Robust Collaborative Learning of Patch-level and Image-level Annotations for Diabetic Retinopathy Grading from Fundus Image

Yehui Yang\(^1\)*, Fangxin Shang\(^1\)*, Binghong Wu\(^1\), Dalu Yang\(^1\), Lei Wang\(^1\), Yanwu Xu\(^1\), Wensheng Zhang\(^2\) and Tianzhu Zhang\(^3\)
\(^1\) Artificial Intelligence Group, Baidu, Beijing, China
\(^2\) Institute of Automation, Chinese Academy of Sciences, Beijing, China
\(^3\) University of Science and Technology of China, Hefei, Anhui, China

Currently, diabetic retinopathy (DR) grading from fundus images has attracted incremental interests in both academic and industrial communities. Most convolutional neural networks (CNNs) based algorithms treat DR grading as a classification task via image-level annotations. However, they have not fully explored the valuable information from the DR-related lesions. In this paper, we present a robust framework, which can collaboratively utilize both patch-level lesion and image-level grade annotations, for DR severity grading. By end-to-end optimizing the entire framework, the fine-grained lesion and image-level grade information can be bidirectionally exchanged to exploit more discriminative features for DR grading. Compared with the recent state-of-the-art algorithms and three over 9-years clinical experienced ophthalmologists, the proposed algorithm shows favorable performance. Testing on the datasets from totally different scenarios and distributions (such as label and camera), our algorithm is proved robust in facing image quality and distribution problems that commonly exist in real-world practice. Extensive ablation studies dissect the proposed framework and indicate the effectiveness and necessity of each motivation. The code and some valuable annotations are now publicly available.

Index Terms—Convolutional neural networks, Diabetic retinopathy, Fundus image, Collaborative learning

![Fig. 1.](Image)

(a) A sample of fundus image with moderate DR, and the arrows indicate some key DR-related lesions. (b) The green boxes are labeled by one of our annotators. The missing-annotated samples, indicated by the red arrows, may confuse the detectors that are trained with the entire images.

I. INTRODUCTION

Diabetes is a universal chronic disease that affects one in every eleven adults worldwide, and approximately 40% to 45% of patients with diabetes may develop diabetic retinopathy (DR) in their lifetime [1], [2], [3], [4]. According to [5], [6], DR is one of the leading causes of irreversible blindness, while most parts of the world are short of qualified ophthalmologists. Therefore, the quick and automatic grade the DR severity is critical and urgent to reduce burdens of the limited ophthalmologists and provide timely morbidity diagnosis for massive patients.

DR grading aims to classify fundus images into different classes in terms of DR severity. According to the *International Clinical Diabetic Retinopathy Disease Severity Scale* [7], DR falls into five severity grades including no DR, mild, moderate, severe, and proliferative. The five grades can also be merged as a binary classification, i.e. No-DR versus DR, or non-referable (no and mild DR) versus referable (moderate and worse DR) [5], [8]. Recently, some researchers trend to leverage the powerful CNNs (convolutional neural networks) for DR grading. Researchers from Google Research use the Inception-v3 [5] to detect referable DR and macular edema. Sankar et al. [9], Alban et al. [10] and Pratt et al. [4] construct multi-class classifiers for DR grading with some popular or their own CNN architectures.

However, the above end-to-end algorithms take the DR grading as a *black box* classification task, which ignores the valuable fine-grained DR-related lesions. It is worth noting that the DR severity grade has strong correlations with different lesion types and combinations [7], such as MA (microaneurysm), hemorrhage and exudate illustrated in Fig. 1(a).

Therefore, some researchers attempt to improve the grading performance by integrating the lesion information. Antal and Hajdu [11] propose an ensemble-based algorithm for MA detection, mapping the fundus images into ‘DR/non-DR’ based on the presence or absence of the MAs. Yang et al. [12] introduce an offline lesion-based weighting scheme to improve the performance of DR grading. Lin et al. [13] propose a similar two-stage framework to integrate the patch-level lesion features and image-level global features with an attention network. Zhou et al. [14] collaboratively optimize the lesion segmentation and DR grading in an adversarial way.

Although above works associate lesion information with DR grading, they still have the following issues:

1) Most of the above works construct a one-way feature transmission from lesion features to DR grades in a two-stage manner, i.e., the lesion-related modules and DR grading modules are trained separately without end-to-end learning.
The lesion and grade features cannot be jointly fine-tuned for the final tasks. In this case, the lesion detectors need to be trained with large amounts of annotated data to provide a precise input for the following grading step; 2) Although [14] can optimize lesion segmentation and DR grading modules in an end-to-end manner, they crave for pixel-level annotations to generate the lesion mask. Obviously, pixel-level annotations are extremely labor-consuming and expensive, especially medical annotations that require the dedication of domain experts.

