Identifying the key components in ResNet-50 for diabetic retinopathy grading from fundus images: a systematic investigation

Yijin Huang[1] 11610128@mail.sustech.edu.cn
[1] *Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China*

Li Lin[1,2] linli@eee.hku.hk
[2] *Department of Electrical and Electronic Engineering, The University of Hong Kong, China*

Pujin Cheng[1] 12032946@mail.sustech.edu.cn
Junyan Lyu[1,3] junyan.lyu@uq.edu.au
[3] *Queensland Brain Institute, The University of Queensland, Australia*

Xiaoying Tang[*1] tangxy@sustech.edu.cn

Editors: Under Review for MIDL 2022

Abstract

Although deep learning based diabetic retinopathy (DR) classification methods typically benefit from well-designed architectures of convolutional neural networks, the training setting also has a non-negligible impact on the prediction performance. The training setting includes various interdependent components, such as objective function, data sampling strategy and data augmentation approach. To identify the key components in a standard deep learning framework (ResNet-50) for DR grading, we systematically analyze the impact of several major components. Extensive experiments are conducted on a publicly-available dataset EyePACS. Based on our observations and an optimal combination of the investigated components, our framework, without any specialized network design, achieves the state-of-the-art result (0.8631 for Kappa) on the EyePACS test set (a total of 42670 fundus images) with only image-level labels. Our codes and pre-trained model are available at https://github.com/YijinHuang/pytorch-classification.

Keywords: Diabetic Retinopathy, Classification, Training Setting, ResNet-50

1. Introduction

Diabetic retinopathy (DR) is one of the microvascular complications of diabetes, causing vision impairment and blindness (Li et al., 2021). The digital color fundus image is the most widely used imaging modality for ophthalmologists to screen and identify the severity of DR. In recent years, deep learning based methods have achieved great success in the field of medical image analysis (Lyu et al., 2019; Araújo et al., 2020; Cheng et al., 2021; Kervadec et al., 2021). In the realm of DR grading, given that lesions are important biomarkers, Attention Fusion Network (Lin et al., 2018) employs a lession detector to predict the probabilities of various lesions and proposes an information fusion method based on an

* Corresponding author

attention mechanism to identify DR. Zoom-in-net (Wang et al., 2017) consists of three sub-networks which respectively localize suspicious regions, analyze lesion patches and classify the image of interest. To enhance the capability of a standard covolutional neural network (CNN), CABNet (He et al., 2020) introduces two extra modules, one for exploring region-wise features for each DR grade and one for generating attention feature maps.

It can be observed that recent progress in automatic DR grading is largely attributed to carefully designed model architecture. Nevertheless, task-specific designs and specialized configurations may limit their transferability and extensibility. Other than model architecture, the training setting is also a key factor affecting the performance of a deep learning method. A variety of interdependent components are typically involved in a training setting, including the design of configurations and empirical decisions of hyper-parameters. Proper training settings can benefit automatic DR grading, while improper ones may damage the grading performance. However, the importance of the training setting has been overlooked or has received less attention in the past few years, especially in the DR grading field. In computer vision, there have been growing efforts in improving the performance of deep learning methods by refining the training setting rather than the network architecture. For example, He et al. (2019) boosts ResNet-50's (He et al., 2016) top-1 validation accuracy from 75.3% to 79.29% on ImageNet (Deng et al., 2009) by applying numerous training procedure refinements. Bochkovskiy et al. (2020) examines various combinations of training configurations such as batch-normalization and residual-connection, and utilizes them to improve the performance of object detection. In the biomedical domain, efforts in this direction have also emerged. For example, Isensee et al. (2021) proposes an efficient deep learning-based segmentation framework for biomedical images, namely nnU-Net, which can automatically and optimally configure its own setting for preprocessing, training and post-processing. In such context, we believe that refining the training setting has a great potential in enhancing the DR grading performance.

