Computers in Biology and Medicine 111 (2019) 103327

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Self-supervised iterative refinement learning for macular OCT volumetric data classification

Jiaming Qiu[a], Yankui Sun[a,b,*]

[a] Department of Computer Science and Technology, Tsinghua University, 30 Shuangqing Road, Haidian District, Beijing, 100084, P.R. China
[b] Guangdong Key Laboratory of Big Data Analysis and Processing, Guangdong 510006, P.R. China

ARTICLE INFO

Keywords:
Optical coherence tomography
Image classification
Deep learning
Self-supervised learning
Convolutional neural network

ABSTRACT

We present self-supervised iterative refinement learning (SIRL) as a pipeline to improve a type of macular optical coherence tomography (OCT) volumetric image classification algorithms. In this type of algorithms, first, two-dimensional (2D) image classification algorithms are applied to each B-scan in an OCT volume, and then B-scan level classification results are combined to obtain the classification result of the volume. Specifically, SIRL consists of repetitive training–sieving–relabeling steps. In the initialization stage, the label of each 2D image is assigned as the label of the volume they belong to, yielding an initial label set. In the training stage, the network is trained using the current label set. In the sieving and relabeling stage, the label of each 2D image is renewed based on the classification result of the trained network, and a new label set is obtained. Experiments are conducted on a clinical dataset and public dataset, on which the performances of the models trained by a normal scheme and our proposed methods are compared under a five-fold cross validation. Our proposed method achieves sensitivity, specificity, and accuracy of 89.74%, 94.87%, and 93.18%, respectively, on the clinical dataset. On the public dataset, the results of the corresponding three metrics are 98.22%, 90.43% and 95.88%. The results demonstrate the effectiveness of our proposed method as an approach to improve the B-scan-classification-based macular OCT volumetric image classification algorithms.

1. Introduction

The macula of retina, located near the center region of the retina of the human eye, is responsible for fine-grained vision. The structure of the macula is highly hierarchical, with elaborated medical definitions of each layer, from the outermost internal limiting membrane (ILM) layer to the innermost retinal pigment epithelium (RPE) layer [1]. Based on clinical experience, damages to this area caused by aging or diabetes, known as age-related macular degeneration (AMD) and diabetic macular edema (DME), have become the main reasons of blindness in adults. To identify these eye diseases with high confidence, effective imaging techniques are indispensable. Optical coherence tomography (OCT), proposed by Huang et al. in the 1990s, is one of the most commonly used three-dimensional (3D) imaging technologies in clinical practice [2]. It generates cross-sectional volumetric images of the retina with high resolution, from which deformations and lesions inside the macula are much easier to be observed.

The objective of identifying ophthalmic diseases from macular OCT volumes can be regarded as a volumetric image classification problem. Current research studies of fully automated OCT volume image classification algorithms can be chiefly categorized into two types. The first type of methods perform classification directly on the volume level. They first attain the global representations of the whole volume, which are then directly classified. Albarrak et al. proposed diagnosing AMD from OCT volumes using decomposition and local binary patterns (LBPs) [3]. Venhuizen et al. proposed extracting the global representations of an OCT volume by clustering and bag-of-word (BOW) models [4]. Fang et al. introduced a feature extraction pipeline by the combination of PCANet [5] and composite kernels [6]. Rasti et al. proposed OCT volume classification using a wavelet-based convolutional neural network (WCNN) and the random forest classifier [7]. Apostolopoulos et al. directly tiled all the B-scans in a volume vertically in a two-dimensional (2D) plane, transforming a volume into a 2D image, following which they classified the tiled images using a 2D CNN [8]. De Fauw et al. also explored classifying the tissue-segmentation map of an OCT volume [9]. Santos et al. proposed a method that extracted features of OCT volumes from the perspective of C-scan using semivariogram and semimadogram functions [10]. Seebock et al. proposed unsupervised learning of the feature representation of retinal OCT volumes by using deep denoising autoencoders to segment

* Corresponding author.
E-mail address: syk@mail.tsinghua.edu.cn (Y. Sun).

https://doi.org/10.1016/j.compbiomed.2019.103327
Received 20 April 2019; Received in revised form 12 June 2019; Accepted 12 June 2019

anomalous regions, which were then clustered [11]. Sun et al. proposed a multiple instance learning-based support vector machine (SVM) to perform volumetric classification using features extracted by histogram of oriented gradient (HOG) and principal component analysis (PCA) [12].

In comparison to the first type, the second type of methods first perform classification on vertical slices of an OCT image volume called the B-scan, and then yield the classification result of the volume based on that of all the B-scans in the volume. Srinivasan et al. proposed classifying B-scans using HOG as the feature extractor and a linear SVM as the classifier [13]. Sun et al. introduced a framework that could be considered as an enhancement of the work of Srinivasan et al. based on a linear SVM [14,15]. Instead of the HOG descriptor, they used sparse coding, dictionary learning, and spatial pyramid matching to perform feature extraction. Karri et al. proposed transfer-learning-based classification of OCT B-scans, where they fine-tuned a pretrained CNN called GoogLeNet on the OCT images as the B-scan classifier [16]. Another transfer-learning-based method was proposed by Kamble et al., in which they fine-tuned Inception-ResNet-v2 on the OCT B-scans to perform DME diagnosis [17]. To obtain the diagnosis result on the volume level, all the aforementioned four methods use majority voting of the B-scan classification results in a volume. Motivated by the mixture of experts (ME) model and concept of multi-scale spatial decomposition (MSSP) in traditional image processing technologies, Rasti et al. explored a multi-scale CNN ensemble structure [18]. To obtain volume-level diagnosis result, they designed a new strategy instead of using majority voting, in which a volume will be tagged as abnormal if the number of B-scans tagged as abnormal is larger than a given threshold. Perdomo et al. proposed a CNN model called OCT-Net to perform DME diagnosis, where the proposed model was used to classify B-scans and majority voting was used to obtain the volume diagnosis result [19]. There are also several researches of OCT image classification algorithms that are only validated on the B-scan level [20,21].