In this paper, we propose a robust end-to-end framework to collaboratively learn from both patch-level lesion and image-level grade annotations (CLPI) for DR grading. The proposed framework mainly consists of a lesion attention generator and a grading module. By training the lesion attention generator with only a few patch-level annotations, the following grading module can achieve the patch-level attention of the input image in a semi-supervised manner. The grading module is to directly predict the DR severity grade based on the lesion attention and the input image. Additionally, our lesion attention generator can be pre-trained with patches which can avoid the missing label problem. As seen in Fig. 1, the missing labels are commonly existing in the annotation of medical images, which may confuse the detectors trained with the entire images [15]. The source code can be found at: https://github.com/clpicode.

The main contributions of this paper can be highlighted as three-folds:

1) We propose a robust collaborative learning framework to integrate the patch-level lesion and image-level grade annotations for DR grading. Experimental comparisons prove CLPI shows outstanding performance against relevant state-of-the-art (SOTA) algorithms. Moreover, the proposed algorithm also achieves comparable performance with three over 9-years clinical experienced ophthalmologists. By training and testing on the datasets from totally different distributions, the proposed CLPI presents robust performance compared to the alternative popular CNN classifiers.

2) We design a novel network architecture, i.e., lesion attention generator, which can generate the patch-level lesion attention map of an entire image with only one forward pass. Experiments prove that our lesion attention generator can effectively improve the performance of DR grading methods. This architecture can be trained with image patches instead of the entire images, which can alleviate missing label problems.

3) Extensive ablation studies have experimentally proved the contributions of lesion features for DR severity grade, as well as the necessity of building a bidirectional way to exchange information in an end-to-end manner between lesion module and grade module.

II. RELATED WORK

Automatic DR grading has attracted tremendous research interests because of its enormous demand. The early attempts for DR grading from fundus images were usually done with two steps: handcrafted feature extraction and classification [16], [17]. Acharya et al. [18] used image-processing techniques to extract features from blood vessels and some key lesions and then classified the fundus images into five grades. Recently, CNN-based algorithms have shown superior performance in computer vision and brought powerful tools for DR assessment [19], [20], [21], [22]. Unlike handcrafted features, these deep networks can automatically learn discriminative features from large scale data.

Since DR severity grades have high correlation with lesions existing in the fundus, some researchers also try to detect some key DR-related lesions [23], [24]. Pixel-level lesion segmentation is a popular way to accurately achieve both the locations and contours of the lesions. Eftekhari et al. [25] aim to segment MAs out of fundus images via a two-step CNN. Chudzik et al. [26] utilize fully convolutional neural networks (FCNNs) for exudate segmentation. Yan et al. [27] propose a mutually local-global algorithm for lesion segmentation based on U-Net [28]. Due to the shortage of pixel-level annotated data, the results of the lesion segmentation approaches are far from promising in practice.

Another way to locate abnormal regions is patch-level lesion detection. Patch-level annotations are relatively easy to obtain compared with pixel-level annotations. Silberman et al. [29] extract SIFT (scale invariant feature transform) features in the image patches and utilized SVM (support vector machine) to distinguish patches with exudates. Haloi et al. [2], [30] achieve promising performance in MA and exudate detection based on sliding windows and CNN classifiers. Van Grinsven et al. [31] propose a selective sampling method for fast hemorrhage detection. Srivastava et al. [32] achieve robust performance in finding MAs and hemorrhages based on multiple kernel learning methods.

Most works take DR grading and lesion detection separately, and only a few approaches integrate both lesion and grade information for DR assessment [11], [12], [13], [14]. Inspired by some recent attention-based methods which integrate local and global features [33], [33], [34], we convert lesion information into an attention map and collaboratively learn the patch-level and image-level features.

Table I lists some recent work for DR grading on some key concepts include 1) whether lesion features are applied for DR grading, 2) whether end-to-end learning is used for collaboratively integrating the lesion and grading features, and 3) the annotation level (patch or pixel) of training samples for achieving lesion features.

| Lesion features for DR grading | End-to-end learning | Annotation level |
|-------------------------------|---------------------|-----------------|
| [3], [5], [10]                | ×                   | -               |
| Yang et al.[12]               | ×                   | patch           |
| Wang et al.[8]                | ×                   | -               |
| Lin et al.[13]                | ✓                   | patch           |
| Zhou et al.[14]               | ✓                   | pixel           |
| CLPI                          | ✓                   | patch           |

1The complete url: https://github.com/PaddlePaddle/Research/tree/master/CV/CLPI-Collaborative-Learning-for-Diabetic-Retinopathy-Grading
According to Table I, CLPI is the first end-to-end algorithm which integrates patch-level lesion and image-level grade annotations for DR grading.

