electronics

MDPI

Article

Automatic Diabetic Retinopathy Grading via Self-Knowledge Distillation

Ling Luo *,†, Dingyu Xue and Xinglong Feng †

College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; xuedingyu@mail.neu.edu.cn (D.X.); fengxinglong@vip.163.com (X.F.)

* Correspondence: lingluo@stumail.neu.edu.cn

† These authors contributed equally to this work.

Received: 17 July 2020; Accepted: 17 August 2020; Published: 19 August 2020

Abstract: Diabetic retinopathy (DR) is a common fundus disease that leads to irreversible blindness, which plagues the working-age population. Automatic medical imaging diagnosis provides a non-invasive method to assist ophthalmologists in timely screening of suspected DR cases, which prevents its further deterioration. However, the state-of-the-art deep-learning-based methods generally have a large amount of model parameters, which makes large-scale clinical deployment a time-consuming task. Moreover, the severity of DR is associated with lesions, and it is difficult for the model to focus on these regions. In this paper, we propose a novel deep-learning technique for grading DR with only image-level supervision. Specifically, we first customize the model with the help of self-knowledge distillation to achieve a trade-off between model performance and time complexity. Secondly, CAM-Attention is used to allow the network to focus on discriminative zone, *e.g.*, microaneurysms, soft/hard exudates, etc.. Considering that directly attaching a classifier after the Side branch will disrupt the hierarchical nature of convolutional neural networks, a Mimicking Module is employed that allows the Side branch to actively mimic the main branch structure. Extensive experiments are conducted on two benchmark datasets, with an AUC of 0.965 and an accuracy of 92.9% for the Messidor dataset and 67.96% accuracy achieved for the challenging IDRID dataset, which demonstrates the superior performance of our proposed method.

Keywords: image classification; convolutional neural network (CNN); diabetic retinopathy (DR); self-knowledge distillation (SKD); attention mechanism

1. Introduction

Diabetic retinopathy (DR) is the predominant manifestation of diabetic microangiopathy, which is one of the complications of diabetes. It is reported that approximately one third of people with diabetes in the United States, Europe and Asia have some degree of DR [1]. It also the major leading cause of blindness and vision defects among working-age adults worldwide [2]. The traditional solution is to have a well-trained clinical ophthalmologist observe fundus imaging and subjectively assess the severity of DR. However, the scarcity of ophthalmologists hinders patients from receiving timely diagnosis and treatment, especially in underdeveloped areas, which eventually leads to irreversible vision loss. With this in mind, an automated computer-aided diagnostic (CAD) system is needed to assist ophthalmologists in the early screening of potential DR, alleviating their labor-intensive workload.

Early research mainly focused on hand-crafted features to represent images, which requires specific domain knowledge. Adarsh et al. [3] used image processing techniques to obtain anatomical and texture features, and then fed them into a multi-class support vector machine (SVM) for classification. In [4], an ensemble-based method for the screening of DR was proposed, which extracted

features from the output of several retinal image processing algorithms, such as image-level, lesion-specific, and anatomical components. Seoud et al. [5] designed a new set of shape features, called Dynamic Shape Features, that do not require precise segmentation of the regions to be classified. These works are effective to a certain extent, but they are very sensitive to noise and artifacts. Moreover, they have domain limitations, which means their generalization is relatively poor.

Recently, convolutional neural networks (CNNs) have demonstrated their outstanding performance in various computer vision tasks, such as image classification [6], semantic segmentation [7], and object detection [8]. As an efficient feature extraction method, CNNs are also applicable to medical imaging, especially DR screening. For example, Zhao et al. [9] presented a model called "BiRA-Net", which employs dual streams to improve performance while introducing attention mechanism for better feature learning. Ni et al. [10] took advantage of the strong correlation of both eyes to improve the prediction accuracy. In addition, selective data sampling is applied to alleviate data imbalance between classes. Considering the "black box" nature of CNN, Wang et al. [11] implemented a visual-interpretable DR grading by introducing the regression activation map after the global average pooling layer. Recently, Li et al. [12] explored the internal relationship between DR and diabetic macular edema (DME) and proposed a novel cross-disease attention network to jointly grading DR and DME. However, the work covered above uses only image-level labels and tends to ignore the impact of lesion-related regions and other prior knowledge on the severity of DR.

In [13], Prasanna et al. pointed out that structures such as microaneurysms, hemorrhages, and soft/hard exudates are closely related to DR and the presence of the above abnormalities determines the DR grade of a patient, as shown in Figure 1. Next, we briefly review some approaches based on DR lesions. Foo et al. [14] used a semi-supervised learning process to obtain segmentation masks, followed by a multitask learning approach to determine DR. Zhao et al. [15] obtained significant performance gains by using vessel priors to guide the attention mechanism of deep-learning architectures. Zhou et al. [16] proposed a collaborative learning method of semi-supervised lesion segmentation and disease grading for medical imaging, where the intervention of the lesion masks can improve the accuracy of classification and enhance the robustness of the model. The drawbacks of these lesion-based approaches [14–16] are that they cannot be trained end-to-end and are too resource-intensive for direct clinical deployment.

