ELLG: Explainable Lesion Learning and Generation for Diabetic Retinopathy Detection

Chenhao Lin[1] , **Jiongli Zhu**[1] , **Chao Shen**[1] , **Pengwei Hu**[2] , **Qian Wang**[3]
[1]School of Cyber Science and Engineering, Xi'an Jiaotong University
[2]IBM Research, China
[3]School of Cyber Science and Engineering, Wuhan University
{linchenhao@, wearyee@stu, chaoshen@mail}.xjtu.edu.cn, hupwei@cn.ibm.com,
qianwang@whu.edu.cn

Abstract

The deep learning based approaches have achieved remarkable success in diabetic retinopathy detection. Due to the accountability in medical diagnosis, the interpretability of computer-aided diagnosis approaches has been investigated recently. However, few of existing approaches make full use of the explainable evidences to improve the diagnosis accuracy. In this paper, we propose an **E**xplainable **L**esion **L**earning and **G**eneration (ELLG) framework to study the interpretability of diabetic retinopathy detection and achieve more accurate diagnosis. We first generate visual explanations for diabetic retinopathy diagnosis using proposed **G**ated **M**ulti-layer **S**aliency **M**ap (GMSM). Then we iteratively generate fundus images with lesions and include them for training to learn more robust lesion features. Our method not only provides more accurate explainable evidences but also addresses the data imbalance problem in diabetic retinopathy detection, therefore results in highly improved detection performance without increasing time-complexity during the inference. The experimental results on two databases demonstrate the efficiency of the proposed approach.

1 Introduction

Automatic digital imaging diagnosis of diabetic retinopathy (DR) has been investigated for many years. Recently, the successful use of deep learning in image classification and detection tasks has inspired researchers to apply Convolutional Neural Networks (CNNs) based approaches on DR detection [Voets *et al.*, 2018; Wan *et al.*, 2018a; Jiang *et al.*, 2019; Costa *et al.*, 2018; Zhu *et al.*, 2019]. Pratt et.al. [Pratt *et al.*, 2016] proposed the five class classification of DR by using specific designed CNN for the first time. In [Wan *et al.*, 2018b], the authors successfully applied transfer learning approach by using several CNN structures including VggNet and ResNet for DR detection. A data-driven approach for DR detection was proposed in [Pires *et al.*, 2019], the authors gradually used data augmentation, multi-resolution training through CNN to improve the detection accuracy. Although these approaches proved their superiority in DR detection, compared with traditional machine learning algorithms, how to explain their decision mechanism is still an open problem.

The interpretability of the detection results is critical because they are highly related to the safety of patients, and it is essential to convince both patients and physicians that the diagnosis is reasonable and trustworthy through illustrating their explanations. Some lesion-level DR classification works have been introduced and illustrated visual explanations of the classification results. In [Yang *et al.*, 2017], a two-stage CNN approach proposed to illustrate lesions in fundus images through an attention mechanism. Similarly, [Wang *et al.*, 2017] also proposed an attention visual understanding of the diabetic retinopathy based on their zoom-in-net. Activation map based methods were also used to illustrate the visual explanations. Ref. [Jiang *et al.*, 2019] combined class activation maps (CAMs) [Zhou *et al.*, 2016] with adaboost to get less biased saliency maps to indicate the lesion positions. Ref. [Wang and Yang, 2018] also provided visual-interpretable feature by adding regression activation map to help localize the discriminative regions of the lesion and show its severity level. More recently, a GAN based method was proposed by [Niu *et al.*, 2019] to synthesize pathological retinal images. The pathological descriptors were extracted by using a DR detection network [Antony and Brüggemann, 2016] trained on a public database [Kaggle, 2016]. Although many works have provided visual explanations for DR detection results through different methods, few of them attempted to make full use of these interpretable detection results to further improve DR detection performance.

Another limitation of applying CNN based approaches on medical imaging diagnosis is the class imbalance problem. The amount of the digital images with lesions is usually not large enough to train robust CNN for accurate DR detection. To address this limitation, previous works [Cao *et al.*, 2018; Qummar *et al.*, 2019] usually applied image preprocessing approach including under and over sampling to balance the data. Such general approaches can alleviate the impact of data imbalance to some extent, while targeted method can be proposed to tackle this problem in DR detection.