In this work, we systematically analyze the influence of several major components of a standard DR classification framework and identify the key elements in the training setting for improving the DR grading performance. The components analyzed in our work are shown in Figure 1. The main contributions of this work can be summarized as follows:

- We examine a collection of designs with respect to the training setting and evaluate them on the most challenging and largest publicly-available fundus image dataset, EyePACS[1]. We analyze and illustrate the impact of each component on the DR grading performance to identify the core ones.
- Based on our observations, we adopt ResNet-50 as the backbone and achieve a quadratically weighted Kappa of 0.8631 on the EyePACS test set, which outperforms many specifically-designed state-of-the-art methods, with only image-level labels.
- We emphasize that the superior performance of our framework is not achieved by a new network architecture, a new objective function nor a new scheme. The key contribution of this work, in a more generalizable sense, is that we outline another direction to improve the performance of deep learning methods for DR grading and

1. https://www.kaggle.com/c/diabetic-retinopathy-detection

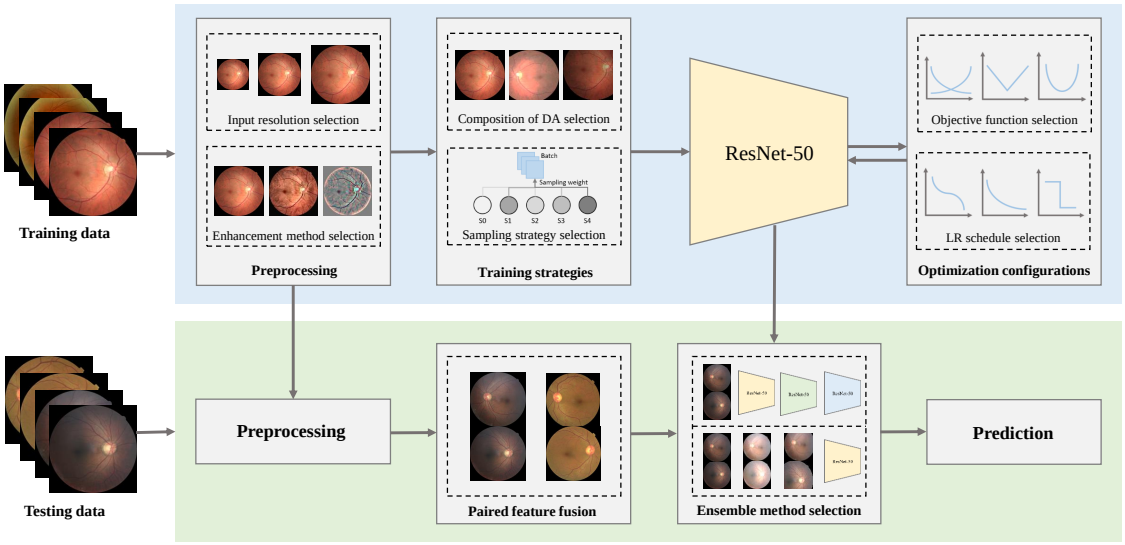


Figure 1: Components analyzed in our deep learning-based DR grading framework.

highlight the importance of training setting refinements in developing deep learning based pipelines.

2. Method

2.1. Dataset description

The EyePACS dataset is the largest publicly-available DR grading dataset released in the Kaggle DR grading competition, consisting of 88702 color fundus images from the left and right eyes of 44351 patients. Images were officially split into 35126/10906/42670 for training/validation/testing. According to the severity of DR, they have also been divided by ophthalmologists into five grades: 0 (normal), 1 (mild DR), 2 (moderate DR), 3 (severe DR), and 4 (proliferative DR) (Lin et al., 2020). The fundus images were acquired under a variety of conditions and from different imaging devices, resulting in variations in image resolution, aspect ratio, intensity, and quality. The class distribution of EyePACS is extremely imbalanced, wherein DR fundus images are dramatically less than normal images.

2.2. Baseline setting

We first specify our baseline for DR grading. In the preprocessing step, for each image, we first identify the smallest rectangle that contains the entire field of view and use the identified rectangle for cropping. After that, we resize each cropped image into 512×512 squares (see Appendix D.1 for details of input resolution selection) and rescale each pixel's intensity value into $[0, 1]$.

A widely used architecture ResNet-50 is employed in this work. We adopt the SGD optimizer with an initial learning rate of 0.001 and Nesterov technique (Nesterov, 1983)

with a momentum factor of 0.9 to train the network. Convolutional layers are initialized with parameters obtained from a ResNet-50 pre-trained on the ImageNet dataset (Deng et al., 2009). We train the model for 25 epochs with a mini-batch size of 16. If not specified, all models are trained with a fixed random seed for fair comparisons. The model having the highest metric on the validation set is selected for testing. The DR grading performance is evaluated using the quadratically weighted Kappa κ (Cohen, 1968), which is an officially-used metric in the Kaggle DR grading competition.