It is noticed that the majority of the second type of methods need the ground-truth labels of each B-scan to conduct supervised learning. Generally, it is impractical to collect the B-scan level ground-truth labels for large-scale OCT datasets; because only experienced ophthalmologists can label this type of images, this prevents the use of crowdsource annotation. Therefore, a common practice is to use volume-level ground-truth labels as the ground-truth labels of each B-scan in a volume. However, the lesions in a retina are usually local, which implies that not all the B-scans in an abnormal volume have lesions. Pathologically, B-scans that without lesions should be labeled as normal even if they are from an abnormal volume, but this is not the case using the aforementioned B-scan labeling rule. Therefore, this common practice might introduce numerous ”dirty labels”. Considering the ”dirty label” issue, in this paper, we propose a self-supervised iterative refinement learning (SIRL) scheme to improve the performance of B-scan-classification-based macular OCT volume diagnosis algorithms.

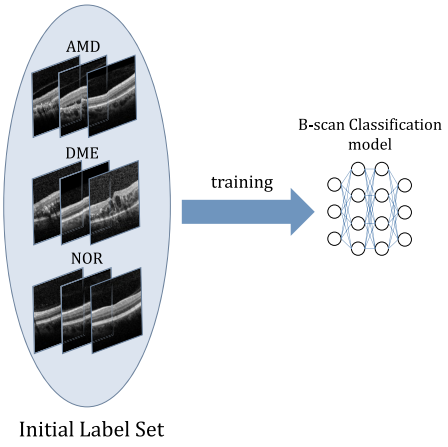
The remainder of this paper is organized as follows. We introduce our method in Sec. 2. In Sec. 3, we describe the experiments conducted on two datasets followed by detailed discussions. The paper is concluded in Sec. 4.

2. Methods

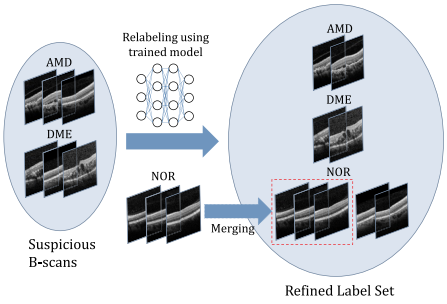
2.1. Self-supervised iterative refinement learning

To alleviate the negative effects caused by the ”dirty label” issue, we propose a new training scheme called SIRL that attempts to determine B-scans with suspicious labels explicitly in the training process and improve the training iteratively. The SIRL scheme can be stated as the following two operations (also see Fig. 1):

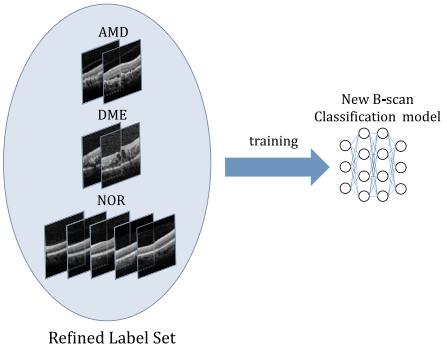
1. Training: A 2D B-scan classifier is trained using the initial label set or label set generated by the relabeling step (see Fig. 1a and c).



(a) Training on the initial label set



(b) Relabeling



(c) Training on the new label set

Fig. 1. Visualized SIRL pipeline: the training–relabeling–retraining process, where the relabeling step and retraining step can repeat several times.

2. Relabeling: A label set is generated based on the classification result of the classifier that trained in the training step (see Fig. 1b).

At the beginning of SIRL, an initial 2D B-scan label set is generated by assigning each B-scan the label of the volume it originates from, following the ”common practice” mentioned in Sec. 1. Then, these two operations proceed iteratively, sequentially, to continuously refine the model and label set. The details of each operation will be discussed in the following subsections.

2.1.1. Training

In the training operation, a 2D B-scan classifier is trained using the initial label set or refined label set. Note that any classifier trained by

supervised learning can be used for SIRL. In this study, we use a CNN as the 2D B-scan classifier, which is a state-of-the-art supervised-learning-based 2D image classification model. Recently, several research works have proved the effectiveness of transfer learning for medical image analysis [16,21,22]. Considering the suggestions from these works, in this study, we also use transfer learning, whose details will be presented in Section 3.3. It is noted that in each step of SIRL, the model is trained from scratch (from the weights obtained by transfer learning) instead of trained from the weights obtained in the previous step.

2.1.2. Relabeling

After training of the network, using the trained model, we sieve all the B-scans in the training set to determine the B-scans whose label might be inconsistent with the volume label with a high probability. Next, we relabel these B-scans based on the prior knowledge contained in the volumetric structure of the data. Specifically, for the B-scans labeled as AMD (extracted from the volumes labeled as AMD), their ground truth labels should be either AMD if they are with lesions, or normal if they are not. The prior is similar for the B-scans labeled as DME. However, for the B-scans labeled as normal, because they are extracted from normal volumes in which no lesions are found, their labels are guaranteed to be correct, and thus, need no sieving. To sieve the B-scans that are mislabeled with a high probability, we apply the following rule to each B-scan labeled as AMD or DME: if its probability for the labeled class is higher than a threshold τ, then its label is kept unchanged, otherwise it is relabeled as normal. After applying the relabeling operation, we obtain a new label set for all the training B-scans, which can be used to retrain the network.

2.1.3. Stopping rule

While applying SIRL, a stopping rule is also needed to determine whether the refining process should stop. For this purpose, we split off a validation set from the training set. After each step, which consists of a training operation and relabeling operation, we perform classification on the validation set using the trained model and one of the volumetric inference strategies mentioned in 2.2. We stop the process when the classification accuracy on the validation set no longer improves. In the test stage, we use the same volume level inference strategy used in the validation stage.