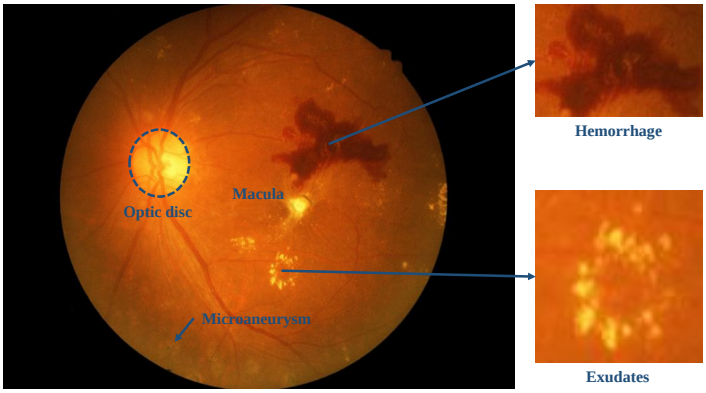

Figure 1. Early pathological signs of DR, such as soft/hard exudates, microaneurysms and hemorrhage. The picture is kindly provided by Messidor database [17], no conflict of interest.

To address the above-mentioned issues, we use a large "teacher network" within the self-knowledge distillation (SKD) [18,19] to guide the compact yet efficient "student network" with only image-level labels, which allows custom pruning of the model according to the actual scenarios in the inference, as shown in Figure 2. Nevertheless, unthinking pruning will disrupt the hierarchical structure of the CNN, so we propose the Mimicking Module (MM) to mitigate it. L_2 loss allows alignment of the block-level outputs between Side branches and the main branch, shortening the spatial distance between them. Furthermore, the off-the-shelf CAM-Attention [20] facilitates the model

to focus on discriminative regions (e.g., lesions), significantly improving the overall performance. For evaluation, we test our method on two publicly available datasets, the Messidor dataset and a new IDRID challenge dataset. Experimental results show that our method outperforms state-of-the-art methods on DR screening. In summary, our contributions of this paper are as follows:

(1) A novel self-knowledge distillation framework is proposed for diabetic retinopathy image grading. It can customize the pruning of the model according to the actual application scenario, which reduces the time delay while not significantly degrading the accuracy.
(2) The introduction of CAM-Attention promotes the model to focus on pathological regions, and the Mimicking Module enables the model to maintain its original hierarchy while pruning. Experimental results confirm that the two proposed modules have a positive effect on the results.
(3) The quantitative and qualitative results on the Messidor and IDRID datasets confirm the effectiveness of the methodology in this paper.

The remainders of this paper are organized as follows. The details of SKD-based DR grading method and its components are presented in Section 2. Section 3 gives experiments on benchmark datasets. Section 4 verifies the effectiveness of each component on the Messidor dataset. Finally, in Section 5, we draw some conclusions.