3. Training setting components

3.1. Objective function

The objective function plays a critical role in deep learning. There are a variety of objective functions that can be used to measure the discrepancy between the predicted probability distribution and the ground truth distribution of the given label. The cross-entropy loss (CE), focal loss (Lin et al., 2017), soft Kappa loss (Fauw, 2015) and regression loss are considered in this work.

The soft Kappa loss (Fauw, 2015) based on the Kappa metric is a typical choice for training a DR grading model. The quadratically-weighted Kappa is sensitive to disagreements in marginal distributions, whereas cross-entropy loss does not take into account the distribution of the predictions and the magnitude of the incorrect predictions. The regression loss also provides a penalty to the distance between prediction and ground truth. Three regression loss functions are examined in this work, namely mean absolute error (MAE), mean square error (MSE), and smooth L1 loss. In the testing phase, the prediction scores are clipped to be between [0, 4] and then simply rounded to integers to serve as the finally predicted grades.

3.2. Learning rate schedule

The learning rate is important in gradient descent methods, which has non-trivial impact on the convergence of the objective function. However, the optimal learning rate may vary at different training phases. Therefore, a learning rate schedule is widely used to adjust the learning rate during training. Three popular schedules are investigated in this work, including the multiple-step decaying schedule, the exponential decaying schedule and the cosine decaying schedule (Loshchilov and Hutter, 2016). Due to the observation that too small learning rates may lead to overfiting of a model at the last few epochs, we set a minimum learning rate for the cosine decaying schedule, namely clipped cosine decaying (CCD).

3.3. Composition of data augmentation

Applying online data augmentation (DA) during training can increase the distribution variability of the input images to improve the generalization capacity and robustness of a model of interest. To systematically study the impact of the composition of data augmentation on DR grading, various popular augmentation operations are considered in this work. For geometric transformations, we apply horizontal and vertical flipping, random rotation, and random cropping. For color transformations, color distortion is a common choice, including

adjustments of brightness, contrast, saturation, and hue. See Appendix A for more details of configurations of augmentation operations.

3.4. Preprocessing

In addition to background removal, two popular preprocessing operations for fundus images are considered in this work, namely Graham processing (Graham, 2015) and contrast limited adaptive histogram equalization (CLAHE) (Huang et al., 2012). Both of them can alleviate the blur, low contrast, and inhomogeneous illumination issues that exist in the EyePACS dataset. More details and representative enhanced images are presented in Appendix C.

3.5. Sampling strategy

As mentioned in section 2.1, EyePACS is an extremely imbalanced dataset. To address this problem, several sampling strategies (Kang et al., 2019; Antony, 2015) for the training set have been proposed to rebalance the data distribution. Three commonly used sampling strategies are examined in this work: (1) instance-balanced sampling samples each data point with an equal probability; (2) class-balanced sampling first selects each class with an equal probability, and then uniformly samples data points from specific classes; (3) progressively-balanced sampling starts with class-balanced sampling and then exponentially moves to instance-balanced sampling.

3.6. Prior knowledge

For medical image analysis, making use of prior knowledge can significantly enhance the performance of deep learning frameworks. In the EyePACS dataset, both the left and right eyes of a patient are provided. Evidence shows that for more than 95% the difference in the DR grade between the left and right eyes is no more than 1 (Wang et al., 2017). As such, to utilize the correlation between the two eyes, we concatenate the feature vectors of both eyes from the global average pooling layer of ResNet-50 and then input it into a paired feature fusion network (PFF). The network consists of three linear layers each followed by a 1D max-pooling layer with a stride of two and rectified linear unit (ReLU).

4. Experimental Results

4.1. Influence of different objective functions

We first evaluate the seven objective functions described in section 3.1. We also evaluate the objective function by combining the Kappa loss and the cross-entropy loss (Fauw, 2015). All objective functions are observed to converge after 25 epochs of training. The validation and test Kappa for applying different loss functions are reported in Table 1. The results demonstrate the focal loss and the combination of the Kappa loss and the cross-entropy loss slightly improve the performance compared to the standard cross-entropy loss. The MSE loss yields a 2.02% improvement over the cross-entropy loss.