2.2. Volume-level inference strategy

To obtain the volume-level classification result, we need a good strategy to combine the B-scan classification results. Majority voting is the most used volume-level inference strategy in previous research. However, because it is not ensured that in an abnormal volume the number of B-scans with lesions will typically be larger than that without lesions, the majority voting strategy might not be the best strategy to use in cooperation with an accurate 2D B-scan classifier. Rasti et al. have proposed the following rule to obtain volume-level results: if more than n% of the B-scans in a volume are predicted abnormal, then the corresponding class of the maximum abnormal probability among the AMD or DME likelihood scores of all the B-scans determines the type of patient retinal disease, otherwise it is predicted as normal [18]. For convenience, we refer this strategy as Infer-strat-A. In this study, we propose the following strategy. For the class with maximum number of predicted B-scans among all the *abnormal* classes, if more than n% of the B-scans in the volume is predicted as this class, where n is a manually selected threshold, then the volume is predicted as this class. Otherwise it is predicted as normal. We refer this strategy as Infer-strat-B, and n as the proportion threshold.

3. Experimental study and results

3.1. Experiment environments

The experiments in this work are conducted on a machine with an Intel Core i7-7700K 4.20 GHz CPU, 32 GB RAM and an NVIDIA Titan X GPU. The operation system of this machine is Ubuntu 16.04. The codes of all the experiments are written in Python 3.5, and we use PyTorch as the deep learning framework for the training and inference of the CNN, powered by the cuDNN Toolbox.

3.2. Datasets

We conduct our experiments on two datasets. The first one is a clinical SD-OCT dataset provided by Beijing Hospital. This dataset consists of 44 AMD volumes, 42 DME volumes, and 43 normal volumes, and the B-scan number inside a volume varies from 12 to 100. All the B-scans of this dataset are kept in their original forms, without any selection or preprocessing. The second one is a publicly available two-class dataset released by the VIP Lab of Duke University [23]. The SD-OCT volumes in this dataset are acquired by the SD-OCT imaging systems from Bioptigen, Inc(Research Triangle Park, NC), located at four clinic sites. It includes 269 AMD volumes and 115 normal volumes, where each volume has about 100 B-scans. For both the clinical dataset and Duke dataset, a preprocessing pipeline, including saturation removal, retinal pigment epithelium (RPE) layer flattening, resizing, and BM3D filtering, is applied to each B-scan in the volumes, as in Ref. [16].

3.3. B-scan classification model

We adopt the ResNet-101 architecture [24] for the B-scan classification. Specifically, we replace the fully connected layer in the original model by a global average pooling layer and a 1×1 convolution layer, and use the pretrained weights on the ImageNet dataset to initialize the model. We use softmax cross-entropy loss and momentum-based stochastic gradient descent (SGD) to train the model, without fixing any layers during training. While training the adopted ResNet-101 model, random horizontal flip and random rotation from -10° to 10° are applied to augment the training data. The learning rate, α, is set as 0.001, and the momentum factor, β, is set as 0.9. We train the model for five epochs, and α is reduced by a factor of 10 at the fourth epoch. Early stopping is also applied to reduce the risk of over-fitting.

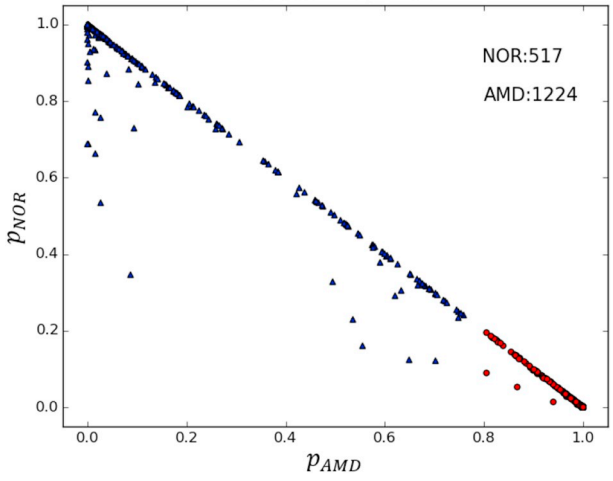
3.4. Evaluation metrics

We use sensitivity, specificity, and accuracy as the performance metrics. In the three-class classification problem, we use the definitions of sensitivity, specificity, and accuracy given in Ref. [6]. It is noted that because our method is not probabilistic on the volume level, i.e., it is unable to provide a confidence probability, we directly compare the sensitivity and specificity of our method to those of other methods instead of applying the receiver operating characteristic (ROC) analysis. We apply a five-fold cross-validation at the volume level for both the datasets. For each method, we repeat the five-fold cross-validation five times and report the mean and standard deviation of each metric. For SIRL, we split off 25% of the training set as the validation set for the use of the stopping rule, i.e., in each fold of the experiment, 60% of the data is used for the training, 20% for the validation, and 20% for the test. For the other compared baseline methods, the entire training set is used for training. The statistical significance of improvement in the metrics is determined using a two-tailed paired t-test with $p = 0.05$.

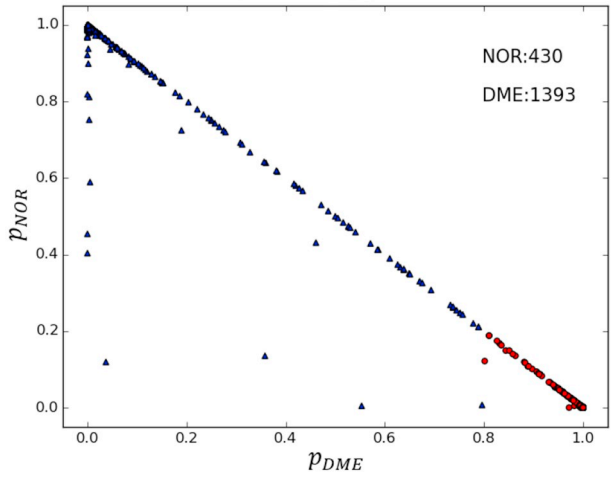
3.5. Experiments on clinical dataset

3.5.1. Quantitative analysis of "dirty labels"

In this study, we conduct a quantitative analysis of the "dirty label"



(a) B-scans labeled as AMD



(b) B-scans labeled as DME

Fig. 2. Distribution of all the training B-scans labeled as abnormal in the probability space. The figures are drawn class-wisely. The B-scans relabeled as "normal" are in blue color, whereas those labeled abnormal are in red color. "NOR" is the abbreviation for "normal". (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

issue. Specifically, we train the adopted ResNet-101 model without SIRL on the training set of a randomly selected fold of the clinical dataset. After training, we re-run the trained network on all the *training B-scans*, which outputs the probability vector, (p_{AMD}, p_{DME}, p_{normal}) for each B-scan, and these vectors can be regarded as points in the probability space. Fig. 2a and Fig. 2b visualize the distributions in the probability space of the training B-scans initially labeled as "AMD" and "DME", respectively.

