Research Article Vol. 13, No. 11 / 1 Nov 2022 / Biomedical Optics Express 5813
Biomedical Optics EXPRESS

Deep attentive convolutional neural network for automatic grading of imbalanced diabetic retinopathy in retinal fundus images

FENG LI,[1,*] SHIQING TANG,[1] YUYANG CHEN,[1] AND HAIDONG ZOU[2,3]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2]Shanghai Eye Disease Prevention & Treatment Center, Shanghai 200040, China
[3]Ophthalmology Center, Shanghai General Hospital, Shanghai 200080, China
[*]lifenggold@163.com

Abstract: Automated fine-grained diabetic retinopathy (DR) grading was of great significance for assisting ophthalmologists in monitoring DR and designing tailored treatments for patients. Nevertheless, it is a challenging task as a result of high intra-class variations, high inter-class similarities, small lesions, and imbalanced data distributions. The pivotal factor for the success in fine-grained DR grading is to discern more subtle associated lesion features, such as microaneurysms (MA), Hemorrhages (HM), soft exudates (SE), and hard exudates (HE). In this paper, we constructed a simple yet effective deep attentive convolutional neural network (DACNN) for DR grading and lesion discovery with only image-wise supervision. Designed as a top-down architecture, our model incorporated stochastic atrous spatial pyramid pooling (sASPP), global attention mechanism (GAM), category attention mechanism (CAM), and learnable connected module (LCM) to better extract lesion-related features and maximize the DR grading performance. To be concrete, we devised sASPP combining randomness with atrous spatial pyramid pooling (ASPP) to accommodate the various scales of the lesions and struggle against the co-adaptation of multiple atrous convolutions. Then, GAM was introduced to extract class-agnostic global attention feature details, whilst CAM was explored for seeking class-specific distinctive region-level lesion feature information and regarding each DR severity grade in an equal way, which tackled the problem of imbalance DR data distributions. Further, the LCM was designed to automatically and adaptively search the optimal connections among layers for better extracting detailed small lesion feature representations. The proposed approach obtained high accuracy of 88.0% and kappa score of 88.6% for multi-class DR grading task on the EyePACS dataset, respectively, while 98.5% AUC, 93.8% accuracy, 87.9% kappa, 90.7% recall, 94.6% precision, and 92.6% F1-score for referral and non-referral classification on the Messidor dataset. Extensive experimental results on three challenging benchmarks demonstrated that the proposed approach achieved competitive performance in DR grading and lesion discovery using retinal fundus images compared with existing cutting-edge methods, and had good generalization capacity for unseen DR datasets. These promising results highlighted its potential as an efficient and reliable tool to assist ophthalmologists in large-scale DR screening.

1. Introduction

Diabetic retinopathy (DR) was one of the most severe complications associated with diabetes caused by small blood vessel damage from high levels of blood glucose in the retina, which could result in progressive vision loss or even irreversible blindness [1,2]. The main pathological features such as microaneurysms (MA), hemorrhages (HM), soft exudates (SE) and hard exudates (HE) were closely related to DR, each of which determined the DR grading in the patients [3]. In terms of type, size, and the number of lesions, DR could be categorized sequentially with

#472176 https://doi.org/10.1364/BOE.472176
Journal © 2022 Received 4 Aug 2022; revised 25 Sep 2022; accepted 6 Oct 2022; published 14 Oct 2022

Biomedical Optics EXPRESS

incremental severity grades into no DR, mild DR, moderate DR, severe DR, and proliferative DR [4]. Early detection and timely treatment were crucial for slowing down the progression of DR and preventing eventual vision loss, and retinal fundus imaging was regarded as a common DR examination method [5,6]. However, detecting different types of tiny lesions and grading the severity of DR by manual visual analysis of retinal fundus images were a highly subjective and time-consuming process, as well as prone to errors, which were often affected by individual bias and clinical experiences. Even among highly skilled ophthalmologists, inter- and intra-grader inconsistencies also occurred. Therefore, it was necessary to develop automatic DR grading approaches to help ophthalmologists in making a timely DR early diagnosis and providing a rationale for further treatment based on its severity, which was of vital importance in clinical practice.

Recently, as artificial intelligence technology develops by leaps and bounds in the area of computer vision and medical imaging community, deep learning, specifically convolutional neural networks (CNNs), had gained great attention and has been proven to be a vigorous tool for automatic DR grading [7]. Deep learning algorithms could automatically learn strong abstract feature representations from a significant amount of training data without hand-crafted explicit lesion features. There mainly existed two types of deep learning methods for DR severity grading, including pixel-level supervision [8,9] and image-level supervision [9–11]. For pixel-level supervision methods, they commonly used lesion information for assisting DR classification. However, these methods required lesion annotations as learning guidance and formulated DR grading and lesion discovery as a two-stage task, which was more intricate compared to a one-stage strategy. In order to alleviate these issues, image-level supervision methods based on retinal fundus images were presented to simultaneously distinguish DR grades and highlight lesion areas. Unfortunately, they tended to pay more attention to the most essential lesion regions and overlook tiny small lesion ones in the fundus images, which may impede the performance of DR grading and lesion location. Moreover, the less distinctive areas detected at a certain severity level could be essential for other severity gradings. Thus, it was desirable to build an effective model to capture more complete lesion features and improve the DR grading performance.

Based on the above discussion, to further discover more detailed small lesion features, promote the performance of DR grading and raise their application in clinical scenarios, we needed to consider the following several aspects. (1) In retinal fundus images, there were many similarities in color and texture among the five DR grades [12,13]. This had an adverse influence on the inter-class diversity and thereby increased the difficulties of accurately discriminating them in the grading task. Therefore, we should build a fine-grained network to capture detailed lesion information from global feature maps, whilst restraining useless information. (2) The relative sparse distribution of lesion regions and the importance of different lesion areas in each image should also be considered. Besides, some lesions were very tiny and only occupied a few pixels. These were prone to be ignored during performing convolution operations, which could damage the final DR severity level. Accordingly, it was demanded to devise an effective model to acquire more complete lesion regions, and evaluate and adaptively merge the contribution of each lesion area. (3) Due to the number of fundus images with severe lesions accounting for a small portion, the data distribution of DR among different grades was seriously imbalanced, which tended to enforce the model to bias on the most important regions and DR grades with more samples. This would impair the performance of lesion location and weaken the model's generalization ability. Hence, it was expected to learn class-specific discriminative features. (4) The high variation in size and position of lesions in retinal fundus images was challenging for DR grading. Moreover, multiple different lesions could be embraced in the fundus image. Even fundus images graded into the same severity level may include inconsistencies in the type and quantity of lesions. Consequently, we needed to consider multi-scale information on lesions, learn the corresponding

Biomedical Optics EXPRESS

lesion characteristics from the diversity of lesion regions, and combine different areas into a complete lesion discovery.

Motivated by the aforementioned observation, we developed a novel deep attentive convolutional neural network (DACNN) to grade DR severity levels and discover associated lesions using only image-level supervision. In specific, we designed stochastic Atrous Spatial Pyramid Pooling (sASPP) based on ASPP to fully extract the global lesion features at multiple scales by using multiple atrous convolutions with different fields of view, and combat the co-adaptation of multiple atrous convolutions. Global Attention Mechanism (GAM) inspired by [14] was applied to the global lesion feature maps to highlight the class-agnostic global attention features. GAM together with sASPP was able to capture more subtle lesion details while eliminating the interference of irrelevant information. Category Attention Mechanism (CAM) focusing on category attention was employed to learn class-specific feature representations, adaptively emphasize and fuse the contribution of each lesion area, and increase the discrepancies in the distance between different stages of DR. Through impactful information emphasized by CAM, more distinctive features for a certain DR severity level with a handful of samples could be learnt, and thereby the issue about the imbalanced data distribution could be addressed. In addition, instead of fixed connections among layers in the deep learning model, we innovatively introduced a learnable connected module (LCM) with a connected weight for each connected layer to automatically and adaptively seek the optimal connections among layers. This strategy could not only strengthen feature propagation but also amalgamate low-level features with high-level semantic information for better parameter reuse, facilitating information flow and enhancing the power of DR grading and lesion location. As a result, by incorporating the above blocks, our network achieved accurate DR severity grades and complete lesion discovery simultaneously. We summed up the main contributions of our study as follows:

1. We constructed a simple yet effective deep attentive convolutional neural network (DACNN) to achieve DR grading and lesion discovery using only image-level supervision by combining sASPP, GAM, CAM, and LCM.

2. In DACNN, we applied complementary GAM and CAM to learn class-agnostic global attention features and class-specific distinguishing region-level semantic features for preserving subtle lesion details and mitigating the problem of imbalanced data distributions.

3. Built upon ASPP, we presented sASPP for capturing lesion features of multiple scales, preventing co-adaptation of the atrous convolutions, and alleviating the overfitting problem. Further, LCM was introduced to effectively boost the model's grading performance, which could automatically perceive the optimal connections among layers during training.

4. Extensive experimental results on different challenging benchmarks containing DDR, EyePACS, and Messidor datasets demonstrated that our approach surpassed current mainstreaming methods and advanced the state-of-the-art performance in the DR grading tasks. Comprehensive ablation studies also verified the importance and necessity of each major component.

2. Related works

In this section, we overviewed recent studies related to DR grading, feature-connected model, and attention mechanism in brief.

