This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

Attention to Lesion: Lesion-Aware Convolutional Neural Network for Retinal Optical Coherence Tomography Image Classification

Leyuan Fang, *Senior Member, IEEE*, Chong Wang, Shutao Li, *Fellow, IEEE*, Hossein Rabbani, *Senior Member, IEEE*, Xiangdong Chen, and Zhimin Liu

***Abstract*—Automatic and accurate classification of retinal optical coherence tomography (OCT) images is essential to assist ophthalmologist in the diagnosis and grading of macular diseases. Clinically, ophthalmologists usually diagnose macular diseases according to the structures of macular lesions, whose morphologies, size, and numbers are important criteria. In this paper, we propose a novel lesion-aware convolutional neural network (LACNN) method for retinal OCT image classification, in which retinal lesions within OCT images are utilized to guide the CNN to achieve more accurate classification. The LACNN simulates the ophthalmologists' diagnosis that focuses on local lesion-related regions when analyzing the OCT image. Specifically, we firstly design a lesion detection network (LDN) to generate a soft attention map from the whole OCT image. The attention map is then incorporated into a classification network to weight the contributions of local convolutional representations. Guided by the lesion attention map, the classification network can utilize the information from local lesion-related regions to further accelerate the network training process and improve the OCT classification. Our experimental results on two clinically acquired OCT datasets demonstrate the effectiveness and efficiency of the proposed LACNN method for retinal OCT image classification.**

***Index Terms*—optical coherence tomography, convolutional neural network, attention network, retinal lesion, image classification.**

I. Introduction

MACULA is the central part of the retina and is mainly responsible for the central vision. The healthiness of macula might be affected by a number of pathologies, including age-related macular degeneration (AMD), choroidal neovascularization (CNV), and diabetic macular edema (DME) [1]–[5]. AMD is a degenerative retinal disease that is the main leading cause of severe visual impairment in older adults [2]. The common clinical characteristic of AMD is the presence of drusen, which is the asymptomatic deposition of extracellular material located between the retinal pigment epithelium (RPE) and the inner collagenous layer of Bruch's membrane [3]. The advanced stage of AMD, also termed as CNV [4], typically results in irreversible macular lesions, such as subretinal and intraretinal fluid accumulation, RPE detachment and fibrotic scars. DME is characterized by fluid-filled cysts and hard exudates within the retina, as well as retinal thickening, which are caused by abnormal leakage from damaged retinal blood vessels [5]. Clinical diagnosis and grading of macular diseases rely on the detection of the macular structures of lesions or abnormalities, (e.g., drusen, scars, fluid, cysts, and exudates), whose morphologies, size, and numbers can be used to determine the types of eye diseases by ophthalmologists. Therefore, it is essential to investigate the macular lesions for the clinical diagnosis and treatment of ophthalmic diseases.

Optical coherence tomography (OCT) allows in vivo 3D cross sectional imaging of human tissue at micrometer resolutions [6]–[8], which has been widely employed in diverse medical applications [9]–[11], especially for diagnostic ophthalmology. High resolution OCT imaging technique enables the sensitive detection of multiple retinal cell layers and quantitative assessment of these macular lesions within retina [12]–[14]. Examples of various macular lesions of retina in spectral-domain OCT (SD-OCT) images are shown in Fig. 1. In the clinical diagnosis, ophthalmologists need to manually identify these lesions at each cross-section of the OCT volume and then make diagnostic decisions of the diseases. Such manual analysis is time-consuming and demanding for expert graders, which is also prone to yield subjective results. Consequently, an automatic and reliable computer assisted OCT image analysis technique is required for efficient diagnosis of eye diseases.

During the past decades, numerous OCT classification algorithms have been developed [15]–[24], which generally consist of the following key components: image preprocessing (e.g., image denosing [25], [26] and curvature correction [19], [27]), feature extraction (e.g., local binary patterns (LBP)

This work was supported in part by the National Natural Science Foundation under Grant No. 61771192, the National Natural Science Foundation for Young Scientist of China under Grant No. 61501180, China Postdoctoral Science Foundation funded Project No. 2017T100597, and National Natural Science Foundation of Hunan Province Grant No. 2018JJ3077. Leyuan Fang and Chong Wang contributed equally to this work.

L. Fang, C. Wang, and S. Li are with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China. (email: fangleyuan@gmail.com; chongwang@hnu.edu.cn; shutao_li@hnu.edu.cn).

H. Rabbani is with the Medical Image & Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, 81745319, Iran. (email: h_rabbani@med.mui.ac.ir).

X. Chen and Z. Liu are with the First Hospital of Hunan University of Chinese Medicine, Department of Ophthalmology, Changsha, 410082, China. (email: 564259166@qq.com; 2548088962@qq.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

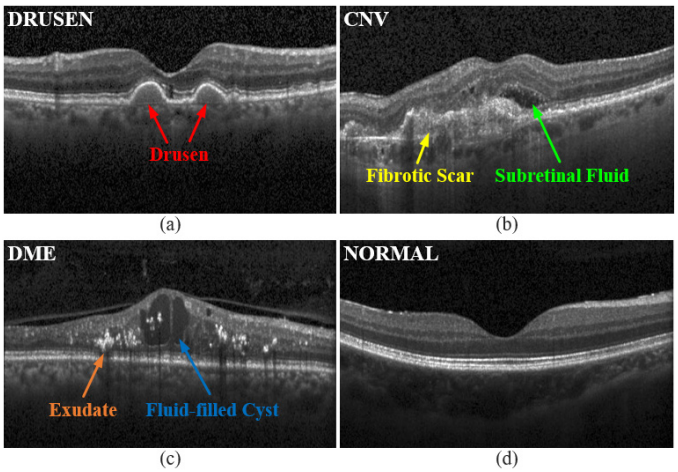

Fig. 1. SD-OCT B-Scans of the retina showing Drusen, CNV, DME and Normal macular, respectively. (a) drusen (red arrows); (b) fibrotic scarring (yellow arrow) and subretinal fluid (green arrow); (c) hard exudates (orange arrow) and fluid-filled cysts (blue arrow); (d) normal retina with clear retinal layer boundaries.

[15], [17], [18], histogram of oriented gradients (HOG) [19], scale-invariant feature transform (SIFT) [21], and bag of words (BoW) descriptor [16]), and classifier design (e.g., support vector machine (SVM) [15], [17], [19], [21], Bayesian classifier [18], and random forest [16], [24]). Very recently, deep learning [28], [29] has demonstrated to be a very powerful technique in the fields of computer vision and pattern recognition. One of the typical deep learning model, called as convolutional neural network (CNN) [28], [30], can automatically learn a hierarchy of abstract features from large training datasets. Several recent works have extended the CNN framework to retinal OCT image analysis, such as retinal layers segmentation [31], [32] and OCT image classification [33]–[37]. In [35], the CNN has been applied to surrogate-assisted OCT images classification. In [36], the wavelet-based CNN is introduced to extract deep wavelet features and then the random forests classifier is utilized for the classification of the macular OCT images. In [37], the U-Net architecture is first used to segment the OCT images, and then the segmented images are used for the diagnosis of retinal diseases. In addition, a new CNN-based technique, known as transfer learning [38], is introduced to discriminate macular diseases with significantly fewer training OCT images, while still demonstrating promising results [39], [40].

Ophthalmologists can easily identify the salient areas (usually corresponding to lesions) in an OCT image mainly due to the attention mechanism of the human visual perception system [41]. The human visual attention mechanism is that one selectively focuses on parts of the visual space to capture salient information where it is needed, instead of processing the whole scene at a time [42]. Based on the feature-integration theory of attention [43], regions in a scene are marked as saliency depending on the difference with their surroundings. Such saliency reflected in OCT images is lesions, which always attract most ophthalmologist's attention during their clinical diagnosis. Therefore, it is natural to incorporate the lesion attention mechanism into the OCT classification models.

In this paper, we propose a novel lesion-aware convolutional neural network (LACNN) method for retinal OCT image classification. Motivated by the ophthalmologist's diagnosis process, we firstly design a lesion detection network (LDN) which can detect various kinds of macular lesions and create the corresponding attention maps. The detected attention maps are used to softly weight the convolutional feature maps of the classification network and enable the proposed LACNN model to focus on the most relevant information, such as macular lesions. Guided by the lesion-related information, the classification network utilizes the information from local lesion-related regions to achieve efficient and accurate OCT classification.