As shown in the two figures, a reasonable portion of the points is far from the point corresponding to the initial class label, which is (1, 0) in both the figures. In Fig. 2a, for 517 out of the 1741 B-scans, accounting for 29.7% of the total number, the model does not fit well; thus, they are relabeled as "normal". In Fig. 2b, this number is 430 out of 1823, accounting for 23.6%. Fig. 3 shows the relabeled results of two selected volumes. The left one is an AMD volume, whereas the right one is a DME volume. Both the volumes consist of 62 B-scans. The red rectangles represent the B-scans that are labeled as abnormal. After relabeling, only nine B-scans in the left one are still tagged as "AMD", whereas in the right one, only nineteen B-scans are tagged as "DME".

3.5.2. Ablation studies

In this section, ablation studies are performed to investigate the effects of several factors, such as, the use of the SIRL training scheme, relabeling threshold τ, and volume-level inference strategy, on the classification performances.

We first establish the baselines without the use of SIRL. Specifically, we use the ResNet-101 models saved after the first step of SIRL as baseline model, i.e., the model trained on 60% of the total data using the initial label set. The performances of the baseline model using different inference strategies is shown in Fig. 4 and detailed in Table 1. From Fig. 4 and Table 1, the baseline model works best using Infer-strat-B with $n = 30$. Under this setting, the sensitivity, specificity, and accuracy of the baseline model are 84.21%, 92.08%, and 89.46%, respectively. While using Infer-strat-A, the model works best with $n = 40$, and the sensitivity, specificity, and accuracy are 83.33%, 91.63%, and 88.84%, respectively. The performance difference between Infer-strat-A and Infer-strat-B is not statistically significant. Using majority voting, the sensitivity, specificity, and accuracy are 81.77%, 90.85%, and 87.80%, respectively, which also do not show significant difference compared to the results of Infer-strat-A and Infer-strat-B.

After establishing the baseline, we evaluate the performance of the SIRL-trained model together with different inference strategies. τ is set as 0.8 in these experiments. The results are shown in Fig. 5 and detailed in Table 2, from which we determine that when SIRL is applied, the best result is an overall sensitivity of 89.14%, specificity of 94.58%, and accuracy of 92.76% using Infer-strat-B with $n = 10$. The improvement compared to the baseline is statistically significant. It is also noticed that unlike the case of the baseline model, Infer-strat-B performs much better than Infer-strat-A. The reason that Infer-strat-B outperforms Infer-strat-A when SIRL is used may be that Infer-strat-A yields the final classification decision based on only one B-scan (the B-scan with the largest probability in the abnormal class). In comparison, Infer-strat-B relies on the distribution of the B-scan level classification results. Because it is not guaranteed that all the B-scans can be relabeled correctly at the end of SIRL, Infer-strat-B may be more robust.

We also evaluate the effect of the relabeling threshold, τ, on the classification performances using the best inference strategies obtained from Table 2. Table 3 shows that the model performs best using $\tau = 0.9$, achieving a sensitivity of 89.74%, a specificity of 94.87%, and an accuracy of 93.18%. However, we also notice that the difference in the classification performances for $\tau = 0.9$ and $\tau = 0.8$ is not statistically significant. Table 4 compares the performances of the baseline model and SIRL-trained model using the best setting of each model. It shows that our proposed method enhances the classification performance without changing the model architecture by a margin of ~5% in the sensitivity, ~2% in the specificity, and ~3% in the accuracy.

3.5.3. Comparison to state-of-the-art methods

We compare our proposed method to the method proposed by Karri et al. [16], which finetunes GoogleNet pretrained on the ImageNet dataset on the OCT image dataset and uses majority voting as the volume-level inference strategy. Their experiments were originally conducted on another public dataset that consisted of 15 AMD volumes, 15 DME volumes, and 15 normal volumes [13]. We make necessary modification to their released MATLAB code so that it can be run on the clinical dataset. The performance results are listed in Table 5. Our proposed method outperforms the method proposed by Karri et al. by a margin of ~4% in the sensitivity, ~2% in the specificity, and ~2% in the accuracy. The improvements in all the three metrics are considered significant under the statistical significance test. The confusion matrices of the two methods are also shown in Fig. 6, where we only show the result of our proposed method using $\tau = 0.8$ for simplicity. In the figure, the numbers are the average of five independent five-fold cross validation experiments. Our method outperforms the baseline method in

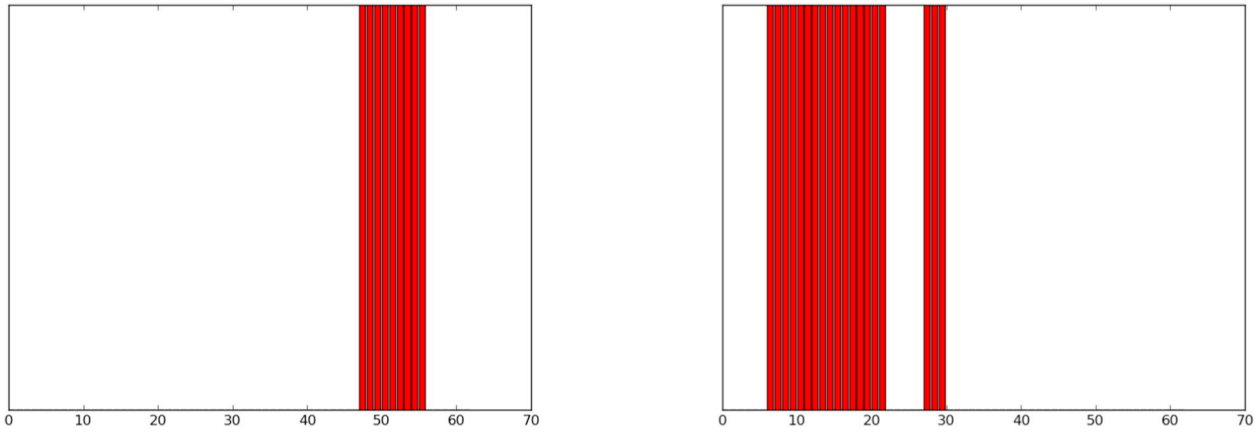


Fig. 3. Relabeling result of two selected volumes, both of which consist of 62 B-scans. The left one is an AMD volume, whereas the right one is a DME volume. The x-axis represents the index of the B-scans, and red areas represent B-scans that are labeled as abnormal. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

each block of the confusion matrix, except for the last two blocks in the last row.

