Identifying the Key Components in ResNet-50 for Diabetic Retinopathy Grading from Fundus Images: A Systematic Investigation

Yijin Huang 1,2, Li Lin 1,3, Pujin Cheng 1, Junyan Lyu 1,4, Roger Tam 2,*, and Xiaoying Tang 1,*

1 Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China; yijinh@student.ubc.ca (Y.H.); linlimir@connect.hku.hk (L.L.);
2 School of Biomedical Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
3 Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
4 Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia
* Correspondence: roger.tam@ubc.ca (R.T.); tangxy@sustech.edu.cn (X.T.)

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 prediction performance. The training setting includes various interdependent components, such as an objective function, a data sampling strategy, and a 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. We demonstrate that (1) the DR grading framework is sensitive to input resolution, objective function, and composition of data augmentation; (2) using mean square error as the loss function can effectively improve the performance with respect to a task-specific evaluation metric, namely the quadratically weighted Kappa; (3) utilizing eye pairs boosts the performance of DR grading and; (4) using data resampling to address the problem of imbalanced data distribution in EyePACS hurts the performance. Based on these observations and an optimal combination of the investigated components, our framework, without any specialized network design, achieves a state-of-the-art result (0.8631 for Kappa) on the EyePACS test set (a total of 42,670 fundus images) with only image-level labels. We also examine the proposed training practices on other fundus datasets and other network architectures to evaluate their generalizability. Our codes and pre-trained model are available online.

Keywords: diabetic retinopathy; classification; training setting

1. Introduction

Diabetic retinopathy (DR) is one of the microvascular complications of diabetes, causing vision impairments and blindness [1,2]. The major pathological signs of DR include hemorrhages, exudates, microaneurysms, and retinal neovascularization. The digital color fundus image is the most widely used imaging modality for ophthalmologists to screen and identify the severity of DR, which can reveal the presence of different lesions. The early diagnosis and timely intervention of DR are of vital importance in preventing patients from vision malfunction. However, due to the rapid increase in the number of patients at risk of developing DR, ophthalmologists in regions with limited medical resources bear a heavy labor-intensive burden in DR screening. As such, developing automated and efficient DR diagnosis and prognosis approaches is urgently needed to reduce the number of untreated patients and the burden on ophthalmic experts.

Based on the type and quantity of lesions in fundus images, DR can be classified into five grades: 0 (normal), 1 (mild DR), 2 (moderate DR), 3 (severe DR), and 4 (proliferative
DR) [3]. Red dot-shaped microaneurysms are the first visible sign of DR, and their presence indicates a mild grade of DR. Red lesions (e.g., hemorrhages) and yellow-white lesions (e.g., hard exudates and soft exudates) have various types of shapes, from tiny points to large patches. A larger amount of such lesions indicate more severe DR grading. Neovascularization, the formation of new retinal vessels in the optic disc or its periphery, is a significant sign of proliferative DR. Figure 1 shows examples of fundus images with different types of lesions.

Various machine learning-based methods [4–7] have been proposed for disease detection. For example, ref. [5] introduces a hybrid ellipse fitting (EF)-based approach for detecting hematological disorders by automatically segmenting blood cells. Ref. [7] performs local binary pattern analysis targeting the texture micro-patterns in fundus images to detect DR. However, these methods often suffer from poor generalization due to their reliance on manually crafted features. In recent years, deep learning-based methods have achieved great success in the field of computer vision. With the capability of highly representative feature extraction, convolutional neural networks (CNNs) have been proposed to tackle different tasks. They have also been widely used in the medical image analysis realm [8–12]. In DR grading, ref. [13] adopts a pre-trained CNN as a feature extractor and re-trains the last fully connected layer for DR detection. Given that lesions are important guidance in DR grading [14], the attention fusion network [15] employs a lesion detector to predict the probabilities of lesions and proposes an information fusion method based on an attention mechanism to identify DR. Zoom-in-net [16] consists of three sub-networks that, respectively, localize suspicious regions, analyze lesion patches and classify the image of interest. To enhance the capability of a standard CNN, CABNet [17] 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, the 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 the deep learning method. A variety of interdependent components are typically involved in a training setting, including the design of configurations (e.g., preprocessing, loss function, sampling strategy, and data augmentation) and empirical decisions of hyperparameters (e.g., input resolution, learning rate, and training epochs). 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 received less attention in the past few years, especially in the DR grading field. In computer vision, there have been growing efforts to improve the performance of deep learning methods by refining the training setting rather than the network architecture. For example, ref. [18] boosts ResNet-50’s [19] top-1 validation accuracy from 75.3% to 79.29% on ImageNet [20] by applying numerous training procedure refinements. Ref. [21] examines combinations of training configurations such as batch-normalization and residual connection, and utilizes them to improve the performance of object detection. Although [18,21] have explored refinements for image classification and object detection tasks, they solely focus on natural images, which may limit their efficacy when applied to medical images. In the biomedical domain, efforts in this direction have also emerged. For example, ref. [22] proposes an efficient deep learning-based segmentation framework for biomedical images, namely nnU-Net, which can automatically and optimally configure its own setting including preprocessing, training, and post-processing. However, nnU-Net is designed for segmentation tasks. In such a context, we believe that refining the training setting has 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. We then evaluate these training practices on multiple datasets and network architectures, with the goal of analyzing their generalizability across both datasets and network architectures. The components analyzed in our work are shown in Figure 2. 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 (https://www.kaggle.com/c/diabetic-retinopathy-detection, accessed on 28 July 2015). We analyze and illustrate the impact of each component on the DR grading performance to identify the core ones.
- By refining several key components, we raise the quadratically weighted Kappa of the plain ResNet-50 [19] from 0.7435 to 0.8631 on the EyePACS test set, which outperforms many specifically designed state-of-the-art methods, with only image-level labels. With a widely used architecture, namely ResNet-50, our framework can serve as a strong, standardized, and scalable DR grading baseline. In other words, other types and directions of most methodological improvements and modifications can be easily incorporated into our framework to further improve the DR grading performance. Our codes and pre-trained model are available at https://github.com/YijinHuang/pytorch-classification (accessed on 1 February 2020).
- We evaluate the proposed training practices on two external retinal fundus datasets and six popular network architectures. Consistent and similar observations on multiple datasets and across different network architectures validate the generalizability and robustness of the proposed training setting refinements and the importance of the identified components in deep learning-based methods for DR grading.
- We emphasize that the superior performance of our framework is not achieved by a new network architecture, a new objective function, or a new scheme. The key contribution of this work, in a more generalizable sense, is that we outline another method to improve the performance of deep learning methods for DR grading and highlight the importance of training setting refinements in developing deep learning-based pipelines. This may also shed new insights into other related fields.