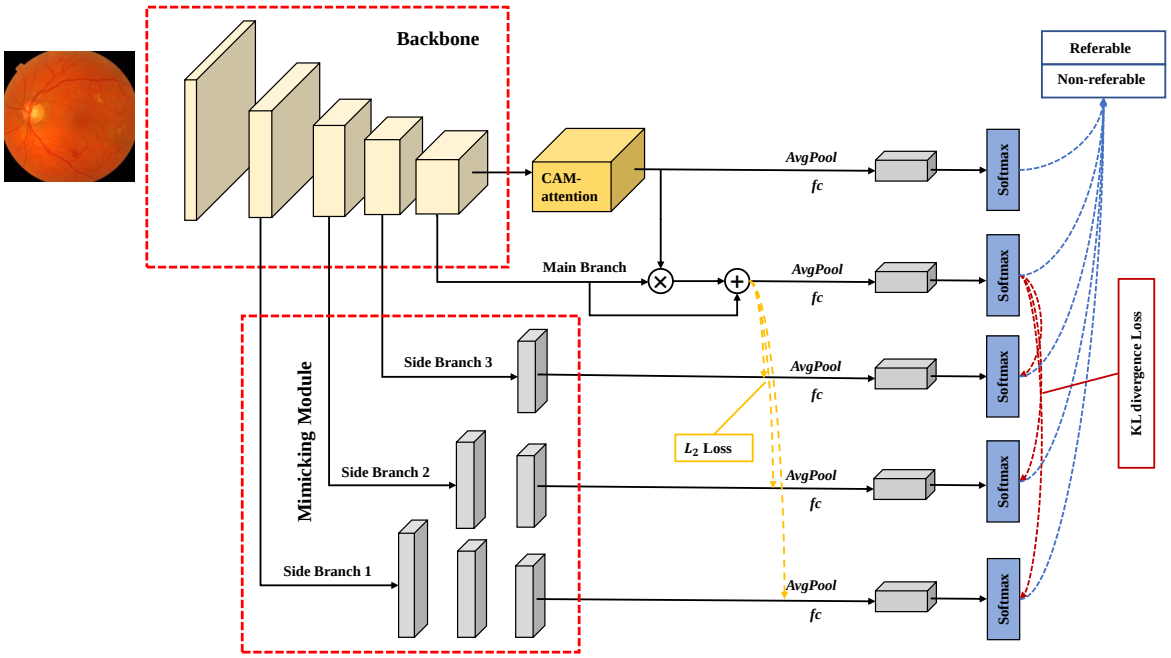

Figure 2. A detailed illustration of our proposed network. In addition to the main branch and the auxiliary attention branch, the proposed framework also has three Side branches attached. Among them, the red dotted box contains multiple groups of *ResBlocks*; *AvgPool* denotes the global average pooling layer; fc denotes the fully connected layer. For binary classification, the images of stage 0 and stage 1 in the Messidor dataset are combined as referable images, and the rest are non-referable images. Backbone and the Mimicking Module will be discussed in Sections 2.3 and 3.2, respectively. Best viewed in color.

2. Methodology

Figure 2 illustrates the overall flowchart of our DR grading method. Our goal is to design a self-knowledge distillation system that integrates scalability and flexibility, which transfers knowledge from an over-parameterized model to compact models, thereby reducing response time to efficiently assist ophthalmologists in the timely diagnosis of potential DR.

2.1. CAM-Attention Module

Although CNN architectures such as ResNet [6] have demonstrated their superior performance on a variety of visual-related tasks, Squeeze-and-Excitation [21] and CBAM [22] components show that by attaching channel or spatial attention components to the backbone, the network can imitate human visual behavior, i.e., focusing on decisive features to achieve outstanding performance gains. Recently, in [20], Fukui et al. extended a response-based visual explanation model named Attention Branch Network (ABN) by introducing attention and perception branches on the basis of Class Activation Mapping (CAM) [23]. Inspired by the work of ABN, we merge it to enhance the "teacher network" representation capabilities while focusing on the discriminative regions. Unlike ABN's approach, CAM-Attention is added after *ResBlock*4 instead of *ResBlock*3, which further reduces time consumption.

Given an input tensor $X_i^0 \in \mathbb{R}^{3\times H_0\times W_0}$ and its corresponding ground truth label $y_i \in \{0, 1, ..., K-1\}$, where i represents the i-th sample and K represents the number of predefined classes. The input tensor first passes through N convolution blocks $\Theta_n^N(\cdot)$ to generate the feature extractor, where intermediate feature maps $X_i^n \in R^{C_n\times H_n\times W_n}$ at the block n can be calculated as $X_i^n = \Theta_n(X_i^{n-1})$. Here, C_n, H_n and W_n represent the number of channels, height and width of the n-th block, respectively. Then, a channel dot-product is performed between the feature extractor and the attention weights to obtain the output of the CAM-Attention $X_i'^n$, which can be formulated as

$$X_i'^n = X_i^n \cdot Atten(X_i^n) + X_i^n \tag{1}$$

where $Atten(\cdot)$ denotes the spatial attention operation. Let $\hat{y}_i^c$ and $\hat{y}_i^m$ be the normalized output logits by $Atten(X_i^n)$ and $X_i'^n$ after passing through the attention branch and the main branch (sequentially traverse a global average pooling layer (GAP), several fully connected layers (FC) and a SoftMax layer, respectively). When the conventional cross-entropy loss L_{CE} is used as the supervision signal, the loss of ABN is as follows

$$L_{ABN} = L_{CE}(\hat{y}^m, y) + \lambda \cdot L_{CE}(\hat{y}^c, y) \tag{2}$$

where λ used to balance them. More details of ABN can be found in [20].

2.2. Self-Knowledge Distillation

Top-performing deep CNN architectures suffer from computational overload, which hinders their further porting to resource-constrained devices. As a trick of model compression, knowledge distillation (KD) [24,25] takes the prediction of probability distribution from a powerful but resource-hungry teacher model as the soft target, combined with one-hot labels to jointly regularize smaller models. However, the paradigm adopted by conventional KD is a two-step optimization, i.e., first training the teacher model and then allowing the learned knowledge flow progressively to the student model by mimicking the probability distribution of the teacher model's output, has the disadvantage of being too costly.

Recently, related work [18,19,26] has shown that teacher and student models can come from the same CNN network, and dynamically transfer knowledge by adding Side classifiers behind some intermediate layers, which is called self-knowledge distillation (SKD). In [27], Lee et al. pointed out that adding auxiliary (Side) classifiers allows the intermediate layer to obtain gradient flows from both the topmost and branch losses, alleviating the "gradient disappearance" problem that occurs in the back propagation of gradients caused by deeper networks, and accelerating the convergence. Let $\hat{y}_j^s$ denotes the j-th Side classifier output, SKD loss can be formulated as

$$L_{SKD} = \sum_j L_{CE}(\hat{y}_j^s, y) + \beta \cdot \sum_j L_{KL}(\hat{y}_j^s, \hat{y}^m) \tag{3}$$

where β is the relative weight between the two loss terms, and L_{KL} represents Kullback–Leibler (KL) divergence between $\hat{y}^s$ and $\hat{y}^m$. Moreover, using a higher value of T in KL results in a softer probability distribution over classes [24].