3.6. Experiments on Duke dataset

We compare our proposed method to several baselines, including the methods proposed by Santos et al. [10] and Karri et al. [16] and the ResNet-101 baseline, on the Duke dataset. The results reported by Santos et al. [10] are obtained by a five-fold cross-validation with 100 repetitions. Similar to the description in the previous section, we make necessary modification to the MATLAB code released by Karri et al. so that it can run on the Duke dataset. Unlike the ResNet-101 baseline in the previous section that trained by 60% of the data in each fold, we train the data with the complete training set (80% of the total data) here. For the ResNet-101 baseline model, majority voting is used as the volume-level inference strategy, whereas for the ResNet-101-SIRL model, Infer-strat-B is used with $n = 10$. The performances of these methods are presented in Table 6. Again, we can see that similar performances are achieved using $\tau = 0.8$ and $\tau = 0.9$ on this dataset, which implies that both these values should be able to be used in practice. Compared to the ResNet-101 baseline trained with the full training set, ResNet-101-SIRL exhibits an improvement of ~4% in the sensitivity, ~1% in the specificity, and ~3% in the accuracy, respectively, where the improvements are considered to be statistically significant. Compared to the method proposed by Santos et al., the sensitivity of our method is ~4% higher, whereas the specificity is ~7% lower. The accuracies of the two methods show no statistically significant differences. Compared to the method proposed by Karri et al., our method shows a decrease of ~1% in the sensitivity, ~5% in the specificity, and ~2% in the accuracy.

3.7. Discussion

In this study, we analyze how the current label-assigning convention may adversely affect the final accuracy of the trained model by neglecting the internal variation in OCT image volumes. To deal with this

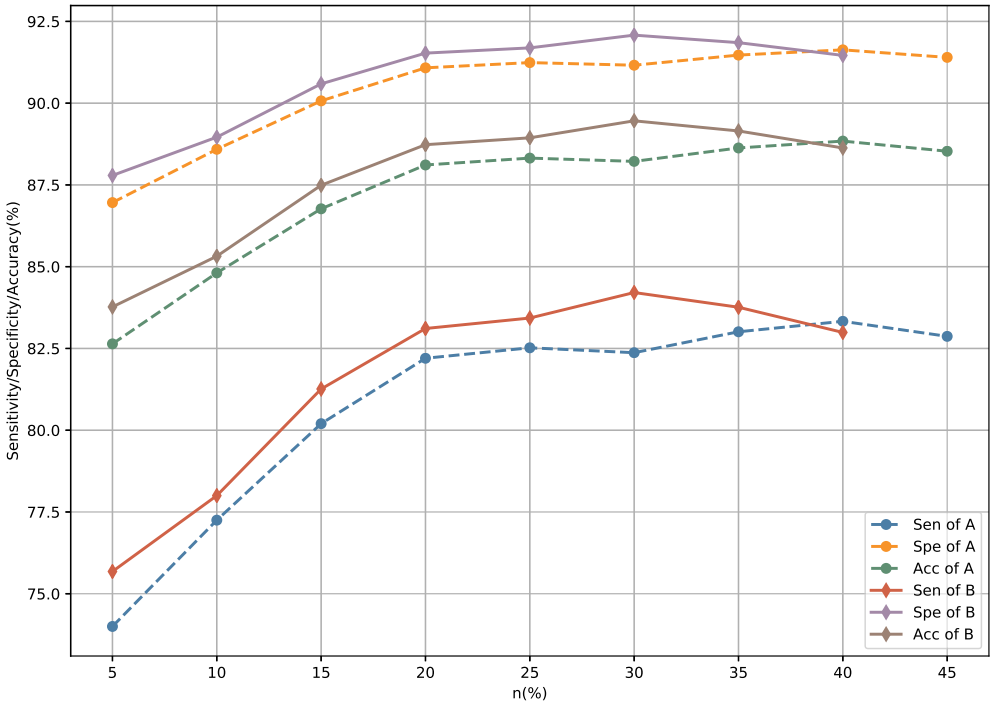


Fig. 4. Performances of the baseline model under different proportion thresholds of Infer-strat-A and Infer-strat-B.

Table 1
Performances of the baseline model under different proportion thresholds of Infer-strat-A and Infer-strat-B.

Model	Inference Strategy	Sensitivity(%)	Specificity(%)	Accuracy(%)
ResNet-101-Baseline	Majority Voting	81.77 ± 3.27	90.85 ± 1.64	87.80 ± 2.18
	5% Infer-strat-A	74.00 ± 1.92	86.96 ± 0.98	82.64 ± 1.29
	10% Infer-strat-A	77.25 ± 2.92	88.59 ± 1.47	84.81 ± 1.95
	15% Infer-strat-A	80.20 ± 3.74	90.07 ± 1.88	86.77 ± 2.50
	20% Infer-strat-A	82.20 ± 3.80	91.08 ± 1.91	88.11 ± 2.55
	25% Infer-strat-A	82.52 ± 3.67	91.24 ± 1.85	88.32 ± 2.46
	30% Infer-strat-A	82.37 ± 3.72	91.16 ± 1.88	88.22 ± 2.50
	35% Infer-strat-A	83.01 ± 3.34	91.47 ± 1.69	88.63 ± 2.24
	40% Infer-strat-A	83.33 ± 3.09	91.63 ± 1.56	88.84 ± 2.06
	45% Infer-strat-A	82.87 ± 2.65	91.40 ± 1.34	88.53 ± 1.77
	5% Infer-strat-B	75.68 ± 3.92	87.79 ± 1.98	83.77 ± 2.61
	10% Infer-strat-B	78.00 ± 2.93	88.96 ± 1.49	85.32 ± 1.95
	15% Infer-strat-B	81.26 ± 3.52	90.59 ± 1.77	87.49 ± 2.34
	20% Infer-strat-B	83.11 ± 3.23	91.53 ± 1.63	88.73 ± 2.15
	25% Infer-strat-B	83.43 ± 2.77	91.69 ± 1.39	88.94 ± 1.84
	30% Infer-strat-B	84.21 ± 2.34	92.08 ± 1.18	89.46 ± 1.55
	35% Infer-strat-B	83.76 ± 2.80	91.85 ± 1.40	89.15 ± 1.85
	40% Infer-strat-B	82.99 ± 3.66	91.46 ± 1.83	88.63 ± 2.42