The remainder of this paper is organized as follows. Section 2 describes the details of our baseline framework, the default training setting, and the evaluation protocol. Descriptions of the investigated components in the training setting are presented in Section 3. Extensive experiments are conducted in Section 4 to evaluate the DR grading performance, the influence of each refinement, and the generalizability of the proposed practices. The discussion and conclusion are, respectively, provided in Sections 5 and 6.
Figure 2. Components analyzed in our deep learning-based DR grading framework. The evaluation process of a framework can be divided into two parts: training (top) and testing (bottom). In the training phase, we first fix the architecture of the selected network (ResNet-50). Then, we examine a collection of designs with respect to the training setting including preprocessing (image resizing and enhancement), training strategies (compositions of data augmentation (DA) and sampling strategies), and optimization configurations (objective functions and learning rate (LR) schedules). In the testing phase, we apply the same preprocessing as in the training phase and employ paired feature fusion to make use of the correlation between the two eyes (the training step of the fusion network is omitted in this figure). Then, we select the best ensemble method for the final prediction.

2. Method

2.1. Dataset

To analyze the components of interest in ResNet-50 and evaluate the performance of models trained with different training settings for DR grading, three widely used retinal datasets (EyePACS, Messidor-2 [23], and DDR [24]) are employed in this work.

EyePACS: The EyePACS dataset is the largest publicly available DR grading dataset released in the Kaggle DR grading competition, consisting of 88,702 color fundus images from the left and right eyes of 44,351 patients. Images were officially split into 35,126/10,906/42,670 fundus images for training/validation/testing. According to the severity of DR, they were also divided by ophthalmologists into the aforementioned five grades. 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 [25]. As shown in Figure 3, the class distribution of EyePACS is extremely imbalanced, wherein DR fundus images are dramatically less than normal images. In this work, the evaluation of each component is mainly performed on EyePACS.

Messidor-2: A total of 1748 fundus images with five-grade annotations and eye pairing are provided in the Messidor-2 dataset. We randomly split the dataset into 1042/176/522 fundus images for training/validation/testing. The main challenge of this dataset lies in the limited number of images for training, and thus we employ this dataset to evaluate the generalization ability of the proposed training practices.

DDR: The DDR dataset consists of 13,673 fundus images with six-class annotations (five DR grades and another “ungradable” class). All ungradable images are excluded, ending up with 6320/2503/3759 for training/validation/testing.
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 $224 \times 224$ squares and rescale each pixel intensity value into $[0, 1]$. Next, we normalize the RGB channels using z-score transformations with the mean and the standard deviations obtained from the entire preprocessed training set. Common random data augmentation operations including horizontal flipping, vertical flipping, and rotation described in Section 3.4 are performed during training.

ResNet-50 [19] is a widely used architecture in the field of deep learning. It has been adopted as a referent architecture for most analyses of training practices [26–28]. ResNet-50 utilizes residual connections to enable the training of very deep neural networks by addressing the vanishing gradient problem, and this strategy has also been adopted in the design of numerous other deep learning models [7,29,30]. Therefore, in this work, ResNet-50 is employed as our baseline model for analyzing different components. We adopt the SGD optimizer with an initial learning rate of 0.001 and Nesterov Accelerated Gradient Descent [31] with a momentum factor of 0.9 to train the network. A weighted decay of 0.0005 is applied for regularization. Convolutional layers are initialized with parameters obtained from a ResNet-50 pre-trained on the ImageNet dataset [20] and the fully connected layer is initialized using He’s initialization method [32]. We train the model for 25 epochs with a mini-batch size of 16 on a single NVIDIA RTX TITAN. All codes are implemented in PyTorch [33]. 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.

2.3. Evaluation Metric

The DR grading performance is evaluated using the quadratically weighted Kappa $\kappa$ [34], which is an officially used metric in the Kaggle DR grading competition. In an ordinal multi-class classification task, given an observed confusion matrix $o$ and an expected matrix $e$, $\kappa$ measures their agreement by quadratically penalizing the distance between the prediction and the ground truth,

$$\kappa = 1 - \frac{\sum_i \sum_j w_{ij} o_{ij}}{\sum_i \sum_j w_{ij} e_{ij}},$$

where $C$ denotes the total number of classes, $w$ is a quadratic weight matrix, and subscripts $i$ and $j$, respectively, denote the row and column indices of the matrices. The weight $w_{ij}$ is defined as $\frac{(i-j)^2}{(C-1)^2}$. $\kappa$ ranges from $-1$ to $1$, with $-1$ and $1$, respectively, indicating total disagreement and complete agreement.
3. Training Setting Components

3.1. 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 $224 \times 224$ input resolution [19]. In ResNet-50, a convolution layer with a kernel size of $7 \times 7$ 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.