2.3. Mimicking Module

For the first time, FitNets [28] introduces hints loss from the teacher hidden layers to guide the training process of the student. Nevertheless, due to the inherent hierarchical representation of CNNs [29] (shallow towards detail and deep towards semantics), blindly attaching classifiers to the middle hidden layers as described in [18] without thinking twice would disrupt this structure.

Our insight comes from the mimicking of teacher's teaching, i.e., students receive what they are taught through stage-wise learning, hence we propose a novel Mimicking Module (MM). More specifically, by attaching thinner (fewer bottlenecks) but the same number of *ResBlocks* as the main branch behind the Side branches, the block-level constraints of the main branch (teacher) are used to allow the Side branch (student) to reach block-alignment and hierarchical information sharing during mimicking, while reducing runtime. Below, we describe the mathematical formulation of it.

Let F^l and F^m be the intermediate layer outputs of the l-th branch and the main branch, respectively. Our optimization target is

$$L_{MM} = \sum_{l} \eta \cdot ||F^l, F^m||_2^2 \tag{4}$$

where $||\cdot||_2^2$ refers to the L_2 norm loss and η is a tunable hyper-parameter. Combining Equations (2)–(4), the optimization objective of the entire network can be written as

$$\arg\min_{W} L_{ABN} + L_{SKD} + L_{MM} \tag{5}$$

where W stands for the weight matrices to be optimized.

3. Experiments

3.1. Datasets Descriptions

Messidor. The Messidor dataset [17] contains 1200 color fundus images with DR and DME annotations, in which DR is classified into four classes according to the severity scale. For a fair comparison with previous works [5,12,30–32], we treat images at levels 0 and 1 as referable and the remainder as non-referable, while using 10-fold cross-validation to verify the effectiveness of the model.

IDRiD. The IDRiD dataset [33] comes from ISBI-2018 Challenge 2 (https://idrid.grand-challenge.org/Grading), with a total of 413 training images and 103 test images. We divide it into five classes according to the organizer's rules, and refer the test set as the validation set to evaluate the experimental results.

3.2. Experimental Setup

Our experiments are conducted using Pytorch toolkit and trained on a single NVIDIA Tesla V100 GPU. By default, we use ResNet18 [6] as the backbone and optimize the network with Adam optimizer [34], accompanied by an initial learning rate of 0.0001. The number of *ResBlocks* for the three Side branches is configured as $\{1, 1, 2\}$, $\{1, 2\}$ and $\{2\}$, respectively. For training, a total of 300 epochs for Messidor and 200 for IDRID, while the batch size is set to 40 for both datasets. Moreover, λ, β and η are empirically set to 0.4, 1, and 1×10^{-7} respectively to ensure gradient equalization. We resize the original images to 224×224, while using simple data augmentation, such as horizontal and vertical flips to increase the diversity of the data. It should be noted that to overcome class imbalance, the number of samples for each class in the training batch is the same (using data

re-sampling). In addition, for an analysis of general dataset, see Appendix A.1. Our code is available at: https://github.com/JACKYLUO1991/DR-Grading.

3.3. Results on Messidor Dataset

To evaluate our training strategy, we follow the SKD training procedure and report the results of the comparison between the main branch without/with the help of CAM-Attention and several existing methods on Messidor dataset.

As shown in Table 1, our method has superior performance over state-of-the-art methods in terms of AUC (Area Under the Receiver Operating Curve), Acc. (accuracy), Pre. (precision) and Rec. (recall *a.k.a.* sensitivity) metrics. The quantitative results can be summarized as follows: (1) our method improves the AUC metric by nearly 10% compared to the method [30] using laborious manual feature extraction; (2) compared with methods such as Zoom-in-net [31] that use additional data to improve performance, we still achieve outstanding results with only Messidor's annotations; (3) in contrast to CANet [12], which uses bulky ResNet50 combined with multitask learning, our method uses lightweight ResNet18 while increasing the AUC, accuracy and precision by 0.3%, 0.3% and 0.4% respectively, falling below the former only in the recall metric; (4) compared to the plain SKD, the SKD with CAM-Attention has significantly improved the performance, such as AUC (0.959 vs. 0.966) and Acc. (91.7% vs. 92.9%) metrics, which reflects the positive effect of focusing on pathological regions over outcomes. From a statistical perspective, we give a 95% confidence interval (CI) for AUC, which ranges from 0.953 to 0.979.

Table 1. Performance comparisons on Messidor dataset. Results are given as the mean or (mean ± std) of 10-fold cross-validation. "†" shows that its results are reproduced from [12] and the remaining values are copied from original papers.