issue, we extend the prevalent single-shot training scheme to a training–sieving–relabeling circulation, where the suspected incorrectly-labeled B-scans are relabeled using the trained model and a new model is trained using the relabeled B-scans. We conduct extensive experiments on two datasets, including one publicly available dataset and one clinical dataset that we directly obtained from the hospital without manual purification.

From the results on the two datasets, we can see that using the same B-scan classification model, the performance typically enhances when the SIRL training scheme is used. In addition, continuity can also be observed from the visualization of the relabeling results in the abnormal volumes, which is consistent with the fact that lesions are 3D tissues that spread through continuous B-scans. Another advantage of our method is that it only adds extra steps in the training stage. For inference, no extra computation is needed.

Table 2
Performances of the SIRL model under different proportion thresholds of Infer-strat-A and Infer-strat-B.

Model	Inference Strategy	Sensitivity(%)	Specificity(%)	Accuracy(%)
ResNet-101-SIRL	Majority Voting	83.27 ± 2.15	91.61 ± 1.07	88.84 ± 1.41
	5% Infer-strat-A	68.11 ± 2.74	84.03 ± 1.39	78.71 ± 1.83
	10% Infer-strat-A	73.83 ± 3.95	86.91 ± 1.98	82.53 ± 2.64
	15% Infer-strat-A	75.76 ± 4.26	87.85 ± 2.13	83.77 ± 2.88
	20% Infer-strat-A	72.91 ± 4.07	86.39 ± 2.06	81.91 ± 2.72
	25% Infer-strat-A	75.29 ± 3.39	87.60 ± 1.70	83.46 ± 2.29
	30% Infer-strat-A	77.12 ± 3.59	88.54 ± 1.80	84.70 ± 2.42
	35% Infer-strat-A	75.57 ± 2.59	87.77 ± 1.29	83.67 ± 1.75
	5% Infer-strat-B	88.36 ± 1.71	94.18 ± 0.85	92.25 ± 1.13
	10% Infer-strat-B	89.14 ± 1.55	94.58 ± 0.77	92.76 ± 1.03
	15% Infer-strat-B	84.98 ± 3.68	92.47 ± 1.86	89.97 ± 2.46
	20% Infer-strat-B	86.97 ± 1.93	93.47 ± 0.97	91.32 ± 1.28
	25% Infer-strat-B	88.09 ± 1.64	94.02 ± 0.84	92.04 ± 1.11
	30% Infer-strat-B	84.38 ± 2.29	92.16 ± 1.16	89.56 ± 1.55
	35% Infer-strat-B	84.51 ± 2.07	92.24 ± 1.04	89.66 ± 1.39

Table 3
Performances of the SIRL model under different values of τ.

Model	Inference Strategy	τ	Sensitivity(%)	Specificity(%)	Accuracy(%)
ResNet-101-SIRL	10% Infer-strat-B	0.96	86.97 ± 0.77	93.49 ± 0.38	91.32 ± 0.51
		0.93	87.76 ± 3.93	93.87 ± 1.95	91.83 ± 2.60
		0.9	89.74 ± 1.51	94.87 ± 0.76	93.18 ± 1.00
		0.8	89.14 ± 1.55	94.58 ± 0.77	92.76 ± 1.03
		0.7	86.69 ± 1.88	93.33 ± 0.96	91.11 ± 1.28
		0.6	87.93 ± 3.13	93.94 ± 1.57	91.94 ± 2.08

On the clinical dataset, our proposed method outperforms the state-of-the-art baselines. However, on the Duke dataset, the performance of our method is not as competitive as some of the state-of-the-art baselines. This might be related to the fact that the ResNet-101 baseline model that we train on the Duke dataset is a relatively weak baseline compared to that in the work of Karri et al. In addition, it might also be related to the fact that the data characteristics of these two datasets are

