Laser Scar Detection in Fundus Images Using Convolutional Neural Networks

Qijie Wei[1,2,3], Xirong Li[1,2(✉)], Hao Wang[1,2], Dayong Ding[3], Weihong Yu[4], and Youxin Chen[4]

[1] Key Lab of DEKE, Renmin University of China, Beijing, China
xirong@ruc.edu.cn
[2] AI & Media Computing Lab, Renmin University of China, Beijing, China
[3] Vistel Inc., Beijing, China
[4] Peking Union Medical College Hospital, Beijing, China

Abstract. In diabetic eye screening programme, a special pathway is designed for those who have received laser photocoagulation treatment. The treatment leaves behind circular or irregular scars in the retina. Laser scar detection in fundus images is thus important for automated DR screening. Despite its importance, the problem is understudied in terms of both datasets and methods. This paper makes the first attempt to detect laser-scar images by deep learning. To that end, we contribute to the community *Fundus10K*, a large-scale expert-labeled dataset for training and evaluating laser scar detectors. We study in this new context major design choices of state-of-the-art Convolutional Neural Networks including Inception-v3, ResNet and DenseNet. For more effective training we exploit transfer learning that passes on trained weights of ImageNet models to their laser-scar counterparts. Experiments on the new dataset shows that our best model detects laser-scar images with sensitivity of 0.962, specificity of 0.999, precision of 0.974 and AP of 0.988 and AUC of 0.999. The same model is tested on the public LMD-BAPT test set, obtaining sensitivity of 0.765, specificity of 1, precision of 1, AP of 0.975 and AUC of 0.991, outperforming the state-of-the-art with a large margin. Data is available at https://github.com/li-xirong/fundus10k/.

Keywords: Laser scar detection · Fundus image · Convolutional Neural Network

1 Introduction

Diabetic retinopathy (DR) refers to damages occurring to retinal blood vessels caused by diabetes mellitus. Since the retina is a very vulnerable tissue, such damages could lead to vision loss or even blindness. DR typically progresses through four stages, *i.e.,* mild nonproliferative DR (NPDR), moderate NPDR, severe NPDR and proliferative DR (PDR) [20]. There are 425 million adults

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11364, pp. 191–206, 2019.
https://doi.org/10.1007/978-3-030-20870-7_12

on the planet that suffer from diabetes[1] and more than one-third of diabetic patients are likely to have DR [12]. To fully carry out eye screening programme, especially for countries of large population, automated screening is an inevitable trend.

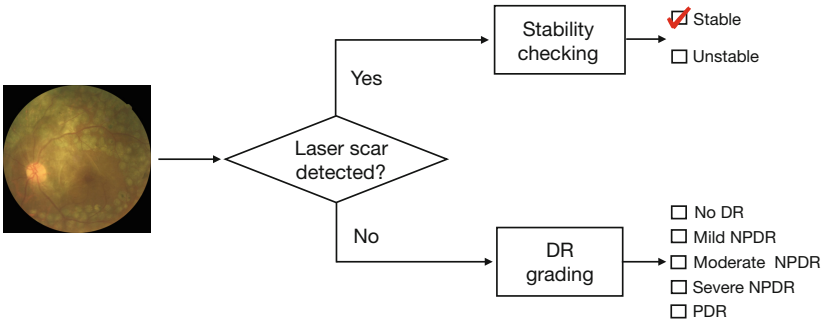


Fig. 1. Diagram of a standard DR screening process, according to the Diabetic Eye Screening Guidance of the NHS, UK [20]. Laser scar detection is an important module for automated DR screening in a real scenario.

(a) No laser scars (b) Fresh laser scars (c) Degenerated laser scars

Fig. 2. Examples of fundus color images of posterior pole, with 45° field of view. Cotton wool spots in (a) resemble fresh laser scars to some extent, while peripapillary atrophy looks like degenerated scars. This paper aims for automated classification of fundus images with and without laser scars. (Color figure online)

Exciting progress is being made on automated DR screening [4,5,14], thanks to the groundbreaking deep learning algorithms for visual recognition [10] and the public availability of large-scale DR-graded data such as Kaggle DR [9]. However, there is an (implicit) condition for these systems to be applicable: *the*

[1] http://www.diabetesatlas.org.

person in consideration has not taken any laser photocoagulation. Laser photocoagulation is a common treatment for severe NPDR and PDR [1], preventing further vision loss by destroying abnormal blood vessels in the retina. According to the Diabetic Eye Screening Guidance of the NHS, UK [20], if there is evidence of previous photocoagulation, judgement should be made differently, see Fig. 1. Due to cauterization of the laser, laser treatment leaves behind circular or irregular scars in the retina, see Fig. 2(b) and (c). Therefore, detecting the presence of laser scars in fundus images is important for automated DR screening in a real scenario.