III. METHODS

In this section, we present the details of the proposed CLPI. As illustrated in Fig. 2, the proposed algorithm can be dissected into the following parts: 1) A lesion attention generator to explore lesion features of an input image; 2) A grading module to classify the DR severity; 3) End-to-end learning details of feature integration for final decision.

A. Lesion Attention Generator

The lesion attention generator aims to explore lesion-related attention maps w.r.t. the input images. Firstly, an input image is split into patches, then a probabilistic vector is generated for each patch according to the lesions existed in the patch. In this paper, the dimension of the vector is 4 corresponding to the normal patches and 3 target lesions including MA, hemorrhage, and hard / soft exudate. Here, normal patches means that the patch without the 3 target lesions. Ideally, the entry in the probabilistic vector with the bigger value indicates that the corresponding lesion will have higher probabilistic exists in the input patch. Finally, the attention map of the entire image is constructed by extend the probabilistic vectors.

A direct way to achieve lesion information in an image is using some SOTA detectors (e.g. Faster-RCNN [35]) for lesion detection. Another way is taking the lesion detection as a patch-based classification problem as Yang et al. [12], and the patches are obtained by sliding windows. However, both the above approaches should be trained offline and cannot be embed in an end-to-end learning framework to fine-tune the lesion attention for adapting the final DR grading.

In our lesion attention generator, the input image splitting and attention map generation are performed within one forward pass, enabling end-to-end training after connecting the lesion attention generator and DR grading (classification) net. The lesion attention generator consists of two parts: the lesion detection backbone and attention map generation. The details are as following:

Lesion Attention Generator Backbone: To detect patch-based lesion information within one forward pass, we design the architecture of the detection backbone as Table II. The activation function between two convolution layers is the recently proposed Mish activation [36], and the batch normalization (BN) [37] is applied before each Mish. A softmax layer follows the last convolution layer which is named as lesion prediction layer.

The motivation of this architecture is to control the receptive field so that each unit in the lesion prediction layer only covers a 68 × 68 region in an input image of size 1024 × 1024 (see Fig. 3). As illustrated in Table II, if the width and height of an input image are 1024 × 1024, the feature map of lesion prediction layer is \( P \in \mathbb{R}^{16 \times 16 \times n} \), wherein each channel \( P_i \in \mathbb{R}^{16 \times 16} \) \((i = 0, 1, ..., n - 1)\) denotes the probabilistic matrix w.r.t. the \(i\)-th class. In this paper, \(n\) is equal to 4 , indicating 3 target lesions plus no target lesion (class 0). By this network design, a single forward pass can obtain the four-dimension probabilistic vectors of 16 × 16 patches in spatial order. This is equivalent to splitting the input image via a 68 × 68 sliding window with the stride 64 then performing forward pass individually.

Moreover, the lesion attention generator can be trained via image patches instead of leveraging an entire image with full lesion annotations. According to the network architecture in Table II, a 1024 × 1024 image will generate a 16 × 16 lesion predictions. Wherein, each prediction corresponding to a 68 × 68 patch in the input image. However, if we only input a 68 × 68 patch, the lesion network backbone will generate a 1 × 1 prediction. This leads to an exciting advantage that our lesion attention generator trained with patches is equivalent to one trained with entire images, which can avoid the confusion brought by the underlying missing labels (see the the experimental evaluation in Section V).

Attention Map Generation: The attention map \( AM \in \mathbb{R}^{W \times H \times n} \) is constructed by expanding the lesion probabilistic matrix \( P \), where \( W \) and \( H \) denote the width and height of the following severity grading net, respectively. In this paper, both \( W \) and \( H \) are set as 512. Each entry of the probability matrix \( P_i \) is expanded to a \( \frac{W}{16} \times \frac{H}{16} \) sub-matrix by duplicating, then each \( P_i \) will generate 16 × 16 sub-matrices. AM is constructed by tiling these sub-matrices together in spatial order. As shown in Fig. 4, the highlights convey the lesion information to construct the imbalanced attention map.

B. Grading Module

The grading module, which contains the neck and head of the CLPI framework, is responsible for grading the severity of DR. In this paper, we set up two types of grading tasks: one is five-grade classification according to the International Clinical Diabetic Retinopathy scale [7], and the other is a binary classification task which wraps the above five grades in non-referable versus referable.

As illustrated in Fig. 2, there exist four key parts in grading module: 1) A classification backbone without fully connected layer (Backbone 2); 2) A shortcut connection to transmit abnormal lesion map close to grading head; 3) A 1 × 1 convolutional layer to integrate the abnormal lesion map and the features from the classification backbone; 4) The global

---

1 If we control the receptive fields of prediction units as 64 × 64 non-overlapped regions, some lesions around the region boundaries may be missed by the detectors. Therefore, the receptive fields are designed with slightly overlapped in the lesion detection backbone.
average pooling (GAP) and softmax layers for predicting the DR grades.