Method	AUC	Acc. (%)	Pre. (%)	Rec. (%)
Pires et al. [30]	0.863	-	-	-
VNXK/LGI [32]	0.887	89.3	-	-
CKML Net/LGI [32]	0.891	89.7	-	-
CANet [12]	0.895	81.0	-	-
Comprehensive CAD [35]	0.910	-	-	-
DSF-RFcara [5]	0.916	-	-	-
Expert [35]	0.940	-	-	-
Multitask net [36] †	0.948	89.9	89.7	85.7
MTMR-Net [37]	0.949	90.3	90.0	86.7
Zoom-in-net [31]	0.957	91.1	-	-
CANet + MultiTask [12]	0.963	92.6	90.6	92.0
SKD w/o CAM-Attention (ours)	0.959	91.7	89.3	87.5
SKD (ours)	0.966 ± 0.02	92.9 ± 3.99	91.0 ± 1.72	91.2 ± 2.18

Figure 3 shows the heatmap generated by the last convolutional layer supported by Grad-CAM [38], where the red highlights indicate regions that the model considers decisive for diagnosis.

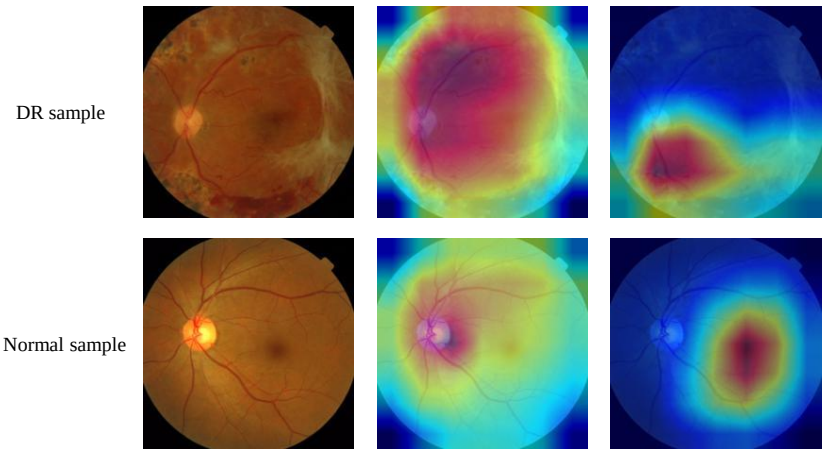

Figure 3. The highlighted regions of the DR decision. From the left to right: the input, highlighting without CAM-Attention and highlighting with CAM-Attention.

3.4. Results on IDRID Dataset

Table 2 summarizes the comparison between our method and those proposed by other challenge participants. To the best of our knowledge, since the competition covers both DR and DME, ref. [39] is the only non-competition research result that provides independent DR grading so far, we list its quantitative results as well. It should be pointed out that our method does not use additional data for pre-training or relying on model ensembles like other solutions.

As can be seen, in the third column of Table 2, our method outperforms the other methods at a smaller input scale, just lower than the solution of Lzyuncc [33] with an input scale of 896 × 896. Moreover, the SKD-refined student (Side branch 1) in the second line has the same classification accuracy (67.96%) as the teacher (main branch), while significantly cutting down the number of parameters, which further confirms the efficiency of our proposed method.

Table 2. Comparison with state-of-the-art results on IDRiD dataset. Our results are bolded in blue. * indicates that the result is obtained from [33]. Consistent with the official evaluation criteria, only the accuracy indicator (unit: %) is given here.

Rank	Method	Main Branch	Side Branch 3	Side Branch 2	Side Branch 1
1	LzyUNCC *	**74.76**	-	-	-
2	SKD(ours)	67.96	66.99	60.19	67.96
3	SUNet [39]	65.06	-	-	-
4	VRT *	59.22	-	-	-
5	Mammoth *	55.34	-	-	-
5	HarangiM1 *	55.34	-	-	-
6	AVSASVA *	54.37	-	-	-
7	HarangiM2 *	47.57	-	-	-

The related confusion matrix of multi-class DR grading is also given which is illustrated in Figure 4. Looking at the confusion matrix, each class is most likely to be predicted correctly except for class 1, which is predominantly classified as class 0. Thus, class 1 is the most difficult to distinguish and its data labeling also confuses experienced ophthalmologists. This problem can potentially be mitigated by using a more powerful network. In [40], Sokolova et al. gave a comprehensive performance measures

for classification tasks. Among them, F1, as the harmonic average of precision and recall, is the most commonly used criterion for multi-classification problems. Mathematically, it can be formulated as:

$$\begin{aligned} macro{-}P &= \frac{1}{m}\sum_{i=1}^{m} P_i \\ macro{-}R &= \frac{1}{m}\sum_{i=1}^{m} R_i \\ macro{-}F_1 &= \frac{2 \times macro{-}P \times macro{-}R}{macro{-}P + macro{-}R} \end{aligned} \tag{6}$$

where m is the number of classes, P_i and R_i denote precision and recall for class i, respectively. Finally 59.98% of F1 can be obtained by calling the scikit-learn library.