2.1. DR grading

Traditional machine learning methods for DR grading adopted common classifiers or their variants, such as support vector machine (SVM) [15], k-nearest neighbor (kNN) [16], random

Biomedical Optics EXPRESS

forest [17] and Gaussian mixture model [15], etc., to carry out automatic taxonomy of severity levels of DR. However, these methods were excessively dependent on hand-crafted, low-level characteristics, which required experienced domain experts and had limited representation capacity. Recently, deep learning algorithms have been widely used in DR grading and made remarkable progress [18]. Instead of hand-crafted features, they leveraged more discriminative deep features to determine DR severity levels. The majority of deep learning methods for distinguishing DR grades may fall into two categories. The first category was to use location information of subtle lesions (namely pixel-level supervision) to assist DR grading [19–25]. For instance, Antal et al. [19] presented an ensemble-based framework for detecting MA while grading DR based on the existence of MA. Dai et al. [25] developed DeepDR to detect retinal lesions and perform early-to-last stages of DR, which consisted of an image quality assessment sub-network, lesion-aware sub-network, and DR grading sub-network. Yet, most of these methods demanded lesion annotation on retinal fundus images, which was extremely expensive. In addition, some methods [19–21] treated lesion location and DR severity grading as two separate tasks, which were more complex 315(pixi6(eparat.751 Tm [(inf)39(or)-19(mation)-315(of)-315(subtle)

Biomedical Optics EXPRESS

maps. In this case, our DACNN integrated feature connections and innovatively designed the LCM.

2.3. Attention mechanism

The attention mechanism played a vital role in capturing fine-grained feature information, and had been widely applied for image classification [35], semantic segmentation [36], speech recognition [37], word representation learning [38], and object detection [39]. In particular, the SE block recalibrated the importance of different channels by means of rescaling and simply employed two fully connected (FC) layers to catch non-linear cross-channel dependency. SENet [40] proposed channel-level dependencies by the usage of an attention-and-gating mechanism, which conducted rescaling to different channels to recalibrate the channel dependency. Non-local network [41] also termed as self-attention mechanism scanned through each element contained in a sequence and updated it via making an aggregation on global information from the whole sequence. CBAM [14] made the rescaling on the importance of different channels and positions. DA-GRU [42] applied the spatial attention mechanism to preprocess data in the encoder stage while employed temporal attention mechanism to obtain the internal relationship of the input information in the time series in the decoder stage, which made the model more focus on useful information at the time scale and the spatial scale and improved the accuracy of SOC estimation. GCNet [43] integrated the Non-local block [41] and SE block [40] to model the global context dependency. CANet [11] integrated the disease-specific attention module and the disease-dependent attention module into a deep network to yield disease-specific and disease-dependent features for jointly grading DR and DME. CABNet [12] combined two complementary the global attention block (GAB) and the category attention block (CAB) to identify small lesion features for DR grading. A lesion-aware transformer (LAT) [9] used a pixel relation-based encoder and a lesion filter-based decoder for DR prediction and lesion discovery at the same time. RTNet [44] considered the intra-class dependencies and inter-class relations among multi-lesion and employed a self-attention transformer and cross-attention transformer to segment the four DR lesions simultaneously. ASPP [45] stemming from spatial pyramid pooling (SPP) allowed the model itself to use multiple atrous convolutions of different rates to extract the optimal representation of features at various scales from the input directly. It could effectively enlarge the field of view in the kernel, without incurring more parameters and computational complexity. Inspired by the above discussion attention mechanism, in this work, our model applied the idea of the attention mechanism to put more emphasis on the DR lesion region and repress the irrelevant lesion information, so that the feature information which contributed to more discriminable feature extractions could be highlighted and the imbalanced data distribution could be handled. Due to the fact that DR severity level classification inherently depended on the global presence of retinal lesions including multi-scale local texture and structures, our DACNN further introduced the ASPP module, and on this basis, we innovatively designed the sASPP module to adapt to fine-grained lesions of various scales, relieve the overfitting problem and avoid the co-adaptation of the atrous convolutions in ASPP.

3. Methods

In this section, we provided details of our DACNN according to network architecture, and core network components. As illustrated in Fig. 1, our DACNN mainly consisted of five major components: Feature extraction module, sASPP, GAM, CAM, and LCM. In particular, we selected ResNet-50 as the backbone network, which was used to extract high-level abstract semantic feature representations from input fundus images. Built upon these high-level features, the sASPP was designed for capturing the lesions of various scales and preventing the co-adaptation of multiple atrous convolutions. The output features of this module were shared in the following modules. The GAM was introduced to learn category-agnostic global channel-level

Biomedical Optics EXPRESS

and spatial-level attention feature maps so as to retain more detailed tiny lesion information and repress the irrelevant noisy information. After that, the CAM was integrated into the network such that discriminative region-level features could be produced in a category-level fashion. Further, the LCM with connected weights was developed to reuse feature maps and search the effective connections between different layers automatically and adaptively. During testing, given an image, a predicted DR severity level and the corresponding lesion activation map would be outputted.

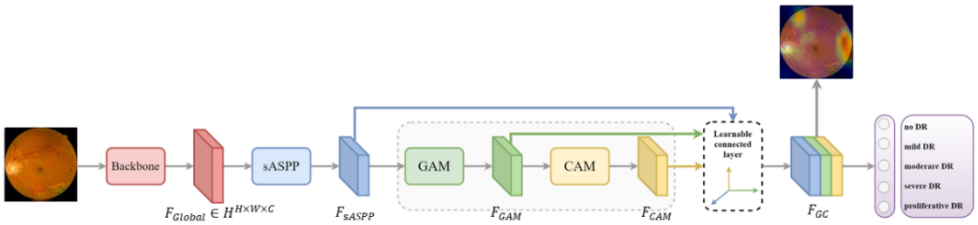


Fig. 1. The overall structure of the proposed DACNN.

3.1. Overview of DACNN

The overall architecture of our DACNN was shown in Fig. 1. We chose to use ResNet-50 pre-trained on ImageNet [46] as the backbone network for feature extraction, which took the fundus image as input and generated the high-level global semantic feature maps $F_{Global} \in R^{H\times W\times C}$, where H, W and C represented height, width, and the number of channels in the feature maps, respectively. Afterward, we fed the resulting feature map into the sASPP module for helping to detect the optimal representation of lesion features with different scales and prevent the co-adaptation of the atrous convolutions. Next, on the feature maps $F_{Global} \in R^{H\times W\times C}$ from the sASPP module, we utilized a 1×1 convolutional layer to decrease the number of channels and yield $F_{\mathrm{Reduce}} \in R^{H\times W\times C'}$, where $C' = C/2$. Then, the GAM took F_{Reduce} as the input and learnt channel-level and spatial-level class-agnostic global attention feature maps F_{GAM} such that subtle lesion details could be preserved and the useless noisy information could be eliminated. After that, F_{GAM} was flowed through CAM, enforcing the model to gather different distinctive region-level feature information for specific DR severity levels, resulting in the output F_{CAM}. The obtained feature maps F_{GAM} and F_{CAM} were complementary to each other, and they were concatenated via LCM. Further, we fed the output feature maps F_{sASPP}, F_{GAM} and F_{CAM} from sASPP, GAM, and CAM into LCM for reusing feature maps and optimizing connections among different layers automatically and adaptively. The output feature map F_{GC} of LCM passed through a global average pooling (GAP) layer and finally fed into a fully connected (FC) layer activated by the softmax function to predict DR severity level with respect to each input fundus image.

3.2. Stochastic atrous spatial pyramid pooling (sASPP)

Typically, there often existed lesion features with various scales in the DR grading task. In order to capture them concurrently, the traditional methods were to either make the resample on the input images for model training, or apply ASPP for extracting multi-scale features from the input directly. However, the former suffered from overwhelming computational cost, and the latter was subject to the co-adaptation problem of multiple atrous convolutions. Motivated by stochastic depth networks [47,48], we introduced randomness into ASPP and designed a novel sASPP to stochastically drop features from each atrous convolution to avoid co-adaptation of atrous convolutions, which could be seen in Fig. 2 (a). The developed sASPP could be regarded as ASPP with varying widths. During the training stage (See Fig. 2 (b)), $b_i \in \{0, 1\}$ was a binary

Biomedical Optics EXPRESS

variable that determined the preservation or neglect of feature maps produced by each atrous convolution. It satisfied the Bernoulli distribution parameterized by probability p_i, which meant that the feature map yielded from each atrous convolution in ASPP would be reserved with a probability of p_i and abandoned with a probability of $1 - p_i$ when b_i was 1 or 0, respectively. Obviously, the second and fourth feature maps were abandoned, and the output feature map a^l in ASPP was equal to $[a_1^l, 0, a_3^l, 0]$. Further, the a^l could be formulated as follows:

$$a^l = [b_1 a_1^l, b_2 a_2^l, \ldots, b_n a_n^l] \tag{1}$$

It was worth noting that the dropout operation in sASPP adopted an independent random variable to the entire feature instead of every pixel within the feature in the original dropout method, which took the complete feature representation from each atrous convolution into account. In sASPP, we introduced a new learnable parameter p_i to control the retained probability of feature maps from each atrous convolution. The larger p_i was, the more contributions made by the feature map obtained from the atrous convolution, and vice versa. In our study, we set p_i as a function of i, which could increase or decrease from p_1 to p_n in a linear manner and be described as the follows:

$$p_i = p_1 + \frac{p_n - p_1}{n - 1} \times (i - 1) \tag{2}$$

In the training stage, we independently sampled n binary variables from Bernoulli distributions with regard to each mini-batch and computed $a^{l+1} = [b_1 a_1^{l+1}, b_2 a_2^{l+1}, \ldots, b_n a_n^{l+1}]$ according to Eq. (1). In the testing stage, we preserved all the feature maps from each atrous convolution as shown in Fig. 2 (c), whose values were scaled by p_i. Herein, sASPP became a determining module employing all the feature maps scaled by the reserved probability. Through the designed sASPP, the discriminative multiple scale lesion-related feature map F_{sASPP} could be obtained from $F_{Global} \in R^{H \times W \times C}$ extracted by the backbone network, and the overfitting problem in the DR grading task could also be effectively alleviated.