In this paper, we propose a framework dubbed ELLG (explainable lesion learning and generation) to address the limitations in DR detection. We summarize our main contributions as follows: 1) Inspired by CAM, We first propose

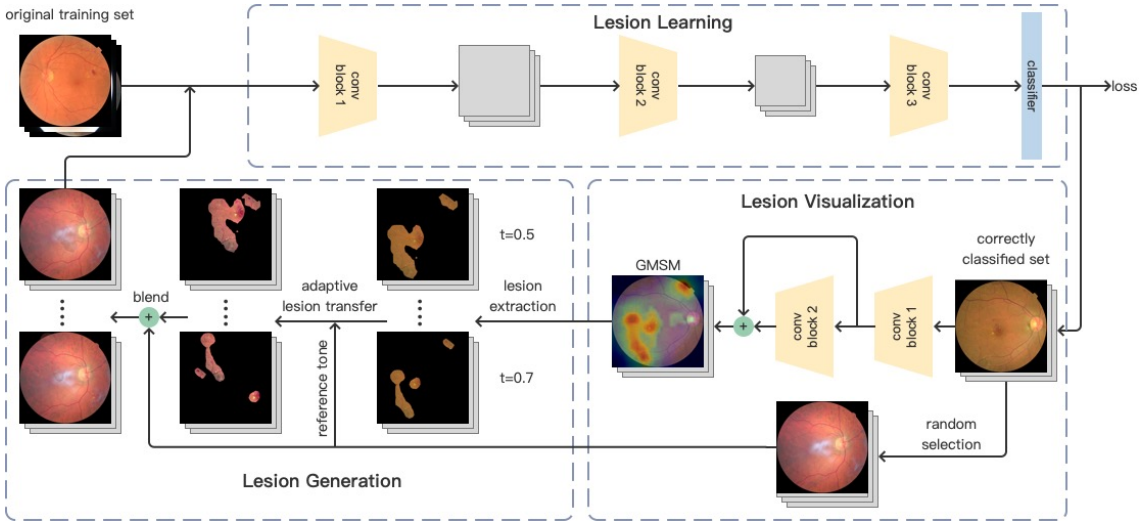


Figure 1: Illustration of the proposed ELLG framework.

a novel GMSM (gated multi-layer saliency map) method to locate the lesions of diabetic retinopathy. Our method illustrating more accurate lesion regions, can generate more reliable visualization of diagnosis. 2) We develop an approach to iteratively generate pathological images and gradually learn more robust lesion features. Our approach takes full advantage of the visual explanations to learn various new generated pathological images and address the common data imbalance problem as well. 3) The proposed method can achieve better detection accuracy and more accurate lesion localization than baseline methods on two public databases, without increasing time complexity during the inference.

2 Methodology

2.1 Overview of ELLG

The framework mainly consists of three parts as shown in Figure 1. Firstly, the Lesion Learning part applies the DR detection net [Antony and Brüggemann, 2016; Niu *et al.*, 2019] to learn the lesions and predict the severity of diabetic retinopathy in the meantime. Then the Lesion Visualization part takes correctly classified images from the DR detection net as the inputs and generates heatmaps for visual explanation by using the proposed GMSM. Based on the visualization results, the Lesion Generation part extracts and adaptively transfers lesions from heatmaps, and then blends them into randomly selected normal images to generate new pathological samples. Finally, the generated samples are iteratively feed into the whole network for more robust lesion learning. In general, the proposed framework iteratively generates new pathological samples and learns the lesion features based on the combination of original and blended images to enhance the model's understanding of lesions.

The DR detection net is widely used in many publications [Niu *et al.*, 2019; Wang and Yang, 2018] as the baseline method. It is a regression model that mainly consists of three convolutional blocks with output sizes of 27×27, 13×13, 6×6, and the subsequent classifier. Each conv block contains several convolution and max pooling layers. It takes 512×512 fundus images as the inputs and outputs one-dimensional diabetic retinopathy results that represent the corresponding severity level.