Despite its importance, the problem of laser scar detection is largely unexplored. Few methods have been developed [3,16,17,19], all relying on hand-crafted visual features. Although fresh laser scars are clearly visible with regular shapes, see Fig. 2(b), they degenerate over time, resulting in irregular boundaries and lower contrast against the background. Moreover, DR lesions such as cotton wool spots resemble fresh scars to some extent, while peripapillary atrophy looks like old scars, as exemplified in Fig. 2(a). All this makes hand-crafted features less effective.

Laser scars are local patterns. They might appear anywhere in a fundus image except for few specific areas including the optic disk, the macular, and main vessels. Meanwhile, they may scatter around a relatively large area. For these two properties we consider a deep Convolutional Neural Network (CNN) appealing for laser scar detection, as the network finds local patterns in its early layers and perform multi-scale analysis in its deeper layers. Probably due to the absence of large-scale labeled laser scar data, we see no effort in developing deep learning techniques for this problem.

In this paper we make the following three contributions.

- First, we present a large-scale dataset consisting of 10,861 fundus images, with expert labels indicating presence or absence of laser scars in each image. The previous largest dataset of this kind has 671 images only[2] [16].
- Second, to the best of our knowledge, this paper is the first deep learning entry for laser scar detection. To reveal what CNNs are the most suited, we systematically investigate major design choices of existing CNNs with good practices identified. In particular, simply and properly adjusting the last pooling layer allows the CNNs to accept input images of a higher resolution, without the need of increasing the number of network parameters.
- Lastly, the proposed deep learning based solution outperforms the best result previously reported on LMD-BAPT [16] (the only public test set). Even though the performance increase can be arguably expected due to the tremendous success of deep learning in varied applications, the optimal use of the technique is task-dependent. By proper adaption of the technique, we establish a new baseline for the task of laser scar detection, which is important for automated diabetic retinopathy screening.

The rest of the paper is organized as follows. We review related work in Sect. 2, followed by a description of the newly constructed dataset in Sect. 3. The

[2] 622 images for training plus 49 images for test [16].

proposed deep learning approach to laser scar detection is depicted in Sect. 4, with its effectiveness verified in Sect. 5. We conclude in Sect. 6.

2 Related Work

There is a paucity of literature on laser scar detection. Dias *et al.* make an initial attempt [3], building a binary classifier with a set of color, focus, contrast and illumination features. A 5-fold cross validation experiment is performed on a dataset composed of 40 fundus images with laser scars and 176 fundus images without laser scars. Syed *et al.* [17] and Tahir *et al.* [19] exploit color, shape and texture based features, with their experiments conducted on a locally gathered dataset consisting of 380 images, among which 51 images have laser scars. More recently, Sousa *et al.* propose to extract geometric, texture, spatial distribution and intensity based features [16], and train a decision tree and a random forest as their laser scar detectors. A common disadvantage of the above methods is their dependency on hand-crafted features which often do not generalize well. Extracting the hand-crafted features involves specifying a number of ad-hoc (and implicit) parameters, making replicability of previous methods extremely difficult, if not impossible. Moreover, previous studies were performed on private datasets, except for [16] where the authors have generously made a training set of 622 images (termed LMD-DRS) and a test set of 49 images (termed LMD-BAPT) publicly accessible. Nevertheless, a large-scale benchmark dataset is missing, making one difficult to understand the state of the art. Probably due to the lack of such a dataset, no effort has ever made to investigate deep learning techniques. We show in Sect. 5 that CNN models trained on the small LMD-DRS dataset do not generalize well.

For deep learning based medical image analysis, some efforts on transfer learning are made [13,15]. Orlando *et al.* adopt two CNNs pre-trained on ImageNet, *i.e.,* OverFeat and VggNet, as feature extractors [13] for glaucoma identification. The CNN weights keep unchanged. For organ localization, Ravishankar *et al.* adopt a CaffeNet pre-trained on ImageNet, reporting that adjusting weights for all layers is better than having weights of some layers fixed [15]. Note that both [13] and [15] use the network architecture of the pre-trained CNNs as is. By contrast, we propose to adjust the last pooling layer. This allows us to double the size of the input, but with the amount of the network parameters unchanged.

3 A Large-Scale Dataset for Laser Scar Detection

In order to construct a large-scale dataset for laser scar detection, we adopted fundus images used in the Kaggle Diabetic Retinopathy Detection task [9]. The Kaggle dataset contains 88,702 fundus color images of posterior pole (with 45° field of view) provided by EyePACS, a free platform for retinopathy screening [2]. To make the subsequent manual labeling manageable, the size of the

Kaggle dataset was reduced to around 11K by random down-sampling. In addition, we gathered from a local hospital 2K fundus color images of posterior pole (also with 45° field of view) of diabetic patients.

For ground-truth labeling, we employed a panel of 45 China licensed ophthalmologists. Each image was assigned to at least three distinct annotators. They were asked to provide a binary label indicting either presence or absence of laser scars in the given image. The number of expert-labeled images was 12,550 in total. As five annotators did not fully complete their assignments, each image has been labeled 2.5 times, approximately. Excluding 1,317 images that were labeled by only one annotator and 372 images receiving diverse labels, we obtained a set of 10,861 expert-labeled images. We term the dataset *Fundus10K*.