3.3. Global attention mechanism (GAM)

Considering convolution operations only own the local perception field, its modeling of dependency was limited, affecting DR grading accuracy. In order to strengthen the model and improve the state-of-the-art, we made an attempt to associate lesion features with the channel and spatial attention modules, constructing the GAM. As shown in Fig. 3, GAM was composed of channel attention and spatial attention, which utilized a single-branch structure contrary to that in CBAM [14]. It used the reduced feature map F_{Reduce} obtained by 1×1 convolutional operation on F_{sASPP} as the input and highlighted category-agnostic global salient features. To be concrete, we first computed the channel-attention feature maps as follows:

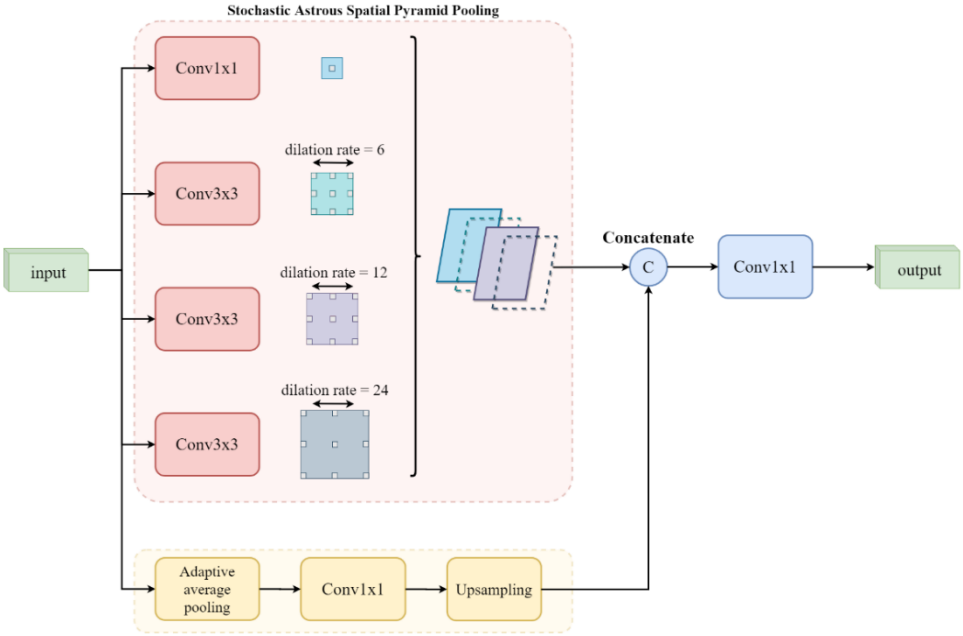
$$F_{Cattention} = (\sigma(Conv(\mathrm{Re}l(Conv(GAP(F_{\mathrm{Reduce}})))))) \otimes F_{\mathrm{Reduce}} \tag{3}$$

where σ and Rel represented the sigmoid activation function and ReLU activation function, respectively. $Conv$ denoted 1×1 convolutional layer. $\otimes$ stood for element-wise multiplication. The channel attention learnt the channel-wise attention weights and put more emphasis on the importance of each feature channel, whilst repressing useless informative channels. Thereafter, the output of GAM namely the spatial attention feature maps F_{GAM} could be calculated as follows:

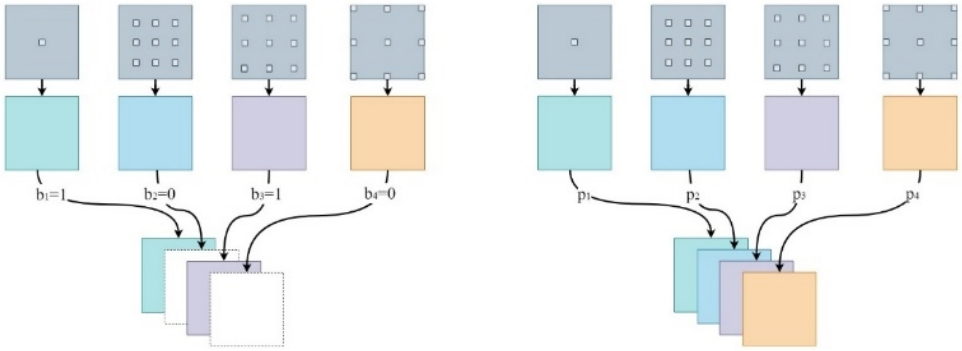
$$F_{GAM} = F_{Cattention} \otimes (\sigma(CGAP(F_{Cattention}))) \tag{4}$$

where $CGAP$ indicated the cross-channel average pooling. The spatial attention learnt spatial attention weights and more focused on the importance of each spatial position, which could make complementation of the above channel-attention. Finally, F_{GAM} was fed into CAM to produce the DR category-specific severity grade attention feature maps.

Biomedical Optics EXPRESS



(a)



(b) (c)

Fig. 2. The schematic illustration of the designed sASPP. (a) sASPP consisted of four atrous convolutions with dilation rates ranging from 6 to 24. The kernel size was 1, 3, 3, and 3 respectively. (b) sASPP in a certain state at the stage of training. Dilation rate and kernel size were the same as (a). b_i denoted whether the feature map existed, where $b_i = 0$ was defined as discard state and vice versa. (c) sASPP at the stage of testing. All the feature maps generated by four atrous convolutions were reserved at the stage of testing in terms of their magnitudes scaled by the retainable probability during the training stage.

Biomedical Optics EXPRESS

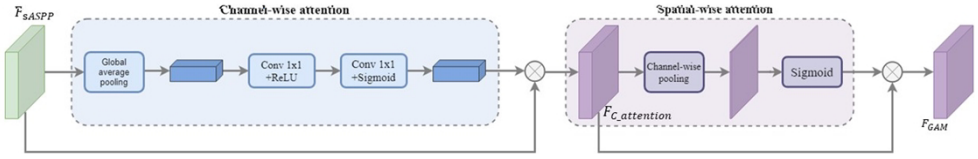


Fig. 3. Description of the Global Attention Mechanism (GAM)

3.4. Category attention mechanism (CAM)

The CNN commonly stacked all the feature maps together indiscriminately, resulting in information confusion among different categories and devoting less attention to the categories with few samples. In this context, we attempted to introduce CAM to learn the attention to the distinctive areas from retinal fundus images in a category-wise way and regard different DR severity grades equally, which was depicted in Fig. 4. Specifically, we firstly fed the output feature map F_{GAM} from GAM into a 1×1 convolutional layer for generating feature maps $F'_{CAM} \in R^{H \times W \times kL}$, where k was the number of channels required to identify discriminative areas relating to each DR category and L denoted the number of severity grades of DR. Then, we calculated the scores $S = \{S_1, S_2, \ldots, S_L\}$ for each DR category in terms of the follows:

$$S_i = \frac{1}{k} \sum_{j=1}^{k} GMP(f'_{i,j}), i \in \{1, 2, \ldots, L\} \tag{5}$$

where GMP indicated global max pooling and $f'_{i,j}$ was the $j-th$ feature map concerning the $i-th$ category from the feature map F'_{CAM}. By S_i, the importance of the feature maps with regard to each DR category could be obtained. Since S_i was not employed as the category attention directly, it was necessary to apply a category-level cross-channel average pooling operation on F'_{CAM} such that the feature maps for each DR severity grade could be derived, which may be represented by

$$F'_{avg-i} = \frac{1}{k} \sum_{j=1}^{k} f'_{i,j}, i \in \{1, 2, \ldots, L\} \tag{6}$$

where F'_{avg-i} represented the semantic feature map regarding the $i-th$ category. Subsequently, we could calculate the class-attention feature map ATT_{CAM} as follows, which emphasized the distinctive and informative lesion areas for DR severity grading:

$$ATT_{CAM} = \frac{1}{L} \sum_{i-1}^{L} S_i F'_{avg-i} \tag{7}$$

At last, we transformed the input feature maps F_{GAM} into the output feature maps F_{CAM} of CAM through ATT_{CAM}, which could be expressed by

$$F_{CAM} = F_{GAM} \otimes ATT_{CAM} \tag{8}$$

In CAM, we assigned a certain amount of feature channels to each DR severity grade, which avoided biasing on the channels and forcing the distance among different DR grades to enlargement. Thereby, it could effectively mitigate the issue of imbalanced data distribution. In addition, CAM could also in-depth dug more category-specific discriminative areas to generate attention features using extremely few parameters, and decreased the feature redundancy.