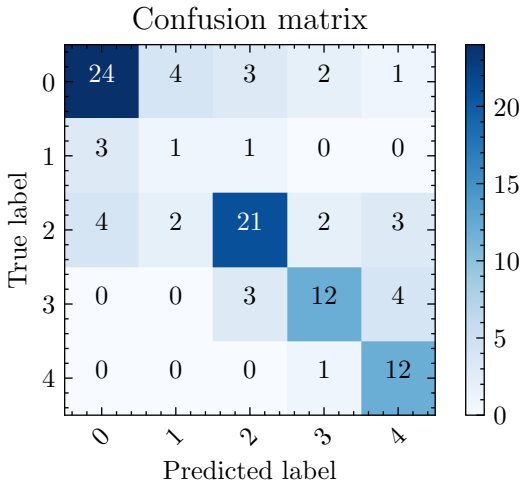

Figure 4. Quantitative results in the form of confusion matrix on IDRID's DR grading dataset with our proposed framework. Horizontal axis indicates the predicted classes and vertical axis indicates the ground truth classes.

4. Analysis and Discussion

4.1. Ablation Studies

Here, fold1 in Messidor dataset is taken as an example to construct ablation experiments for several factors to measure their contributions towards our remarkable results. We regard the main branch without CAM-Attention (CA) as the baseline, and then add the CA module and the MM module in order, as well as training with SKD.

From Table 3 we can conclude that (i) the addition of the CA module promotes the model to focus on pathological features, and the accuracy is improved by 1.66% relative to the baseline, which is also consistent with the conclusion in Section 3.3; (ii) all the four classifiers outperform the baseline in terms of accuracy by virtue of the MM's hierarchical mimicry mechanism and co-optimization, which leverages weight sharing to facilitate performance of the primary task, and (iii) with the help of SKD, students progressively approximate the distribution of the teacher. In particular, the performance of Side branch 1 (94.17% with SKD) is equivalent to that of the teacher without SKD, which verifies the effectiveness of SKD's training strategy.

Table 3. Ablation experiments performed on Messidor dataset according to the accuracy metric (unit: %).

Method	Main Branch	Side Branch 3	Side Branch 2	Side Branch 1
baseline	91.67	-	-	-
baseline + CA	93.33	-	-	-
baseline + CA + MM	94.17	92.50	93.33	93.33
baseline + CA + MM + SKD	95.83	95.83	95.00	94.17

4.2. Efficiency of the Network

One way to get the speed of a model is to simply calculate how many computations it does, which also eliminates the performance difference caused by a specific GPU model or other equipped hardware resources. We typically count this as FLOPs (floating-point operations) and it is inversely proportional to time-consuming.

As shown in Figure 5, the dynamic adaptation of the inference time can be achieved according to the actual scenario, which highlights the advantages of SKD. In particular, compared to the main branch, the Side branch 1 has increased operating efficiency by 1.4 times. It is reasonable to believe that this gap can be more prominent when the model is deeper and the teacher-student block-level compression ratio is increased. Moreover, this technology can be further applied to portable devices to improve their execution efficiency.

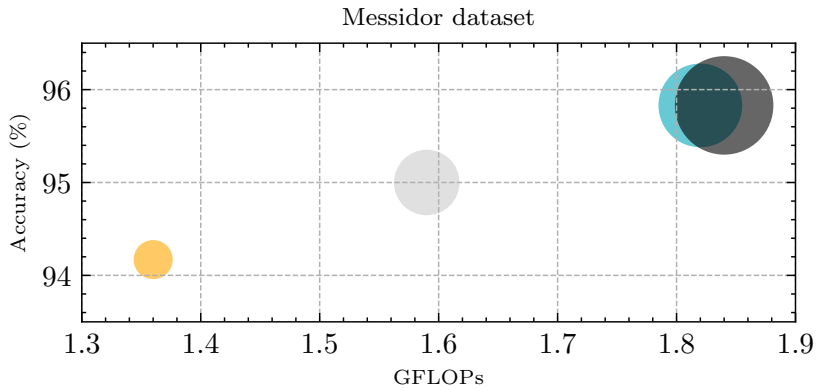

Figure 5. Accuracy-FLOPs curve on Messidor dataset. The circles plotted represent Side branch 1, Side branch 2, Side branch 3, and main branch in order of increasing radius.

4.3. Discussion on Free Parameters Selection

Currently, for the free parameters in Equations (2)–(4), we empirically assign them, with the core idea of unifying the loss terms to the same order of magnitude. Specifically, different loss functions in the same task have very different scales, so it is necessary to consider unifying these scales with weights. Generally, the gradient size of different loss functions is different in the process of model convergence, and the sensitivity to different learning rates is also differentiated. Adjusting different losses to the same order of magnitude can prevent the loss of small gradients from being dominated by the loss of large ones, so that the learned features have better generalization ability. However, the limitation of manual tuning is that it requires repeated trial and error to obtain the optimal value, and the process is usually very cumbersome. Moreover, the results are often sub-optimal.