We split Fundus10K into three disjoint subsets as follows. We first constructed a hold-out test set by randomly sampling 20% of the images. The remaining data is split at random into a training set of 7,602 images and a validation set of 1,086 images. Table 1 shows data statistics.

Table 1. Laser-scar datasets used in this work. We have constructed a large-scale dataset of 10,861 fundus images with expert annotations. A hold-out test set is constructed by randomly sampling 20% of the images. We term the set Test-2k. In addition, we include LMD-BAPT [16], the only public test set, as our second test set.

	Our contribution (10,861 images)			LMD-BAPT from [16]
	Training (70%)	*Validation (10%)*	*Testing (20%)*	
No. images	7,602	1,086	2,173	49
No. images with laser scars	282	42	80	34

4 Our Approach

We aim to build a CNN that predicts if any laser scar is present in a given fundus image. For a formal description, let x be a specific image and $y \in \{0, 1\}$ as a binary label, where $y = 1$ indicates the image contains laser scars and 0 otherwise. We define $p(y = 1|x)$ as a probabilistic output of the classifier, larger values indicating higher chances of laser scar occurrence. Such a soft classification enables laser scar detection in multiple scenarios. By specifying a particular operating point on a precision-recall curve, one can aim for either high-recall (sensitivity) or high-precision detection. We simply use 0.5 as a decision threshold, *i.e.,* test images having $p(y = 1|x) > 0.5$ will be classified as having laser scars, unless stated otherwise. Also, one might employ $p(y = 1|x)$ as a ranking criterion to retrieve laser-scar images from a large unlabeled collection.

4.1 CNNs for Laser Scar Detection

We express a CNN implementation of $p(y|x)$ as

$$p(y|x) := softmax(\cdots CNN(x)), \tag{1}$$

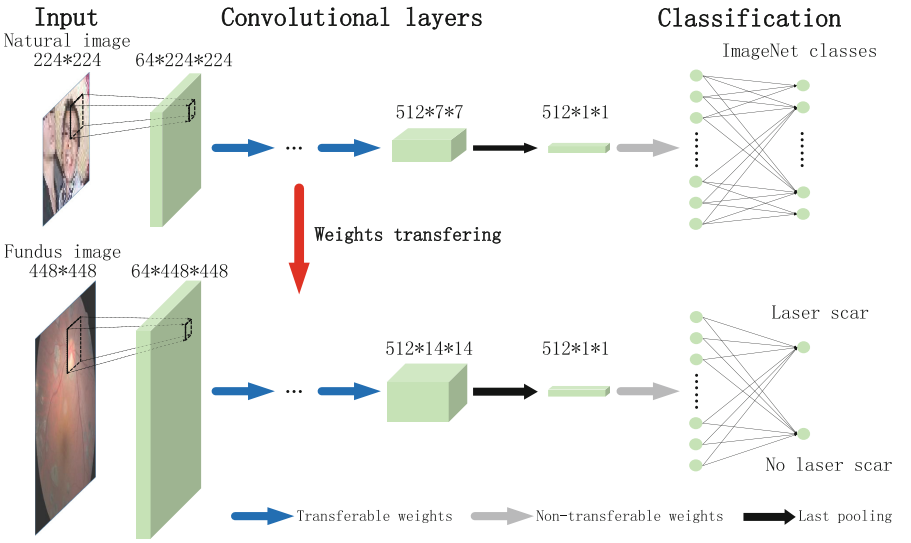


Fig. 3. Diagram of transferring weights from a trained ImageNet CNN model for laser scar detection. The convolutional layers of our laser scar detector are initialized by the corresponding weights from a pre-trained (ResNet-18) model. By adjusting the last pooling layer, we use a double-sized input (448×448), without increasing the number of trainable parameters. Best viewed in color.

where $\cdots CNN$ indicates stacked CNN layers that take a fix-sized RGB image as input and sequentially produce an array of feature maps that describe the visual content at multiple scales. The $softmax$ module employs fully connected layers to convert the last feature map (and optionally some preceding feature maps if skip connections are used) to final predictions.

Choices of CNNs. It remains open what CNN architectures are suited for laser scar detection. Hence, instead of inventing new architectures, we investigate existing options. We consider Inception-V3 [18], ResNet [7], and DenseNet [8], for their state-of-the-art performance on the ImageNet visual recognition task. To reveal the influence of the network depth on the performance, we exploit ResNet-18, ResNet-34, ResNet-50, DenseNet-121, DenseNet-169 and DenseNet-201. The three DenseNet networks have nearly the same architecture, with deeper network repeating a common convolutional block for more times. The case is similar in ResNet, except that ResNet-50 uses so-called BottleNeck convolutional blocks which are deeper but with less parameters.