2.2 Gated Multi-layer Saliency Map

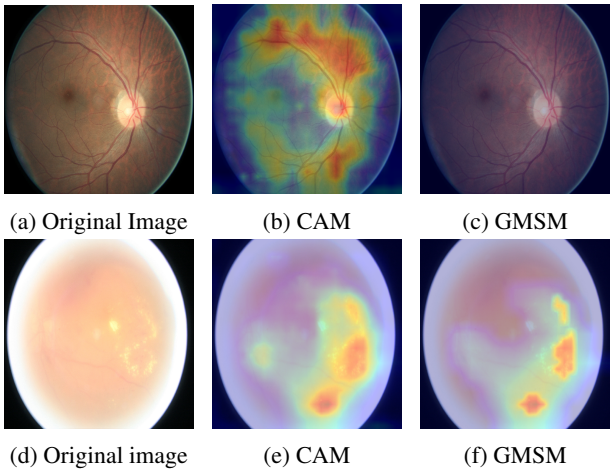
Inspired by CAM, we attempted to generate visual explanations based on the activation maps. Class activation mapping algorithms, including CAM, Grad-CAM and Grad-CAM++, are designed for classification models, in which only the result of a specific class is used. As for regression models like the DR detection net we used, instead of using a global average pooling layer proposed in CAM which may impact the detection accuracy, we preserved the original structure of the DR detection net and treated the output cell as the result of a specific class in a classification model without softmax layer. Therefore, the saliency map L can be calculated by:

$$L_{ij} = relu\left(\sum_k w_k \cdot A_{ij}\right) \tag{1}$$

where w_k represents the weights calculated by the gradients (more details in [Chattopadhay *et al.*, 2018]), and A represents the activation map. Considering that different fundus images have different structures, some normal regions may also have a slight fluctuation and then become non-zero in the saliency map. Especially, some tissues like optic papilla in the normal images will cause the fluctuation and be largely magnified after the normalization by using the original CAM, as shown in Figure 2b. To reduce the influence of these slight fluctuations and get more accurate saliency maps, we proposed a gate relu mechanism designed as follows:

$$L_{ij} = max\left(\sum_k w_k \cdot A_{ij}, t\right) \tag{2}$$

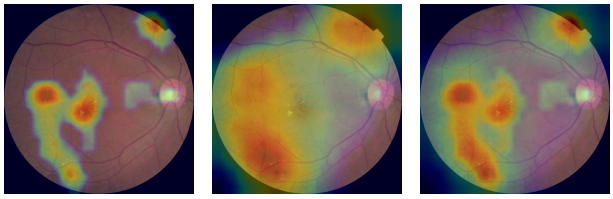
In addition, as shown in Figure 3 a b, because the lower layers with high resolution contained more location and detail



(a) Original Image (b) CAM (c) GMSM

(d) Original image (e) CAM (f) GMSM

Figure 2: Comparison between heatmaps generated from CAM and our GMSM. The first row uses a level 0 severe (normal) image, and the second row uses a level 2 severe image.



(a) (b) (c)

Figure 3: Heatmaps generated from (a) lower layer, (b) higher layer,(c) multi-fused layer.

information, the saliency maps of lower layers had higher accuracy in lesion location but lower confidence. It resulted in that the results from lower layer were overconfident in some positions and might miss some mild lesion regions, while the results from higher layer were more reliable but imprecise in localization. Therefore, we fused saliency maps from two different layers to acquire a more robust saliency map called gated multi-layer saliency map (GMSM):

$$L_{ij} = \frac{1}{2} \cdot \left(norm\left(L_{l_{ij}}\right) + norm\left(L_{h_{ij}}\right)\right) \tag{3}$$

L_l and L_h represent saliency map generated from lower-layer and higher-layer conv blocks calculated by Eq. (2) respectively, and both of them are normalized to the same range before averaging. As shown in Figure 3 c, by using the proposed GMSM, more accurate lesions can be located and illustrated.

2.3 Lesion Generation and Iterative Learning

Like other medical imaging diagnosis tasks, the DR detection also has the imbalanced data problem. Its database is highly imbalanced with 73.48% of level 0 severe, i.e. normal fundus images. In addition, from the pathological point of view, the location and pattern of lesions can be various and stochastic. Therefore, we proposed to generate random lesions and feed the network more unseen samples. This process not only addressed the imbalance problem but also made the network learn more robust lesion features of diabetic retinopathy.