3.2. Loss Function

The objective function plays a critical role in deep learning. Let $D = \{(x_i, y_i), i = 1, ..., N\}$ denote the training set, where $x_i$ is the input image and $y_i$ is the corresponding ground truth label. There are a variety of objective functions that can be used to measure the discrepancy between the predicted probability distribution $\hat{y}_i$ and the ground truth distribution $\tilde{y}_i$ (one-hot encoded $y_i$) of the given label.

3.2.1. Cross-Entropy Loss

The cross-entropy loss is the most commonly used loss function for classification tasks, which is the negative log-likelihood of a Bernoulli or categorical distribution,

$$ CE(\tilde{y}, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \tilde{y}_i \log(\hat{y}_i). \tag{2} $$

3.2.2. Focal Loss

The focal loss was initially proposed in RetinaNet [35], which introduces a modulating factor into cross-entropy to down-weight the loss of well-classified samples, giving more attention to challenging and misclassified ones. The focal loss is widely used to address the class imbalance problem in training deep neural networks. As mentioned before, EyePACS is an extremely imbalanced dataset with the number of images per class ranging from 25,810 to 708. Therefore, the focal loss is applied for better feature learning with samples from the minority classes. The focal loss is defined as

$$ FL(\tilde{y}, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \tilde{y}_i (1 - \hat{y}_i)^\gamma \log(\hat{y}_i), \tag{3} $$

where $\gamma$ is a hyperparameter. When the predicted probability $\hat{y}_i$ is small, the modulating factor $(1 - \hat{y}_i)^\gamma$ is close to 1. When $\hat{y}_i$ is large, this factor goes to 0 to down-weight the corresponding loss.

3.2.3. Kappa Loss

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. Therefore, the soft Kappa loss [36,37] based on the Kappa metric is another common choice for training the DR grading model,

$$ KL(y, \hat{y}) = 1 - \frac{o(y, \hat{y})}{e(y, \hat{y})}, \tag{4} $$

where

$$ o(y, \hat{y}) = \sum_{i,n} \left(\frac{y_i - n}{C - 1}\right)^2 \hat{y}_i, \tag{5} $$

and

$$ e(y, \hat{y}) = \sum_{i,n} \left(\frac{y_i - n}{C - 1}\right)^2. $$
\[ e(y, \hat{y}) = \sum_{m,n} \frac{(m-n)^2}{(C-1)^2} (\sum_{i} \mathbb{1}_{|n| = y_i})(\sum_{j} \hat{y}_{j,m}), \tag{6} \]

where \( C \) is the number of classes, \( \hat{y}_{j,k} (k \in [1, C]) \) is the predicted probability of the \( k \)-th class of \( \hat{y} \), and \( \mathbb{1}_{|n| = y_i} \) is an indicator function equaling 1 if \( n = y_i \) and otherwise 0. As suggested by a previous work [37], combining the Kappa loss with the standard cross-entropy loss can stabilize the gradient at the beginning of training to achieve better prediction performance.

### 3.2.4. Regression Loss

In addition to Kappa loss, the regression loss also provides a penalty to the distance between prediction and ground truth. When a regression loss is applied, the softmax activation of the fully connected layer is removed and the output dimension is set to be 1 to produce a prediction score \( \hat{y}_i \) for the DR grade. Three regression loss functions are considered in this work, namely L1 loss (Mean Absolute Error, MAE), L2 loss (Mean Square Error, MSE), and smooth L1 loss (SmoothL1), which are, respectively, defined as

\[
\text{MAE}(y_i, \hat{y}_i) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{7} \]

\[
\text{MSE}(y_i, \hat{y}_i) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{8} \]

\[
\text{SmoothL1}(y_i, \hat{y}_i) = \begin{cases} 
0.5(y_i - \hat{y}_i)^2, & \text{if } |y_i - \hat{y}_i| < 1 \\
|y_i - \hat{y}_i| - 0.5, & \text{otherwise.}
\end{cases} \tag{9} \]

In the testing phase, the prediction scores are clipped to be between [0, 4] and then simply rounded to integers to serve as the final predicted grades.

### 3.3. Learning Rate Schedule

The learning rate is important in gradient descent methods, which has a 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. Multiple-step decaying, exponential decaying, and cosine decaying [38] are popular learning rate adjustment strategies in deep learning. Specifically, the multiple-step decaying schedule decreases the learning rate by a constant factor at specific training epochs. The exponential decaying schedule exponentially decreases the learning rate by \( \gamma \) at every epoch, namely

\[ \eta_t = \gamma^t \eta_0, \tag{10} \]

where \( \eta_t \) is the learning rate at epoch \( t \). A typical choice of \( \gamma \) is 0.9. The cosine decaying schedule decreases the learning rate following the cosine function. Given a total number of training epochs \( T \), the learning rate in the cosine decaying schedule is defined as

\[ \eta_t = \frac{1}{2} \left( 1 + \cos \left( \frac{t\pi}{T} \right) \right) \eta_0. \tag{11} \]

The setting of the cosine decaying schedules is independent of the number of epochs, making them more flexible than other schedules.

### 3.4. Composition of Data Augmentation

Applying online data augmentation 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, as shown in Figure 4, 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. Moreover, Krizhevsky color augmentation [39] is evaluated in our experiments, which has been suggested to be effective by the group that ranked third place in the Kaggle DR grading competition [40].