4.2 Transfer Learning

Instead of training CNNs from scratch, we aim for a better starting point by transferring weights from their counterparts pre-trained on ImageNet. A

straightforward strategy is to follow exactly the same configuration as the existing models, by enforcing the size of the input image to be the de facto 224×224. This strategy is unlikely to be optimal, because a fundus image has a much larger resolution than a consumer picture. Since laser scars are not so big, resolution may affect the ability of the CNN to distinguish them. However, a double-sized input means the feature maps will be four times as large as the original ones. Consequently, the amount of parameters in the first fully connected layer will increase substantially. We consider a simple yet effective trick: adjusting the last pooling layer to maintain the size of the last feature map. The adjustment varies over CNNs. As for ResNet, DenseNet and Inception-v3, they all use global average pooling as their last pooling layer. So we double the pooling size, from 7×7 to 14×14 for ResNet and DenseNet and from 12×12 to 24×24 for Inception-v3. We refer to Fig. 3 for a conceptual illustration.

5 Evaluation

5.1 Experimental Setup

Training Procedure. We use SGD with a mini-batch of 20, a weight decay factor of 1×10^{-4}, and a momentum of 0.95. The initial learning rate is set to be 1×10^{-3}. Validation is performed every 800 batches. An early stop occurs when the validation performance does not improve in 10 consecutive validation steps. For data augmentation we perform random rotation, crop, flip and changes in brightness, saturation and contrast. As the two classes are highly imbalanced, we over sample the laser-scar images to make each batch nearly balanced.

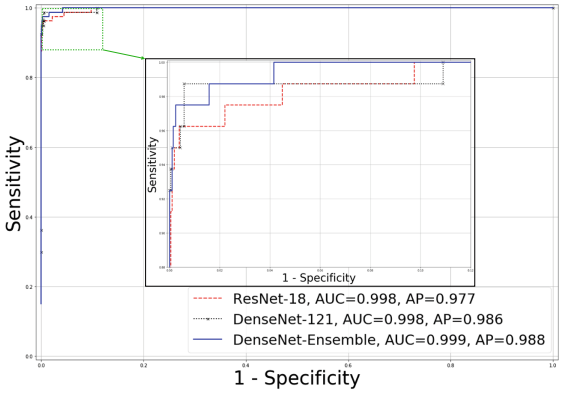
CNN Ensemble. As the SGD based training yields models of slightly different performance, for each CNN we repeat the training procedure three times and use model averaging to obtain its final prediction. Considering that CNNs with varied depth might be complementary to each other, we further investigate two ensembled CNNs, namely ResNet-Ensemble which equally combines ResNet-18, ResNet-34 and ResNet-50 and DenseNet-Ensemble which combines the three variants of DenseNet.

Performance Metrics. We report *Sensitivity*, *Specificity* and AUC as commonly used as performance metrics of a screening or diagnostic test. In particular, we consider laser-scar images as positive instances, and images without laser cars as negatives. As such, *Sensitivity* is defined as the number of correctly detected positives divided by the number of true positives. *Specificity* is defined as the number of correctly detected negatives divided by the number of true negatives. AUC is the area under a receiver operating characteristic (ROC) curve. Given an extremely imbalanced test set, like our Test-2k with only 3.68% positive examples, AUC and *Specificity* tend to be high and less discriminative. Under this circumstance, *Precision* and *Average Precision* (AP) are better metrics. *Precision* is defined as the number of correctly detected positives divided by the number of images detected as positives. AP is a rank-based measure [11]. Higher numbers of the five metrics indicate better performance.

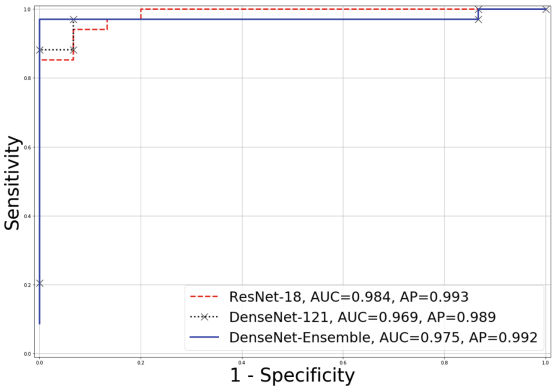
5.2 Experiments

All models are evaluated on our test set of 2,173 images (which we term Test-2k for the ease of reference), unless otherwise stated.

Experiment 1. Choice of CNNs. Table 2 shows performance of different CNNs. Concerning the network architecture, DenseNet leads the performance in terms of AP, followed by ResNet and Inception-v3. Concerning the individual models, the best overall performance of DenseNet-121 suggests that this CNN strikes a proper balance between model capability and learnability for laser scar detection. Its performance can be further improved by model ensembling, as shown in Table 2 and Fig. 4.



(a) Test set: Test-2k



(b) Test set: LMD-BAPT

Fig. 4. ROC curves of ResNet-18, DenseNet-121 and DenseNet-Ensemble on the two test sets. Best viewed in color.