Lesion Patch Extraction

To acquire lesion patches without manual annotation, we designed binary masks M to locate the regions with large saliency values. We first normalized the saliency map L to $[0, 1]$ and get L_{norm}, then we set thresholds τ which were randomly selected within a certain range. The binary masks M were set to 1 for those spatial positions with value larger than τ in L_{norm}:

$$M_{ij} = \begin{cases} 1, & L_{norm_{ij}} > \tau \\ 0, & otherwise \end{cases} \tag{4}$$

We then applied these masks over the original image and get corresponding lesion patches.

Adaptive Lesion Transfer

The proposed adaptive lesion transfer approach mainly includes two parts: random shifting and adaptive color transfer. To make the model learn the real features of lesions rather than focusing on absolute lesion positions which may bring overfitting problem, before blending them into another image, we keep randomly shifting a single lesion patch in the loops(100 in our setting) until there is no overlap with other patches or the loop is over.

Noticing that different fundus images have different tones, simply blending randomly shifted lesions patches into another fundus image may cause color difference problem. For example, even a normal patch of a dark-tone fundus image (like Figure 2 a) may be recognized as retinal hemorrhage if it is mixed up with a light-tone fundus image (like Figure 2 d), which can seriously impact the training process. To address this problem, we applied the color transfer algorithm in [Reinhard *et al.*, 2001], which converts the tone of the source image to the tone of the target image. The main idea of the transformation is to make the source image to have the same mean and variance as the target image in the $l\alpha\beta$ space.

To enhance the linear representation in-between training examples, following [Zhang *et al.*, 2017], mixup with a random ratio was applied:

$$I_{b_{ij}} = \begin{cases} r \cdot I_{s_{ij}} + (1 - r) \cdot I_{o_{ij}}, & I_{s_{ij}} \neq 0 \\ I_{o_{ij}}, & otherwise \end{cases} \tag{5}$$

where I_s and I_o denotes the adaptively transferred lesion image and another randomly selected fundus image (from the correctly classified set), and r is the blending ratio. Correspondingly, the label of the blended image was also averaged with the same ratio:

$$l_b = classify(r \cdot l_s + (1 - r) \cdot l_o) \tag{6}$$

where $classify()$ is the function that maps the predicted severity values within a certain range to their corresponding levels. After that, the generated fundus images were fed into the DR-detection net for iteratively training and more robust lesion features were expected to be learned.

3 Experiments

3.1 Experimental Settings

The conventional DR detection database [Kaggle, 2016] (*Database1*) containing 35k training fundus images has been used for the evaluation. Additionally, we have used a recent released database [Kaggle, 2019] (*Database2*) which has 3.6k training fundus images, as a supplementary to further validate the effectiveness of the proposed ELLG approach. Following previous work, 10% data from the training set has been split as the validation set. The kappa score and accuracy have been calculated for the evaluation. As in DR-Detection-Net, the fundus images have been resized to 512, 256 and 128 respectively for the three-part training. The method in [Antony and Brüggemann, 2016; Niu *et al.*, 2019] has been used as the baseline. To present the effectiveness of the proposed method more intuitively, we did not apply any special ensemble or fusion method.

For the database1, the network was trained on original data for 250 epochs and then finetuned on original and combined data alternately for 4 rounds with 30 epochs per round. Batch size is 48. In each round, the learning rate starts at 3e-5, then decays to 3e-6 and 3e-7 after each 10 epochs. For the database2, the learning rate is 8e-5, and the batch size is 64. The models are trained for 30 epochs. The hyper parameter t in Eq. (2) is set to 1.00, and τ in Eq. (4) is limited in [0.5,0.8].

3.2 Experimental Results

For *Database1*, as shown in Table 1, our proposed ELLG has achieved better performance than previous SOTA non-ensemble methods on validation, private and public datasets.