In this paper, we use DenseNet-121 [38] as classification backbone by removing the fully connected layers. The abnormal lesion map transmitted by the shortcut is $1 - P_0$, wherein $P_0$ is the probability matrix of class 0 (no target lesion) from the lesion prediction layer. As shown in Fig. 2, we stack the abnormal attention map with the feature maps of the classification backbone to collaboratively learn both detection and classification nets.

The motivations of the shortcut between the lesion prediction layer and the output of the classification backbone are: (1) to directly back-propagate gradient to the lesion attention generator to avoid the potential gradient-vanishing; (2) to provide the semi-supervised lesion-related information closer to the prediction layer to improve the DR grading performance.

C. End-to-end Collaborative Learning of CLPI Framework

As shown in Fig. 2, a weighted attention map $I_{AM} \in \mathbb{R}^{W \times H \times n \times c}$ is constructed to feed the grading module. Let $I_{re} \in \mathbb{R}^{W \times H \times c}$ denotes the resized input image ($c$ is the channel of the input images). The weighted attention map can be calculated as

\[ I_k^{AM} = (AM_i + 1) \odot I_{re}^j, \quad s.t., k = 0, 1, ..., n \times c - 1, \]
\[ i = 0, 1, ..., n - 1, \]
\[ j = 0, 1, ..., c - 1, \]

(1)

where $I_k^{AM}$ denotes the $k$-th channel of $I^{AM}$. $AM_i$ is the $i$-th channel of the attention map, and $I_{re}^j$ denotes the $j$-th channel of $I^{re}$. $\odot$ means the element-wise product. 1 is a constant matrix with the same size of $AM_i$. It is used to prevent information dropout when the probabilities in $I_k^{AM}$ are small.

Since the entries in the attention map $AM_i$ implicitly carry the probabilities of having the $i$-th lesion in the patches, an image patch has a specific lesion with a higher probability will get a higher weight in the element-wise production. As seen in Fig. 4, imbalanced attentions are covered on the input image by highlighting the lesion patches.

For collaboratively training the CLPI framework, first, we pre-train the lesion attention generator with patches. Then the entire CLPI framework is put into end-to-end training by using only image-level DR grade labels. The classification backbone
of the grading module can also be pre-trained for speeding up the convergence.

The loss function for both pre-training of lesion attention generator and the end-to-end training of the entire framework is cross-entropy loss:

$$CE(x_i, y_i, S) = - \sum_{c=1}^{M} 1(y_i = c) \log(S(c|x_i)))$$

(2)

where \(x_i\) and \(y_i\) denote the \(i\)-th input image/patch and the corresponding label respectively, and \(S\) is the softmax output of the CNNs classifiers. \(M\) is the class number, i.e., \(M = 4\) for pre-train lesion attention generator, \(M = 2\) and \(5\) for binary and multiple DR grading respectively. \(1(.)\) is an indicator function that is equal to 1 when \(y_i = c\), and \(S(c|x_i))\) denotes the output probability of the unit w.r.t. class \(c\).

IV. EXPERIMENTAL RESULTS

A. Datasets

Lesion Dataset: We use the public available IDRiD Dataset [39] that provides 81 fundus images (54 for training and 27 for testing) with pixel-level annotations of lesions including MAs, haemorrhages and exudates. Since we only need patch-level testing) with pixel-level annotations of lesions including MAs, [39] that provides 3M of the CNNs classifiers.

S\times x\) is cross-entropy loss generator and the end-to-end training of the entire framework the convergence.

We use the public available Lesion Dataset:

A. Datasets

- **Messidor-2 Dataset** [40] contains 1,200 fundus images from three French hospitals. However, their severity grade only has four levels, which is slightly different from the five-level international standard [7].
- **Messidor-1 Dataset** is an extension of the original Messidor-1 dataset, which contains 1748 eye fundus images, and each image is classified into one of the five DR grades according to [7].
- **LIQ-EyePACs** is a subset of the EyePACS dataset [41] for evaluating the robustness of the grading methods, which contains some low image quality samples. Since there are plenty number of label biases in the original EyePACS dataset according to our cooperative ophthalmologists, we turn to **LIQ-EyePACs** under the following guidelines: 1) The labels of **LIQ-EyePACs** are rechecked of our cooperative ophthalmologists. 2) The quality of images is relatively low which contains noises like under/over-exposure and out-of-focus problem. These noises are commonly encountered in real-world practice. The detailed label distribution is in Table IV.
- **Private Datasets**: Our fundus images are collected from over 20 hospitals. With the help of more than 30 licensed ophthalmologists, the **private** DR grade dataset has 36,270 samples that contain one of five DR severity grade labels. To improve the overall annotation efficiency, each fundus image is first classified into the referable or non-referable DR by at least three licensed ophthalmologists, the binary groundtruths of the images are given by the majority voting. Based on the binary DR grade, three well-trained annotators or one licensed ophthalmologist will label the images with one of the five fine-grained DR grade according to [7].