Experiment 2. CNN Initialization Strategies. In order to justify the effectiveness of transfer learning described in Sect. 4.2, we compare CNNs trained with randomly initialized weights and the same models but with their initial weights transfered from their ImageNet counterparts. For random initialization, the weights are initialized using Gaussian distribution with zero-mean and variance calculated according to [6]. We found that when randomly initialized, CNNs with an input size of 448×448 did not converge. So this experiment uses a smaller input size of 224×224. To reduce redundancy, we show only the results of the ResNet series and Inception-v3 in Table 3. DenseNet has similar results. Transfer learning not only leads to better models but also reduces the training time by around 50%.

Table 2. Performance of different CNNs. The input size of each CNN is 448×448, with its initial weights transfered from the corresponding ImageNet model.

Model	Sensitivity	Specificity	Precision	AP	AUC
ResNet-18	0.938	0.998	0.949	0.977	0.998
ResNet-34	0.950	0.998	0.938	0.973	0.997
ResNet-50	0.950	0.996	0.905	0.976	0.996
ResNet-Ensemble	0.950	0.997	0.927	0.977	0.998
Inception-v3	0.938	**0.999**	0.962	0.968	0.996
DenseNet-121	0.938	**0.999**	**0.974**	0.986	0.998
DenseNet-169	0.950	0.997	0.916	0.979	0.998
DenseNet-201	**0.962**	**0.999**	0.962	0.987	**0.999**
DenseNet-Ensemble	0.950	**0.999**	**0.974**	**0.988**	**0.999**

Table 3. Performance of CNNs trained with and without transfer learning, respectively. Note that the input size of each CNN is 224×224. Weight transferring consistently improves the performance.

Model	Initialization	Sensitivity	Specificity	Precision	AP	AUC
ResNet-18	Random	0.812	0.991	0.774	0.882	0.980
	Transfer	**0.913**	**0.994**	**0.849**	**0.954**	**0.992**
ResNet-34	Random	**0.887**	0.990	0.780	0.905	0.988
	Transfer	**0.887**	**0.998**	**0.947**	**0.962**	**0.997**
ResNet-50	Random	0.850	0.991	0.791	0.888	0.989
	Transfer	**0.900**	**0.997**	**0.923**	**0.957**	**0.997**
Inception-v3	Random	0.762	0.998	0.938	0.894	0.977
	Transfer	**0.887**	**0.999**	**0.959**	**0.969**	**0.998**

Experiment 3. The Influence of CNN Input Size. Table 4 shows performance of CNNs trained with two input sizes, *i.e.,* 224×224 and 448×448, separately. Using the larger input is more beneficial for less deeper models. Compare ResNet-18 and ResNet-50 for instance. For ResNet-18, increasing the input size lifts its AP from 0.954 to 0.977, while the corresponding number of ResNet-50 increases from 0.959 to 0.976. Enlarging the input further, say up to 896×896, gives a marginal improvement at the cost of much increased GPU memory. Hence, we do not go further in this direction. Additionally, we observe that among the five performance metrics, Precision and AP are more discriminative than the other three.

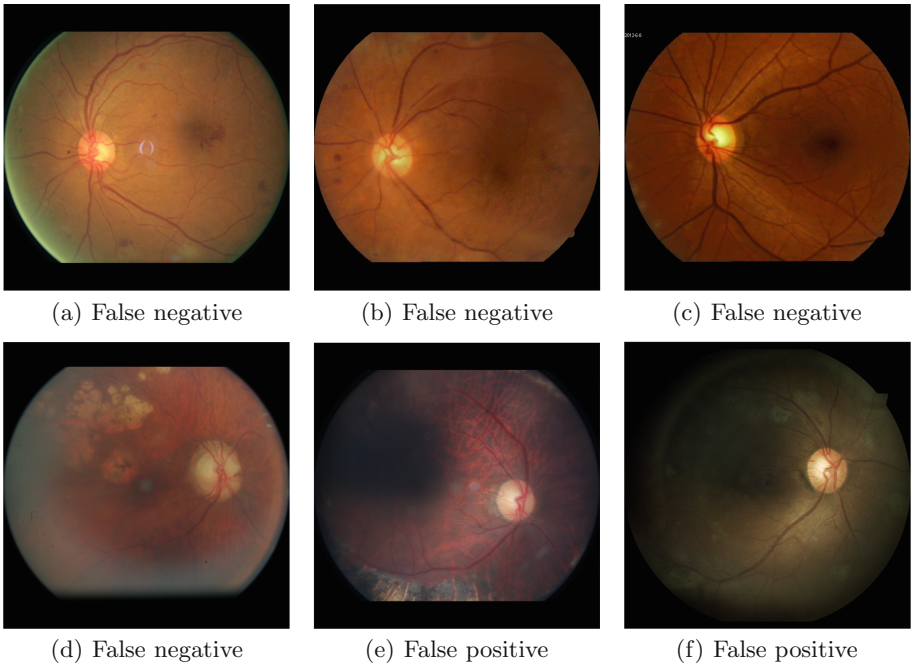
Table 4. Performance of CNNs with two different input sizes. The initial weights of each CNN is passed from the corresponding ImageNet model. Given the same CNN architecture, the larger input tends to be more helpful for less deeper models.