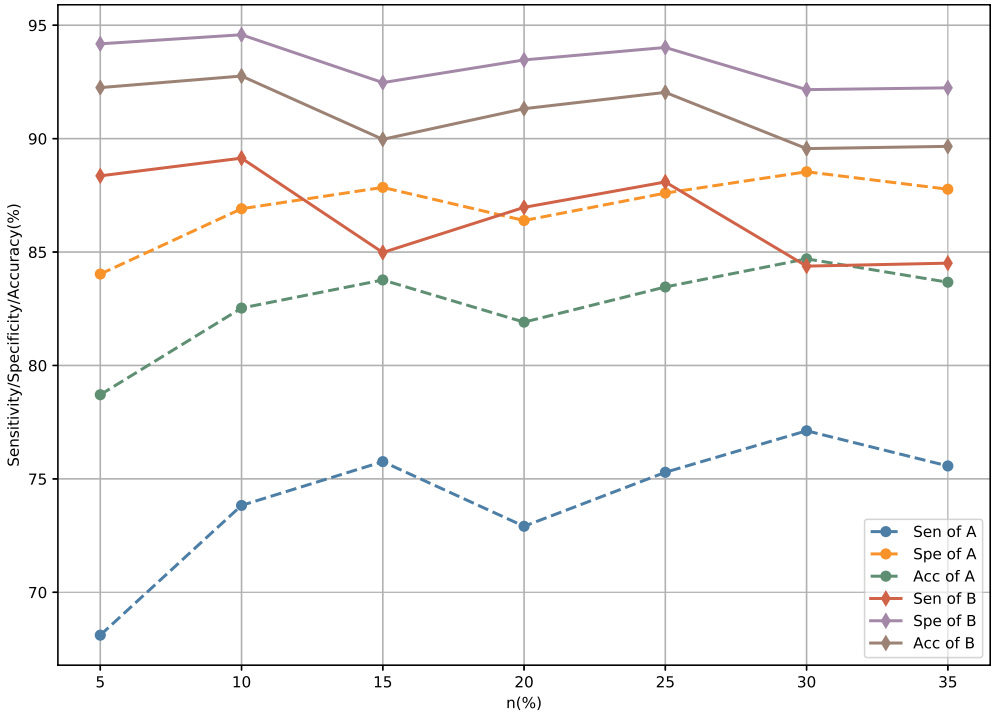


Fig. 5. Performances of the SIRL-trained model under different proportion thresholds of Infer-strat-A and Infer-strat-B.

Table 4
Comparison of the baseline model and SIRL model.

Method	Sensitivity(%)	Specificity(%)	Accuracy(%)
ResNet-101-Baseline + 30% Infer-strat-B	84.21 ± 2.34	92.08 ± 1.18	89.46 ± 1.55
ResNet-101-SIRL(τ = 0.8) + 10% Infer-strat-B	89.14 ± 1.55	94.58 ± 0.77	92.76 ± 1.03
ResNet-101-SIRL(τ = 0.9) + 10% Infer-strat-B	89.74 ± 1.51	94.87 ± 0.76	93.18 ± 1.00

Table 5
Comparison with the method proposed by Karri et al. [16] on the clinical dataset.

Method	Sensitivity(%)	Specificity(%)	Accuracy(%)
Karri et al. [16]	85.78 ± 0.63	92.86 ± 0.31	90.49 ± 0.41
ResNet-101-SIRL(τ = 0.8) + 10% Infer-strat-B	89.14 ± 1.55	94.58 ± 0.77	92.76 ± 1.03
ResNet-101-SIRL(τ = 0.9) + 10% Infer-strat-B	89.74 ± 1.51	94.87 ± 0.76	93.18 ± 1.00

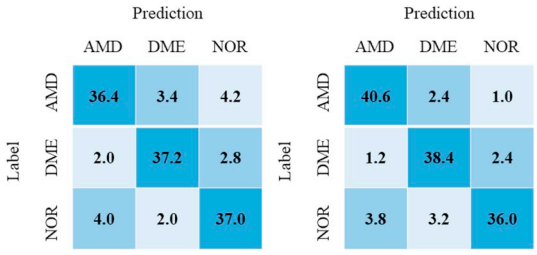


Fig. 6. Confusion matrices of the method proposed by Karri et al. (GoogLeNet with majority voting, left) and our proposed method (ResNet-101-SIRL using τ = 0.8 and 10% Infer-strat-B, right) on the clinical dataset. The numbers in the confusion matrices are the average number of five independent experiments. NOR is the abbreviation for normal.

Table 6
Sensitivities, specificities, and accuracies of the different methods on the Duke dataset.

Method	Sensitivity(%)	Specificity(%)	Accuracy(%)
Semivariogram + SVM [10]	94.20 ± 3.10	97.50 ± 3.20	95.20 ± 2.30
Karri et al. [16]	98.88 ± 0.24	95.48 ± 1.15	97.86 ± 0.48
ResNet-101	93.60 ± 2.47	89.21 ± 6.79	92.28 ± 1.56
ResNet-101-SIRL (τ = 0.8) + 10% Infer-strat-B	98.22 ± 0.72	90.43 ± 2.36	95.88 ± 0.19
ResNet-101-SIRL (τ = 0.9) + 10% Infer-strat-B	96.88 ± 0.96	92.17 ± 1.82	95.47 ± 0.48

different. In the clinical dataset, the lesions are much apparent, which helps improving the accuracy of the label refining process, whereas in the Duke dataset, the lesions are relatively hard to detect, which is even harder after denoising. Therefore, better relabeling schemes are expected to be explored in future works. Moreover, it is noted that in this study, we also find that a volume-level inference strategy better than majority voting can be used when the labels are refined. It is also noted that the method proposed by Sun et al., which had similar concepts as ours, achieved an accuracy of 94.4% on the Duke dataset [12], but they used a different train/test set separation.

4. Conclusion

In this paper, we propose the SIRL training scheme to deal with the dirty label issue introduced by the rule of labeling a B-scan with its volume label. This rule is a well-known setting for training 2D B-scan classifiers when only the volume-level label is accessible. By introducing an innovative training–sieving–relabeling pipeline, SIRL attempts to identify normal B-scans in the abnormal volumes in a self-supervised manner, thereby improving the label quality. Because SIRL only appears in the training stage, no extra time cost will be introduced in the inference stage, which may help building efficient computer-aided macular disease diagnosis systems for clinical usage. We also propose a volume-level inference strategy to substitute the majority voting strategy. Experiments are conducted on both public and clinical datasets to demonstrate the advantage of our proposed method.

Conflicts of interest

None Declared.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61671272, the Opening Project of Guangdong Province Key Laboratory of Big Data Analysis and Processing under Grant No. 201803, and the National Key Research Program of China under Grant No. 2016YFB1000600. We thank the editor-in-chief and anonymous reviewers for their valuable suggestions for this paper.