To demonstrate the influence of different objective functions on the distribution of predictions, we present the confusion matrics of the test set for the cross-entropy loss and the MSE loss in Figure 2. As shown in Figure 2, the prediction-versus-ground truth distance

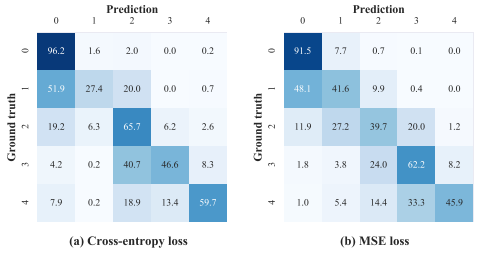


Figure 2: Confusion matrices

Loss	Val Kappa	Test Kappa
CE	0.8054	0.8032
Focal (γ=2)	0.8079	0.8059
Kappa	0.7818	0.7775
Kappa + CE	0.8047	0.8050
MAE	0.7655	0.7679
Smooth L1	0.8094	0.8117
MSE	**0.8207**	**0.8235**

Table 1: DR grading performance.

from using MSE is smaller than that from using cross-entropy. That is, the predictions from the model using the MSE loss as the objective function show more diagonal tendency.

4.2. Influence of different learning rate schedules

Further on we study the influence of different learning rate schedules. For the multiple-step decaying schedule, we decrease the learning rate by 0.1 at epoch 15 and epoch 20. For the exponential decaying schedule, we set the decay factor to be 0.9. All experiments are conducted using the baseline setting with the MSE loss. The minimum learning rate is set to be 10^{-4} in clipped cosine decaying. The experimental results are shown in Table 2. The results demonstrate that the clipped cosine decaying schedule gives the highest improvement of 0.37% in the test Kappa.

Schedule	Validation Kappa	Test Kappa
Constant	0.8207	0.8235
Multiple Steps [15, 20]	0.8297	0.8264
Exponential (p=0.9)	0.8214	0.8185
Cosine	**0.8269**	0.8267
Clipped Cosine (η_{min}=1e-4)	0.8258	**0.8272**

Table 2: DR grading performance of models using different learning rate schedules.

4.3. Influence of different compositions of data augmentation

We evaluate ResNet-50 with different compositions of data augmentation. In addition to flipping and rotation in the baseline setting, we consider random cropping and color jitter. We also evaluate the model trained without any data augmentation. All experiments are based on the best setting from previous evaluations. As shown in Table 3, even a simple composition of geometric data augmentation operations (the third row of Table 3) in the baseline setting can provide a significant improvement of 3.49% on the test Kappa. The best test Kappa of 0.8310 is achieved by applying the composition of flipping, rotation, cropping, and color jitter for data augmentation during training. We adopt this composition in our subsequent experiments.

Flipping	Rotation	Cropping	Color Jitter	Validation Kappa	Test Kappa
				0.7913	0.7923
✓				0.8124	0.8125
✓	✓			0.8258	0.8272
✓		✓		0.8194	0.8217
✓			✓	0.8129	0.8167
✓			✓	0.8082	0.8159
✓	✓	✓		0.8276	0.8247
✓	✓	✓	✓	**0.8307**	**0.8310**

Table 3: DR grading performance of models using different compositions of DA.

4.4. Influence of different preprocessing methods

Two popular image enhancement methods are evaluated in our study, Graham processing and CLAHE. Both of them have been suggested to be beneficial for DR identification (Yang et al., 2017; Sahu et al., 2019). However, we observe that they are not helpful for DR grading in our framework. Based on the previous optimal combination, using Graham processing or CLAHE, the test performance is dropped to 0.8260 and 0.8238 respectively.

4.5. Influence of different sampling strategies

Further, we concern about the influence of different sampling strategies. To alleviate the imbalance issue in EyePACS, except for the instance-balanced sampling, the class-balanced sampling and the progressively-balanced sampling are examined in the training phase. However, because we repeatedly sample data points from the minority classes at each epoch, severe overfitting is observed and results in poor performance on the validation set. Further analysis can be found in Appendix B.