![Figure 4. Illustration of common data augmentation operations.](image)

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 $[0.7, 1.3]$, we then resize this region back to 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.

3.5. Preprocessing

In addition to background removal, two popular preprocessing operations for fundus images are considered in this work, namely Graham processing [41] and contrast limited adaptive histogram equalization (CLAHE) [42]. Both of them can alleviate the blur, low contrast, and inhomogeneous illumination issues that exist in the EyePACS dataset.

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 [43,44] to remove image variations due to different lighting conditions or imaging devices. Given a fundus image $I$, the processed image $\hat{I}$ after Graham is obtained by

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

where $G(\theta)$ is a 2D Gaussian filter with a standard deviation $\theta$, $*$ is the convolution operator, and $\alpha$, $\beta$, $\gamma$ are weighting factors. Following [44], $\theta$, $\alpha$, $\beta$, and $\gamma$ are, respectively, set as 10, 4, $-4$, and 128. As shown in Figure 5, 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) [45], which has also been widely used to process fundus images and has been suggested to be able to highlight lesions [46–48]. HE improves the image contrast by spreading out the most frequently occurring 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 Figure 5.

3.6. Sampling Strategy

As mentioned in Section 2.2, EyePACS is an extremely imbalanced dataset. To address this problem, several sampling strategies [40,49] 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. In this case, the class with more samples than the others can be dominant in the training phase, leading to model bias during testing; (2) Class-balanced sampling first selects each class with an equal probability and then uniformly samples data points from specific classes. In this way, samples in the minority classes are given more attention for better representation learning; (3) Progressively balanced sampling starts with class-balanced sampling and then exponentially moves to instance-balanced sampling. Please note that we follow the interpolation strategy adopted by [40] instead of the one presented by [49], which linearly interpolates the sampling weight from instance-balanced sampling to class-balanced sampling. Specifically, the sampling weight in this work is defined as

$$p_{i}^{PB}(t) = \alpha t p_{i}^{CB} + (1 - \alpha t) p_{i}^{IB},$$

where $p_{i}^{PB}$, $p_{i}^{CB}$ and $p_{i}^{IB}$ are sampling weights in progressively balanced, class-balanced, and instance-balanced sampling, $t$ indexes the training epoch and $\alpha$ is a hyperparameter that controls the change rate.

![Image](image.png)

**Figure 5.** Representative enhanced fundus images using Graham processing and CLAHE.

### 3.7. Prior Knowledge

For medical image analysis, 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 [16]. Moreover, as demonstrated in Figure 6, the quality of the left and right fields of an eye pair may be different, and it is difficult to identify the grade of a fundus image with poor quality. In this case, information on the eye on the other side may greatly benefit the estimation of the grade of the poor one.

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. The network consists of three linear layers each followed by a 1D max-pooling layer with a stride of 2 and rectified linear unit (ReLU). Considering that the grading criterion for left and right eyes is the same, the feature fusion network only outputs the prediction for one eye and then changes the order of the two feature vectors during concatenation for the prediction of the other eye.
3.8. Ensembling

Ensemble methods [50] 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 [39, 51], 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 [52, 53], 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 final predicted one.

4. Experimental Results

4.1. Influence of Different Input Resolutions

First, we study the influence of different input resolutions using the default setting specified in Section 2.2. The experimental results are shown in Table 1. 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 $128 \times 128$ to $512 \times 512$. Increasing the resolution to $1024 \times 1024$ further improves the test Kappa by another 1.32% but with a large computational cost increase of 64.84 G floating-point operations (FLOPs). Considering the trade-off between performance and computational cost, the $512 \times 512$ input resolution is adopted for all our subsequent experiments.

| Resolution | Training Time | FLOPs | Validation Kappa | Test Kappa |
|------------|---------------|-------|------------------|------------|
| $128 \times 128$ | 1 h 54 m | 1.35 G | 0.6535 | 0.6388 |
| $256 \times 256$ | 2 h 19 m | 5.40 G | 0.7563 | 0.7435 |
| $512 \times 512$ | 5 h 16 m | 21.61 G | 0.8054 | 0.8032 |
| $768 \times 768$ | 11 h 15 m | 48.63 G | 0.8176 | 0.8137 |
| $1024 \times 1024$ | 11 h 46 m (2 GPUs) | 86.45 G | 0.8187 | 0.8164 |
4.2. Influence of Different Objective Functions

We further evaluate the seven objective functions described in Section 3.2. We also evaluate the objective function by combining the Kappa loss and the cross-entropy loss [37]. All objective functions are observed to converge after 25 epochs of training. The validation and test Kappa scores for applying different loss functions are reported in Table 2. 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 observation that using the Kappa loss alone makes the training process unstable and results in inferior performance is consistent with that reported in [37]. The MSE loss takes into account the distance between the prediction and the ground truth, yielding a 2.02% improvement compared to the cross-entropy loss. It gives more penalties for outliers than the MAE loss and the smooth L1 loss, making itself have the highest validation and test Kappa among all the objective functions we consider.

To demonstrate the influence of different objective functions on the distribution of predictions, we present the confusion matrices of the test set for the cross-entropy loss and the MSE loss in Figure 7. Considering the imbalanced distribution of different classes in EyePACS, we normalize the matrices by dividing each value by the sum of its corresponding row. As shown in Figure 7, although employing the MSE loss does not improve the performance of correctly discriminating each category, the prediction-versus-ground truth distance from using MSE is smaller than that from using cross-entropy (e.g., 7.9% of proliferative DR images (Grade 4) are predicted to be normal when using the cross-entropy loss, while only 1.0% when using the MSE loss). That is, the predictions from the model using the MSE loss as the objective function show more diagonal tendency compared to those using the cross-entropy loss, which contributes to the improvement in the Kappa metric. This diagonal tendency is important for DR grading in clinical practice because even if the diagnosis is wrong we expect our prediction to be at least close to the correct one.