Biomedical Optics EXPRESS

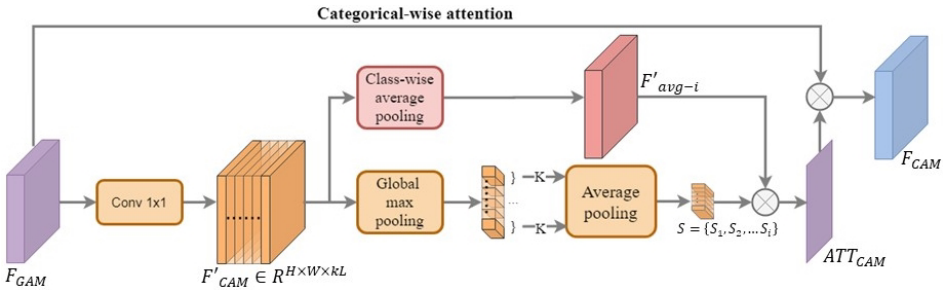


Fig. 4. Illustration of Category Attention Mechanism (CAM)

3.5. *Learnable connected module (LCM)*

With the depth of CNN increasing, the gradient vanishing or gradient exploding may happen. To strengthen feature propagation and encourage parameter reuse, feature connection was a good candidate approach. In previous connected structures, they were commonly fixed or handcrafted, which led to insufficient utilization of the information in feature maps. This occurred could be owing to ineffective modeling of the inference among the connected layers. In this respect, we developed LCM and tried to introduce it into our DACNN, as presented in Fig. 1. Concretely, we applied batch normalization (BN) to normalize the feature maps F_{sASPP}, F_{GAM}, and F_{CAM} from sASPP, GAM, and CAM, respectively, whilst generating corresponding learnable connected weights. We could make a description of the whole learnable operation using the identity function as the following equation:

$$y_i = \alpha u_i + \beta v_i + \gamma \omega_i \tag{9}$$

where y_i indicated the $i - th$ pixel that resulted from three different input feature maps, u_i, v_i, and ω_i. u, v, and ω denoted three under-connected layers or modules. α, β, and γ corresponded to the connected scalar weights shared across all channels. Assuming that the input feature maps were given by a 3-D tensor, they would be assigned the corresponding connected weights which could be automatically learnt and adaptive to these input features. Let $\alpha + \beta + \gamma = 1$, and $\alpha, \beta, \gamma \in [0, 1]$, we defined the following equation:

$$\alpha = \frac{e^{\lambda_\alpha}}{e^{\lambda_\alpha} + e^{\lambda_\beta} + e^{\lambda_\gamma}} \tag{10}$$

where λ_α was a control parameter. From Eq. (10), it could be observed that α was calculated by use of a softmax function with λ_α, which could be automatically learnt via standard back-propagation. In a similar way, β and γ were defined by using another two control parameters λ_β and λ_γ, respectively. Note that each term (such as αu_i, βv, and $\gamma \omega_i$) in the Eq. (9) needed to be activated through a nonlinear function including BN [49] and PReLUs [50]. The proposed LCM was able to be trained for finding the optimal connections among different under-connected layers or modules and adapted to the input feature maps. Meanwhile, it inherited the merits in previous fixed connect architecture, strengthening the feature information from different connected layers or modules reusing and facilitating the information flow. It was worth mentioning that the devised simple LCM could be used as a universal plug and play CNN module, which could be utilized directly in any existing feature connection schemes by taking their original counterparts in place of connection components.

Adam was a versatile algorithm for first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower-order moments, which computed individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. It integrated the computation of first and second moments with the bias

Biomedical Optics EXPRESS

correction term, and used momentum and adaptive learning rates to speed up convergence. In Adam, the magnitudes of parameter updates were invariant to diagonal rescaling of the gradients, while the stepsizes were approximately bounded by the stepsize hyperparameter. It combined the advantages of AdaGrad and RMSProp to deal with sparse gradients and non-stationary objectives, which was well-suited for non-convex optimization problems with large datasets and/or high-dimensional parameter spaces. Swarm intelligent algorithm was the collective behavior of decentralized, self-organized systems with scalability, adaptability, and collective robustness, which simulated the mutual behavior in groups of insects or animals to discourse a widespread range of difficult optimization problems under stationary environments. It was typically made up of a population of simple agents interacting locally with each other and with their environment. Among these agents, heuristic information was exchanged in the form of local interaction generating the behavior of adaptive search and resulting in global optimization. The key elements of swarm intelligent algorithm contained: 1) A large number of simple processing elements work without supervision. 2) Neighbourhood communication. 3) Though convergence was guaranteed, the time to convergence was uncertain. However, swarm intelligent algorithms were not appropriate for time-critical applications and their parameters were problem-dependent, whilst suffering from a stagnation situation or a premature convergence to a local optimum. Different from Adam or swarm intelligent methods, we developed an innovative learnable connected module (LCM) for better feature detail extractions related with DR lesion, which not only connected different modules in a feed-forward manner but also perceived the optimal connections for each connected module leading to automatically and adaptively learning the connections among different modules. In LCM, we directly connected the output of stochastic atrous spatial pyramid pooling (sASPP), global attention mechanism (GAM) and category attention mechanism (CAM) modules, and introduced a learnable connected weight for each connected module that was simple scalar variable and shared across all channels. During training, these connected weights were automatically learned and adaptive to the data by standard back-propagation with the help of Adam optimizer. In this fashion, the feature information from different modules could be reused and the information propagation could be strengthened, which overcome the drawbacks of previous fixed connection scheme. In addition, the designed LCM was simple and effective, and could be easily utilized in existing feature connection schemes as a universal plug and play CNN module by substituting their original counterparts for connection components.

3.6. Lesion visualization

Aiming to gain an insight into lesion features, we used Grad-CAM technique [51] to generate heatmaps for visualizing the lesion-relevant areas for each DR severity prediction. Firstly, we calculated the gradient $\partial y^c / \partial F_{ij}^k$ of the score y^c of any DR grade c relative to the feature map F^k of LCM and then undertook a global average pooling operation on the dimensions of height and width, obtaining the corresponding weight scores which could be calculated by the following equation.

$$\varphi_k^c = \frac{1}{z} \sum_i \sum_j \frac{\partial y^c}{\partial F_{ij}^k} \tag{11}$$

In the above equation, the weight scores φ_k^c reflected the contribution of the feature map to the DR grade prediction. After eliminating the influence of negative values by use of the ReLU activation function, we could obtain the final lesion region visualization results for DR grading, which could be given by

$$L_{Grad-CAM}^c = \mathrm{Re}LU\left(\sum_k \varphi_k^c F^k\right) \tag{12}$$

Biomedical Optics EXPRESS

4. Experiments and results

In this section, we firstly described three challenging benchmark containing DDR, EyePACS and Messidor datasets, evaluation indicators, and implementation details. Then, the quantitative and qualitative results of the proposed method were given and analyzed in detail.

4.1. Datasets

We carried out experiments on three benchmarks containing DDR [52], EyePACS [53], and Messidor [54] datasets.

DDR dataset comprised 13,673 color fundus images from 9,598 patients with an average age of 54. These fundus images were acquired by various fundus cameras with a 45-degree field of view. It provided three types of annotations, including image-level annotations for DR grading, pixel-level annotations for DR lesion-related segmentation, and bounding-box annotations for lesion detection associated with DR. In addition to five stages specified by the international clinical diabetic retinopathy (ICDR) disease severity scale [55], ungradable was also considered in this dataset. All 13,673 images were graded by ophthalmologists, while 757 gradable images were selected to perform lesion-related pixel-wise annotations (e.g., 486 EX annotations, 601 HE annotations, 570 MA annotations, and 239 SE annotations) and bounding-box annotations. The entire dataset was divided into the training set (6,320 images), validation set (2,503 images), and testing set (3,759 images) at a 5:2:3 ratio. It was worth noting that this dataset was mainly adopted to train our model for verifying its generalization performance. Besides, we implemented ablation studies in this dataset to investigate the effectiveness of each key component of our DACNN.

EyePACS dataset contained 35,126 training images, 10,906 validation images and 42,670 testing images. All fundus images were taken by using different types of cameras with a diversity of lighting conditions and weak annotation quality. Their grading scale had five grades from 0 to 4, which was in accordance with the ICDR disease severity scale. In this dataset, there existed some images with artifacts, out of focus, under- and over-exposed. In this dataset, we compared the proposed DACNN with recent state-of-the-art deep learning methods for multi-class DR grading task, and evaluated its generalization performance.

Messidor dataset was made up of 1,200 color fundus images (540 normal images and 660 abnormal images) captured using 8 bits per color plane at 1440×960, 2240×1488 or 2304×1536 pixels. The grading labels of each image in this dataset were provided, where DR was graded into four severity levels and diabetic macular edema (DME) was assigned into three categories. Nevertheless, their grading scale was slightly different from ICDR protocol [54]. Hence, for a fair comparison, we only performed a binary classification for making a distinction between non-referable (grade 0 and grade 1) and referable (grade 2 and grade 3), adhering to previous works [56,57].

4.2. Evaluation metrics

The performance of our DACNN in the binary classification task distinguishing between the referral and non-referral was measured by using accuracy (Acc), precision (Pre), recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). For the five DR categories, we utilized Acc and the quadratic weighted kappa, which could effectively reflect the model's performance on the unbalanced dataset. All statical analyses in this study were implemented by Python version 3.7.1, Scikit-learn version 0.20.0, Pandas version 0.23.4, and NumPy version 1.15.4.

Biomedical Optics EXPRESS

4.3. Implementation details

We implemented the proposed DACNN with Keras using TensorFlow backend. All experiments were conducted on the workstation equipped with NVIDIA RTX 2080Ti GPUs. By default, our network took fundus images with a resolution of 512×512 as input to mitigate the computational burden. The initial learning rate and weight decay were set to 0.005 and 0.00001, respectively. In addition, the batch size was set to 16 and k was assigned as 5. In our experiments compared with the state-of-the-art methods, we conducted binary-class DR grading task on the Messidor dataset, while performed multi-class DR severity grading on the EyePACS dataset. For binary DR grading task, the training and test images were 600 and 300, respectively. These test images constituted an experimental dataset for evaluating binary-class DR grading performance of our model. The model was trained for approximate 40 epochs with the Adam optimizer and binary cross-entropy loss function. In the case of multi-class DR grading, we used 35,126 images as training dataset to train our model, and 42,670 images as experimental dataset to test its performance. Similarly, we finished the training process at nearly 40 epochs with the Adam optimizer and multi-class cross-entropy loss function. In generalization experiments across DR datasets, we utilized 13,673 images from the DDR dataset as exclusively training dataset, and 42,670 images from the EyePACS dataset as experimental dataset for assessing the model's generalizability capability. The whole training process was stopped when 40 epochs were reached by using the Adam optimizer and cross-entropy loss function.

