degeneration," *Primary Care*, vol. 42, no. 3, pp. 377–391, Sep 2015.

[2] S. Pershing, E. A. Enns, B. Matesic, D. K. Owens, and J. D. Goldhaber-Fiebert, "Cost-effectiveness of treatment of diabetic macular edema," *Ann. Internal Med.*, vol. 160, no. 1, pp. 18–29, 2014.

[3] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. ODonoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, p. 1342, 2018.

[4] L. G. Fritsche, W. Chen, M. Schu, B. L. Yaspan, Y. Yu, G. Thorleifsson, D. J. Zack, S. Arakawa, V. Cipriani, S. Ripke *et al.*, "Seven new loci associated with age-related macular degeneration," *Nature genetics*, vol. 45, no. 4, p. 433, 2013.

[5] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, and T. F. *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[6] P. P. Srinivasan, L. A. Kim, P. S. Metgtu, S. W. Cousins, G. M. Comer, J. A. Izatt, and S. Fariu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomed. Opt. Exp.*, vol. 5, no. 10, pp. 3568–3577, 2014.

[7] Y. Wang, Y. Zhang, Z. Yao, R. Zhao, and F. Zhou, "Machine learning based detection of age-related macular degeneration (amd) and diabetic macular edema (dme) from optical coherence tomography (oct) images," *Biomed Opt Express*, vol. 7, no. 12, pp. 4928–4940, Dec 2016.

[8] Y. Sun, S. Li, and Z. Sun, "Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning," *J. Biomed. Opt.*, vol. 22, no. 1, p. 16012, Jan 2017.

[9] Y. Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman, and J. M. Rehg, "Automated macular pathology diagnosis in retinal oct images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding," *Med. Image Anal.*, vol. 15, no. 5, pp. 748–759, 2011.

[10] R. Rasti, A. Mehridehnavi, H. Rabbani, and F. Hajizadeh, "Wavelet-based convolutional mixture of experts model: An application to automatic diagnosis of abnormal macula in retinal optical coherence tomography images," in *Proc. 10th Iranian Conf. Machine Vision and Image Processing (MVIP)*, pp. 192–196.

[11] L. Huang, X. He, L. Fang, H. Rabbani, and X. Chen, "Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network," *IEEE Signal Process. Lett.*, vol. 26, no. 7, pp. 1026–1030, July 2019.

[12] B. Khagi, C. G. Lee, and G. Kwon, "Alzheimers disease classification from brain mri based on transfer learning from cnn," in *2018 11th Biomedical Engineering International Conference (BMEiCON)*, Nov 2018, pp. 1–4.

[13] Y. Rong, D. Xiang, W. Zhu, K. Yu, F. Shi, Z. Fan, and X. Chen, "Surrogate-assisted retinal oct image classification based on convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 253–263, Jan 2019.

[14] R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh, "Macular oct classification using a multi-scale convolutional neural network ensemble," *IEEE Trans. Med. Imag.*, vol. 37, no. 4, pp. 1024–1034, Apr 2018.

[15] S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomed. optics express*, vol. 8, no. 2, pp. 579–592, 2017.

[16] Q. Ji, W. He, J. Huang, and Y. Sun, "Efficient deep learning-based automated pathology identification in retinal optical coherence tomography images," *Algorithms*, vol. 11, no. 6, 2018.

[17] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottomup and top-down attention for image captioning and vqa," *CoRR*, vol. abs/1707.07998, 2017. [Online]. Available: http://arxiv.org/abs/1707.07998

[18] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention mode," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1181–1185, Aug 2018.

[19] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, 2018.

[20] Y. Xiang, Q. Chen, X. Wang, and Y. Qin, "Answer selection in community question answering via attentive neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 505–509, Apr 2017.

[21] Z. Zhang, P. Chen, M. Sapkota, and L. Yang, "Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references," in *Int. Conf. Med. Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 320–328.

[22] J. Schlemper, O. Oktay, M. Schaap, M. P. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *CoRR*, vol. abs/1808.08114, 2018. [Online]. Available: http://arxiv.org/abs/1808.08114

[23] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *CoRR*, vol. abs/1801.09927, 2018. [Online]. Available: http://arxiv.org/abs/1801.09927

[24] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, "Ultrasound image segmentation: A deeply supervised network with attention to boundaries," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1637–1648, Jun 2019.

[25] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018. [Online]. Available: http://arxiv.org/abs/1804.03999

[26] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE Trans. Med. Imag.*, Feb 2019.

[27] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *arXiv preprint arXiv:1809.02983*, 2018.

[28] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 299–307.

[29] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Aˆ 2-nets: Double attention networks," in *Adv. Neural Inf. Process. Sys.*, 2018, pp. 352–361.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit., CVPR*, 2016, pp. 770–778.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, Dec 2015.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Sys.*, 2017, pp. 5998–6008.

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, Sep 2015, pp. 41.1–41.12.