Model	Input size	Sensitivity	Specificity	Precision	AP	AUC
ResNet-18	224×224	0.913	0.994	0.849	0.954	0.992
	448×448	**0.938**	**0.998**	**0.949**	**0.977**	**0.998**
ResNet-34	224×224	0.887	**0.998**	**0.947**	0.962	**0.997**
	448×448	**0.950**	**0.998**	0.938	**0.973**	**0.997**
ResNet-50	224×224	0.900	**0.997**	**0.923**	0.959	**0.997**
	448×448	**0.950**	0.996	0.905	**0.976**	0.996
ResNet-Ensemble	224×224	0.925	**0.998**	**0.937**	0.964	0.996
	448×448	**0.950**	0.997	0.927	**0.977**	**0.998**
Inception-v3	224×224	0.887	**0.999**	0.959	**0.969**	**0.998**
	448×448	**0.938**	**0.999**	**0.962**	0.968	0.996
DenseNet-121	224×224	0.913	0.998	0.948	0.963	0.993
	448×448	**0.938**	**0.999**	**0.974**	**0.986**	**0.998**
DenseNet-169	224×224	0.925	**0.999**	**0.961**	0.970	0.994
	448×448	**0.950**	0.997	0.916	**0.979**	**0.998**
DenseNet-201	224×224	0.913	0.996	0.901	0.973	0.997
	448×448	**0.962**	**0.999**	**0.962**	**0.987**	**0.999**
DenseNet-Ensemble	224×224	0.913	**0.999**	0.961	0.970	0.997
	448×448	**0.950**	**0.999**	**0.974**	**0.988**	**0.999**

Experiment 4. Comparison to the State-of-the-Art. For existing methods [3,17,19], their code and data are not publicly available. As they rely heavily on low-level image processing with implementation details not clearly documented, it is difficult to replicate the methods with the same preciseness as intended by their developers. So we do not include them for comparison. Alternatively, we

Table 5. Laser scar detection performance on LMD-BAPT. High AP indicates the sensitivity of our CNN models can be further optimized, see also ROC curves in Fig. 4.

Model	Sensitivity	Specificity	Precision	AP	AUC
Decision tree [16]	0.618	0.933	–	–	–
Random forest (500 trees) [16]	0.676	0.867	–	–	–
This paper					
Fine-tuned ResNet-18	0.765	0.933	0.963	0.955	0.878
ResNet-18	0.706	**1.0**	**1.0**	**0.993**	**0.984**
DenseNet-121	0.765	**1.0**	**1.0**	0.989	0.969
DenseNet-Ensemble	0.765	**1.0**	**1.0**	0.992	0.975
DenseNet-Ensemble (*decision threshold*: 0.216)	**0.971**	**1.0**	**1.0**	0.992	0.975



(a) False negative (b) False negative (c) False negative

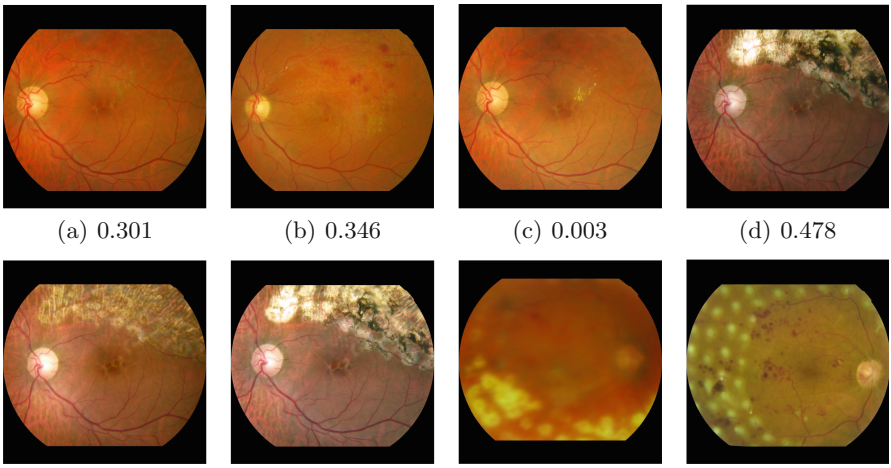
(d) False negative (e) False positive (f) False positive

Fig. 5. Misclassification by DenseNet-Ensemble on the Test-2k test set. False negative images (a)–(d) are severely degenerated laser scars. False positive image (e) is peripapillary atrophy visually similar to laser scars, while False positive image (f) is affected by dirty marks from camera lens. Best viewed digitally in close-up.

add a fine-tuning baseline that uses pre-trained ResNet-18 as feature extractor, as used by [13] for glaucoma identification. To the best of our knowledge, LMD-DRS and LMD-BAPT from [16] are the only two laser-scar datasets that are publicly accessible, with LMD-BAPT as a test set. As Table 5 shows, our CNN models surpass the state-of-the-art. Recall that our decision threshold is 0.5.