Note that, the strategy of attention-based CNN has been previously applied to several other medical image analysis problems [44]–[48]. In this paper, we propose a LACNN method for retinal OCT image classification, in which the macular lesions detected by the LDN are used to guide the classification network to focus on more discriminative information from local lesion-related regions. Specifically, the main contributions of our paper are described as follows.

1. We propose an attention-based CNN method for retinal OCT image classification. To the best of our knowledge, this is the first time that an attention model is introduced in the field of OCT image analysis.
2. In order to utilize the lesion information, the LACNN proposes a lesion-attention module, which can effectively enhance the features from local lesion-related regions while still preserving the meaningful structures in global OCT images.
3. Without any image preprocessing tricks, the LACNN method achieves desirable classification results, which can be considered as an effective computer-aided diagnosis tool for clinically OCT-based eye disease diagnosis.

The remainder of this paper is organized as follows: Section II describes related works, including the CNN model and CNN based OCT classification method. The proposed LACNN method is introduced in Section III. Experimental results on clinical OCT datasets are shown in Section IV. Section V concludes this paper and suggests future works.

II. Related Work

A. CNN Model

CNN has been successfully applied to a variety of computer vision and image processing applications, including image classification [30], [49], object detection [50], [51], and image segmentation [52], [53]. The typical CNN comprises several convolutional (Conv) layers, nonlinearity, pooling layers, and fully connected (FC) layers, which are described as follows.

The convolutional layer is the most important component of CNN. For each convolutional layer, a set of two-dimensional (2D) kernels are learned to express local spatial connectivity with the previous layer, which generates multiple convolutional feature maps by conducting the convolution operation between the input signals with these kernels:

$$\mathbf{X}_j^l = \sum_i \sigma(\mathbf{X}_i^{l-1} * \mathbf{w}_{i,j}^l + \mathbf{b}_j^l), \quad (1)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

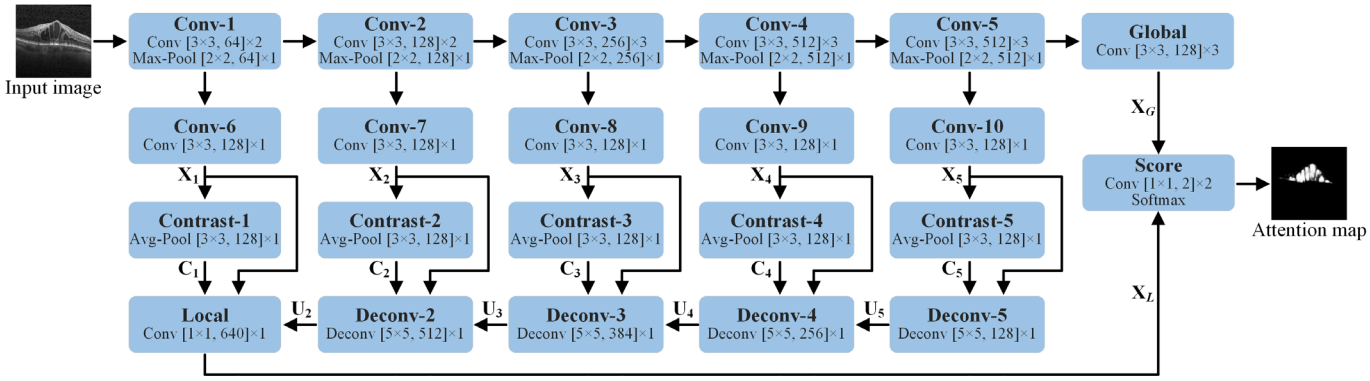

Fig. 2. Outline of the LDN for macular lesions detection. Values inside the bracket refer to the kernel size and the number of feature maps, and value outside the bracket represents the number of stacked layers with same structures. For example, there are two convolutional layers with same structures (kernel size 3×3, 64 feature maps) and one max-pooling layer (kernel size 2×2, 64 feature maps) in Conv-1 block. The Conv, Contrast and Deconv represent the convolution block, local contrast processing block, and deconvolution block, respectively.

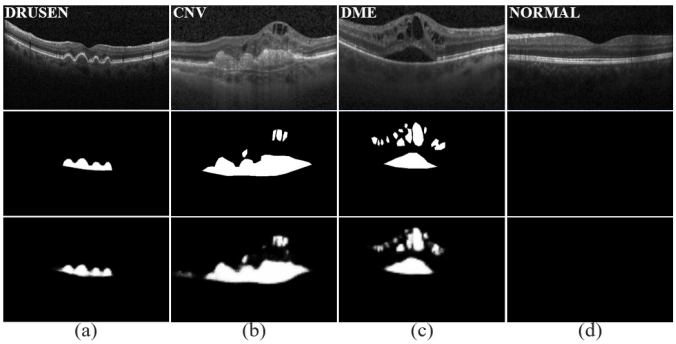

Fig. 3. The lesion detection results of LDN. Top row: original retinal OCT images of various macular diseases. Middle row: The legion-related regions coarsely annotated by ophthalmologists. Bottom row: The attention maps generated by the LDN. Note that, the LDN can effectively detect various kinds of lesions, such as (a) drusen, (b) scars, (c) cysts and subretinal fluid, and (d) no response for normal macula.

where $*$ represents the 2D convolution operation, $\mathbf{w}_{i,j}^{l}$ and $\mathbf{b}_{j}^{l}$ are the learnable kernel weight matrix and additive basis, respectively. $\mathbf{X}_{j}^{l} \in \mathbb{R}^{M\times N\times C}$ denotes the output feature maps of the convolutional layer l, which is interpreted as a set of $M\times N$ local descriptors of C channels. $\mathbf{X}_{i}^{l-1}$ refers to the input feature maps of the previous layer l-1. Usually, each kernel is convolved with all the feature maps of the previous layer and creates a 2D output. Outputs of all kernels are stacked together to create the 3D output feature map. Convolutional layers are usually followed by a non-linear activation function σ. The rectifier linear units (ReLU) [30] is the mostly used activation function due to its fast convergence, which is defined as:

$$\sigma(x)=\max(x,0). \tag{2}$$

To reduce computation complexity and improve translation invariance, a pooling layer is often applied after convolutional layers, which fuses nearby spatial information in the same feature map with the max or averaging operations [29]. By stacking a series of convolutional layers interleaved with non-linearity and pooling layers, CNN is capable of capturing hierarchical representations as powerful image descriptions.

After several convolutional and pooling layers, one or more fully connected layers are used to integrate all features of the previous layer into a feature vector. Finally, softmax function [28] is used to transform the last feature vector into probability distribution of outputs.

B. CNN Based OCT Classification

In ophthalmology, CNN has been recently applied for the detection of diabetic retinopathy from fundus photographs [54], segmentation of retinal layers [31], and quantification of macular fluid in OCT images [55]. For the retinal OCT classification, CNN is expected to extract high-level features from original OCT images [33], [39], [40]. Formally, given an input OCT image $\mathbf{x}$, the output of CNN can be computed by a series of convolutional, pooling, and FC operations:

$$\boldsymbol{a}_k(\mathbf{x}) = f(\mathbf{W}\mathbf{x}+\mathbf{b}), \tag{3}$$

where f represents a composite function, which is obtained by multiple linear and nonlinear operations, and $\boldsymbol{a}_k$ denotes the class score for class k. $\mathbf{W}$ and $\mathbf{b}$ are the weight and bias parameters, respectively. As described in Section II-A, the softmax function is used to produce the probability distribution of outputs that the image $\mathbf{x}$ belongs to each class:

$$p(k \mid \mathbf{x}) = \frac{e^{\boldsymbol{a}_k}}{\sum_{k=1}^{K} e^{\boldsymbol{a}_k}}, \tag{4}$$

where K is the number of the output classes. CNN is usually trained with backpropagation rule [56] and stochastic gradient descent (SGD) algorithm by minimizing the cross-entropy error between the probabilistic outputs and the one-hot labels:

$$L_{CLS} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} \mathrm{I}(l_i = k)\log p(k \mid \mathbf{x}_i) + \lambda \left\|\mathbf{W}\right\|_2, \tag{5}$$

where m is the number of samples in per mini-batch, l_i denotes the class label of image $\mathbf{x}_i$, $\mathrm{I}(\cdot)$ is an indicator function which equals one if l_i is equal to k, and $\left\|\mathbf{W}\right\|_2$ is the regularization term of weights with the decay factor λ. Once the optimization is completed, the trained CNN can be used to predict the label of test image $\mathbf{x}^*$ based on the maximum probability $p(k \mid \mathbf{x}^*)$:

$$Class(x^*) = \arg\max_{k=1,2,\ldots,K} p(k \mid \mathbf{x}^*). \tag{6}$$

III. Proposed LACNN Method for OCT Classification

Clinically, ophthalmologists usually diagnose macular diseases according to the structures of different macular lesions. Motivated by the ophthalmologist's diagnosis process, we propose the lesion-aware convolution neural network (LACNN) method for retinal OCT image classification. In this section, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