The patch-level annotations of IDRiD and the name list of **LIQ-EyePACs** will be provided upon request.

| **Distribution of LIQ-EyePACs and Private DR grade dataset.** |
|-----------------|----------------|-------------|------------|----------|
| **None** | **Mild** | **Moderate** | **Severe** | **PDR** |
| **LIQ-EyePACs** | 7,286 | 675 | 1,507 | 247 | 285 |
| **Private** | 19,826 | 3,220 | 9,760 | 2,069 | 1,395 |

B. Implementation details

Data Preprocessing and Augmentation: We subtract all the images by the local average color to highlight effective details on the fundus image. We also apply some commonly used data augmentation strategies in the training process, including randomly crop with scale=(0.9, 1.1), rotation=(0, 180), horizontal and vertical flip (p=0.5), and random rotation with the rotate degree range in (0, 180).

Training Tips: The lesion attention generator is pre-trained with the patches from IDRid dataset.

To train the final framework in an end-to-end manner, the gradients are only back-propagated from the shortcut to fine-tune the lesion attention generator. The grade information from the shortcut can be more directly transmitted through the shortcut than from the deep grading net.

Evaluation Metrics: For five-grades classification, we use Cohen Kappa values which is a commonly used metric to measure the agreement between the predictions and the reference grades. Kappa values vary between 0 (random agreement between raters) and 1 (complete agreement between raters) [42]. For the binary classification, we evaluate with AUC (area under the ROC curve) metric.

C. The Effectiveness and Robustness of CLPI in DR Grading

Comparison with SOTA algorithms: Table V lists the comparison between CLPI and some recent SOTA papers, including VNXK [43], CKML [43] Zoom-in-Net [8], AFN [13], and Semi+Adv [14]. Wherein, VNXK and CKML are similar methods with different kernel strategies. Zoom-in-Net explores the suspicious regions in fundus image in an unsupervised way for DR grading. AFN transmits lesion attention for DR grading in an offline way. Semi+Adv utilizes pixel-level lesion information to improve DR grading performance. All the algorithms are under the same evaluation protocols as presented in Zoom-in-Net and Semi+Adv. We can see that our CLPI achieves outstanding performance by comparing to the SOTA algorithms in most cases. Moreover, trained with the same lesion and grade datasets in IDRiD and Messidor-1, the CLPI achieves comparable performance with [14] that utilizes pixel-level lesion annotations.

Comparison with Senior Ophthalmologists:
In this section, we select the common annotations (989 fundus images from our *Private* dataset) of three senior ophthalmologists, whose clinic experience over 23, 13 and 9 years respectively, as DR testing set. All the three ophthalmologists label the testing images as referable or non-referable DR, and the groundtruths are achieved by majority voting. The sensitivities and specificities of the three ophthalmologists are shown in Fig. 5 as well as the ROC (receiver operating characteristic) curve of CLPI. We can see that CLPI achieves comparable performance with the senior ophthalmologists in detecting referable DR.

Similarly, *Messidor-2* is split into training, validation and testing sets by 6:2:2, and the last five rows in Table VI are the results of the models that only trained with *Messidor-2* datasets.

We can see that the performance of all the methods decrease in the testing sets from different distributions. Meanwhile, the CLPI outperforms the alternative algorithms by a larger margin in these cases, which proves the robustness of CLPI in facing different practical circumstances.

In *Messidor-1* dataset, each image is annotated into one of four DR severity grade, which is different from the annotation standard compare to the datasets listed in Table VI. Therefore, the results on *Messidor-1* are not listed in table.

**D. Ablation Studies**

In this section, we experimentally prove some key concerns of CLPI from the following aspects: 1) Our lesion attention generator is effective in exploring valuable lesion information for improving the DR grading performance. 2) It is necessary to optimize CLPI in an end-to-end manner, which can collaboratively build bidirectional information exchange between image-level grade and five-grained lesion features, to achieve more promising results.

**The Effectiveness of lesion attentions generator.** To prove that our lesion attention generator is helpful for DR grading, we train some popular classification architectures directly with our weighted attention map $I_A$ (see Fig. 2) instead of the fundus images. Wherein, the $I_A$ are achieved off-line by our lesion attention generator. Fig. 6 illustrates the results of the classification architectures trained with private dataset and tested on *Messidor-2* dataset. All the methods are pre-trained with the patches from *IDRiD* dataset except CLPI-. CLPI- is the proposed framework without patch-level pre-training, named as CLPI-.

| Table V | COMPARISON OF CLPI WITH THE SOTA ALGORITHMS IN *Messidor-1* DATASET. |
|---------|-------------------------------------------------|
|          | VNXK [43] | CKML [43] | Zoom-in-Net [8] | AFN [13] | Semi-Adv [14] | CLPI |
| AUC of Non-referable / Referable DR | 0.887 | 0.891 | 0.957 | 0.968 | 0.976 | **0.985** |
| AUC of No-DR / DR | 0.870 | 0.862 | 0.921 | - | 0.943 | **0.959** |

**The robustness of CLPI Against Commonly Existed Challenges In Real-world Practice:**

In real-world practice, compare to the training set, the testing images may be generated from totally different camera brands and label distributions. Additionally, the quality of the testing images is usual various in terms of the photographers and shooting environments. These issues bring inevitable challenges to the robustness of all the DR grading methods in reality.