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

Supplementary

SUPP A

Fig. 1 shows the attention maps obtained from Conv4_2b layer of network for a sample of each disease taken from NEH database. The convergence pattern of the network can be observed from the given figure highlighting the relevant regions for each case. Fig. 2 illustrates the difference between the attention maps of the fine-tuned ResNet50 (without attention) network and the multi-level DAM architecture. It can be inferred from the figure that the maps of our method is more populated towards the macular part of the scan and has focus towards the higher entropy regions.

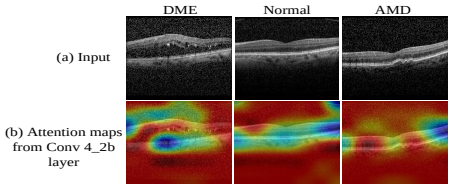

Fig. 1. Attention maps for different samples of NEH database from Conv4_2b layer of the network. Here, blue color denotes the highest attention while red denotes the lowest attention.

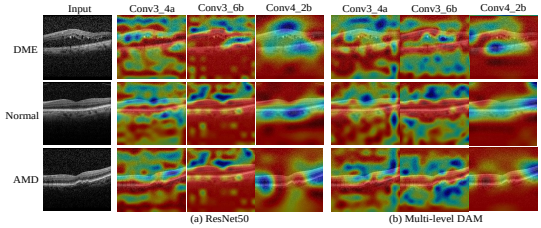

Fig. 2. Comparison of attention maps of different samples of NEH database obtained from ResNet50 and multi-level DAM networks. Here, blue color denotes the highest attention while red denotes the lowest attention.

SUPP B

A comparison of the dual attention mechanism with a single self-attention module based network has also been presented in Fig. 3 as well as in Tables I and II. It can be observed from the given figure that the dual attention mechanism produces more focussed maps than the self-attention based network hence illustrating the effectiveness of our proposed method. Tables I and II depict marginal improvement in the performance in case of dual attention technique for 5-fold CV protocol and good improvement for LPO protocol on both the databases.

TABLE I
COMPARISON OF MULTI-LEVEL DAM WITH A SINGLE ATTENTION BASED MODEL FOR 5-FOLD CV PROTOCOL.

Networks	Duke Database			NEH Database		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Multi-level DAM	99.97	99.97	99.97	99.62	99.60	99.62
With self attention module only	99.91	99.91	99.91	99.60	99.60	99.60

TABLE II
COMPARISON OF MULTI-LEVEL DAM WITH A SINGLE ATTENTION BASED MODEL FOR ONE TEST CASE OF LPO PROTOCL

Networks	Duke Database			NEH Database		
	Accuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
Multi-level DAM	97.38	97.41	97.59	97.91	98.14	97.13
With self attention module only	94.38	94.52	94.84	96.86	96.93	95.62

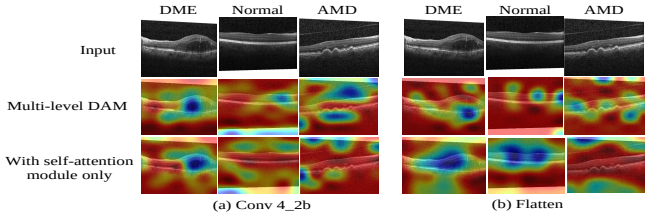

Fig. 3. Comparison of attention maps of different samples of Duke database obtained from multi-level DAM and single self-attention module based networks. Here, blue color denotes the highest attention while red denotes the lowest attention.

SUPP C

We have carried out experiments to study the effect of resizing the scans on the classification performance which is reported here in Table III. The resizing leads to a marginal decrease in performance of the network but gives slight reduction in number of model parameters. Hence, there is a slight trade-off between obtained performance and the computation involved with the change in size of the input images.

TABLE III
PERFORMANCE OF MULTI-LEVEL DAM WITH DIFFERENT DIMENSIONS OF INPUT OCT SCANS.

Input Size	Model Parameters (in millions)	Duke Database			NEH Database		
		Acuuracy (%)	Precision (%)	Recall (%)	Accuracy (%)	Precision (%)	Recall (%)
512×512	23.56	100	100	100	99.86	99.86	99.86
384×384	23.55	100	100	100	99.72	99.72	99.72
224×224	23.54	99.97	99.97	99.97	99.62	99.60	99.62
198×198	23.54	99.97	99.97	99.97	99.10	99.16	99.10

SUPP D

Cross-database analysis of the proposed model has been shown in Table IV. It can be noted that the performance on Duke database when the network is trained on NEH database is better, however, vice-versa case does not hold true. The overall performance is far below that of the well-practised protocols and requires new research direction for developing generic network models.

TABLE IV
CROSS DATABASE ANALYSIS OF THE PROPOSED NETWORK.

Performance Metrics	Trained of Duke tested on NEH	Trained on NEH tested on Duke
Accuracy (%)	43.59	66.58
Precision (%)	51.48	82.07
Recall (%)	46.74	61.10