![Confusion matrices](image)

**Figure 7.** Confusion matrices from models, respectively, using the cross-entropy loss and the MSE loss as the objective function. All values in the confusion matrices are normalized.

4.3. Influence of Different Learning Rate Schedules

Further on, we study the influence of different learning rate schedules. All experiments are conducted using the baseline setting with the 512 × 512 input resolution and the MSE loss. The experimental results are shown in Table 3. The results demonstrate that except for the exponential decaying schedule, all schedules improve the Kappa on both the validation and test sets and the cosine decaying schedule gives the highest improvement of 0.32% in the test Kappa. A plausible reason for the performance drop caused by the exponential decaying schedule is that the learning rate decreases too fast at the beginning of training. Therefore, the initial learning rate should be carefully tuned when the exponential decaying schedule is employed.
Table 2. DR grading performance of models using different objective functions on EyePACS. $\gamma$ is empirically set to be 2 for the focal loss.

| Loss                  | Validation Kappa | Test Kappa |
|-----------------------|------------------|------------|
| Cross Entropy (CE)    | 0.8054           | 0.8032     |
| Focal ($\gamma = 2$) | 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 3. DR grading performance of models using different learning rate schedules on EyePACS. We set the initial learning rate to 0.001 in all experiments. 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 $\gamma$ to 0.9.

| 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     |

4.4. 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, color jitter, and Krizhevsky color augmentation. 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 4, even a simple composition of geometric data augmentation operations (the third row of Table 4) in the baseline setting can provide a significant improvement of 3.49% on the test Kappa. Each data augmentation operation combined with flipping can improve the corresponding model’s performance. However, the composition of all data augmentation operations considered in this work degrades the DR grading performance because too strong transformations may shift the distribution of the training data far away from the original one. Therefore, we do not simultaneously employ the two color transformations. 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 following experiments.

Table 4. DR grading performance of models using different compositions of data augmentation on EyePACS.

| Flipping | Rotation | Cropping | Color Jitter | Krizhevsky | Val. 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** |
| ✓        | ✓        | ✓        | ✓            | ✓          | **0.8308** | 0.8277     |
| ✓        | ✓        | ✓        | ✓            | ✓          | 0.8247     | 0.8252     |
4.5. 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 [44,47]. Although lesions become more recognizable with the application of the two preprocessing methods, they are not helpful for DR grading. As shown in Table 5, our framework with the Graham method achieves a 0.8227 test Kappa, which is lower than the default setting by about 0.5%. Applying CLAHE also hurts the performance of our framework, decreasing the test Kappa by about 0.7%. Unexpected noise and artifacts introduced by the preprocessing may be a cause of performance degradation in our experiments. As such, no image enhancement is applied in our following experiments.

Table 5. DR grading performance on EyePACS with different preprocessing methods. Our default preprocessing setting consists of background removal and image resizing. The parameters used in the Graham method are set following [44]. The clipping value and tile grid size of CLAHE are, respectively, set to be 3 and 8.

| Preprocessing          | Validation Kappa | Test Kappa |
|------------------------|------------------|------------|
| Default                | 0.8307           | 0.8310     |
| Default + Graham [41]  | 0.8262           | 0.8260     |
| Default + CLAHE [42]   | 0.8243           | 0.8238     |

4.6. Influence of Different Sampling Strategies

Further, we are concerned about the influence of different sampling strategies. To alleviate the imbalance issue in EyePACS, class-balanced sampling, and progressively balanced sampling are conducted in the training phase. However, as illustrated in Figure 8, because we repeatedly sample data points from the minority classes at each epoch, overfitting results in poor performance on the validation set. 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 strategies may be more effective in cases where the training set is imbalanced and the test set is balanced [49].

4.7. Influence of Feature Fusion of Paired Eyes

We evaluate the improvement resulting from utilizing the correlation between the paired two eyes for DR grading. The best model from previous evaluations is fixed and adopted to generate a feature vector of each fundus image. A simple paired feature fusion network described in Section 3.7 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 6, 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.

4.8. Influence of Different Ensemble Methods

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 amount 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.
Figure 8. 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.

Table 6. The performance of models on EyePACS for stacking refinements one by one. The first row is the result of the baseline we describe in Section 2.2. HR, MSE, CD, DA, PFF, and ENS, respectively, denote the application of high resolution, MSE loss, cosine decaying schedule, data augmentation, paired feature fusion, and ensemble of multiple models.

| HR  | MSE  | CD  | DA  | PFF  | ENS  | Validation Kappa | Test Kappa | Δ Test Kappa |
|-----|------|-----|-----|------|------|------------------|------------|--------------|
| ✓   | ✓    | ✓   | ✓   | ✓    | ✓    | 0.7563           | 0.7435     | 0%           |
| ✓   | ✓    | ✓   | ✓   | ✓    | ✓    | 0.8054           | 0.8032     | +5.97%       |
| ✓   | ✓    | ✓   | ✓   | ✓    | ✓    | 0.8207           | 0.8235     | +2.03%       |
| ✓   | ✓    | ✓   | ✓   | ✓    | ✓    | 0.8258           | 0.8267     | +0.32%       |
| ✓   | ✓    | ✓   | ✓   | ✓    | ✓    | 0.8307           | 0.8310     | +0.43%       |
| ✓   | ✓    | ✓   | ✓   | ✓    | ✓    | 0.8597           | 0.8581     | +2.71%       |
| ✓   | ✓    | ✓   | ✓   | ✓    | ✓    | **0.8660**       | **0.8631** | +0.50%       |

Table 7. The performance of models with different ensemble methods on EyePACS.

| # 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** |

4.9. 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 6. We observe three significant improvements that stand out from that table. First, increasing the input resolution from $224 \times 224$ to $512 \times 512$ gives the highest improvement of 5.97%. Then, the choice of the MSE loss and utilization of the eye pair fusion, respectively, improve the test Kappa by another 2.03% and 2.71%. Additional improvements of 0.32%, 0.43%, and 0.5% on the test Kappa are obtained by applying a cosine decaying schedule, data augmentation, and ensemble (multiple models). Note that the incremental results alone do not completely reflect the importance of different components. The baseline configuration may also affect the corresponding improvements. In Figure 9, 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.