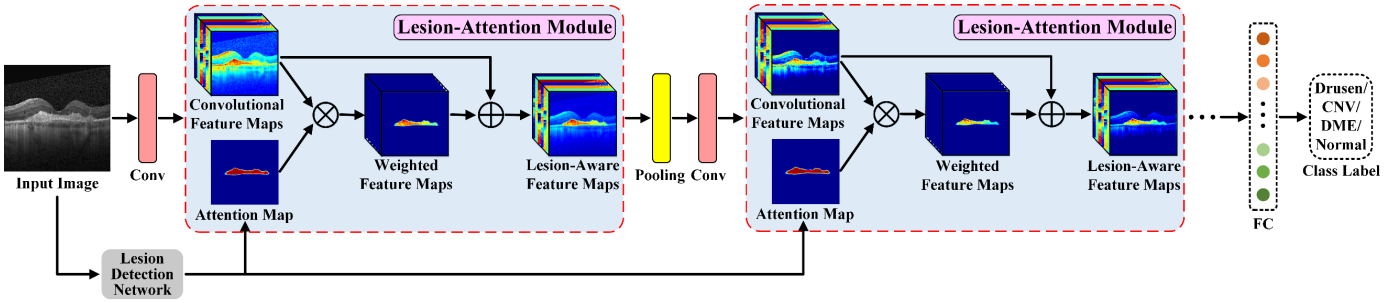

Fig. 4. The framework of LACNN for retinal OCT image classification.

introduce the proposed LACNN method, which utilizes the information from local lesion-related regions to improve the retinal OCT classification. We will first describe the lesion detection network (LDN) for macular lesions detection in Section III-A. Then, we illustrate the architecture of LACNN in Section III-B. Section III-C presents the training process of LACNN.

A. Lesion Detection Network

To utilize the local discriminative information of various macular lesions for the retinal OCT classification, we firstly need to accurately detect them from OCT images. However, as indicated in [15], manually designed features or rules are hard to detect and identify the complex structures of lesions due to the co-existence of multiple lesions, high variabilities of lesion in shape, size and magnitude, and the existence of blood vessels. Therefore, we design a deep learning based lesion detection network (called as LDN) to effectively detect various kinds of macular lesions. The LDN aims to obtain attention map when we train the network for the task of lesions detection. In this way, the network's prediction is based on the macular lesions, which we expect the network to focus on. We achieve this by making the network's attention trainable in an end-to-end fashion. The outline of LDN is shown in Fig. 2 and is described below in detail.

The LDN is a CNN model for the salient lesions detection [57], which consists of convolution blocks, local contrast processing blocks, and deconvolution blocks. All of the blocks are organized in a 4 × 5 grid, in which each column of the grid processes resolution-specific features.

The first row of the LDN comprises five convolution blocks derived from VGG16 [58] (Conv-1 to Conv-5), and these blocks contain two or three convolutional layers with 3 × 3 kernel followed by one max pooling layer which down-samples the feature maps by a factor of 2 (e.g., 224 × 224, 112 × 112, ..., 7 × 7). The rightmost block of the first row generates global feature $\mathbf{X}_G$ representing the global context of input image by using three convolutional layers with 3 × 3 kernel.

The second row is a set of five convolution blocks (Conv-6 to Conv-10), which are connected to the Conv-1 to Conv-5 blocks in the first row, respectively. The aim of these blocks is to extract multi-level local convolutional features $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_5\}$. Each of these convolution blocks has a 3 × 3 kernel with 128 channels.

The third row comprises five local contrast processing blocks (Contrast-1 to Contrast-5). Since the macular lesions are usually different with their surrounding areas in contrast, and result in distinctive features of lesions within CNN, the local contrast processing block is adopted along each resolution level to capture the local contrast information and improve the capability of lesions detection. This is achieved by subtracting the value of 3 × 3 average pooling from its local convolutional feature maps:

$$\mathbf{C}_i = \mathbf{X}_i - \mathrm{AvgPool}(\mathbf{X}_i), \tag{7}$$

where $\mathbf{C}_i$ (i = 1, 2, …, 5) denotes the local contrast features and AvgPool(·) is a 3 × 3 average pooling operation.

The last row has a set of deconvolution blocks (Deconv-2 to Deconv-5), which are used to increase the spatial resolution of the feature maps from 7 × 7 (bottom right) to 112 × 112 (bottom left). The deconvolution block firstly concatenates the information of local convolutional feature $\mathbf{X}_i$, local contrast feature $\mathbf{C}_i$, and the up-sampled feature $\mathbf{U}_{i+1}$ of previous deconvolution block. The combined feature is then fed into a deconvolutional layer, which increases the spatial resolution of feature maps by a factor of 2:

$$\mathbf{U}_i = \mathrm{Deconv}(\mathbf{X}_i, \mathbf{C}_i, \mathbf{U}_{i+1}), \tag{8}$$

where $\mathbf{U}_i$ is the resulting up-sampled features and Deconv(·) represents a deconvolutional layer with a stride of 2 and a 5 × 5 kernel. The leftmost block of the last row extracts the final local feature $\mathbf{X}_L$ by using a convolutional layer with a 1 × 1 kernel.

The extracted global feature $\mathbf{X}_G$ and local feature $\mathbf{X}_L$ are separately mapped into categorical scores by a convolutional layer with 1 × 1 kernel and 2 channels in the Score block. The obtained scores are fused together by means of element-wise summation and finally incorporated into a softmax layer to compute the salient lesions probability. For an input OCT image, the LDN outputs an attention map at half of the input resolution (we resize back to input size with bilinear interpolation), in which each spatial location represents the probability for each pixel of input image belonging to lesions or not. The loss function for LDN is defined as the binary cross-entropy loss:

$$L_{LDN} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{c\in\{0,1\}} \mathrm{I}(y_j = c)\log p(\hat{y}_j = c), \tag{9}$$

where n is the total number of pixels of input image, y_j and $\hat{y}_j$ are the truth and predicted labels at the pixel location j, respectively. The LDN is trained by using a small amount of OCT images with extra lesion-level annotations to control the attention map learning process. Some lesion detection results of the LDN are illustrated in Fig. 3. As can be observed, the LDN is able to detect various kinds of complex lesions which are close to their ground truths, while still has no responses on normal retina OCT image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

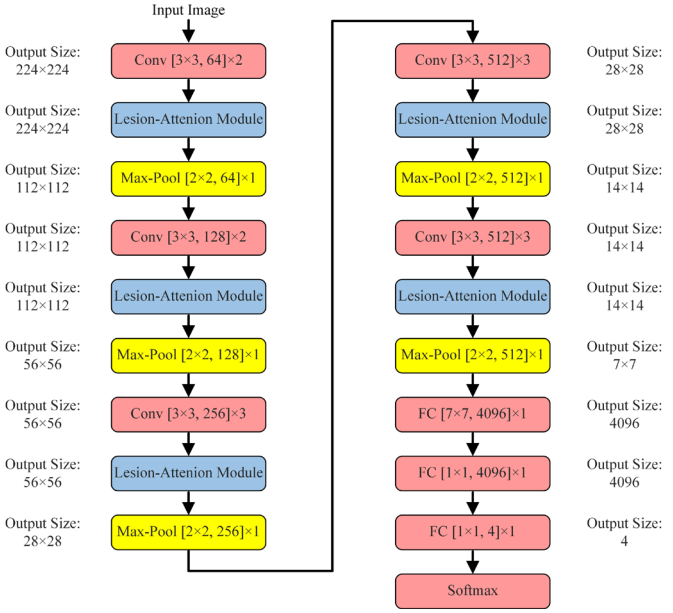

Fig. 5. The detailed architecture of LACNN. Values inside the bracket refer to the kernel size and the number of feature maps, and value outside the bracket represents the number of stacked layers with same structures.