To evaluate the robustness of different approaches, Table VI lists the results of the models trained with one dataset while tested on another dataset with a totally different distribution and collected environment. *LIQ-EyePACs* is the dataset with low quality images that contains various noises including illumination, blur, artifacts problems. All the some popular CNNs architectures in Table VI, including DenseNet121[38], ResNet50 [44] and Inception-V4 [45]), are pre-trained with the same lesion patches samples as CLPI for a fair comparison.

In the first several rows in Table VI, our private dataset is randomly split into training, validation and testing sets by approximate 6:2:2. All the models are only trained with our private training dataset and tested on the other datasets.
CLPI and CLPI- proves the merits of exploring lesion features for DR grading.

The necessity of the end-to-end collaborative learning scheme. In this part, we set up ablation studies from both quantitative and qualitative aspects.

Table VII records the quantitative ablation studies. Wherein, **CLPI without end-to-end learning** means that training the CLPI by fixing the pre-trained lesion attention generator. In this way, the information is only one-way transmit from the lesion attention generator to the grading module, therefore the lesion attention generator cannot be optimized by the image-level grade annotations. **CLPI without shortcut** denotes that training CLPI framework without the lesion shortcut in Fig. 2, i.e., the gradient are back-propagated through the deep classification backbone to the lesion attention generator. As seen in Table VII, the outperformance of CLPI over CLPI without end-to-end learning quantitatively reveals that end-to-end learning is necessary to fully integrate the lesion and grade information. In additional, the comparison between CLPI and CLPI without shortcut indicates the motivation of introducing the shortcut is effective and reasonable.

Fig. 7 illustrates qualitative comparisons in an interpretable way. Wherein, the first column shows the groundtruth of our lesion annotation in terms of MA, hemorrhage and exudate. The second column records the detection result of our lesion attention generator, and the highlighted boundingboxes are predicted to contain at least one of the three lesions. The third and fourth columns are the class activation map (CAM) [46] heatmaps of CLPI- (without lesion pre-training) and CLPI respectively, and the redder regions indicate the higher abnormal probabilities. CAM is an interpretable way to visualize the class-related heatmaps in the CNNs-based classifiers. All the CAMs are extracted from the last convolution layer of the networks, and readers can refer [46] for more details.

Additionally, we can see most lesion regions are caught by both the CAMs of grading net and CLPI in Fig. 7, which reveals that the lesions have high-relevance with the final decision of the CNNs-based grading networks. This study proves the correlation between the lesion and grade in an experimental aspect, which confirms the reasonability of exploiting lesion information for automatically DR grading.

To sum up both the quantitative and qualitative ablation study: by end-to-end optimizing the final framework to collaboratively build bidirectional information exchange, CLPI achieves more discriminative features to improve the performance of DR grading.

V. DISCUSSION

Evaluated by the above experiments and ablation studies, the effectiveness and robustness of CLPI has been proved. By introducing a few patch-level lesion annotations, the proposed framework can collaboratively integrate both lesion and image-level grade information to achieve promising results. The proposed lesion attention generator and shortcut connection are carefully designed to facilitate the end-to-end training of the framework.

Additionally, when our lesion attention generator is served as a lesion detector, we find that it is more robust to missing labels compare to some SOTA detectors including Fast-RCNN [35], YOLO-V3 [47], SSD [48] and Faster-RCNN + ResNeXt10 backbone + deformable convolution network [49]. In Fig. 8, we randomly discard some lesion annotations in the training dataset, and record the performance reduction rates.