![Figure 9. 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.](image)

4.10. 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 or annotations are tabulated in Table 8. Our proposed method, without any fancy technique, outperforms previous state-of-the-art results by 0.91% in terms of the test Kappa.

We then visualize our results using Grad-CAM [54]. As illustrated in Figure 10, representative results of four eye pairs corresponding to the four DR grades from 1 to 4 are provided. It reveals that our method’s performance in DR grading may be a result of its ability to recognize different signs of DR, namely lesions. We observe that the region of the heatmap in a severe DR image is usually larger than that in a mild one because the amount of lesions to some degree reflects the DR grade and the lesions are what the network focuses on.
Figure 10. Visualization results from GradCAM. Representative eye pairs of four grades (mild DR, moderate DR, severe DR, and proliferate DR) are presented from top to bottom. The intensity of the heatmap indicates the importance of each pixel in the corresponding image for making the prediction.

Table 8. Comparisons with state-of-the-art methods on EyePACS with only image-level labels. Symbol ‘-’ indicates the backbone of the method is designed by the corresponding authors. The results listed in the first three rows denote the top-3 entries on Kaggle’s challenge.

| Method       | Backbone | Test Kappa |
|--------------|----------|------------|
| Min-Pooling  | -        | 0.8490     |
| o_O          | -        | 0.8450     |
| RG           | -        | 0.8390     |
| Zoom-in Net [16] | -        | 0.8540     |
| AFN [15]    | -        | 0.8590     |
| CABNet [17] | ResNet-50 | 0.8456     |
| Ours        | ResNet-50 | 0.8581     |
| Ours (ensemble) | ResNet-50 | 0.8631     |

4.11. Generalization Ability of the Refinements

To evaluate the generalization ability of the proposed training setting refinements, two external retinal fundus datasets, Messidor-2 and DDR, are adopted to validate the models using the same training practices. As shown in Table 9, the improvements from each component on these two datasets are in line with the results on EyePACS. Increasing
the image resolution, applying the MSE loss, and the utilization of the eye pair fusion
contribute significant improvements in the test Kappa scores. Incremental improvements
are also observed from the learning rate schedule, data augmentation, and ensemble. Note
that pair feature fusion is not utilized in the DDR dataset because eye pair labels are not
available for that dataset. We observe that the key refinements we have identified for
ResNet-50-based DR grading are shared across different datasets, such as the penalty to
the distance between prediction and ground truth provided by the MSE loss is important
for improving the Kappa metric. These consistent results demonstrate that the proposed
training setting refinements can be generalized to other retinal datasets.

Table 9. The DR grading performance on Messidor-2 and DDR datasets. Paired feature fusion is not
feasible for the DDR dataset because eye pair information is not available for that dataset. HR, MSE,
CD, DA, and PFF, respectively, denote the application of high resolution, MSE loss, cosine decaying
schedule, data augmentation, and paired feature fusion.

| HR | MSE | CD | DA | PFF | Messidor-2 Test Kappa | ∆ Kappa | DDR Test Kappa | ∆ Kappa |
|----|-----|----|----|-----|-----------------------|---------|----------------|---------|
| ✓  | ✓   | ✓  | ✓  | ✓   | 0.7036                | 0%      | 0.7680          | 0%      |
| ✓  | ✓   | ✓  | ✓  | ✓   | 0.7683                | +6.47%  | 0.7870          | +1.90%  |
| ✓  | ✓   | ✓  | ✓  | ✓   | 0.7768                | +0.85%  | 0.8000          | +1.30%  |
| ✓  | ✓   | ✓  | ✓  | ✓   | 0.7864                | +0.96%  | 0.8056          | +0.56%  |
| ✓  | ✓   | ✓  | ✓  | ✓   | 0.7980                | +1.16%  | 0.8326          | +2.70%  |
| ✓  | ✓   | ✓  | ✓  | ✓   | 0.8205                | +2.25%  | -               | -       |