B. The Architecture of LACNN

The architecture of LACNN is illustrated in Fig. 4. Basically, the LACNN is constructed by a series of lesion-attention modules, convolutional layers, and pooling layers. In each lesion-attention module, the feature maps from previous convolutional layers and attention map from LDN are the inputs, and lesion-aware feature maps are the output. The lesion-attention module is detailed as below.

Given an OCT image $\mathbf{x}$, $\boldsymbol{G}(\mathbf{x})$ is denoted as the convolutional feature maps from a convolutional layer. To exploit the information of lesions to guide the OCT classification network, we use the attention map $\boldsymbol{W}_i$ generated by LDN to softly weight the representation $\boldsymbol{G}(\mathbf{x})$, and obtain a local compact representation $\boldsymbol{L}(\mathbf{x})$:

$$\boldsymbol{L}_{i,c}(\mathbf{x}) = \boldsymbol{G}_{i,c}(\mathbf{x}) \otimes \boldsymbol{W}_i \,, \tag{10}$$

where i ranges over all spatial positions, $c \in \{1, 2, ..., C\}$ is the index of the feature map channel, and $\otimes$ denotes the element-wise product. As in (10), the attention map can be considered as a feature selector, which enhances the discrimination of meaningful features and suppresses trivial features from convolutional layers. Note that, since the pooling layer among different modules will change the spatial dimension of convolutional feature maps, we down-sample the attention map to match the spatial resolution of the representation $\boldsymbol{G}(\mathbf{x})$ in each module.

Furthermore, the local weighted CNN feature maps (corresponding to different kinds of macular lesions) are integrated with the original CNN feature maps, which has the following two reasons. Firstly, even though lesions are very important for diagnosing the diseases, other information from the whole OCT images (e.g., lesions in different positions, morphology of retinal layers, and layer thickness) may still be beneficial to the OCT image classification. Secondly, since no lesions exist in normal macula, the attention maps with values close to zero (see Fig. 3 (d)) can reduce the values of feature maps to zeros, and lead to uniform categorical probability distribution in (4), thus losing the discriminative ability for the whole network. Therefore, similar to the residual learning [59], we introduce an identical mapping structure to fuse the local weighted CNN feature maps with original CNN feature maps. Specifically, we calculate outputs $\boldsymbol{H}(\mathbf{x})$ of the lesion-attention module as follows:

$$\boldsymbol{H}_{i,c}(\mathbf{x}) = \boldsymbol{L}_{i,c}(\mathbf{x}) + \boldsymbol{G}_{i,c}(\mathbf{x}) \,, \tag{11}$$

where the operation $\boldsymbol{L}(\mathbf{x}) + \boldsymbol{G}(\mathbf{x})$ is performed by a shortcut connection and element-wise summation. The obtained $\boldsymbol{H}(\mathbf{x})$ consists of the information from both local lesion-related features and global OCT image, which is expected to improve the OCT classification. Compared with the typical CNN, the main contribution of the proposed LACNN lies on our lesion-attention module, which can extract discriminative and meaningful features while still preserve the useful structures in global OCT images.

Adopting a series of lesion-attention modules can gradually refine the feature maps, and thus CNN features become more and more discriminative for specific lesions as network goes deeper. The final class decisions for LACNN with input OCT image $\mathbf{x}$ can be obtained according to (6). Note that, our lesion-attention module can be easily adapted to any state-of-the-art network structures. In our work, we use VGG16 [58] as baseline network to construct the LACNN architecture. Specifically, we incorporate the lesion-attention module into VGG16 before each pooling layer since the Max-pooling operation will further preserve discriminative features. The detailed architecture of LACNN is shown in Fig. 5. In addition, we have also attempted to insert lesion-attention module into all layers of VGG16 (also detailed in the experimental part), which does not show significant performance improvement, but greatly creating more computation cost.

C. Training Strategy of LACNN

In this paper, we adopt a two-stage training scheme for LACNN, and the overall training procedure is presented below. In each stage, we use the model with the highest accuracy on the validation set for testing.

Stage I. We firstly independently train the LDN using a small amount of roughly annotated lesion-level OCT datasets to minimize the loss L_{LDN} in (9).

Stage II. Once the LDN is optimized in **Stage I**, the weight parameters of the LDN are fixed when training LACNN. When the attention map is obtained by LDN, we feed it into the LACNN and finally train the LACNN by minimizing the classification loss L_{CLS} according to (5).

IV. Experimental Results

A. Clinical Datasets

To evaluate the effectiveness of the proposed LACNN [1] method, we have validated it on two publicly available OCT datasets (UCSD dataset [39] and NEH dataset [34]).

The UCSD dataset is used for 2D slice-based classification and composed of 84484 OCT B-scans (8866 drusen, 37455

[1]Source code of the proposed LACNN method will be freely released on the website: https://sites.google.com/site/leyuanfang/home.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

CNV, 11598 DME, and 26565 normal) acquired from 4686 patients at the Shiley Eye Institute of the University of California San Diego (UCSD). All of the images (Spectralis OCT, Heidelberg Engineering, Germany) were selected from retrospective cohorts of adult patients without exclusion criteria based on age, gender, or race. The final OCT scans are obtained by a horizontal foveal cut from original standard image format according to manufacturer's software and instructions. More details about the dataset can be found in [39]. To develop the LDN for macular lesion detection, we randomly select 2500 B-scans (500 drusen, 500 CNV, 500 DME, and 1000 normal) from the UCSD dataset to establish the lesion-level datasets. Since our LDN is designed to automatically discover macular lesions instead of accurately segmenting them, accurate pixel-level annotations for lesions are not required. In the lesion-level datasets, each OCT B-scan is coarsely annotated to delineate the lesion regions (using bounding polygons) by two ophthalmologists from the department of ophthalmology at the first affiliated hospital of Hunan university of Chinese medicine (HUCM). Fig. 3 (the second row) displays some examples of the annotated lesions from each class B-scan. Note that, the images in lesion-level datasets are excluded from the UCSD datasets when training the LACNN.

The NEH dataset is used for 3D volume-based classification and comprises 148 SD-OCT volumes (48 dry AMD, 50 DME, and 50 normal), acquired by Heidelberg SD-OCT imaging systems at Noor Eye Hospital in Tehran (NEH). Each volume contains a number of B-scans (ranging from 19 to 61). The original axial resolutions of the B-scans are 3.5 μm with the scan-dimension of 8.9 × 7.4 mm^2. More details about the scanning protocols for this dataset can be found in [34].

B. Experimental Settings

The LDN is optimized by using Adam optimizer [60] training on randomly selected samples from the lesion-level datasets. The network is individually trained until convergence is reached. Convergence is determined by calculating the performance on the independent validation set (100 B-scans randomly selected from the lesion-level datasets) at regular intervals during training. At each iteration, the mini-batch size is set to one B-scan. The number of epoch is set to 10 and the initial learning rate was set to 10^{-5}. To improve robustness of the network, we perform data augmentation strategies on each B-scan of the lesion-level datasets by horizontally flipping and rotating between +15 and -15 degrees.

For training the LACNN, we optimize the network using Adam optimizer with batch of 24 images per step and an initial learning rate of 10^{-5}. The weight decay factor λ in (5) is set to 0.0002. At each iteration, the loss of the model was recorded, and at every 10 iterations, the performance of the neural network was assessed using an independent validation set. The maximum number of iterations is chosen to 10 epochs. The training is stopped when the accuracy on the validation set no longer increased. During training of both LACNN and LDN, we resize the original OCT images to 224 × 224 as input. The Xavier algorithm [61] is used for weight initialization of the two networks.