### Table VI

| Training sets | Testing sets | Private | Messidor-2 | LIQ-EyePACS |
|--------------|--------------|---------|-----------|-------------|
|              | Methods      | Five-grade (Kappa) | Binary (AUC) | Five-grade (Kappa) | Binary (AUC) | Five-grade (Kappa) | Binary (AUC) |
| Private      | ResNet-50    | 0.885 0.980         | 0.655 0.874     | 0.629 0.844       |
|              | DenseNet-121 | 0.886 0.979         | 0.645 0.880     | 0.614 0.848       |
|              | Inception-V4 | 0.809 0.980         | 0.667 0.887     | 0.625 0.849       |
|              | CLPI         | 0.908 0.983         | 0.703 0.946     | 0.692 0.916       |
| Messidor-2   | ResNet-50    | 0.662 0.916         | 0.793 0.945     | 0.492 0.840       |
|              | DenseNet-121 | 0.565 0.923         | 0.794 0.945     | 0.396 0.826       |
|              | Inception-V4 | 0.497 0.811         | 0.803 0.954     | 0.312 0.685       |
|              | CLPI         | 0.838 0.969         | 0.832 0.975     | 0.514 0.877       |

Fig. 7. The qualitative comparison of abnormal region locating. By collaboratively integrating lesion and grade annotations, CLPI can focus on the lesion regions more accurately.
(\textit{performance after discard} / \textit{performance with full annotations}). In this case, the discarded annotations will turn to missing-annotated samples for the detectors trained with entire images, but the negative patches for training our lesion detection net can be collected from the No-DR images to avoid the missing labels. Therefore, the SOTA detectors are sensitive to the discard of the annotations while the proposed detection architectures show relatively robust performance.

Although the proposed lesion attention generator can effectively exploit lesion information and the architecture is the robust to missing labels, it can only be served as attention generator rather than a lesion detector. Because the proposals generation scheme of the architecture do not take scale into consideration, and do not have a regression module to get precise bounding box. Besides, we only cover three key lesions among a dozen of DR-related lesions. However, with limited lesion annotations, our lesion attention generator is elegant and qualified to achieve valuable attention for improving the performance of DR grading.

CLPI and Semi + Adv [14] have similar motivation to collaboratively integrate the lesion and grade information. However, the two methods are totally different in architecture designing. As a result, CLPI achieves comparable performance compared to [14], while only need patch-level lesion annotations instead of pixel-level masks used by [14]. It is worth noting that the patch-level annotations are much easier to be achieved as well as time and labor saving compare to pixel-level annotations.

Moreover, the proposed framework can also be extended to other classification tasks that the objects in the image are related to the image category, such as scene classification and other disease grading tasks based on medical images. We have not evaluated our framework on these tasks in this paper, and it’s our future work to make CLPI more general.

### VI. Conclusion

In this paper, we have proposed a collaboratively learning framework to integrate patch-level lesion and image-level grade features for robust DR grading. On one hand, the lesion attention generator provides valuable semi-supervised lesion attention for DR grading. On the other hand, the grade supervision is back-propagated to optimize the attention map. Trained with patches, our lesion attention generator can detect lesions over an entire image in only one forward pass which facilitates the end-to-end learning of the entire framework. Extensive experimental comparisons have proved the proposed CLPI achieves comparable performance with SOTA algorithms as well as senior ophthalmologists. The robustness of our method was also proved by evaluating the DR grading methods under various challenges in real-world scenario. Ablation studies have shown the effectiveness of the lesion attention scheme as well as the advantages of the end-to-end collaborative learning of CLPI. There still exist lots of meaningful open issues, such as precisely detect more types of lesions, and extent CLPI framework to other applications expect DR grading.

### REFERENCES

[1] R. Gargeya and T. Leng, “Automated identification of diabetic retinopathy using deep learning,” pp. 962–969, 2017.
[2] M. Haloi, “Improved microaneurysm detection using deep neural networks,” arXiv preprint arXiv: 1505.04424v2, 2015.
[3] J. Shaw, R. Sicree, and P. Zimmet, “Global estimates of the prevalence of diabetes for 2010 and 2030,” Diabetes research and clinical practice, vol. 87, pp. 4–14, 11 2009.
[4] H. Pratt, F. Coenen, D. Broadbent, S. Harding, and Y. Zheng, “Convolutional neural networks for diabetic retinopathy,” Procedia Computer Science, vol. 90, pp. 200–205, 12 2016.
[5] V. Gulshan, L. Peng, M. Coram, M. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. Nelson, J. Mega, and D. Webster, “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.
[6] I. Kocur and S. Resnikoff, “Visual impairment and blindness in Europe and their prevention,” British Journal of Ophthalmology, vol. 86, no. 7, pp. 716–722, 2002.
[7] A. A. of Ophthalmology, “International clinical diabetic retinopathy disease severity scale,” in url http://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf, 2012.
[8] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, “Zoom-in-net: Deep mining lesions for diabetic retinopathy detection,” in MICCAI, 2017.
[9] M. Sankar, B. Krishnan, and R. Parvathi, “Earliest diabetic retinopathy classification using deep convolution neural networks.pdf,” International Journal of Advanced Engineering Technology, 2016.
[10] M. Alban and T. Gilligan, “Automated detection of diabetic retinopathy using fluorescein angiography photographs,” Report of standford education, 2016.
[11] B. Antal and A. Hajdu, “An ensemble-based system for microaneurysm detection and diabetic retinopathy grading,” IEEE Transactions on Biomedical Engineering, vol. 59, pp. 1720–6, 04 2012.
12] Y. Yang, T. Li, W. Li, and et al., “Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks,” in MICCAI, 2017.