4.6. Influence of feature fusion of paired eyes

We evaluate the improvement resulted from utilizing the correlation between the paired two eyes for DR grading. The best model from previous evaluations is fixed and adopted to generate feature vector of each fundus image. The simple paired feature fusion network is trained for 20 epochs with a batch size of 64. The learning rate is set to be 0.02 without any decaying schedule. As shown in Table 4, paired feature fusion improves the validation Kappa by 2.90% and the test Kappa by 2.71%, demonstrating the importance of the eye pair correlation to DR grading.

4.7. Comparison of the importance of all components

Finally, we investigate and compare the importance of all considered components in our DR grading task. We quantify the improvement from each component by applying them one by one, the results of which are shown in Table 4. We observe two significant improvements outstand from that table. First, the choice of the MSE loss and utilization of the eye pair fusion respectively improve the test Kappa by 2.03% and 2.71%. Additional improvements of 0.37% and 0.38% on the test Kappa are obtained by applying clipped cosine decaying

MSE	CCD	DA	PFF	Val Kappa	Test Kappa	Δ
				0.8054	0.8032	0%
✓				0.8207	0.8235	+2.03%
✓	✓			0.8258	0.8272	+2.40%
✓	✓	✓		0.8307	0.8310	+2.78%
✓	✓	✓	✓	**0.8597**	**0.8581**	+5.49%

Table 4: The performance of models for stacking refinements one by one.

Method	Test Kappa
Min-Pooling	0.8490
o_O	0.8450
RG	0.8390
Zoom-in Net	0.8540
CABNet	0.8456
Ours	0.8581
Ours (ensemble)	**0.8631**

Table 5: Comparisons with other methods with only image-level labels.

schedule and data augmentation. A total of 5.49% improvement of Kappa is achieved by combining all of these refinements. In addition, we analyze the influence of input resolution and ensemble methods in Appendix D. A further study on significance of difference choices in each component is provided in Appendix E.

4.8. Comparison with state-of-the-art

To assess the performance of our framework that incorporates the optimal set of all components investigated in this work, comparisons between the proposed method and previously-reported state-of-the-art ones without any utilization of additional datasets nor annotations are tabulated in Table 5. The results listed in the first three rows denote the top-3 entries on Kaggle's challenge. Zoom-in Net and CABNet with ResNet-50 backbone are compared. Our proposed method, without any fancy technique, outperforms previous state-of-the-art results by 0.91% in terms of the test Kappa.

5. Conclusion

In this work, we systematically investigate several important components in CNN for improving the performance of ResNet-50 based DR grading. Extensive experiments on the publicly-available EyePACS dataset are conducted to evaluate the influence of different selections for each component. Finally, based on our findings, a simple yet effective framework for DR grading is proposed. Our study can be summarized as below.

- We raise the ResNet-50 Kappa metric from 0.8032 to 0.8631 on the EyePACS dataset, outperforming other specially-designed DR grading methods.
- Achieving state-of-the-art performance without any network architecture modification, we emphasize the importance of training setting refining in the development of deep learning based frameworks.
- Our codes and pre-trained model are publicly accessible. We believe our simple yet effective framework can serve as a strong, standardized, and scalable baseline for further studies and developments of DR grading algorithms.

Acknowledgments

The authors would like to thank Meng Li from Zhongshan Ophthalmic Centre of Sun Yat-sen University as well as Yue Zhang from the University of Hong Kong for their help on this work. This study was supported by the National Natural Science Foundation of China (62071210), the Shenzhen Basic Research Program (JCYJ20190809120205578), the National Key R&D Program of China (2017YFC0112404), and the High-level University Fund (G02236002).

References

Mathis Antony. Team o_o solution summary, 2015. URL https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15617#latest-373487.

Teresa Araújo, Guilherme Aresta, Luís Mendonça, Susana Penas, Carolina Maia, Ângela Carneiro, Ana Maria Mendonça, and Aurélio Campilho. Dr— graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Medical Image Analysis*, 63:101715, 2020.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18, 2004.

Pujin Cheng, Li Lin, Yijin Huang, Junyan Lyu, and Xiaoying Tang. I-secret: Importance-guided fundus image enhancement via semi-supervised contrastive constraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–96. Springer, 2021.

Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

Niladri Sekhar Datta, Himadri Sekhar Dutta, Mallika De, and Saurajeet Mondal. An effective approach: image quality enhancement for microaneurysms detection of non-dilated retinal fundus image. *Procedia Technology*, 10:731–737, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jeffrey De Fauw. Detecting diabetic retinopathy in eye images, 2015. URL http://defauw.ai/diabetic-retinopathy-detection/.

Ben Graham. Kaggle diabetic retinopathy detection competition report. *University of Warwick*, 2015.

Along He, Tao Li, Ning Li, Kai Wang, and Huazhu Fu. Cabnet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

Kai-Qi Huang, Qiao Wang, and Zhen-Yang Wu. Natural color image enhancement and evaluation algorithm based on human visual system. *Computer Vision and Image Understanding*, 103(1):52–63, 2006.

Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE transactions on image processing*, 22(3):1032–1041, 2012.

Yijin Huang, Li Lin, Meng Li, Jiewei Wu, Pujin Cheng, Kai Wang, Jin Yuan, and Xiaoying Tang. Automated hemorrhage detection from coarsely annotated fundus images in diabetic retinopathy. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1369–1372. IEEE, 2020.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. *Medical Image Analysis*, 67:101851, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Tao Li, Wang Bo, Chunyu Hu, Hong Kang, Hanruo Liu, Kai Wang, and Huazhu Fu. Applications of deep learning in fundus images: A review. *Medical Image Analysis*, page 101971, 2021.

Li Lin, Meng Li, Yijin Huang, Pujin Cheng, Honghui Xia, Kai Wang, Jin Yuan, and Xiaoying Tang. The sustech-sysu dataset for automated exudate detection and diabetic retinopathy grading. *Scientific Data*, 7(1):1–10, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Zhiwen Lin, Ruoqian Guo, Yanjie Wang, Bian Wu, Tingting Chen, Wenzhe Wang, Danny Z Chen, and Jian Wu. A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 74–82. Springer, 2018.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Junyan Lyu, Pujin Cheng, and Xiaoying Tang. Fundus image based retinal vessel segmentation utilizing a fast and accurate fully convolutional network. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 112–120. Springer, 2019.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

Gwenolé Quellec, Katia Charrière, Yassine Boudi, Béatrice Cochener, and Mathieu Lamard. Deep image mining for diabetic retinopathy screening. *Medical image analysis*, 39:178–193, 2017.

Sima Sahu, Amit Kumar Singh, SP Ghrera, Mohamed Elhoseny, et al. An approach for de-noising and contrast enhancement of retinal fundus image using clahe. *Optics & Laser Technology*, 110:87–98, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Zhe Wang, Yanxin Yin, Jianping Shi, Wei Fang, Hongsheng Li, and Xiaogang Wang. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 267–275. Springer, 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. In *International conference on medical image computing and computer-assisted intervention*, pages 533–540. Springer, 2017.

Appendix A. Data Augmentation Details

For the cropping operation, we randomly crop a rectangular region the size of which is randomly sampled in [1/1.15, 1.15] times the original one and the aspect ratio is randomly sampled in [0.7, 1.3], and then we resize this region back to be of the original size. Horizontal and vertical flipping is applied with a probability of 0.5. The color distortion operation adjusts the brightness, contrast, and saturation of the images with a random factor in [-0.2, 0.2] and the hue with a random factor in [-0.1, 0.1]. The rotation operation randomly rotates each image of interest by an arbitrary angle.

Appendix B. Overfitting Caused by Sampling Strategies

As illustrated in Fig. 3, the gap between the training Kappa and the validation Kappa increases as the probability of sampling the minority classes increases. Instance-balanced sampling, a strategy that we most commonly use, achieves the highest validation Kappa at the end of the training. A plausible reason for this result is that the class distribution of the training set is consistent with that of the validation set as well as those of real-world datasets. The class-based sampling strategy may be more effective in cases where the training set is imbalanced and the test set is balanced (Kang et al., 2019).