References

[1] C. Gerth, R.J. Zawadzki, J.S. Werner, E. Héon, Retinal microstructure in patients with EFEMP1 retinal dystrophy evaluated by fourier domain OCT, Eye 23 (2) (2009) 480–483, https://doi.org/10.1038/eye.2008.251 2009.
[2] D. Huang, E.A. Swanson, C.P. Lin, J.S. Schuman, W.G. Stinson, W. Chang, M.R. Hee, T. Flotte, K. Gregory, C.A. Puliafito, J.G. Fujimoto, Optical coherence tomography, Science 254 (5035) (1991) 1178–1181 1991.
[3] A. Albarrak, F. Coenen, Y. Zheng, Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction, The 17th Annual Conference in Medical Image Understanding and Analysis (MIUA), BMVA, Birmingham, UK, 2013, p. 2013.
[4] F.G. Venhuizen, B. van Ginneken, B. Bloemen, M.J.J.P. van Grinsven, R. Philipsen, C. Hoyng, T. Theelen, C.I. Snchez, Automated age-related macular degeneration classification in OCT using unsupervised feature learning, Medical Imaging 2015: Computer-Aided Diagnosis, International Society For Optics and Photonics, Orlando, Florida, United States, 2015, p. 2015.
[5] T.H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PCANet: a simple deep learning baseline for image classification? IEEE Trans. Image Process. 24 (12) (2015) 5017–5032, https://doi.org/10.1109/TIP.2015.2475625 2015.
[6] L. Fang, C. Wang, S. Li, J. Yan, X. Chen, H. Rabbani, Automatic classification of retinal three-dimensional optical coherence tomography images using principal component analysis network with composite kernels, J. Biomed. Opt. 22 (11) (2017) 116011, https://doi.org/10.1117/1.JBO.22.11.116011 2017.
[7] R. Rasti, A. Mehridehnavi, H. Rabbani, F. Hajizadeh, Automatic diagnosis of abnormal macula in retinal optical coherence tomography images using wavelet-based convolutional neural network features and random forests classifier, J. Biomed. Opt. 23 (3) (2018) 0350052018.
[8] S. Apostolopoulos, C. Ciller, S.I. De Zanet, S. Wolf, R. Sznitman, RetiNet: Automatic AMD Identification in OCT Volumetric Data, (2016) arXiv:1610.03628.
[9] J. De Fauw, J.R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C.O. Hughes, R. Raine, J. Hughes, D.A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P.T. Khaw, M. Suleyman, J. Cornebise, P.A. Keane, O. Ronneberger, Clinically applicable deep learning for

diagnosis and referral in retinal disease, Nat. Med. 24 (9) (2018) 1342 2018.
[10] A.M. Santos, A.C. Paiva, A.P.M. Santos, S.A.T. Mpinda, D.L. Gomes, A.C. Silva, G. Braz, J.D.S. de Almeida, M. Gattass, Semivariogram and semimadogram functions as descriptors for AMD diagnosis on SD-OCT topographic maps using support vector machine, Biomed. Eng. Online 17 (1) (2018) 160 2018.
[11] P. Seebock, S.M. Waldstein, S. Klimscha, H. Bogunovic, T. Schlegl, B.S. Gerendas, R. Donner, U. Schmidt-Erfurth, G. Langs, Unsupervised identification of disease marker candidates in retinal OCT imaging data, IEEE Trans. Med. Imaging 38 (4) (2019) 1037–1047, https://doi.org/10.1109/TMI.2018.2877080 2019.
[12] W. Sun, X. Liu, Z. Yang, Automated detection of age-related macular degeneration in OCT images using multiple instance learning, Ninth International Conference on Digital Image Processing (ICDIP 2017), International Society for Optics and Photonics, Hong Kong, China, 2017, p. 2017, , https://doi.org/10.1117/12.2282522.
[13] P.P. Srinivasan, L.A. Kim, P.S. Mettu, S.W. Cousins, G.M. Comer, J.A. Izatt, S. Farsiu, Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images, Biomed. Opt. Express 5 (10) (2014) 3568–3577, https://doi.org/10.1364/BOE.5.003568 2014.
[14] Y. Sun, S. Li, Z. Sun, Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning, J. Biomed. Opt. 22 (1) (2017) 016012, , https://doi.org/10.1117/1.JBO.22.1.016012 2017.
[15] Z. Sun, Y. Sun, Automatic detection of retinal regions using fully convolutional networks for diagnosis of abnormal maculae in optical coherence tomography images, J. Biomed. Opt. 24 (5) (2019) 056003, , https://doi.org/10.1117/1.JBO.24.5.056003 2019.
[16] S.P.K. Karri, D.E. Chakraborty, J. Chatterjee, Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration, Biomed. Opt. Express 8 (2) (2017) 579–592 2017.
[17] R.M. Kamble, G.C.Y. Chan, O. Perdomo, F.A. Gonz, M. Kokare, Automated diabetic macular edema (DME) analysis using fine tuning with inception-resnet-v2 on OCT images, 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), IEEE, Sarawak, Malaysia, Malaysia, 2018, p. 2018.
[18] R. Rasti, H. Rabbani, A. Mehridehnavi, F. Hajizadeh, Macular OCT classification using a multi-scale convolutional neural network ensemble, IEEE Trans. Med. Imaging 37 (4) (2018) 1024–1034, https://doi.org/10.1109/TMI.2017.2780115 2018.
[19] O. Perdomo, S. Otálora, F.A. González, F. Meriaudeau, H. Müller, OCT-Net, A convolutional network for automatic classification of normal and diabetic macular edema using SD-OCT volumes, 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1423–1426 2018.
[20] Y. Rong, D. Xiang, W. Zhu, K. Yu, F. Shi, Z. Fan, S. Member, X. Chen, Surrogate-assisted retinal OCT image classification based on convolutional neural networks, IEEE Journal of Biomedical and Health Informatics 23 (1) (2019) 253–263, https://doi.org/10.1109/JBHI.2018.2795545 2019.
[21] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M.K. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V.A. Huu, C. Wen, E.D. Zhang, C.L. Zhang, O. Li, X. Wang, M.A. Singer, X. Sun, J. Xu, A. Tafreshi, M.A. Lewis, H. Xia, K. Zhang, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1124, https://doi.org/10.1016/j.cell.2018.02.010 e9 (2018).
[22] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (5) (2016) 1299–1312, https://doi.org/10.1109/TMI.2016.2535302 2016.
[23] S. Farsiu, S.J. Chiu, R.V. O'Connell, F.A. Folgar, E. Yuan, J.A. Izatt, C.A. Toth, Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography, Ophthalmology 121 (1) (2014) 162–172, https://doi.org/10.1016/j.ophtha.2013.07.013 2014.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, Nevada, USA, 2016, p. 2016.