As indicated by the ROC curve of DenseNet-Ensemble in Fig. 4, an adjusted threshold of 0.216 would yield sensitivity of 0.941 and specificity of 1.

For a more intuitive understanding, all misclassification by DenseNet-Ensemble are shown in Figs. 5 and 6. In particular, the number of misclassified images is six and eight on our test set and LMD-BAPT, respectively. Misclassification is largely due to severe degeneration of laser scars, making them either mostly invisible or visually similar to peripapillary atrophy.



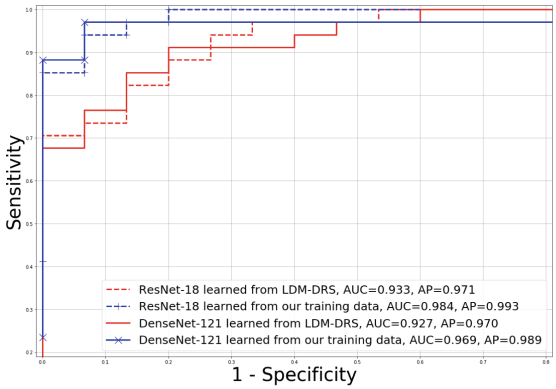
(a) 0.301 (b) 0.346 (c) 0.003 (d) 0.478

(e) 0.365 (f) 0.216 (g) 0.270 (h) 0.290

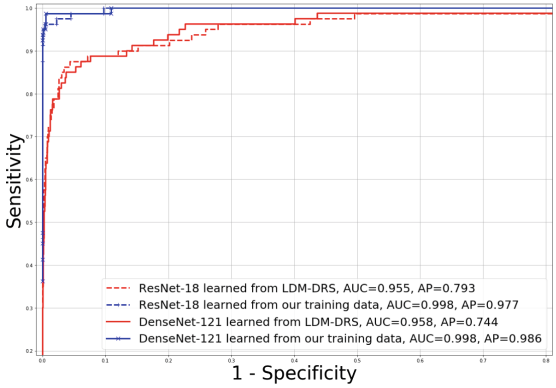
Fig. 6. Misclassification by DenseNet-Ensemble on the LMD-BAPT test set, all false negatives given 0.5 as decision threshold. Score below each image is $p(y = 1|x)$ by DenseNet-Ensemble. Image (a)–(c) are over degenerated, (d)–(f) have large laser scars visually similar to peripapillary atrophy, (g) is fresh laser scars, while (h) is out of focus and obscured.

To further justify the necessity of the newly constructed dataset, we have also trained models using LMD-DRS [16], the only training set that is publicly accessible. As ROC curves in Fig. 7 show, our training data results in better models on both test sets. Moreover, for the models trained on LMD-DRS, a clear drop of their AP scores indicate that they do not generalize well across datasets.

Discussion. As we have noted, given a highly imbalanced test set AUC is less informative. Despite the high AUC of 0.99, for key metrics such as Sensitivity, Precision and AP, there remains room for improvement. The number of positive



(a) Test set: LMD-BAPT



(b) Test set: Test-2k

Fig. 7. ROC curves of ResNet-18 and DenseNet-121 learned from our dataset and the LMD-DRS dataset, respectively. Best viewed in color.

images in Test-2k is only 80, meaning the differences shown between compared methods are only in few misclassified images. This might weaken our conclusion. To resolve the concern, with much efforts we collected 80 new positive images from our hospital partners, and added them to Test-2k. On the expanded test set, which we term Test-2k$^+$, we re-test the previously trained models. The new results are given in Table 6. Similar conclusions can be drawn as those in Experiment 1. That is, DenseNet performs better than ResNet and Inception-v3 in terms of the overall performance, which can be further improved by model ensembling. Miss detections of the newly added test images by DenseNet-Ensemble are shown in Fig. 8.