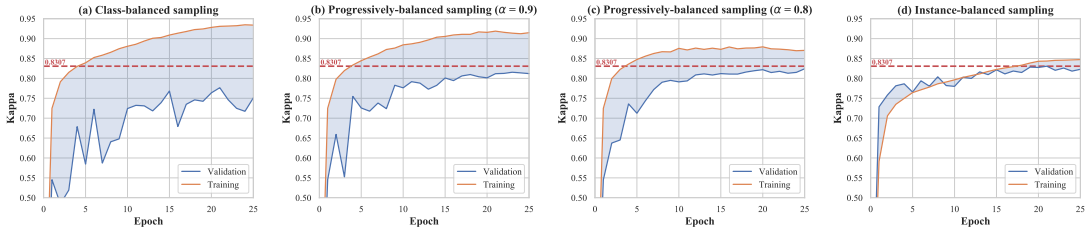


Figure 3: The performance of models using different sampling strategies for training. The dotted red line represents the best validation Kappa among these four experiments, which is achieved by instance-balanced sampling.

Appendix C. Preprocessing Details

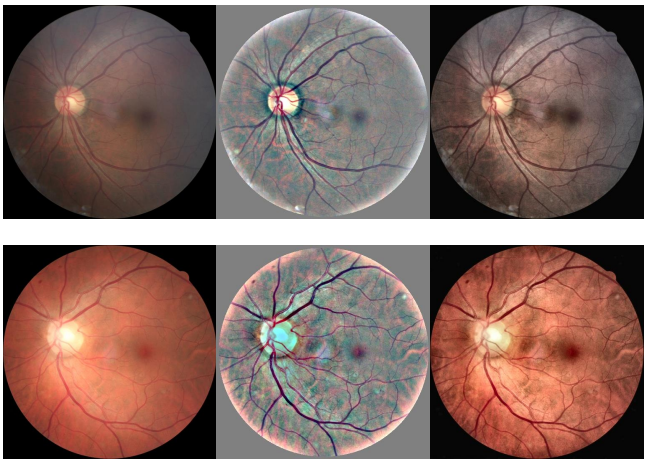
The Graham method was proposed by B. Graham the winner of the Kaggle DR grading competition. This preprocessing method has also been used in many previous works (Quellec et al., 2017; Yang et al., 2017) to remove image variations due to different lighting conditions or imaging devices. Given a fundus image $\mathbf{I}$, the processed image $\hat{\mathbf{I}}$ after Graham is obtained by

$$\hat{\mathbf{I}} = \alpha\mathbf{I} + \beta\mathbf{G}(\theta) * \mathbf{I} + \gamma, \tag{1}$$

where $\mathbf{G}(\theta)$ is a 2D Gaussian filter with a standard deviation θ, $*$ is the convolution operator, and α, β, γ are weighting factors. Following (Yang et al., 2017), θ, α, β, and γ

are respectively set as 10, 4, -4, and 128. As shown in Fig. 4, all images are normalized to be relatively consistent with each other and vessels as well as lesions are particularly highlighted after Graham processing.

CLAHE is a contrast enhancement method based on Histogram Equalization (HE) (Huang et al., 2006), which has also been widely used to process fundus images and has been suggested to be able to highlight lesions (Huang et al., 2020; Sahu et al., 2019; Datta et al., 2013). HE improves the image contrast by spreading out the most frequently-occurred intensity values in the histogram, but it amplifies noise as well. CLAHE was proposed to prevent an over-amplification of noise by clipping the histogram at a predefined value. Representative enhanced images via CLAHE are also illustrated in Fig. 4.



(a) Original (b) Graham (c) CLAHE

Figure 4: Representative enhanced fundus images using Graham processing and CLAHE.

Appendix D. Additional Experimental Results

D.1. Input resolution

The resolution of the input image has a direct impact on the DR grading performance. Generally, ResNet-50 is designed for images of 224×224 input resolution (He et al., 2016). In ResNet-50, a convolution layer with a kernel size of 7×7 and a stride of 2 followed by a max-pooling layer is applied to dramatically downsample the input image first. Therefore, using images with very small input resolution may lose key features for DR grading, such as tiny lesions. In contrast, a network fed with large resolution images can extract more fine-grained and dense features at the cost of a smaller receptive field and a higher computational cost. In this work, a range of resolutions is evaluated to identify the trade-off.

The experimental results are shown in Table 6. As suggested by the results, DR grading benefits from larger input resolutions at the cost of higher training and inference computational expenses. A significant performance improvement of 16.42% in the test Kappa is obtained by increasing the resolution from 128×128 to 512×512. Increasing the resolu-

tion to 1024×1024 further improves the test Kappa by another 1.32% but with a large computational cost increase of 64.84G floating-point operations (FLOPs). Considering the trade-off between performance and computational cost, the 512×512 input resolution is adopted for all our subsequent experiments.