In order to avoid overfitting, we took several approaches as follows: 1) In DACNN, we designed stochastic Atrous Spatial Pyramid Pooling (sASPP) module based on atrous spatial pyramid pooling (ASPP), and the random operation added in sASPP could be used as an effective regularization term to help alleviate the overfitting. 2) We replaced ReLU function between two fully connected layers with PReLU function to adaptively learn the parameters of rectifiers and reduced overfitting risk. Meanwhile, fine-tuning ResNet-50 backbone network pre-trained on the ImageNet also mitigated the overfitting problem. 3) In addition, we also adopted the techniques of dropout and batch-normalization in our DACNN, and the feature dropping operation and batch-normalization operation during training could reduce overfitting resulting from insufficient training data but too many features. 4) During training, we further used an early stopping criterion to determine the optimized number of iterations to prevent the model from overfitting.

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

For the purpose of evaluating the grading performance of our model, we performed binary-class and multi-class tasks on the Messidor dataset and EyePACS dataset, respectively, and compared it with various current state-of-the-art deep learning methods, all of which were commonly used in DR grading tasks. A total of four representative cutting-edge methods namely CANet [11], CABNet [12], DR|GRADUTE [58], and the newly proposed method Lesion-Aware Transformer (LAT) [9] were used as competing methods. The quantitative comparison results of our approach and other cutting-edge methods were provided in Table 1. As can be seen from Table 1, for binary-class task discriminating referral from non-referral, our approach achieved the best results (AUC: 98.5%, Acc: 93.8%, Pre: 94.6%, F1-score: 92.6%, and Kappa: 87.9%) on the Messidor dataset among these competing methods, even if some methods [9] applied additional lesion information to assist DR classification. For example, compared with the state-of-the-art method LAT, our model had a performance gain of 0.6% and 2.8% over it on AUC and Kappa metrics, respectively. In addition, by comparison with CABNet, our DACNN achieved a substantial increase of 1.6%, 0.7%, 1.7%, and 1.1% in terms of AUC, Acc, Pre, and F1-score, respectively, while significantly surpassing CANet with 2.2%, 1.2%, 4.0%, and 1.3% improvements. In the case of a multi-class DR grading task on the EyePACS dataset, our DACNN still achieved remarkably improved performances using only image-level labels compared with other competing methods. Specifically, our DACNN obtained 1.8% Acc and 1.8% Kappa performance gains

Biomedical Optics EXPRESS

over the recent CABNet, respectively. Compared with the state-of-the-art LAT, our method still performed tolerably and outperformed it (Kappa: 0.886 vs 0.884) by a small margin. Further, we had supplemented three classical approaches (including ResNet-50, MobileNet-1.0, and Inception-v3) in the DR grading field, and reported the quantitative comparison results of our method and these methods. As seen in Table 1, it was very clear that our model outperformed these classical approaches by a large margin for DR binary-class and multi-class grading tasks in all metrics. These quantitative experimental results indicated that our DACNN performed best among recent state-of-the-art deep learning methods on DR severity grading task. The excellent performance of our model mainly benefitted from its strong discriminating power of subtle multi-scale lesion features by integrating GAM, CAM, sASPP, and LCM. This enabled our model to pay more attention to various small lesion regions and effectively learn their representative lesion features, thereby greatly enhancing the performance of comprehensive DR grading. At last, we also listed the number of the parameters of different methods, as summarized in Table 1. It could be observed that our approach produced significant performance improvements over other methods at the cost of increasing a few extra parameters.

Table 1. The quantitative comparisons of our approach and other state-of-the-art methods for binary-class and multi-class DR grading tasks on Messidor and EyePACS datasets. "-" denotes no results reported in corresponding works. The best results were shown in bold.[a]

Methods	Parameters	Messidor						EyePACS	
		AUC	Acc	Pre	Recall	F1-score	Kappa	Kappa	Acc
Our DACNN	157.86M	**0.985**	**0.938**	**0.946**	0.907	**0.926**	**0.879**	**0.886**	**0.880**
CABNet [12]	25.19M	0.969	0.931	0.929	0.902	0.915	-	0.868	0.862
LAT [9]	75.5M	0.979	-	-	-	-	0.851	0.884	-
CANet [11]	29.03M	0.963	0.926	0.906	**0.920**	0.913	-	-	-
DR|GRADUATE [58]	7.82M	0.910	0.912	0.933	0.614	0.741	0.710	0.740	0.536
ResNet-50 [30]	25.54M	0.880	0.929	0.857	0.796	0.826	0.781	0.653	0.815
MobileNet-1.0 [59]	3.2M	0.867	0.927	0.872	0.764	0.815	0.769	0.513	0.798
Inception-v3 [60]	21.77M	0.876	0.932	0.883	0.780	0.828	0.786	0.657	0.824

[a]AUC: the area under the receiver operating characteristic (ROC) curve; Acc: accuracy; Pre: precision

4.5. Ablation studies

In this section, we performed a series of ablation experiments in the DDR dataset to deeper insight into the effects of each major component of our model, as shown in Table 2. Results and analysis were as follows:

1. Effectiveness of GAM: To explore the effectiveness of the GAM, we derived two baselines: the baseline No.1 (backbone only) versus No.2 (baseline + GAM) in Table 2. We could clearly observe that the performance in accuracy and Kappa score was consistently improved (about 1.1% and 0.8%, respectively), which meant that introducing the GAM module could enable our model to accurately distinguish true DR lesion-related regions.

2. Effectiveness of CAM: We investigated the importance of CAM. From Table 2, we could find that No.3 (baseline + GAM + CAM) got a performance boost relative to the No.2 (baseline + GAM), upgrading to the values of 0.872 and 0.915 in terms of Acc and Kappa score. This indicated that CAM could be applied for guiding to deeply mine more discriminative lesion-related features, which was helpful in improving the DR severity grading performance.

Biomedical Optics EXPRESS

3. Effectiveness of GAM and CAM: We also investigated the contributions of the combination of the GAM and CAM components. As illustrated in Table 2, No.3 performed better than other settings (such as, No.1~No.2) in most indicators. In contrast to No.1, No.3 could obtain higher performance gains by a large margin, exceeding No.1 by 1.4% and 1.0% in Acc and Kappa score, respectively. These improvements demonstrated that GAM together with CAM could significantly boost the model's performance and show that they complemented each other.

4. Effectiveness of LCM: To understand the relative importance of LCM, we integrated LCM into No.3, producing No.4 (baseline + GAM + CAM + LCM). From these results in Table 2 (No.4 versus No.3), it was noted that No.4 using the feature connections generated by LCM outperformed No.3 without the feature connections. This may be ascribed to the fact that LCM learned optimal control parameters adaptively by standard back-propagation, strengthened image feature propagation and parameter reutilization, and thereby generated more robust features to adapt to lesion appearance variations.

5. Effectiveness of sASPP: We further validated the effectiveness of sASPP module. From Table 2 (No.5 versus No.4), it could be seen that sASPP scheme led to an increase in Acc and Kappa score, from 0.873 to 0.889, and 0.921 to 0.930, respectively, proving its effectiveness and necessity in improving performance. This superiority mainly benefitted from the ability to capture multi-scale lesion features associated with DR, and the randomness helping to alleviate the overfitting, in sASPP, which was favorable in the DR grading task.

6. Effectiveness of Different k in DACNN: We made an analysis on the impact of different hyper-parameter k which described the number of feature channels for each stage of DR severity grades. The influence of various k on DR grading performance was displayed in Supplement 1. These results indicated that with the value of k increasing, the performance of DR grading was gradually improved. When $k = 5$, the best result (Acc: 0.889 and Kappa: 0.930) for DR grading was obtained. Nevertheless, when the score of k was further increased, the degradation of grading performance occurred. We argued that this was mainly because the overfitting and feature redundancy in our model may exist when the value of k was over 5. Accordingly, we assigned k as 5 in our model for achieving better performance.

7. Effectiveness of our DACNN on Different Imbalanced Data Distributions: For the purpose of evaluating the effectiveness of our DACNN on different imbalanced data distributions, we decreased the amount of the training images relevant to the categories with less images. In the DDR dataset, the number of samples graded as DR 0 and DR 2 was more, whereas those assigned as DR 1, DR 3, and DR 4 were less. We defined the imbalanced ratio as M/L. M represented the number of samples in the union of DR 0 and DR 2 with more samples. Similarly, L denoted the number of samples in the union of DR 1, DR 3, and DR 4 with less samples. We kept the amount of training images in DR 0 and DR 2 fixed, and adjusted the imbalanced ratios to 5:1, 6:1, and 7:1, in a bid to train model. For carrying out a fair comparison, the same training images were adopted for the baseline and our DACNN. In Supplement 1, it was obviously observed that as the imbalanced ratio increased, the DR grading performance degraded for both baseline and our DACNN. Yet, by comparison with the baseline, our DACNN had a smaller drop. As a specific, when the imbalanced ratio was 7:1, the performance of baseline dropped by approximately 3.5% in Acc, whereas our model only slightly descended by 2.5%. This demonstrated the superiority of our DACNN on the imbalanced data distribution.

8. Effectiveness of Different Backbones: For demonstrating the generalization of the aforementioned modules, we utilized several state-of-the-art backbone architectures, mainly

Biomedical Optics EXPRESS

including MobileNet [59], Inception-v3 [60], and DenseNet [34]. From Supplement 1, we could obviously observe that the baseline models integrating all the above-mentioned modules were able to achieve consistent substantial improvement, while ResNet-50 with all modules gained the best results in the DDR dataset. Note that the improvement for DenseNet-121 was particularly obvious in performance by 11.3% and 15.1% in Acc and Kappa score, respectively. In addition, they also brought a large performance boost for MobileNet-1.0, with 5.3% Acc and 5.6% Kappa score improvements. These results implied that the developed GAM, CAM, sASPP, and LCM could be utilized in a wide range of backbone networks and consistently promote the performance of DR grading.