Classification performance is evaluated based on the following evaluation metrics: accuracy (ACC), sensitivity (SE), precision (PR), specificity (SP), and area under the ROC curves (AUC), each of which are computed for independent class. In our four-class classification problem, the sensitivity for independent class is the prediction accuracy, and the specificity is defined in the same way for each class label, where the negative samples are the samples not in the considered class. Due to the imbalance of samples among different classes, the overall sensitivity (OS), overall precision (OP), and overall accuracy (OA) are also computed. The OA is defined as:

$$\text{Overall Accuracy} = \frac{\text{correctly classified samples}}{\text{total number of samples}}. \quad (12)$$

The proposed LACNN method is implemented using the Tensorflow framework [62] with NVIDIA Cuda v8.0 and cuDNN v5.1 accelerated library, and is coded in Python and MATLAB. All experiments are performed under an Ubuntu 16.04 operating system on a machine with CPU Intel Core i7-7700K 3.60 GHz, GPU NVIDIA GeForce 1080 Ti, and 16 GB of RAM.

C. Compared Methods

For the retinal OCT image classification problem, the proposed LACNN method is compared with other well-known classification methods: HOG-SVM [19], Transfer learning [39], VGG16 network [33], and MCME [34]. The HOG-SVM is a machine learning based method which utilizes the multiscale histogram of oriented gradients (HOG) descriptor to extract the feature vector from each B-scan and trains multiple binary SVMs with linear kernel for OCT classification. The other three approaches are based on deep leaning. By using the InceptionV3 architecture pretrained on the ImageNet dataset [30], the transfer learning method freezes all of the convolutional layers and only retrains the last fully connected layer to recognize OCT images among drusen, CNV, DME and normal macula. The VGG16 is a plain network for comparison purpose, which consists of multiple convolutional and pooling layers, and three fully connected layers. The MCME method utilizes the multi-scale convolutional mixture of expert ensemble model for 3-D OCT classification. In our experiments, the hyper-parameters for VGG16 were set to the same as that of LACNN (batch size = 24, initial learning rate = 10^{-5}, weight decay factor λ = 0.0002, number of training epochs = 10), as described in Section IV-B. We implemented the transfer learning method based on the public codes[2], in which the number of training epochs was set to 10 for a fair comparison with other methods, and all the other hyper-parameters were kept unchanged (batch size = 256, initial learning rate = 10^{-3}).

D. Results on UCSD Dataset

We firstly validated the proposed LACNN method on the UCSD dataset. In our experiments, we sequentially divided the whole UCSD datasets into η subsets. In each experiment, LACNN model was trained with one subset and tested on the remaining $(\eta - 1)$ subsets. The experiments were repeated η times with each of the η subsets used exactly once as the training set and the final experimental results were averaged over all the experiments.

Table I shows the quantitative results of different methods on the UCSD dataset (η = 6). As can be observed, deep learning

[2] Codes can be downloaded at: https://data.mendeley.com/datasets/rscbjbr9sj/2

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

TABLE I
CLASSIFICATION RESULTS (IN PERCENTAGE) ON UCSD DATASET. EACH OF THE FOUR ROWS GIVES THE CLASSIFICATION METRICS FOR INDIVIDUAL CLASS DRUSEN, CNV, DME AND NORMAL, RESPECTIVELY. THE BEST RESULTS IN THIS TABLE ARE LABELED IN BOLD

Methods	Classes	ACC	SE	PR	SP	AUC	OA	OS	OP
HOG-SVM [19]	Drusen	90.2 ± 0.4	29.5 ± 3.5	52.6 ± 4.0	**97.0** ± 0.8	81.3 ± 0.9	78.1 ± 0.7	65.3 ± 0.9	71.8 ± 1.3
	CNV	85.7 ± 0.2	87.6 ± 2.1	82.0 ± 1.1	84.0 ± 1.5	92.2 ± 0.2			
	DME	91.4 ± 0.2	53.8 ± 2.1	74.6 ± 0.8	97.2 ± 0.2	87.3 ± 0.5			
	Normal	89.1 ± 0.8	90.4 ± 1.2	78.1 ± 2.0	88.4 ± 1.4	94.6 ± 0.4			
Transfer Learning [39]	Drusen	87.2 ± 0.7	70.7 ± 1.6	42.0 ± 1.7	89.1 ± 0.9	89.2 ± 0.7	79.5 ± 1.0	77.9 ± 0.4	73.1 ± 1.1
	CNV	86.9 ± 1.1	76.2 ± 3.3	**93.9** ± 0.7	95.9 ± 0.7	96.1 ± 0.3			
	DME	91.6 ± 0.7	75.5 ± 3.2	66.9 ± 4.4	94.1 ± 1.3	93.8 ± 0.5			
	Normal	93.3 ± 0.2	88.9 ± 1.3	89.5 ± 1.0	95.2 ± 0.6	98.1 ± 0.1			
VGG16 [58]	Drusen	90.7 ± 1.0	54.7 ± 6.5	54.5 ± 5.3	94.7 ± 1.7	88.7 ± 0.7	83.2 ± 1.2	76.2 ± 0.7	76.4 ± 1.5
	CNV	91.0 ± 1.2	86.6 ± 3.9	93.2 ± 1.5	94.7 ± 1.5	97.2 ± 0.3			
	DME	92.8 ± 0.5	70.9 ± 4.7	74.6 ± 4.7	96.2 ± 1.1	93.6 ± 0.6			
	Normal	91.8 ± 0.3	92.6 ± 3.4	83.3 ± 2.6	91.5 ± 2.0	97.2 ± 0.2			
LACNN	Drusen	**93.6** ± 1.4	**72.5** ± 7.9	**70.0** ± 5.7	95.9 ± 2.1	**93.4** ± 1.5	**90.1** ± 1.4	**86.8** ± 1.3	**86.2** ± 2.3
	CNV	**92.7** ± 1.5	**89.8** ± 4.5	93.5 ± 1.3	**95.1** ± 1.6	**97.7** ± 0.5			
	DME	**96.6** ± 0.2	**87.5** ± 1.5	**86.4** ± 1.6	**98.0** ± 0.3	**97.4** ± 0.4			
	Normal	**97.4** ± 0.2	**97.3** ± 1.0	**94.8** ± 1.1	**97.4** ± 0.5	**99.2** ± 0.2			

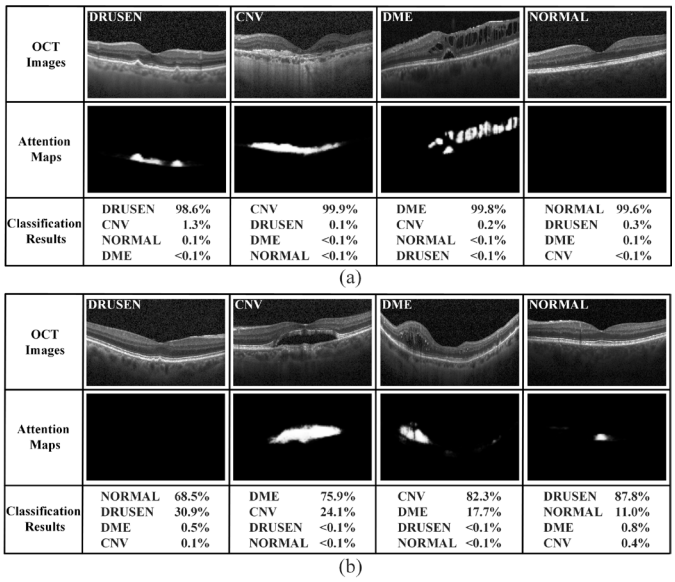

Fig. 6. Examples of classification results of the proposed LACNN. We present the predicted categories and the corresponding probability scores. (a) Correctly classified cases; (b) Inaccurately classified cases.

based methods (e.g., transfer learning, VGG16, and LACNN) achieve a better performance than HOG-SVM method. In addition, by introducing lesion-attention modules, the LACNN achieves much improvement in terms of all quantitative metrics compared with the transfer learning and plain VGG16 network. Specifically, when using one sixth of the OCT data (about 13000 images) for training, we achieve an OA of 90.1%. For each individual class of the Drusen, CNV, DME and Normal, the gains (on the sensitivities) of the LACNN over its counterpart without lesions attention are about 17.8%, 3.2%, 16.6% and 4.7%, respectively, which demonstrates the effectiveness of utilizing macular lesions information to guide CNN for OCT classification. Moreover, the LACNN also outperforms the transfer learning method with the improvements of about 10% in OA, and significant improvement of ACC, SE and SP for individual class can also