We also evaluate our proposed training settings on EyePACS using different backbones. Some popular model architectures are considered in this work, including a lightweight model MobileNet [55], a deeper model ResNet-101, and two ResNet variants DenseNet-121 [29], ResNeXt-50 [30]. We also look into recently developed transformer-based architectures, including the small-scale Visual Transformer (ViT-S) [56] and small-scale hybrid Visual Transformer (ViT-HS) [57]. Because the architecture of visual transformers is largely different from that of CNNs, we adopt alternative training hyperparameters for our two ViT architectures following [58]. As shown in Table 10, the consistent improvements from the investigated training practices, exerted on DR grading performance, reveal that the proposed practices can be generalized to different network architectures. We observe higher test Kappa scores for network architectures with more advanced designs or higher capacities. Notably, using cosine decaying as a learning rate schedule does not work well on ResNet-101 or ViT-S. The reason may be due to the fact that our proposed refinements and configurations are determined empirically based on ResNet-50, and thus they may not necessarily be optimal for all other network architectures under consideration. Furthermore, we observe that cosine decaying is effective for all architectures without any other refinements, indicating that the order of stacking refinements may also affect the observed contribution of each component. With that being said, we show that our configurations can be a good starting point for tuning training strategies for DR grading.
Table 10. The DR grading performance on EyePACS using different network architectures. Underlining indicates that the improvement from the corresponding new component on that specific architecture is not consistent with that on ResNet-50. HR, MSE, CD, DA, and PFF, respectively, denote the application of high resolution, MSE loss, cosine decaying schedule, data augmentation, and paired feature fusion. MNet, D-121, RX-50, R-101, ViT-S, ViT-HS, respectively, denote MobileNet, DenseNet-121, ResNeXt-50, ResNet-101, small-scale Visual Transformer, small-scale hybrid Visual Transformer. $\kappa$ denotes Kappa score.

| Test Kappa | MNet | D-121 | RX-50 | R-101 | ViT-S | ViT-HS | Avg. $\Delta\kappa$ |
|------------|------|-------|-------|-------|-------|-------|---------------------|
| $\checkmark$ | 0.7517 | 0.7442 | 0.7395 | 0.7414 | 0.6797 | 0.7168 | 0%                 |
| $\checkmark$ | 0.7979 | 0.8046 | 0.8020 | 0.8075 | 0.7864 | 0.8073 | +7.20%             |
| $\checkmark$ $\checkmark$ | 0.8117 | 0.8158 | 0.8189 | 0.8228 | 0.8056 | 0.8256 | +1.57%             |
| $\checkmark$ $\checkmark$ $\checkmark$ | 0.8118 | 0.8255 | 0.8217 | 0.8193 | 0.8019 | 0.8257 | +0.09%             |
| $\checkmark$ $\checkmark$ $\checkmark$ $\checkmark$ | 0.8226 | 0.8336 | 0.8362 | 0.8267 | 0.8215 | 0.8356 | +1.17%             |
| $\checkmark$ $\checkmark$ $\checkmark$ $\checkmark$ $\checkmark$ | 0.8515 | 0.8558 | 0.8566 | 0.8528 | 0.8360 | 0.8524 | +2.15%             |

5. Discussion

Recently, deep learning methods have exhibited great performance on the DR grading task, but there is a trend that deep neural networks today become very large and highly sophisticated, making them difficult to be transferred and extended. Inspired by [59], who states that the exact architecture is not the most important determinant in obtaining a good solution, we present a simple but effective framework without any dazzling design in the network architecture itself. Our proposed framework outperforms several state-of-the-art specifically designed approaches tested on the EyePACS dataset. The promising performance of our proposed framework comes from the right choices of the input resolution, the objective function, the learning rate schedule, the composition of data augmentation, the utilization of the eye pair, and the ensemble of multiple models. We also show that some popular techniques for fundus image-related tasks are not always beneficial for DR grading, such as image enhancement approaches and re-sampling strategies.

In this work, we focus on improving the DR grading performance of ResNet-50 on the EyePACS dataset. All refinements and configurations are determined empirically under that specific setting. Although we demonstrate that our refinements can generalize well to other network architectures and are robust across different datasets, our proposed solutions for DR grading may be still dependent on the property of the specific dataset of interest and the specific network of interest. In other words, our empirically selected parameters may not be the best for other neural network architectures or datasets. For example, the learning rate and its schedule need to be adjusted accordingly to identify the optimal solutions for frameworks using other types of neural networks as the backbones. The data augmentation composition may also need to be modified and the paired feature fusion strategy may not always be applicable to other DR grading datasets, such as the DDR dataset. Nevertheless, we show that our framework and the empirically selected parameters can be a good starting point for the trial-and-error process during method design.

Our framework still has considerable room for improvement. In addition to the components we analyzed, there are other major components in deep learning-based frameworks that are also worthy of being systematically investigated and refined. For example, regularization techniques, such as L1/L2 regularization and dropout [60], are essential to control the complexity of a model of interest to avoid overfitting, which may also affect the DR grading performance. In addition, how we combine different refinements and the order of stacking those different refinements may also have non-trivial impacts on the DR grading performance.

Recently, many specifically designed components have been proposed to further improve the performance of deep learning-based methods using fundus images. Although they go beyond the scope of this work, those specifically designed components may have
great potential in enhancing the performance of DR grading. For example, image quality is an important factor affecting the diagnoses of different ophthalmic diseases. Therefore, image quality enhancement [25,61] may serve as a preprocessing method to improve the DR grading performance. Another direction of improvement relates to the class imbalance issue of the EyePACS dataset. In this work, simple weighted resampling methods [49] are investigated, and the observed overfitting results indicate that these simple resampling methods are of limited help in improving the DR grading performance. Recently, a sophisticated sampling method, Balanced-MixUp [62], has been proposed for imbalanced medical image classification tasks. In Balanced-MixUp, a more balanced training distribution is produced based on the MixUp regularization method [63], and promising results have been reported on the DR grading task. Finally, more advanced data augmentation approaches, such as generative adversarial network-based augmentation approaches [64], may be worthy of exploration to further boost the DR grading performance.

6. Conclusions

In this work, we systematically investigate several important components in deep convolutional neural networks for improving the performance of ResNet-50-based DR grading. Specifically, the input resolution, objective function, learning rate schedule, data augmentation, preprocessing, data sampling strategy, prior knowledge, and ensemble method are looked into in our study. 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. The experimental results yielded from this study are summarized below.