Table 2. The ablation studies of our DACNN in the DDR dataset. The best results were shown in bold.[a]

Method	Acc	Kappa
(No.1) Baseline (Resnet-50)	0.858	0.905
(No.2) Baseline + GAM	0.869	0.913
(No.3) Baseline + GAM + CAM	0.872	0.915
(No.4) Baseline + GAM + CAM + LCM	0.873	0.921
(No.5) Baseline + GAM + CAM + LCM + sASPP	**0.889**	**0.930**

[a]Acc: accuracy

4.6. Generalization across DR datasets

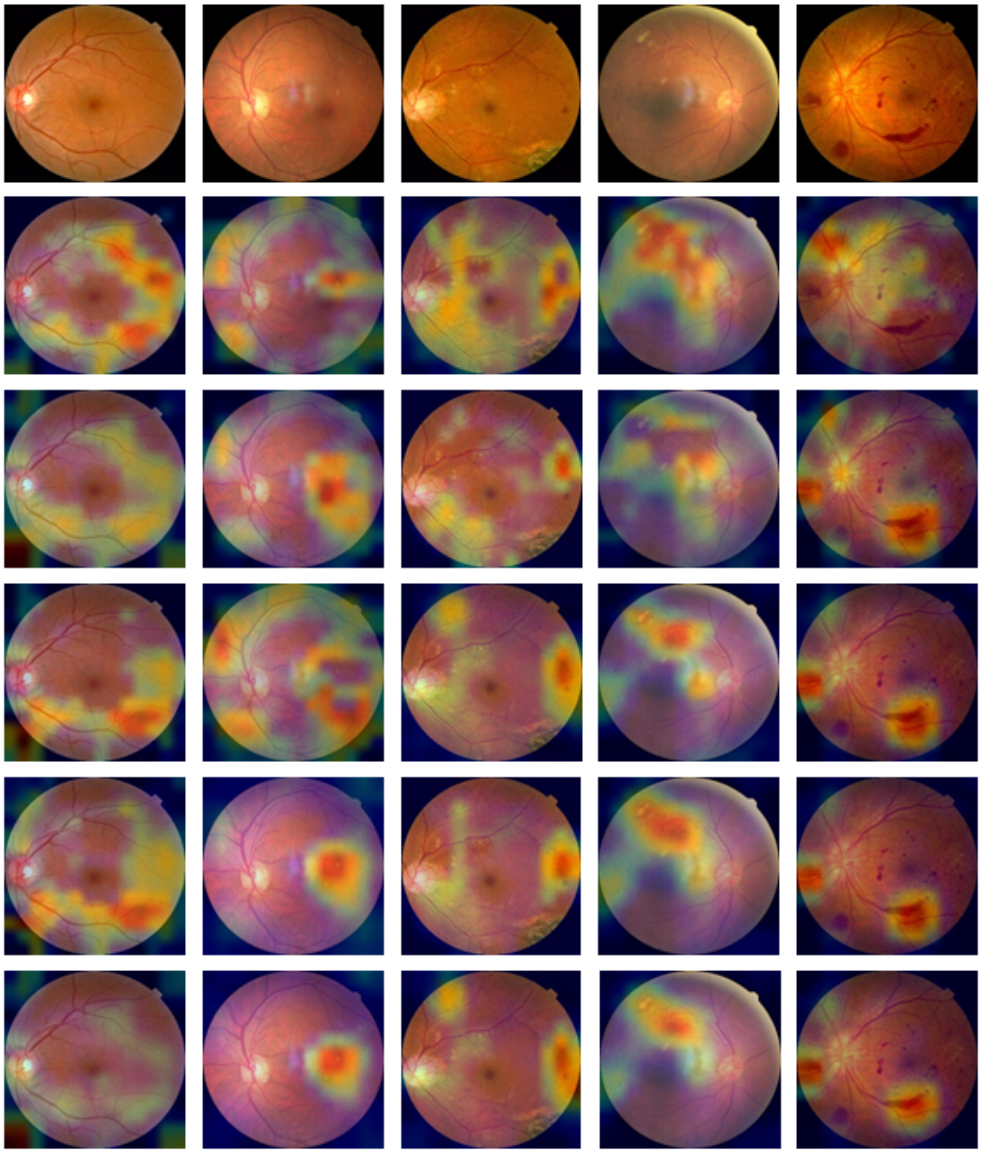
As for practical clinical application, realizing the generalization across domains under diverse imaging conditions was challenging yet meaningful. With the goal of testing the generalizability capability of our model, we further investigated its performance by using the images from the DDR dataset for exclusively training and the publicly accessible EyePACS dataset for external validation. Table 3 provided the comparison results of different methods on the generalization. Among these results, our model performed the best by means of narrowing down the gap between fundus photographs under different conditions, with an AUC of 96.1%, an Acc of 86.5%, and a kappa score of 88.1%. For instance, our method achieved a large performance boost than the DeepDR, with 2.4% AUC improvement for identifying referral and non-referral. Similarly, it remarkably exceeded the AFN by 2.2% on Kappa score and approached to that generated by the LAT. In contrast to the CABNet, our method slightly outperformed it by 0.3% and 1.3% on Acc and Kappa indicators, respectively. However, the DR grading performance of our model had a slight drop in the EyePACS dataset than that yielded in the DDR dataset (Acc:0.865 vs 0.889, Kappa:0.881 vs 0.930), but still exhibited outstanding results. This happened could be explained by the fact that the fundus images from the EyePACS dataset were captured by different types of

Table 3. Generalization comparisons of our DACNN and other advanced methods on EyePACS dataset. "-" denotes no results reported in corresponding works. The best results were shown in bold.[a]

Method	AUC	Acc	Kappa
AFN [61]	-	-	0.859
LAT [9]	-	-	**0.884**
CABNet [12]	-	0.862	0.868
DeepDR [25]	0.937	-	-
Our DACNN	**0.961**	**0.865**	0.881

[a]AUC: the area under the receiver operating characteristic (ROC) curve; Acc: accuracy

Biomedical Optics EXPRESS



(a) no DR (b) mild DR (c) moderate DR (d) severe DR (e) proliferative DR

Fig. 5. Visualization results between baseline, sASPP, GAM, CAM, and LCM. The six rows denoted the original images, heatmaps of sASPP, GAM, CAM, and LCM, respectively.

cameras and their image qualities were relatively low which contained some noises like uneven illumination, artifacts, out of focus, under- and over-exposed. The above results indicated that our model owed good generalizability to unseen DR datasets, and had a great potential to be incorporated in clinical settings to complement and strengthen existing DR screening programs by virtual of saving resources and increasing workflow efficiency.

Biomedical Optics EXPRESS

4.7. Lesion visualization

To better gain insight into the model and understand the effectiveness of major modules in our model, we used Grad-CAM [51] to generate activation maps from different modules to visualize discriminative image regions for qualitative analysis. Some representative samples were as illustrated in Fig. 5. We took the heatmaps of baseline (with only backbone), baseline with sASPP, baseline with sASPP and GAM, baseline integrating sASPP, GAM, and CAM, and baseline incorporating sASPP, GAM, CAM and LCM to present the improvements of different components of our model. In Fig. 5, from left to right in the top row, we provided five images relative to the five severity levels (from DR 0 to DR 4) from the DDR dataset. The second row corresponded to the heatmaps generated by the baseline. The third row described the visualization results by the baseline with sASPP. The fourth row displayed the results produced by the baseline with sASPP and GAM. The fifth row provided the heatmaps resulting from the baseline integrating sASPP, GAM, and CAM. The last row indicated activation maps acquired by the baseline incorporating sASPP, GAM, CAM, and LCM. We could see from Fig. 5 that for the baseline, it neglected some critical information and emphasized some unrelated areas. Compared with the results obtained by the baseline, sASPP obviously refined global features but still only identify a small part of the lesion without fully covering the corresponding DR lesion areas. GAM highlighted global attention maps rather than region-level features, and still generated some irrelevant characteristics. Although CAM could recalibrate the global attention maps from GAM to focus on some obvious discriminative lesion areas, it still located few useless feature regions. After being refined by LCM, our model could recognize fine-grained discriminative lesion regions and more precisely fine-tune the location of suspicious small lesion areas, which could be helpful for the clinical DR diagnosis. Through the above analysis, the outstanding ability of our model to capture small lesion features could be demonstrated qualitatively.

5. Discussion

With the development of deep learning technology, the performance on automated DR gradeability assessment had been greatly improved. Nevertheless, as for fine-grained DR grading, it was still challenging as a result of small lesions such as MA, HM, SE, and HE that were difficult to capture using traditional CNNs, and an imbalanced DR data distribution. The key to resolve fine-grained DR grading was to identify more subtle distinctive DR lesion features and tackle the problem of imbalanced data distribution. In this work, we presented and validated a novel DACNN for automated fine-grained DR grading and lesion discovery only by using image-level supervision. It combined sASPP, GAM, CAM and LCM into a unified network frame to learn multi-scale subtle discriminative lesion features and cope with the issue of imbalanced data distributions. Further, the suspicious-looking lesion regions involving DR in retinal fundus images were generated by means of the DACNN, which increased the model's interpretability. Comprehensive experimental results on three public challenging datasets verified that our method significantly improved the performance on DR grading and lesion discovery by comparison with current state-of-the-art models trained using different supervision ways, and had good generalizability, which could potentially be incorporated in clinical workflow for complementing and enhancing existing DR screening programs as well as assisting primary care physicians or ophthalmologists in making better diagnosis objectively and rapidly.