TABLE II
CLASSIFICATION RESULTS (IN PERCENTAGE) TOGETHER WITH AVERAGE TEST TIME (IN SECOND/VOLUME) ON NEH DATASET. THE BEST RESULTS IN THIS TABLE ARE LABELED IN BOLD

Methods	OS	OP	F_1-score	AUC	Test Time
Transfer Learning [39]	84.33 ± 6.19	87.00 ± 5.99	84.20 ± 6.20	94.51 ± 3.47	**0.0009**
MCME (l_3-l_2-l_1) [34]	97.30 ± 2.49	97.76 ± 2.11	97.38 ± 2.45	99.30 ± 1.39	0.0074
MCME (l_4-l_3-l_2-l_1) [34]	**99.36** ± 1.33	**99.39** ± 1.21	**99.34** ± 1.34	**99.80** ± 1.19	0.0082
LACNN	99.33 ± 1.49	**99.39** ± 1.36	99.33 ± 1.49	99.40 ± 1.34	0.0206

be observed. This demonstrates that the LACNN can exploit the discriminative lesion features to achieve high performance when only very limited number of samples is available.

Fig. 6 shows examples of correctly (a) and inaccurately (b) classified cases, together with the corresponding probability score for each class. From Fig. 6 (a), we can see that guided by the lesions information the LACNN can correctly classify OCT images with high confidence scores. From Fig. 6(b), we can find that some misclassification cases happen between Drusen (with small lesion areas) and Normal. The reason might be that it is particularly difficult for the LDN to detect the lesion of drusen with small areas. Another misclassification cases happen between CNV (with lesions of subretinal fluid) and DME, because both of them contain severe fluid accumulation with similar visual characteristics.

It should be noted that the average training and test time for LACNN is about 0.0095 and 0.0012 second per B-scan, respectively. For the plain VGG16 network, the training and test time is about 0.0087 and 0.0011 second per B-scan, respectively. This shows that the time complexity of LACNN is similar to that of VGG16, but the LACNN can achieve 8.3% performance gain (in OA) compared with the VGG16.

E. Results on NEH Dataset

We also evaluated the proposed LACNN method on the NEH dataset. In order to achieve fair comparison with [34], we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

TABLE III
CLASSIFICATION RESULTS (OA IN PERCENTAGE) OF DIFFERENT LESION ATTENTION STRATEGIES. THE BEST RESULT IN THIS TABLE IS LABELED IN BOLD

Methods	LACNN-1	LACNN-2	LACNN-3	LACNN-4	LACNN-5	LACNN-A	VGG16	LACNN
OA	89.2 ± 1.2	88.9 ± 1.2	88.0 ± 1.1	87.8 ± 0.8	85.4 ± 2.1	87.3 ± 1.9	83.2 ± 1.2	**90.1** ± 1.4

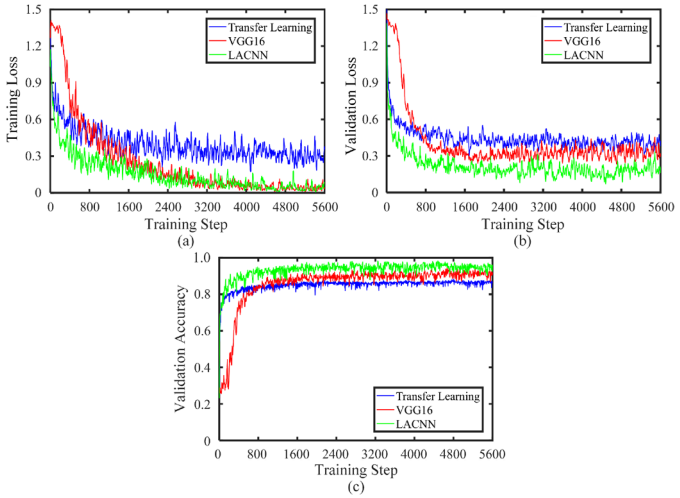

Fig. 7. Comparison of the learning efficiency between three CNN models (Transfer learning, VGG16, and LACNN). (a) Training loss; (b) Validation loss; (c) Validation accuracy.

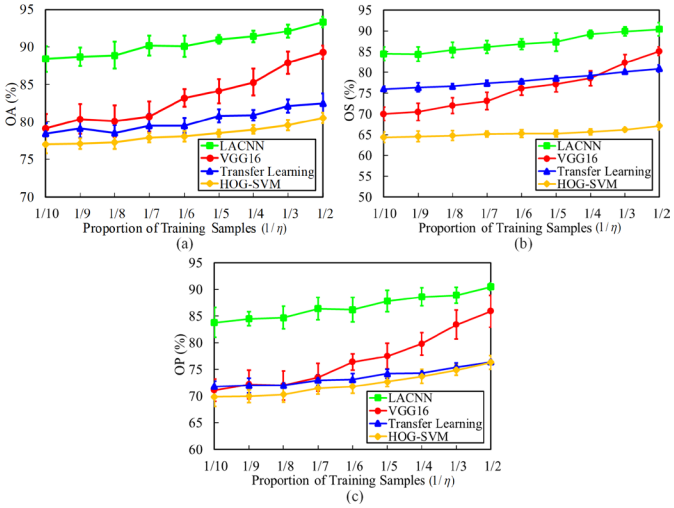

Fig. 8. Effect of different numbers of training samples for HOG-SVM, Transfer Learning, VGG16, and LACNN on UCSD dataset.

performed 5-fold cross validation and utilized the diagnostic strategy adopted in [34]. In our experiments, the diagnostic thresholds for the transfer learning, MCME, and LACNN methods are all chosen to 25%, which can almost obtain their best results. We augmented the NEH dataset by horizontally flipping and rotating between +15 and -15 degrees to alleviate overfitting. In addition, a dropout factor of 70% was used in the first two FC layers of LACNN and the weight decay factor λ for LACNN was set to 0.002. Other experimental parameters for LACNN and transfer learning methods (e.g., batch size, learning rate, number of training epochs) were kept unchanged as described in Section IV-B.

The classification results are shown in Table II. As can be observed, the proposed LACNN method outperforms the transfer learning method and the simple version (l_3-l_2-l_1) of MCME method with the improvements of about 15% and 2% in OS, respectively. In addition, the LACNN method achieves competitive performance in comparison with the best version (l_4-l_3-l_2-l_1) of MCME method, which demonstrates the effectiveness of utilizing macular lesions information to guide CNN for the classification of 3-D OCT volumes. Although the time complexity of LACNN is higher than that of the other evaluated methods, the proposed LACNN is an end-to-end model, and the test time of 0.0206 second per OCT volume can still meet the requirement of the clinical diagnosis for the eye diseases. Note that, in this experiment, we directly transfer the LDN (trained on lesion-level samples from UCSD dataset) to create attention maps for OCT images in NEH dataset, and higher performance of LACNN can be expected if more macular lesions from NEH dataset are utilized in the training process of LDN.

F. Learning Efficiency during Training

In this section, the learning efficiency of different deep learning models are compared on UCSD dataset. To achieve a fair comparison, the mini-batch size in each method is set to 24 and the number of maximum epochs is selected to be 10 to ensure reliable convergence for each method. Other parameters (e.g., learning rate) are kept as the same as that in Section IV-B. Similar to [39], additional 1000 B-scans (250 Drusen, 250 CNV, 250 DME and 250 Normal) independent of the training dataset are used as validation set. Fig. 7 depicts the learning process of three models in terms of (a) training loss, (b) validation loss, and (c) validation accuracy over the training steps. Plots are normalized with a smoothing factor of 0.6 to clearly visualize the trends.

In comparison to VGG16 network, the training and validation losses of LACNN start with a rapid descent and then reach the convergence soon, which demonstrates that the proposed LACNN method can significantly accelerate the network training process. This also shows that the detected macular lesion information can guide the network to pay more attention to important and discriminative features while ignoring insignificant and trivial features. In addition, we can observe that the LACNN also learns faster than transfer leaning methods and exhibits considerably lower validation error, verifying that the information introduced by lesion-attention modules is more beneficial to the OCT classification, compared with the transferred knowledge from natural image domain.