- We raised the ResNet-50 Kappa metric from 0.7435 to 0.8631 on the EyePACS dataset, outperforming other specially designed DR grading methods. The generalization ability of the proposed training practices was successfully established on two external retinal fundus datasets and six other types of network architectures.
- Achieving state-of-the-art performance without any network architecture modification, we emphasized the importance of training setting refining in the development of deep learning-based frameworks.
- Our codes and pre-trained model are publicly accessible at https://github.com/YijinHuang/pytorch-classification (accessed on 1 February 2020). 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.

Author Contributions: Conceptualization, Y.H. and X.T.; methodology, Y.H.; validation, Y.H., L.L. and P.C.; investigation, Y.H. and J.L.; data curation, Y.H. and L.L.; writing—original draft preparation, Y.H.; writing—review and editing, X.T., R.T. and J.L.; supervision, X.T. and R.T.; funding acquisition, X.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Shenzhen Basic Research Program (JCYJ20190809120205578); the National Natural Science Foundation of China (62071210); the Shenzhen Science and Technology Program (RCYX20210609103056042); the Shenzhen Basic Research Program (JCYJ20200925153847004); the Shenzhen Science and Technology Innovation Committee (KCYX2020122117340001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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.

Conflicts of Interest: The authors declare no conflict of interest.
25. Cheng, P.; Lin, L.; Huang, Y.; Lyu, J.; Tang, X. I-secret: Importance-guided fundus image enhancement via semi-supervised contrastive constraining. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 87–96.

26. Wightman, R.; Touvron, H.; Jégou, H. Resnet strikes back: An improved training procedure in timm. arXiv 2021, arXiv:2110.00476.

27. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choie, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 6023–6032.

28. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.

29. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

30. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.

31. Nesterov, Y.E. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. SIAM J. Optim. 1983, 269, 543–547.

32. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.

33. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

34. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 1968, 70, 213. [CrossRef]

35. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

36. de la Torre, J.; Puig, D.; Valls, A. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. Pattern Recognit. Lett. 2018, 105, 144–154. [CrossRef]

37. Fawu, J.D. Detecting Diabetic Retinopathy in Eye Images. Available online: http://defawu.ai/diabetic-retinopathy-detection/ (accessed on 28 July 2015).

38. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. arXiv 2016, arXiv:1608.03983.

39. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 2012, 25, 1097–1105. [CrossRef]

40. Antony, M. Team o_O Solution Summary. Available online: https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15617#latest-373487 (accessed on 29 July 2015).

41. Graham, B. Kaggle Diabetic Retinopathy Detection Competition Report; University of Warwick: Coventry, UK, 2015.

42. Huang, S.C.; Cheng, F.C.; Chiu, Y.S. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. IEEE Trans. Image Process. 2012, 22, 1032–1041. [CrossRef]

43. Quellec, G.; Charrière, K.; Boudi, Y.; Cochener, B.; Lamard, M. Deep image mining for diabetic retinopathy screening. Med. Image Anal. 2017, 39, 178–193. [CrossRef] [PubMed]

44. Yang, Y.; Li, T.; Li, W.; Wu, H.; Fan, W.; Zhang, W. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 11–13 September 2018; Springer: Berlin/Heidelberg, Germany, 2017; pp. 533–540.

45. Huang, K.Q.; Wang, Q.; Wu, Z.Y. Natural color image enhancement and evaluation algorithm based on human visual system. Comput. Vis. Image Underst. 2006, 103, 52–63. [CrossRef]

46. Huang, Y.; Lin, L.; Li, M.; Wu, J.; Cheng, P.; Wang, K.; Yuan, J.; Tang, X. Automated hemorrhage detection from coarsely annotated fundus images in diabetic retinopathy. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1369–1372.

47. Sahu, S.; Singh, A.K.; Ghreira, S.; Elhoseny, M. An approach for de-noising and contrast enhancement of retinal fundus images using CLAHE. Opt. Laser Technol. 2019, 110, 87–98.

48. Datta, N.S.; Dutta, H.S.; De, M.; Mondal, S. An effective approach: Image quality enhancement for microaneurysms detection of non-dilated retinal fundus image. Procedia Technol. 2013, 10, 731–737. [CrossRef]

49. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. arXiv 2019, arXiv:1910.09217.

50. Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. J. Artif. Intell. Res. 1999, 11, 169–198. [CrossRef]

51. Caruana, R.; Niculescu-Mizil, A.; Crew, G.; Ksikes, A. Ensemble selection from libraries of models. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 18.

52. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
53. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

54. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.

55. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv 2017, arXiv:1704.04861.

56. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.

57. Steiner, A.; Kolesnikov, A.; Zhai, X.; Wightman, R.; Uszkoreit, J.; Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. arXiv 2021, arXiv:2106.10270.

58. Yu, S.; Ma, K.; Bi, Q.; Bian, C.; Ning, M.; He, N.; Li, Y.; Liu, H.; Zheng, Y. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 45–54.

59. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. Med. Image Anal. 2017, 42, 60–88. [CrossRef] [PubMed]

60. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 2014, 15, 1929–1958.

61. Zhao, H.; Yang, B.; Cao, L.; Li, H. Data-driven enhancement of blurry retinal images via generative adversarial networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 75–83.

62. Galdran, A.; Carneiro, G.; González Ballester, M.A. Balanced-MixUp for Highly Imbalanced Medical Image Classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 323–333.

63. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. arXiv 2017, arXiv:1710.09412.

64. Zhou, Y.; Wang, B.; He, X.; Cui, S.; Shao, L. DR-GAN: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. IEEE J. Biomed. Health Inform. 2020, 26, 56–66. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.