In previous studies, the majority of deep learning methods [19,20] graded the severity of DR with the assistance of location information of lesion in a two-stage way, termed as pixel-level supervision. It allowed the network to learn more information from corresponding lesion location. However, these two-stage deep learning methods required annotations of lesion locations in retinal fundus images according to expert knowledges or output of lesion segmentation, and errors introduced by annotations or segmentation would affect the performance of subsequent DR grading. Moreover, the structures of these networks were commonly highly complex. In addition,

Biomedical Optics EXPRESS

some approaches [24,25] used multi-branch subnetworks to classify different stages of DR grades and identify retinal lesions. They formulated lesion location and DR grading into different individual tasks, which increased the model's complexity and demanded higher computation costs. On the contrary, there also existed some deep learning methods for DR grading with image-level supervision [11,12]. Yet, these methods had a common point that they suffered from locality of convolution operations and limited the receptive field of the models, which brought great challenges to capture small/tiny discriminative lesion features at various scales. Unlike these methods, instead, we tried to develop a simple yet effective DACNN integrating sASPP, GAM, CAM and LCM so as to seek deeply more discriminative and representative semantic feature information associated with DR for enhancing DR grading capacity with only image-level supervision. It was important to underline that the designed sASPP, GAM, CAM and LCM were effective and universal, and could be easily utilized in a wide range of backbone networks as a universal plug and play modules and substantially boost the performance in fine-grained DR grading task. In the current study, our DACNN was able to gain excellent results with the overall accuracy of 0.889, and Kappa score of 0.930 for multi-class DR grading task in the DDR dataset. Moreover, the AUC, Acc and Kappa values in the Messidor dataset reached up to 0.985, 0.938 and 0.879, respectively. These were supported by the ablation study and comparison results with state-of-the-art methods, presented in the above results section.

In our work, we quantitatively compared our DACNN with recent mainstream deep learning methods for DR grading, and analyzed the influence of each main module on the DACNN, as presented in Table 1 and Table 2. The results fully indicated that the presented method was able to perform better than these current state-of-the-art methods, and each of the devised modules was effective. On the other hand, only giving an accurate DR severity grade prediction was insufficient for real-world clinical application. If we were capability of providing some evidences on how the proposed model made certain predictions, this would offer ophthalmologists better help and confidence on the prediction DR severity grade. To this end, we performed lesion visualization with the Grad-CAM technique [51], which could highlight the lesion regions within the input retinal fundus images when predicting DR severity grade. From Fig. 5, we could intuitively observe that our developed DACNN located accurately lesion areas corresponding DR. Owing to integrated sASPP, GAM, CAM and LCM, our model could automatically evolve with the help of these modules and discovery more small discriminative lesion regions. The good performance generated by our DACNN could be attributed to the following reasons. First, the input of sASPP in DR grading task was the high-level abstract semantic features extracted from ResNet-50 backbone which could fully represent lesions, while sASPP was able to capture multi-scale lesion features and attempted to combine different feature maps produced from multiple atrous convolutions adaptively to avoid their co-adaptation. Second, by merging GAM and CAM modules that were complementary each other, the more tiny lesion feature details could be obtained. Third, LCM could automatically and adaptively learn optimal connections among different layers to strengthen feature propagation and fusion. To sum up, combing sASPP, GAM, CAM, and LCM enabled our model to search diversity of target DR lesion regions in an optimal manner, adaptively aggregate contextual information and corresponding lesion features associated with these target areas, leading to feature discriminability boost and substantial improving performance of DR lesion discovery.

In real world settings, medical images often displayed variations in appearance under various imaging conditions, and domain shift between different datasets occurred. Consequently, it was challenging but meaningful to achieve good generalization among multiple domains. From generalization comparisons of different methods from the DDR dataset to the EyePACS dataset (See Table 3), it could be observed that our methods on the EyePACS dataset also achieved nearly best results (AUC: 0.961, Acc: 0.865, and Kappa: 0.881) in AUC, Acc, and Kappa metrics through cutting down the gap among images from diversity of conditions. In spite of the fact

Biomedical Optics EXPRESS

that our model obtained an almost equal performance as LAT [9] on EyePACS dataset (Kappa: 0.881 versus 0.884), we could notice that it achieved much better results on the Messidor dataset (Kappa:0.879 versus 0.851). This suggested that the developed method manifested more robust and generalization ability than existing advanced approaches when directly applied to other datasets. Additionally, it must be emphasized also that our model's performance decreased slightly when training was implemented in the DDR dataset, and testing on another EyePACS dataset. This was mainly due to the class imbalance among different DR grades and large variation in illumination, resolution, intensity and quality, on the EyePACS dataset. Such simple yet effective method could be applied on diverse retinal fundus images captured by different cameras, facilitating its adaptability. Notwithstanding a slight drop, the performance of the proposed DACNN was still adequate to be considered a viable DR screening method to some extent.

Despite our model manifesting impressive performance, it had still several limitations. First, we only used image-wise supervision to train the model, leading to challenge to find the accurate tiny lesion location. One feasible solution is to attempt to integrate lesion segmentation information into our model to help improve the capacity in more tiny lesion discovery. Second, in current study, we only investigated the model's generalization ability in the EyePACS datast. In the future, we will apply it on different publicly available DR datasets and real-world clinical DR datasets to fully evaluate its generalization performance. At last, the number of samples in mild, severe and proliferative DR was insufficient, which may have a negative influence on model's performance. As a future research, we will introduce generative adversarial networks (GANs) to synthesize more high-quality retinal fundus photographs for model training so that the DR grading performance could be further boosted.

6. Conclusion

In this paper, a simple yet effective DACNN combining sASPP, GAM, CAM, and LCM was developed to predict DR severity grade in retinal fundus images. The proposed DACNN could be trained in an end-to-end fashion for gathering discriminative lesion features and performing fine-grained DR grading. Specifically, we integrated stochastic operation into ASPP, designing sASPP to extract multi-scale lesion features and prevent the co-adaptation of multiple atrous convolutions in ASPP. Then, GAM was introduced to capture class-agnostic global attention features and retain lesion details, while incorporating CAM to learn class-specific features and enlarge the distance among different stages of DR severity grades. Further, LCM was presented with the goal of adaptively searching the optimal connections from different layers, and strengthening feature information propagation and reuse. Extensive experiments on different datasets demonstrated that our DACNN could manifest remarkable performance in DR grading tasks, and had a good generalization. Through ablation study, the effectiveness of the critical module in our DACNN was also clearly shed light on. From our experiments, we could conclude that our DACNN was able to become a good alternative for assisting ophthalmologists in making better diagnoses and treatment of DR patients. It was sufficiently feasible and could be potentially extended to the detection of other diseases using corresponding medical images, such as chest X-ray, MRI, and CT.

Funding. Key Technologies Research and Development Program (2020YFC2008704); National Natural Science Foundation of China (51675321).

Acknowledgements. We thank Shanghai Eye Disease Prevention and Treatment Center and Shanghai General Hospital for their invaluable help.

Disclosures. The authors declare no conflict of interest related to this article.

Data availability. The DDR dataset in this paper is publicly available in Ref. [62]. The EyePACS dataset in this paper is publicly available in Ref. [63]. The Messidor dataset in this paper is publicly available in Ref. [64].

Biomedical Optics EXPRESS

Supplemental document. See Supplement 1 for supporting content.