G. Effect of Different Numbers of Training Samples

In this section, the effect of different numbers of training and test samples on each method is analyzed on UCSD dataset. The parameters for all the methods are kept the same as that in Section IV-B. In this experiment, the overall accuracy, overall sensitivity, and overall precision are used here, and η varies from 2 to 10 (from 1/2 to 1/10 of the whole samples for training) for the UCSD dataset. The results are shown in Fig. 8. As can be observed, the performances of all methods generally improve as the numbers of training samples increase. More importantly, the proposed LACNN method consistently provides superior performances over the other compared

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2019.2898414, IEEE Transactions on Medical Imaging

TABLE IV
CLASSIFICATION RESULTS (IN PERCENTAGE) OF THE LACNN INTEGRATED WITH DIFFERENT CNN ARCHITECTURES. THE BEST RESULTS IN THIS TABLE ARE LABELED IN BOLD

Methods	Classes	ACC	SE	PR	SP	AUC	OA	OS	OP
AlexNet [30]	Drusen	90.0 ± 1.9	58.5 ± 7.3	51.5 ± 5.2	93.9 ± 2.9	87.8 ± 2.7	81.5 ± 2.8	75.2 ± 2.7	76.9 ± 1.4
	CNV	90.4 ± 1.3	84.5 ± 4.3	94.0 ± 2.2	95.3 ± 1.4	96.9 ± 0.6			
	DME	92.2 ± 1.1	69.8 ± 4.8	75.9 ± 4.7	95.7 ± 1.7	93.0 ± 1.3			
	Normal	90.9 ± 1.3	91.5 ± 7.8	86.2 ± 3.3	90.7 ± 3.4	97.0 ± 0.4			
LACNN-AlexNet	Drusen	93.5 ± 1.3	74.7 ± 5.3	67.4 ± 9.4	95.6 ± 2.3	94.3 ± 0.6	89.4 ± 1.8	85.5 ± 1.3	84.8 ± 2.5
	CNV	93.0 ± 1.6	89.4 ± 4.9	94.8 ± 1.4	95.9 ± 1.3	97.9 ± 0.4			
	DME	95.6 ± 0.6	82.6 ± 5.3	85.0 ± 5.9	97.6 ± 1.1	96.3 ± 0.5			
	Normal	96.2 ± 0.4	96.3 ± 1.3	92.1 ± 1.2	96.2 ± 0.7	98.8 ± 0.2			
Inception [63]	Drusen	93.7 ± 0.7	77.3 ± 9.7	66.3 ± 4.9	95.5 ± 1.1	94.8 ± 1.0	87.8 ± 2.5	85.5 ± 1.5	83.3 ± 2.7
	CNV	92.8 ± 1.6	88.4 ± 5.1	95.5 ± 1.9	96.4 ± 1.8	98.1 ± 0.3			
	DME	94.6 ± 2.3	82.9 ± 5.8	80.2 ± 5.2	96.4 ± 3.3	96.8 ± 0.8			
	Normal	95.1 ± 0.8	93.3 ± 4.4	91.3 ± 3.2	95.9 ± 1.8	98.8 ± 0.3			
LACNN-Inception	Drusen	**94.9** ± 1.1	**78.8** ± 5.7	**68.8** ± 4.3	**96.7** ± 2.1	**96.0** ± 0.7	**92.1** ± 0.9	**89.3** ± 1.6	**87.6** ± 2.9
	CNV	**94.5** ± 1.0	**92.3** ± 3.6	**96.9** ± 1.3	**96.4** ± 1.5	**98.6** ± 0.4			
	DME	**97.1** ± 0.5	**88.9** ± 1.2	**90.1** ± 3.6	**98.3** ± 0.5	**97.9** ± 0.4			
	Normal	**97.5** ± 0.3	**97.2** ± 1.1	**94.5** ± 1.7	**97.7** ± 0.6	**99.3** ± 0.2			

methods under all the different numbers of training samples. In particular, when small number of training samples are used, our proposed LACNN method has more advantages over other methods. When we use a half of data (almost 42000 B-scans) for training, we achieve an average overall accuracy of 93.3%.

H. Comparison of Different Lesion Attention Strategies

In this subsection, we compare different lesion attention methods to prove the effectiveness of the proposed attention strategy. Table III shows the overall accuracy values obtained by different attention methods on UCSD dataset. Here, the LACNN-*i* (i = 1, 2 … 5) refers to method that only inserts the lesion-attention module before *i*-th pooling layer of VGG16 network, respectively. The LACNN-A represents method that inserts the lesion-attention module after all of convolutional and pooling layers. In this experiment, we set η to 6.

From Table III, we can see that adopting lesion-attention modules can improve the classification performance to some extents, compared with the plain VGG16 network, and the proposed strategy (in LACNN) indeed outperforms other methods. In addition, we can also observe that integration of lesion information in shallow layers performs better than that in deep layers. Our explanation is that shallow layers capture fine features with rich spatial information while deeper layers encode abstract features with less detailed structural information. Therefore, the shallow features weighted by attention maps are more discriminative for classification. However, inserting too many lesion-attention modules may conversely bring in redundant information, which greatly degrades the performance and increases the computational cost (e.g., LACNN-A).

I. Integration with Different Architectures

We next investigate the effect of integrating the lesion-attention modules with other state-of-the-art CNN architectures, e.g., AlexNet [30] and InceptionV3 [63] on UCSD dataset. When the baseline network is the AlexNet, we include the lesion-attention modules before the pool1, pool2, conv4, conv5, and pool5 layers. When using InceptionV3 as baseline network, we incorporate the lesion-attention modules into InceptionV3 before each inception block. By making these changes for above CNN architectures, we can construct LACNN-AlexNet and LACNN-Inception models, respectively. In this experiment, η is also set to 6 and the LACNN-AlexNet is trained using Adam optimizer with initial learning rate of 10^{-4} and mini-batch size of 32. For the LACNN-Inception network, optimization is performed using Adam with initial learning rate of 10^{-4} and a mini-batch size of 12 images due to limitation of computation memory. For each considered CNN model, the original OCT images were resized to the input size of each model.

The results on UCSD dataset are shown in Table IV. As can be seen, significant performance improvements can be achieved when lesion-attention modules are incorporated into both architectures. In particular, the LACNN-AlexNet achieves an overall accuracy of 89.4%, which is superior to its plain counterpart AlexNet (75.2%). The LACNN-Inception outperforms the Inception by a margin of 4.3% (OA) at minimal increases in computational cost. The results show that our LACNN can be effectively applied on different CNN architectures. We also note that the performance growth of the LACNN-Inception over its plain counterpart is less than that of the LACNN-AlexNet. One of the reasons might be that the Inception network is deeper and wider than AlexNet, and thus it will be harder to be optimized.

V. CONCLUSIONS

In this paper, we proposed a novel lesion-aware convolutional neural network (LACNN) method for retinal OCT image classification. Compared with the previous networks, the proposed LACNN adopts attention mechanism to focus on salient macular lesion-related regions within an OCT image. In addition, guided by the lesion-related information, the classification network can utilize the information from local lesion-related regions to achieve more efficient and accurate OCT classification. The experimental results on two clinical OCT datasets demonstrated the superiority of the proposed

LACNN method over several well-known classical and deep learning methods.

In this paper, we utilized the information of various kinds of macular lesions for improving retinal OCT image classification. However, the proposed LACNN framework can be easily applied to any other complex macular diseases, such as macular hole (MH), which is characterized by full- or partial-thickness macular hole located in macula. Therefore, in our future works, we will investigate the applicability of the proposed LACNN model for classifying OCT images from other pathologies (e.g., macular hole, macular telangiectasia, and central serous retinopathy).

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Editors and the three Anonymous Reviewers for their valuable comments and suggestions, which greatly helped us to improve the technical quality and presentation of our manuscript.

REFERENCES

[1] D. Pascolini and S. P. Mariotti, "Global estimates of visual impairment: 2010," *Br. J. Ophthalmol.*, vol. 96, no. 5, pp. 614-628, 2012.