Method	Dataset	Kappa	Accuracy
[Ghosh *et al.*, 2017]	val	0.7400	-
[Krishnan *et al.*, 2018]	val	0.7600	0.7610
[Kwasigroch *et al.*, 2018]	val	0.7760	0.5080
Baseline[Niu *et al.*, 2019]	val	0.8030	0.7983
ELLG	val	**0.8123**	**0.8129**
Baseline[Niu *et al.*, 2019]	private	0.8054	-
ELLG	private	**0.8126**	-
Baseline[Niu *et al.*, 2019]	public	0.8093	-
ELLG	public	**0.8172**	-

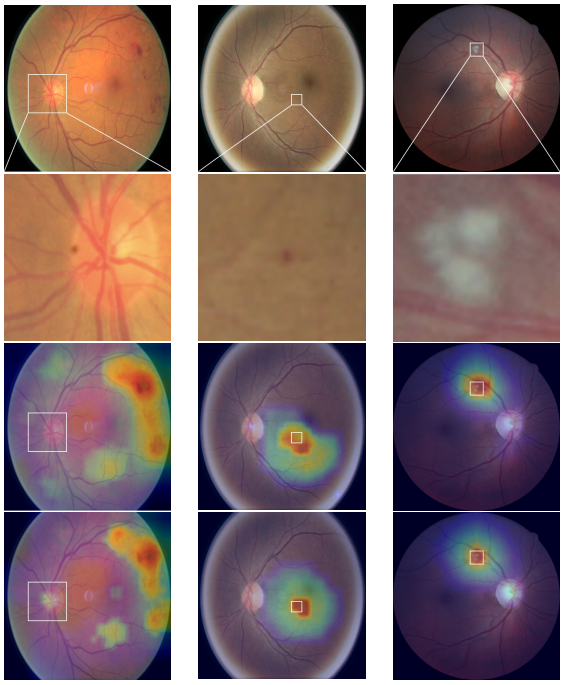
Table 1: Comparative results on *Database1*

The *Database2* has also been used for the evaluation. We fine-tuned the baseline model from *Database1* by using both baseline and the proposed method. It can be seen from Table 2, the proposed approach has better learning and adaptive ability on the data from different domain.

Model	Kappa	Accuracy
Baseline[Niu *et al.*, 2019]	0.8963	0.7830
ELLG	**0.9203**	**0.8269**

Table 2: Comparative results on *Database2*

In addition, we illustrated comparative visualization results generated by using baseline and the proposed ELLG respectively which are shown in Figure 4. Specially, in Figure 4 a, the lesions (rectangle) can be recognized by ELLG only. It can be attributed to our iterative lesion generation and learning mechanism. Besides, in Figure 4 b and c, the peak regions of GMSMs generated from baseline model have a slight deviation from the lesion and are less concentrated compared to those generated from the ELLG model, which means our ELLG model is more sensitive to the lesions detection.



(a) (b) (c)

Figure 4: First and second row are original images, and their lesion details. The third and fourth row present GMSMs generated from the baseline model and ELLG model respectively.

4 Conclusion

In this paper, we have proposed the explainable lesion learning and generation (ELLG) framework to study the interpretability of DR detection, address the data imbalance problem, and achieve more accurate DR diagnosis. To illustrate more reliable visual explanations for DR diagnosis, we have generated heatmaps by using the proposed Gated Multilayer Saliency Map (GMSM). Based on the GMSMs, lesion patches have been extracted and then adaptively transferred and blended with other fundus images. Then the generated and original images have been used together to train the model and to learn more robust lesion feature. The comparative experimental results on two databases have validated the effectiveness of the proposed method for DR detection.

5 Acknowledgement

This research is supported by the National Key Research and Development Program of China (2020AAA0107700) and the National Natural Science Foundation of China (62006181, 61822309, 61703301, U1736205).

References

[Antony and Brüggemann, 2016] M. Antony and S. Brüggemann. Team o_o solution for the kaggle diabetic retinopathy detection challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15807, 2016.

[Cao *et al.*, 2018] Peng Cao, Fulong Ren, Chao Wan, Jinzhu Yang, and Osmar Zaiane. Efficient multi-kernel multi-instance learning using weakly supervised and imbalanced data for diabetic retinopathy diagnosis. *Computerized Medical Imaging and Graphics*, 69:112–124, 2018.