Resolution	Training time	FLOPs	Validation Kappa	Test Kappa
128×128	1h 54m	1.35G	0.6535	0.6388
256×256	2h 19m	5.40G	0.7563	0.7435
512×512	5h 16m	21.61G	0.8054	0.8032
768×768	11h 15m	48.63G	0.8176	0.8137
1024×1024	11h 46m (2 GPUs)	86.45G	0.8187	0.8164

Table 6: DR grading performance with different input resolutions. Two GPUs are used to train the model with 1024×1024 input resolution due to the CUDA memory limitation.

D.2. Ensembling

Ensemble methods (Opitz and Maclin, 1999) are widely used in data science competitions to achieve better performance. The variance in the predictions and the generalization errors can be considerably reduced by combining predictions from multiple models or inputs. However, ensembling too many models can be computationally expensive and the performance gains may diminish with the increasing number of models. To make our proposed pipeline generalizable, two simple ensemble methods are considered: 1) for the ensemble method that uses multiple models (Krizhevsky et al., 2012; Caruana et al., 2004), we average the predictions from models trained with different random seeds. In this way, the datasets have different sampling orders and different data augmentation parameters to train each model, resulting in differently trained models for ensembling, 2) for the ensemble method that uses multiple views (Simonyan and Zisserman, 2014; Szegedy et al., 2016), we first generate different image views via random flipping and rotation (test-time augmentation). Then these views including the original one are input into a single model to generate each view's DR grade score. We then use the averaged score as the finally predicted one.

We also evaluate the impact of the number of input views for the ensemble method of multiple views and the number of models for the ensemble method of multiple models. The experimental results are tabulated in Table 7. We observe that as the number of models increases, both the test Kappa and the validation Kappa steadily increase. Unsurprisingly, the computational cost also monotonically increases with the number of ensembling. For the ensemble method that uses multiple models, the performance gain from increasing the number of models diminishes in the end and the best test Kappa is achieved by using 10 models.

# views / models	Multiple views		Multiple models	
	Validation Kappa	Test Kappa	Validation Kappa	Test Kappa
1	0.8597	0.8581	0.8597	0.8581
2	0.8611	0.8593	0.8622	0.8596
3	0.8608	0.8601	0.8635	0.8615
5	0.8607	0.8609	0.8644	0.8617
10	**0.8633**	0.8603	**0.8660**	**0.8631**
15	0.8631	**0.8611**	0.8653	**0.8631**

Table 7: The performance of models with different ensemble methods.

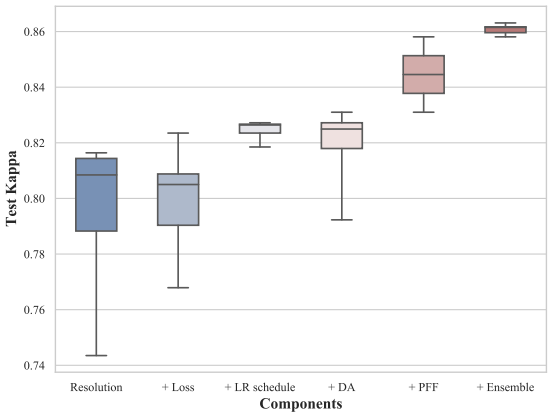


Figure 5: Box plots of the test Kappa of all experiments in this work. The experiments in each column are set up based on the best model considering all its left components. DA and PFF denote the experiment results of different compositions of data augmentation and applying paired feature fusion or not.

Appendix E. Variance in Difference Choices of Each Component

The incremental results alone do not completely reflect the importance of different components. The baseline configuration may also affect the corresponding improvements. In Fig. 5, we present the ranges and standard deviations of all experiments in this work. If the range of a box is large, it indicates that the results of different choices of this component vary significantly. The top bar of the box represents the highest test Kappa that can be achieved by specifically refining the corresponding component. Obviously, a bad choice of either resolution, objective function or data augmentation may lead to a great performance drop. Applying a learning rate schedule and ensembling can both provide steady improvements but using different schedules or ensemble methods does not significantly change the DR grading result.