[2] D. C. Neely, K. J. Bray, C. E. Huisingh, M. E. Clark, G. J. McGwin, and C. Owsley, "Prevalence of undiagnosed age-related macular degeneration in primary eye care," *JAMA Ophthalmol.*, vol. 135, no. 6, pp. 570-575, 2017.

[3] A. Abdelsalam, L. Del Priore, and M. A. Zarbin, "Drusen in age-related macular degeneration: pathogenesis, natural course, and laser photocoagulation–induced regression," *Surv. Ophthalmol.*, vol. 44, no. 1, pp. 1-29, 1999.

[4] K. B. Freund, L. A. Yannuzzi, and J. A. Sorenson, "Age-related macular degeneration and choroidal neovascularization," *Am. J. Ophthalmol.*, vol. 115, no. 6, pp. 786-791, 1993.

[5] F. E. Hirai, M. D. Knudtson, B. E. Klein, and R. Klein, "Clinically significant macular edema and survival in type 1 and type 2 diabetes," *Am. J. Ophthalmol.*, vol. 145, no. 4, pp. 700-706, 2008.

[6] D. Huang *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178-1181, 1991.

[7] N. Nassif *et al.*, "In vivo high-resolution video-rate spectral-domain optical coherence tomography of the human retina and optic nerve," *Opt. Exp.*, vol. 12, no. 3, pp. 367-376, 2004.

[8] W. Drexler and J. G. Fujimoto, "State-of-the-art retinal optical coherence tomography," *Progress Retinal Eye Res.*, vol. 27, no. 1, pp. 45-88, 2008.

[9] S. Bhat, I. V. Larina, K. V. Larin, M. E. Dickinson, and M. Liebling, "4D reconstruction of the beating embryonic heart from two orthogonal sets of parallel optical coherence tomography slice-sequences," *IEEE Trans. Med. Imag.*, vol. 32, no. 3, pp. 578-788, 2013.

[10] M. E. Brezinski and J. G. Fujimoto, "Optical coherence tomography: high-resolution imaging in nontransparent tissue," *IEEE J. Sel. Topics Quantum Electron.*, vol. 5, no. 4, pp. 1185-1192, 1999.

[11] N. D. Gladkova *et al.*, "In vivo optical coherence tomography imaging of human skin: norm and pathology," *Skin. Res. Technol.*, vol. 6, no. 1, pp. 6-16, 2000.

[12] C. A. Puliafito *et al.*, "Imaging of macular diseases with optical coherence tomography," *Ophthalmology*, vol. 102, no. 2, pp. 217-229, 1995.

[13] V. J. Srinivasan *et al.*, "High-definition and 3-dimensional imaging of macular pathologies with high-speed ultrahigh-resolution optical coherence tomography," *Ophthalmology*, vol. 113, no. 11, pp. 2054-2065, 2006.

[14] M. R. Hee *et al.*, "Optical coherence tomography of age-related macular degeneration and choroidal neovascularization," *Ophthalmology*, vol. 103, no. 8, pp. 1260-1270, 1996.

[15] 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.

[16] F. G. Venhuizen *et al.*, "Automated staging of age-related macular degeneration using optical coherence tomography," *Inv. Ophthalmol. Vis. Sci.*, vol. 58, no. 4, pp. 2318-2328, 2017.

[17] G. Lemaitre *et al.*, "Classification of SD-OCT volumes using local binary patterns: experimental validation for DME detection," *J. Ophthalmol.*, vol. 10, no. 12, pp. 329-346, 2016.

[18] A. Albarrak, F. Coenen, and Y. Zheng, "Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction," in *Proc. 17th Conf. Med. Image Understanding Anal.*, 2013, pp. 59-64.

[19] P. P. Srinivasan *et al.*, "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.

[20] L. Fang, C. Wang, S. Li, J. Yan, X. Chen, and H. Rabbani, "Automatic classification of retinal three-dimensional optical coherence tomography images using principal component analysis network with composite kernels," *J. Biomed. Opt.*, vol. 22, no. 11, pp. 11-16, 2017.

[21] 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, pp. 12-21, 2017.

[22] 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. Exp.*, vol. 7, no. 12, pp. 4928-4940, 2016.

[23] S. Farsiu *et al.*, "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography," *Ophthalmology*, vol. 121, no. 1, pp. 162-172, 2014.

[24] F. G. Venhuizen *et al.*, "Automated age-related macular degeneration classification in OCT using unsupervised feature learning," in *Proc. SPIE Med. Imag.*, 2015, pp. 9411-9417.

[25] L. Fang, S. Li, D. Cunefare, and S. Farsiu, "Segmentation based sparse reconstruction of optical coherence tomography images," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 407-421, 2017.

[26] R. Kafieh, H. Rabbani, and I. Selesnick, "Three dimensional data-driven multi scale atomic representation of optical coherence tomography," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1042-1062, 2015.

[27] R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Curvature correction of retinal OCTs using graph-based geometry detection," *Phys. Med. Biol.*, vol. 58, no. 9, pp. 2925-2938, 2013.

[28] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097-1105.

[31] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomed. Opt. Exp.*, vol. 8, no. 5, pp. 2732-2744, 2017.

[32] A. G. Roy *et al.*, "ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Exp.*, vol. 8, no. 8, pp. 3627-3642, 2017.

[33] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration OCT images," *Ophthalmol. Retina*, vol. 1, no. 4, pp. 322-327, 2017.

[34] 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, 2017.

[35] Y. Rong *et al.*, "Surrogate-assisted retinal OCT image classification based on convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 253-263, 2019.

[36] R. Rasti, A. Mehridehnavi, H. Rabbani, and F. Hajizadeh, "Automatic diagnosis of abnormal macula in retinal optical coherence tomography images using wavelet-based convolutional neural network features and random forests classifier," *J. Biomed. Opt.*, vol. 23, no. 3, pp. 35-45, 2018.

[37] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Med.*, vol. 24, no. 9, pp. 1342-1350, 2018.

[38] S. J. Pan and Q. A. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data*, vol. 22, no. 10, pp. 1345-1359, 2010.

[39] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131, 2018.
[40] S. P. 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. Opt. Exp.*, vol. 8, no. 2, pp. 579-592, 2017.
[41] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 193-222, 1995.
[42] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognit.*, vol. 7, no. 1-3, pp. 17-42, 2000.
[43]A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97-136, 1980.
[44] M. Ghafoorian *et al.*, "Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin," *Neuroimage Clin.*, vol. 14, no. 3, pp. 391-399, 2017.
[45] B. E. Bejnordi *et al.*, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *J. Med. Imag.*, vol. 4, no. 4, pp. 44-54, 2017.
[46] A. Lisowska *et al.*, "Context-aware convolutional neural networks for stroke sign detection in non-contrast CT scans," in *Proc. 21th Conf. Med. Image Understanding Anal.*, 2017, pp. 494-505.
[47] K. Kushibar *et al.*, "Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features," *Med. Image Anal.*, vol. 48, no. 6, pp. 177-186, 2018.
[48] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "DCAN: Deep contour-aware networks for object instance segmentation from histology images," *Med. Image Anal.*, vol. 36, no. 4, pp. 135-146, 2017.
[49] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207-1216, 2016.
[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2015, pp. 91-99.
[51] D. Qi *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182-1195, 2016.
[52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431-3440.
[53] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240-1251, 2016.
[54] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402-2410, 2016.
[55] T. Schlegl *et al.*, "Fully automated detection and quantification of macular fluid in OCT using deep learning," *Ophthalmology*, vol. 125, no. 4, pp. 549-558, 2017.
[56] Y. LeCun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Adv. Neural Inform. Process. Syst.*, 1990, pp. 396-404.
[57] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6593-6601.
[58] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," 2015, [Online]. Available: https://arxiv.org/abs/1409.1556
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.
[60] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization," 2014, [Online]. Available: https://arxiv.org/abs/1412.6980
[61] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Conf. Artif. Intell. Statist.*, 2010, pp. 249-256.
[62] M. Abadi *et al.* "TensorFlow: A system for large-scale machine learning," 2016, [Online]. Available: https://arxiv.org/abs/1605.08695
[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818-2826.