13] Z. Lin, R. Gao, Y. Wang, and et al., “A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion,” in MICCAI, 2018.

14] Y. Zhou, X. He, L. Huang, and et al., “Collaborative learning of semi-supervised segmentation and classification for medical images,” in CVPR, 2019.

15] M. Xu, Y. Bai, B. Ghanem et al., “Missing labels in object detection,” in CVPR, 2019.

16] A. Pinz, S. Bengsger, P. Datlinger, and A. Kruger, “Mapping the human retina,” IEEE Transactions on Medical Imaging, vol. 17, pp. 606–19, 09 1998.

17] A. Sopharak, B. Uyyanonvara, and S. Barman, “Automatic exudate detection from non-dilated diabetic retinopathy retinal images using fuzzy c-means clustering,” Sensors, vol. 9, pp. 2148–61, 03 2009.

18] U. R. Acharya, C. Lim, E. Ng, C. Chee, and T. Tamura, “Computer-based detection of diabetes retinopathy stages using digital fundus images,” Proceedings of the Institution of Mechanical Engineers, Part H, Journal of engineering in medicine, vol. 223, pp. 545–53, 08 2009.

19] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in NeuralIPs, 2012.

20] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Hasakova, and M. Prounott, “Deep learning algorithm predicts diabetic retinopathy progression in individual patients,” NPJ digital medicine, vol. 2, no. 1, pp. 1–9, 2019.

21] K. B. Nielsen, M. L. Lautrup, J. K. Andersen, T. R. Savarimuthu, and J. Grauslund, “Deep learning-based algorithms in screening of diabetic retinopathy: A systematic review of diagnostic performance,” Ophthalmology Retina, 2018.

22] X. Li, X. Hu, L. Yu, L. Zhu, C. Fu, and P. Heng, “Cacnet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading,” IEEE Transactions on Medical Imaging, vol. 39, no. 5, pp. 1483–1493, 2020.

23] K. M. Adal, P. G. van Etten, J. P. Martinez, K. W. Rouwen, K. A. Vermeser, and L. J. van Vliet, “A automated system for the detection and classification of retinal changes due to red lesions in longitudinal fundus images,” IEEE Transactions on Biomedical Engineering, vol. 65, no. 6, pp. 1382–1390, 2018.

24] C. Playout, R. Duval, and F. Cheriet, “A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images,” IEEE Transactions on Medical Imaging, vol. 38, no. 10, pp. 2434–2444, 2019.

25] N. Eftekhar, H.-R. Pourreza, M. Masoudi, K. Ghiasi-Shirazi, and E. Saeedi, “Microaneurysm detection in fundus images using a two-step convolutional neural network,” Biomedical engineering online, vol. 18, no. 1, p. 67, 2019.

26] P. Chudzik, S. Majumdar, F. Caliva, B. Al-Diri, and A. Hunter, “Exudate segmentation using fully convolutional neural networks and inception modules,” in SPIE, Medical Imaging, 03 2018, p. 105.

27] Z. Yan, X. Han, C. Wang, Y. Qiu, Z. Xiong, and S. Cui, “Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images,” in arXiv preprint arXiv:1901.06047, 2019.

28] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in MICCAI, 2015.

29] N. Silberman, K. Aehrich, R. Fergus, and L. Subramanian, “Case for automated detection of diabetic retinopathy,” in AAAI, 01 2010.

30] M. Haloi, S. Dandapat, and R. Sinha, “A gaussian scale space approach for exudates detection, classification and severity prediction,” arXiv preprint arXiv:1505.00737, 05 2015.

31] M. J. Van Grinsven, B. van Ginneken, C. B. Hoyng, and T. Theelen, “Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images,” IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1273–1284, 2016.

32] R. Srivastava, L. Duan, D. Wong, J. Liu, and T.-Y. Wong, “Detecting retinal microaneurysms and hemorrhages with robustness to the presence of blood vessels,” Computer Methods and Programs in Biomedicine, vol. 138, 10 2016.

33] S. Chen, B. Wang, X. Tan, and X. Hu, “Embedding attention and residual network for accurate salient object detection,” IEEE Transactions on Cybernetics, vol. 50, no. 5, pp. 2050–2062, 2020.

34] H. Fu, Y. Xu, S. Lin, D. W. K. Wong, M. Baskaran, M. Mahesh, T. Aung, and J. Liu, “Angle-closure detection in anterior segment oct based on multilevel deep network,” IEEE Transactions on Cybernetics, vol. 50, no. 7, pp. 3358–3366, 2020.