There are two works worthy of further investigation: one comes from the literature [41], and its basic idea is to estimate the uncertainty of each loss item. Specifically, each loss is divided by the uncertainty, which is basically equivalent to automatically reducing the weight of the corresponding loss. The other comes from an open source project (https://github.com/ultralytics/yolov5) that uses genetic algorithms to search for parameters, which is more efficient than grid search. Since our work

focuses on the proposed SKD distillation method, we will search for free parameters as the direction of future work.

5. Conclusions

In this paper, for diabetic retinopathy grading, we first introduce the CAM-Attention that allows the model to focus on discriminative regions to obtain a powerful teacher network. Then, a training strategy called self-knowledge distillation (SKD) is presented, which enables dynamic adjustment of inference time while improving performance. Finally, considering that attaching classifiers directly after the sharing layers would disrupt the hierarchical consistency between the teacher and students, we propose a Mimicking Module. Experimental results demonstrate that the proposed SKD could boost the performance of the student significantly.

This work can be further applied to resource-constrained devices, e.g., mobile phones, to reduce model inference latency without significant performance degradation. In addition, for automatic medical image screening, our work can relieve the fatigue of ophthalmologists while quickly obtaining diagnosis results. On the other hand, the limitation of this research lies in the optimization of hyper-parameters, which is currently optimized only by manual tuning. Our next work will introduce the genetic algorithm mentioned in [8] to select hyper-parameters. In our future work, we will also focus on semi-supervised as well as weakly supervised learning to eliminate the system's strong dependence on label data, while using graph neural network (GNN) for modeling.

Author Contributions: Conceptualization, L.L. and D.X.; methodology, L.L. and D.X.; software, L.L.; validation, L.L. and X.F.; formal analysis, L.L.; investigation, L.L. and X.F.; resources, D.X.; data curation, L.L.; writing—original draft preparation, L.L.; writing—review and editing, D.X.; visualization, L.L.; supervision, D.X.; project administration, D.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DR	Diabetic retinopathy
DME	Diabetic macular edema
KD	Knowledge distillation
MM	The Mimicking Module
CNN	Convolution neural network
KL	Kullback–Leibler divergence
GNN	Graph neural network
CI	Confidence interval
FLOPs	Floating-point operations

Appendix A

Appendix A.1. Results on CIFAR-100

To further confirm the generalizability of our proposed method, we construct experiments on CIFAR-100 dataset [42]. CIFAR-100 contains 50,000 training sets and 10,000 test sets for a total of 100 classes, with an image size of 32 × 32 pixels. For the fairness of the experiment, data preprocessing, training parameters selection and hyper-parameters configuration are carried out according to [18]. DR classification transfer is done based on ResNet18 and is regarded as the baseline. The left value of the slash comes from [18], and the right value is the result we reproduced.

From Table A1, we can see the advantages of the method in this paper, especially on Side branch 1, which can achieve a balance between speed and accuracy. The method in [18] directly attaches a classifier after a certain Side branch, which on the one hand destroys the consistency of

hierarchical features in CNNs, and on the other hand causes significant performance degradation due to cutting off most of the high-level feature layers. In contrast, our method brings considerable gains (especially an increase of 8.61% in accuracy on Side branch 1), reflecting the powerful feature mapping ability of the Mimicking Module. In addition, multiple branches have been improved in terms of accuracy, indicating that adding CAM-Attention improves the performance of the teacher (main branch) to assist students (Side branches) in learning. These experiments confirm the importance of the knowledge interaction process in promoting the efficiency of sub-branch and improving the baseline performance of a single CNN model.

Table A1. Accuracy (%) comparison on CIFAR-100 dataset.

Method	Main Branch	Side Branch 3	Side Branch 2	Side Branch 1
baseline	77.09/75.61	-	-	-
SKD [18]	78.64	78.23	74.57	67.85
ours	79.01	78.79	77.85	76.46

References

1. Li, T.; Gao, Y.; Wang, K.; Guo, S.; Liu, H.; Kang, H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.* **2019**, *501*, 511–522. [CrossRef]
2. Zheng, Y.; He, M.; Congdon, N. The worldwide epidemic of diabetic retinopathy. *Indian J. Ophthalmol.* **2012**, *60*, 428. [PubMed]
3. Adarsh, P.; Jeyakumari, D. Multiclass SVM-based automated diagnosis of diabetic retinopathy. In Proceedings of the International Conference on Communication and Signal Processing, Melmaruvathur, India, 3–5 April 2013; pp. 206–210.
4. Antal, B.; Hajdu, A. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowl.-Based Syst.* **2014**, *60*, 20–27. [CrossRef]
5. Seoud, L.; Hurtut, T.; Chelbi, J.; Cheriet, F.; Langlois, J.P. Red lesion detection using dynamic shape features for diabetic retinopathy screening. *IEEE Trans. Med. Imaging* **2015**, *35*, 1116–1126. [CrossRef] [PubMed]
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, LV, USA, 26 June–1 July 2016; pp. 770–778.
7. Luo, L.; Xue, D.; Feng, X. EHANet: An Effective Hierarchical Aggregation Network for Face Parsing. *Appl. Sci.* **2020**, *10*, 3135. [CrossRef]
8. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
9. Zhao, Z.; Zhang, K.; Hao, X.; Tian, J.; Chua, M.C.H.; Chen, L.; Xu, X. Bira-net: Bilinear attention net for diabetic retinopathy grading. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1385–1389.
10. Ni, J.; Chen, Q.; Liu C.; Wang, H.; Cao, Y.; Liu, B. An Effective CNN Approach for Diabetic Retinopathy Stage Classification with Dual Inputs and Selective Data Sampling. In Proceedings of the IEEE International Conference on Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1578–1584.
11. Wang, Z.; Yang, J. Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. In Proceedings of the Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, NO, USA, 2–7 February 2018; pp. 514–521.
12. Li, X.; Hu, X.; Yu, L.; Zhu, L.; Fu, C.W.; Heng, P.A. CANet: Cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading. *IEEE Trans. Med. Imaging* **2020**, *39*, 1483–1493. [CrossRef] [PubMed]
13. Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Meriaudeau, F. Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data* **2018**, *3*, 25. [CrossRef]