References

1. D. A. Antonetti, P. S. Silva, and A. W. Stitt, "Current understanding of the molecular and cellular pathology of diabetic retinopathy," Nat. Rev. Endocrinol. **17**(4), 195–206 (2021).
2. S. Rego, M. Dutra-Medeiros, F. Soares, and M. Monteiro-Soares, "Screening for diabetic retinopathy using an automated diagnostic system based on deep learning: diagnostic accuracy assessment," Ophthalmologica **244**(3), 250–257 (2021).
3. M. S. Farooq, A. Arooj, R. Alroobaea, A. M. Baqasah, M. Y. Jabarulla, D. Singh, and R. Sardar, "Untangling computer-aided diagnostic system for screening diabetic retinopathy based on deep learning techniques," Sensors **22**(19), C1 (2022).
4. D. S. W. Ting, C. Y. L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. S. Yeo, and S. Y. Lee, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," JAMA **318**(22), 2211–2223 (2017).
5. E. Ipp, D. Liljenquist, B. Bode, V. N. Shah, S. Silverstein, C. D. Regillo, J. I. Lim, S. Sadda, A. Domalpally, G. Gray, M. Bhaskaranand, C. Ramachandra, and K. Solanki, "Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy," Jama Netw. Open. **4**(11), e2134254 (2021).
6. M. R. Islam, L. F. Abdulrazak, M. Nahiduzzaman, M. O. F. Goni, M. S. Anower, M. Ahsan, J. Haider, and M. Kowalski, "Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images," Comput. Biol. Med. **146**, 105602 (2022).
7. X. F. Wang, M. Xu, J. C. Zhang, L. Jiang, L. Li, M. X. He, N. L. Wang, H. R. Liu, and Z. L. Wang, "Joint Learning of Multi-Level Tasks for Diabetic Retinopathy Grading on Low-Resolution Fundus Images," IEEE J. Biomed. Health Inform. **26**(5), 2216–2227 (2022).
8. N. Eftekhari, H. R. Pourreza, M. Masoudi, K. Ghiasi-Shirazi, and E. Saeedi, "Microaneurysm detection in fundus images using a two-step convolutional neural network," Biomed. Eng. Online. **18**(1), 67 (2019).
9. R. Sun, Y. H. Li, T. Z. Zhang, Z. D. Mao, F. Wu, and Y. D. Zhang, "Lesion-Aware Transformers for Diabetic Retinopathy Grading," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), pp. 10933–10942.
10. R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," Ophthalmology **124**(7), 962–969 (2017).
11. X. M. Li, X. W. Hu, L. Q. Yu, L. Zhu, C. W. Fu, and P. A. Heng, "CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," IEEE Trans. Med. Imaging **39**(5), 1483–1493 (2020).
12. A. L. He, T. Li, N. Li, K. Wang, and H. Z. Fu, "CABNet: category attention block for imbalanced diabetic retinopathy grading," IEEE Trans. Med. Imaging **40**(1), 143–153 (2021).
13. R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumki, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," Ophthalmology **126**(4), 552–564 (2019).
14. S. H. Woo, J. Park, J.Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *15th European Conference on Computer Vision*, (2018), pp. 3–19.
15. M. Hardas, S. Mathur, A. Bhaskar, and M. Kalla, "Retinal fundus image classification for diabetic retinopathy using SVM predictions," Phys. Eng. Sci. Med. **45**(3), 781–791 (2022).
16. L. B. Frazao, N. Theera-Umpon, and S. Auephanwiriyakul, "Diagnosis of diabetic retinopathy based on holistic texture and local retinal features," Inf. Sci. **475**, 44–66 (2019).
17. C. Pratheeba and N.N. Singh, "A Novel Approach for Detection of Hard Exudates Using Random Forest Classifier," J. Med. Syst. **43**(7), 180 (2019).
18. M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey," IEEE Access **10**, 28642–28655 (2022).
19. B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," IEEE Trans. Biomed. Eng. **59**(6), 1720–1726 (2012).
20. K. Y. Lin, W. H. Hsih, Y. B. Lin, C. Y. Wen, and T. J. Chang, "Update in the epidemiology, risk factors, screening, and treatment of diabetic retinopathy," J. Diabetes Invest. **12**(8), 1322–1325 (2021).
21. Y. H. Yang, T. Li, W. S. Li, H. S. Wu, W. Fan, and W. S. Zhang, "Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks," in the *International Conference on Medical Image Computing and Computer-assisted Intervention*, (Springer, 2017), pp. 533–540.
22. Y. Zhou, X. D. He, L. Huang, L. Liu, F. Zhu, S. S. Cui, and L. Shao, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *2019 IEEE/CVF Computer Vision and Pattern Recognition Conference*, (2019), pp. 2079–2088.
23. H. Wang, G. H. Yuan, X. G. Zhao, L. B. Peng, Z. R. Wang, Y. M. He, C. Qu, and Z. M. Peng, "Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening," Comput. Meth. Prog. Bio. **191**, 105398 (2020).

Biomedical Optics EXPRESS

24. Z. Wu, G. L. Shi, Y. Chen, F. Shi, X. J. Chen, G. Coatrieux, J. Yang, L. M. Luo, and S. Li, "Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network," Artif. Intell. Med. **108**, 101936 (2020).
25. L. Dai, L. Wu, H. T. Li, C. Cai, Q. Wu, H. Y. Kong, R. H. Liu, X. N. Wang, X. H. Hou, and Y. X. Liu, "A deep learning system for detecting diabetic retinopathy across the disease spectrum," Nat. Commun. **12**(1), 1–11 (2021).
26. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Megan, and R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," JAMA **316**(22), 2402–2410 (2016).
27. M. D. Abramoff, Y. Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning," Invest. Ophthalmol. Visual Sci. **57**(13), 5200–5206 (2016).
28. F. Li, Y. G. Wang, T. Y. Xu, L. Dong, L. Yan, M. S. Jiang, X. D. A. Zhang, H. Jiang, Z. Z. Wu, and H. D. Zou, "Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs," Eye **36**(7), 1433–1441 (2022).
29. Y. C. Lai, F. Fan, Q. S. Wu, W. C. Ke, P. X. Liao, Z. H. Deng, H. Chen, and Y. Zhang, "LCANet: Learnable Connected Attention Network for Human Identification Using Dental Images," IEEE Trans. Med. Imaging **40**(3), 905–915 (2021).
30. K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *the Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), pp. 770–778.
31. S. N. Xie, R. Girshick, P. Dollar, Z. W. Tu, and K. M. He, "Aggregated Residual Transformations for Deep Neural Networks," in the *Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), pp. 5987–5995.
32. R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway Networks," arXiv preprint arXiv:1505.00387, 2015.
33. G. Larsson, M. Maire, and G. Shakhnarovich, "FractalNet: Ultra-Deep Neural Networks without Residuals," arXiv preprint arXiv:1605.07648, 2016.
34. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *the Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), pp. 2261–2269.
35. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," Med. Image Anal. **53**, 197–207 (2019).
36. A. Sinha and J. Dolz, "Multi-Scale Self-Guided Attention for Medical Image Segmentation," IEEE J. Biomed. Health Inform. **25**(1), 121–130 (2021).
37. Mustaqeem and S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," Appl. Soft Comput. **102**(1), 107101 (2021).
38. L. L. Gao, Z. Guo, H. W. Zhang, X. Xu, and H. T. Shen, "Video Captioning With Attention-Based LSTM and Semantic Consistency," IEEE Trans. Multimedia **19**(9), 2045–2055 (2017).
39. Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection," IEEE Trans. Image Process **30**, 7012–7024 (2021).
40. J. Hu, L. Shen, S. Albanie, G. Sun, and E. H. Wu, "Squeeze-and-Excitation Networks," IEEE Trans. Pattern Anal. **42**(8), 2011–2023 (2020).
41. Z. Zhu, M. D. Xu, S. Bai, T. T. Huang, and X. Bai, "Asymmetric Non-Local Neural Networks for Semantic Segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), pp. 593–602.
42. K. Yang, Y.G. Tang, S.J. Zhang, and Z. Zhang, "A deep learning approach to state of charge estimation of lithium-ion batteries based on dual-stage attention mechanism," Energy **244**(1), 1–11 (2022).
43. Y. Cao, J. R. Xu, S. Lin, F. Y. Wei, and H. Hu, "GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), pp. 1971–1980.
44. S. Q. Huang, J. A. Li, Y. Z. Xiao, N. Shen, and T. F. Xu, "RTNet: Relation Transformer Network for Diabetic Retinopathy Multi-Lesion Segmentation," IEEE Trans. Med. Imaging **41**(6), 1596–1607 (2022).
45. L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Trans. Pattern Anal. **40**(4), 834–848 (2018).
46. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM **60**(6), 84–90 (2017).
47. J. J. Hu, Y. Y. Chen, and Z. Yi, "Automated segmentation of macular edema in OCT using deep neural networks," Med. Image Anal. **55**, 216–227 (2019).
48. G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep Networks with Stochastic Depth," in *14th European Conference on Computer Vision (ECCV)*, (2016), **9908**, pp. 646–661.
49. S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *32nd International Conference on Machine Learning*, (2015), 37, pp. 448–456.
50. K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), pp. 1026–1034.

Biomedical Optics EXPRESS

51. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *16th IEEE International Conference on Computer Vision (ICCV)*, (2017), pp. 618–626.
52. T. Li, Y. Q. Gao, K. Wang, S. Guo, H. R. Liu, and H. Kang, "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening," Inf. Sci. **501**, 511–522 (2019).
53. J. Cuadros and G. Bresnick, "EyePACS: an adaptable telemedicine system for diabetic retinopathy screening," J. Diabetes Sci. Technol. **3**(3), 509–516 (2009).
54. E. Decenciere, X. W. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, J. R. Ordonez-Varela, P. Massin, A. Erginay, B. Charton, and J. C. Klein, "Feedback on a publicly distributed image database: The Messidor database," Image Anal. Stereol. **33**(3), 231–234 (2014).
55. C. P. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, and J. T. Verdaguer, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," Ophthalmology **110**(9), 1677–1682 (2003).
56. Y. Zhou, B. Y. Wang, L. Huang, S. S. Cui, and L. Shao, "A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading, and Transferability," IEEE Trans. Med. Imaging **40**(3), 818–828 (2021).
57. F. Y. Tang, P. Luenam, A. R. Ran, A. A. Quadeer, R. Raman, P. Sen, R. Khan, A. Giridhar, S. Haridas, M. Iglicki, D. Zur, A. Loewenstein, H. P. Negri, S. Szeto, B. K. Y. Lam, C. C. Tham, S. Sivaprasad, M. McKay, and C. Y. Cheung, "Detection of Diabetic Retinopathy from Ultra-Widefield Scanning Laser Ophthalmoscope Images: A Multicenter Deep Learning Analysis," Ophthalmol. Retina. **5**(11), 1097–1106 (2021).
58. T. Araujo, G. Aresta, L. Mendonca, S. Penas, C. Maia, A. Carneiro, A. M. Mendonça, and A. Campilho, "DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images," Med. Image Anal. **63**, 101715 (2020).
59. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. J. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, (2018), pp. 1–9.
60. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), pp. 2818–2826.
61. Z. W. Lin, R. Q. Guo, Y. J. Wang, B. Wu, T. T. Chen, W. Z. Wang, D. Z. Chen, and J. Wu, "A Framework for Identifying Diabetic Retinopathy Based on Anti-noise Detection and Attention-Based Fusion," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, (Springer International Publishing, 2018), 74–82.
62. T. Li, Y. Q. Gao, K. Wang, S. Guo, H. R. Liu, and H. Kang, "OIA-DDR," GitHub, 2019, https://github.com/nkicsl/DDR-dataset
63. Kaggle Competition, "Diabetic Retinopathy Detection," Kaggle repository, 2015, https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data
64. Messidor program partners, "Messidor," ADCIS repository, 2018, https://www.adcis.net/en/third-party/messidor/