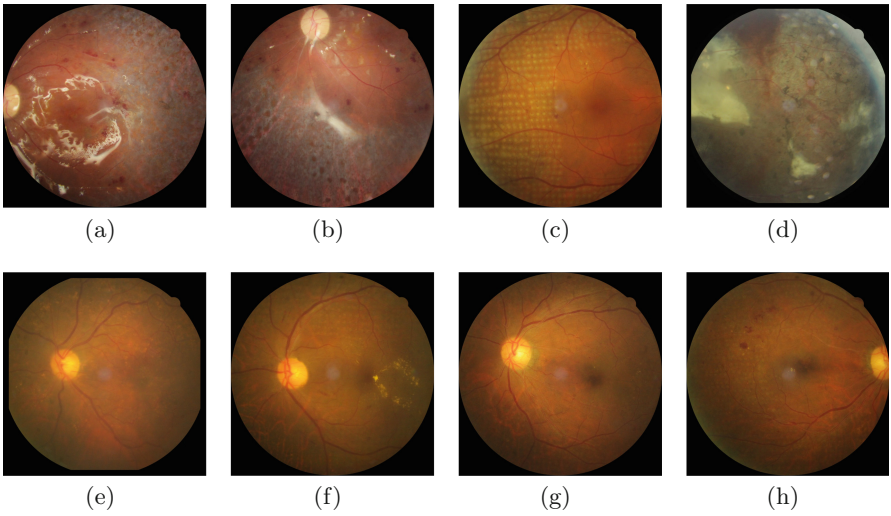


Fig. 8. All eight miss detections in the 80 newly added positive examples, made by DenseNet-Ensemble. Best viewed digitally in close-up.

Table 6. Performance of different CNNs on the expanded Test-2k$^+$ test set. The input size of each CNN is 448×448, with its initial weights transfered from the corresponding ImageNet model. The default decision threshold of 0.5 is used.

Model	Sensitivity	Specificity	Precision	AP	AUC
ResNet-18	0.925	0.997	0.974	0.981	**0.998**
ResNet-34	0.919	0.993	0.967	0.973	0.996
ResNet-50	0.919	0.996	0.948	0.974	0.995
ResNet-Ensemble	0.925	0.997	0.974	0.981	**0.998**
Inception-v3	0.919	**0.999**	0.980	0.968	0.994
DenseNet-121	0.919	**0.999**	**0.987**	0.982	0.998
DenseNet-169	0.931	0.997	0.955	0.977	0.996
DenseNet-201	**0.938**	**0.999**	0.980	**0.983**	**0.998**
DenseNet-Ensemble	0.925	**0.999**	**0.987**	**0.983**	**0.998**

6 Conclusions

We present the first deep learning approach to laser scar detection in fundus images. By performing transfer learning on the recent DenseNet models, a highly effective laser scar detector is developed. The detector obtains Precision of 1.0 and AP of 0.993 on the public LMD-BAPT test set, and Precision of 0.974 and AP of 0.988 on the newly built Test-2k test set. The success of our deep learning approach is largely attributed to the large expert-labeled laser-scar dataset proposed by this work.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61672523), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 18XNLG19). The authors thank anonymous reviewers for their feedbacks.

References

1. AAO: Diabetic retinopathy ppp - updated 2017 (2017). https://www.aao.org/preferred-practice-pattern/diabetic-retinopathy-ppp-updated-2017
2. Cuadros, J., Bresnick, G.: EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. JDST **3**(3), 509–516 (2009)
3. Dias, J., Oliveira, C., da Silva Cruz, L.: Detection of laser marks in retinal images. In: CBMS (2013)
4. Gargeya, R., Leng, T.: Automated identification of diabetic retinopathy using deep learning. Ophthalmology **124**(7), 962–969 (2017)
5. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: ICCV (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Huang, G., Liu, Z., Weinberger, K., van der Maaten, L.: Densely connected convolutional networks. In: CVPR (2017)
9. Kaggle: Diabetic retinopathy detection (2015). https://www.kaggle.com/c/diabetic-retinopathy-detection
10. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
11. Li, X., Uricchio, T., Ballan, L., Bertini, M., Snoek, C., Del Bimbo, A.: Socializing the semantic gap: a comparative survey on image tag assignment, refinement and retrieval. ACM Comput. Surv. **49**(1), 14:1–14:39 (2016)
12. Liu, Y., et al.: Prevalence of diabetic retinopathy among 13473 patients with diabetesmellitus in China: a cross-sectional epidemiological survey in sixprovinces. BMJ Open **7**(1), e013199 (2017)
13. Orlando, J., Prokofyeva, E., del Fresno, M., Blaschko, M.: Convolutional neural network transfer for automated glaucoma identification. In: ISMIPA (2017)
14. Pratt, H., Coenen, F., Broadbent, D., Harding, S., Zheng, Y.: Convolutional neural networks for diabetic retinopathy. Procedia Comput. Sci. **90**, 200–205 (2016)
15. Ravishankar, H., et al.: Understanding the mechanisms of deep transfer learning for medical images. In: Carneiro, G., et al. (eds.) LABELS/DLMIA -2016. LNCS, vol. 10008, pp. 188–196. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_20
16. Sousa, J., Oliveira, C., Silva Cruz, L.: Automatic detection of laser marks in retinal digital fundus images. In: EUSIPCO (2016)
17. Syed, A., Akbar, M., Akram, M., Fatima, J.: Automated laser mark segmentation from colored retinal images. In: INMIC (2014)

18. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
19. Tahir, F., Akram, M., Abbass, M., Khan, A.: Laser marks detection from fundus images. In: HIS (2014)
20. Taylor, D.: Diabetic eye screening revised grading definitions (2012). http://bmec.swbh.nhs.uk/wp-content/uploads/2013/03/Diabetic-Screening-Service-Revised-Grading-Definitions-November-2012.pdf