14. Foo, A.; Hsu, W.; Lee, M.L.; Lim, G.; Wong, T.Y. Multi-Task Learning for Diabetic Retinopathy Grading and Lesion Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–20 February 2020; pp. 13267–13272.
15. Zhao, M.; Hamarneh, G. Retinal image classification via vasculature-guided sequential attention. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27 October–3 November 2019.
16. Zhou, Y.; He, X.; Huang, L.; Liu, L.; Zhu, F.; Cui, S.; Shao, L. Collaborative learning of semi-supervised segmentation and classification for medical images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, LA, USA, 16–19 June 2019; pp. 2079–2088.
17. Decencière, E.; Zhang, X.; Cazuguel, G.; Lay, B.; Cochener, B.; Trone, C.; Gaun, P.; Ordonez, R.; Massin, P.; Erginay, A.; et al. Feedback on a publicly distributed image database: The Messidor database. *Image Anal. Stereol.* **2014**, *33*, 231–234. [CrossRef]
18. Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; Ma, K. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 3713–3722.
19. Li, D.; Chen, Q. Dynamic Hierarchical Mimicking Towards Consistent Optimization Objectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 7642–7651.
20. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, LA, USA, 16–19 June 2019; pp. 10705–10714.
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
22. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, LV, USA, 26 June–1 July 2016; pp. 2921–2929.
24. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
25. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 535–541.
26. Kim, I.; Baek, W.; Kim, S. Spatially Attentive Output Layer for Image Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9533–9542.
27. Lee, C.Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In Proceedings of the Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
28. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
29. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
30. Pires, R.; Avila, S.; Jelinek, H.F.; Wainer, J.; Valle, E.; Rocha, A. Beyond lesion-based diabetic retinopathy: A direct approach for referral. *IEEE J. Biomed. Health Inform.* **2015**, *21*, 193–200. [CrossRef] [PubMed]
31. Wang, Z.; Yin, Y.; Shi, J.; Fang, W.; Li, H.; Wang, X. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 6–11 October 2017; pp. 267–275.
32. Vo, H.H.; Verma, A. New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In Proceedings of the IEEE International Symposium on Multimedia (ISM), San Jose, CA, USA, 11–13 December 2016; pp. 209–215.
33. Porwal, P.; Pachade, S.; Kokare, M.; Deshmukh, G.; Son, J.; Bae, W.; Liu, L.; Wang, J.; Liu, X.; Gao, L.; et al. IDRiD: Diabetic Retinopathy–Segmentation and Grading Challenge. *Med. Image Anal.* **2020**, *59*, 101561. [CrossRef] [PubMed]
34. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

35. Sánchez, C.I.; Niemeijer, M.; Dumitrescu, A.V.; Suttorp-Schulten, M.S.; Abramoff, M.D.; van Ginneken, B. Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. *Invest. Ophthalmol. Visual Sci.* **2011**, *52*, 4866–4871. [CrossRef] [PubMed]
36. Chen, Q.; Peng, Y.; Keenan, T.; Dharssi, S.; Agro, E. A multi-task deep learning model for the classification of Age-related Macular Degeneration. *arXiv* **2018**, arXiv:1812.00422.
37. Liu, L.; Dou, Q.; Chen, H.; Qin, J.; Heng, P.A. Multi-task deep model with margin ranking loss for lung nodule analysis. *IEEE Trans. Med. Imaging* **2019**, *39*, 718–728. [CrossRef] [PubMed]
38. 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.
39. Tu, Z.; Gao, S.; Zhou, K.; Chen, X.; Fu, H.; Gu, Z.; Chen, J.; Yu, Z.; Liu, J. SUNet: A Lesion Regularized Model for Simultaneous Diabetic Retinopathy and Diabetic Macular Edema Grading. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 4–7 April 2020; pp. 1378–1382.
40. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
41. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7482–7491.
42. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. Available online: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (accessed on 17 June 2020).