[Chattopadhay *et al.*, 2018] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[Costa *et al.*, 2018] Pedro Costa, Adrian Galdran, Asim Smailagic, and Aurélio Campilho. A weakly-supervised framework for interpretable diabetic retinopathy detection on retinal images. *IEEE Access*, PP:1–1, 03 2018.

[Ghosh *et al.*, 2017] Ratul Ghosh, Kuntal Ghosh, and Sanjit Maitra. Automatic detection and classification of diabetic retinopathy stages using cnn. In *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 550–554. IEEE, 2017.

[Jiang *et al.*, 2019] Hongyang Jiang, Kang Yang, Mengdi Gao, Dongdong Zhang, He Ma, and Wei Qian. An interpretable ensemble deep learning model for diabetic retinopathy disease classification. volume 2019, pages 2045–2048, 07 2019.

[Kaggle, 2016] Kaggle. Kaggle diabetic retinopathy detection challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection, 2016.

[Kaggle, 2019] Kaggle. Kaggle aptos 2019 blindness detection challenge. https://www.kaggle.com/c/aptos2019-blindness-detection, 2019.

[Krishnan *et al.*, 2018] Arvind Sai Krishnan, Vilas Bhat, Pravin Bhaskar Ramteke, Shashidhar G Koolagudi, et al. A transfer learning approach for diabetic retinopathy classification using deep convolutional neural networks. In *2018 15th IEEE India Council International Conference (INDICON)*, pages 1–6. IEEE, 2018.

[Kwasigroch *et al.*, 2018] Arkadiusz Kwasigroch, Bartlomiej Jarzembinski, and Michal Grochowski. Deep cnn based decision support system for detection and assessing the stage of diabetic retinopathy. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 111–116. IEEE, 2018.

[Niu *et al.*, 2019] Yuhao Niu, Lin Gu, Feng Lu, Feifan Lv, Zongji Wang, Imari Sato, Zijian Zhang, Yangyan Xiao, Xunzhang Dai, and Tingting Cheng. Pathological evidence exploration in deep retinal image diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1093–1101, 2019.

[Pires *et al.*, 2019] Ramon Pires, Sandra Avila, Jacques Wainer, Eduardo Valle, Michael D Abramoff, and Anderson Rocha. A data-driven approach to referable diabetic retinopathy detection. *Artificial intelligence in medicine*, 96:93–106, 2019.

[Pratt *et al.*, 2016] Harry Pratt, Frans Coenen, Deborah M Broadbent, Simon P Harding, and Yalin Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science*, 90:200–205, 2016.

[Qummar *et al.*, 2019] Sehrish Qummar, Fiaz Gul Khan, Sajid Shah, Ahmad Khan, Shahaboddin Shamshirband, Zia Ur Rehman, Iftikhar Ahmed Khan, and Waqas Jadoon. A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access*, 7:150530–150539, 2019.

[Reinhard *et al.*, 2001] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[Voets *et al.*, 2018] Mike Voets, Kajsa Møllersen, and Lars Bongo. Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PloS one*, 14, 03 2018.

[Wan *et al.*, 2018a] Shaohua Wan, Yan Liang, and Yin Zhang. Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Computers and Electrical Engineering*, 72:274 – 282, 2018.

[Wan *et al.*, 2018b] Shaohua Wan, Yan Liang, and Yin Zhang. Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Computers & Electrical Engineering*, 72:274–282, 2018.

[Wang and Yang, 2018] Zhiguang Wang and Jianbo Yang. Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Wang *et al.*, 2017] Zhe Wang, Yanxin Yin, Jianping Shi, Wei Fang, Hongsheng Li, and Xiaogang Wang. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. 06 2017.

[Yang *et al.*, 2017] Yehui Yang, Tao Li, Wensi Li, Haishan Wu, Wei Fan, and Wensheng Zhang. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. 05 2017.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization", booktitle = "the ieee conference on computer vision and pattern recognition (cvpr). June 2016.

[Zhu *et al.*, 2019] Cheng-Zhang Zhu, Rong Hu, Bei-Ji Zou, Rong-Chang Zhao, Chang-Long Chen, and Ya-Long Xiao. Automatic diabetic retinopathy screening via cascaded

framework based on image- and lesion-level features fusion. *Journal of Computer Science and Technology*, 34(6):1307–1318, 2019.