Neurophotonics

Neurophotonics.SPIEDigitalLibrary.org

Artificial intelligence deep learning algorithm for discriminating ungradable optical coherence tomography three-dimensional volumetric optic disc scans

An Ran Ran
Jian Shi
Amanda K. Ngai
Wai-Yin Chan
Poemen P. Chan
Alvin L. Young
Hon-Wah Yung
Clement C. Tham
Carol Y. Cheung

SPIE.

An Ran Ran, Jian Shi, Amanda K. Ngai, Wai-Yin Chan, Poemen P. Chan, Alvin L. Young, Hon-Wah Yung, Clement C. Tham, Carol Y. Cheung, "Artificial intelligence deep learning algorithm for discriminating ungradable optical coherence tomography three-dimensional volumetric optic disc scans," *Neurophoton.* **6**(4), 041110 (2019), doi: 10.1117/1.NPh.6.4.041110.

Artificial intelligence deep learning algorithm for discriminating ungradable optical coherence tomography three-dimensional volumetric optic disc scans

An Ran Ran,[a] Jian Shi,[a] Amanda K. Ngai,[a] Wai-Yin Chan,[a] Poemen P. Chan,[a,b] Alvin L. Young,[c] Hon-Wah Yung,[d] Clement C. Tham,[a,b,*,†] and Carol Y. Cheung[a,*,†]

[a]The Chinese University of Hong Kong, Department of Ophthalmology and Visual Sciences, Hong Kong Special Administrative Region, China
[b]Hong Kong Eye Hospital, Hong Kong Special Administrative Region, China
[c]Prince of Wales Hospital, Hong Kong Special Administrative Region, China
[d]Tuen Mun Eye Center, Hong Kong Special Administrative Region, China

Abstract. Spectral-domain optical coherence tomography (SDOCT) is a noncontact and noninvasive imaging technology offering three-dimensional (3-D), objective, and quantitative assessment of optic nerve head (ONH) in human eyes *in vivo*. The image quality of SDOCT scans is crucial for an accurate and reliable interpretation of ONH structure and for further detection of diseases. Traditionally, signal strength (SS) is used as an index to include or exclude SDOCT scans for further analysis. However, it is insufficient to assess other image quality issues such as off-centration, out of registration, missing data, motion artifacts, mirror artifacts, or blurriness, which require specialized knowledge in SDOCT for such assessment. We proposed a deep learning system (DLS) as an automated tool for filtering out ungradable SDOCT volumes. In total, 5599 SDOCT ONH volumes were collected for training (80%) and primary validation (20%). Other 711 and 298 volumes from two independent datasets, respectively, were used for external validation. An SDOCT volume was labeled as ungradable when SS was <5 or when any artifacts influenced the measurement circle or >25% of the peripheral area. Artifacts included (1) off-centration, (2) out of registration, (3) missing signal, (4) motion artifacts, (5) mirror artifacts, and (6) blurriness. An SDOCT volume was labeled as gradable when SS was ≥5, and there was an absence of any artifacts or artifacts only influenced <25% peripheral area but not the retinal nerve fiber layer calculation circle. We developed and validated a 3-D DLS based on squeeze-and-excitation ResNeXt blocks and experimented with different training strategies. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy were calculated to evaluate the performance. Heatmaps were generated by gradient-weighted class activation map. Our findings show that the presented DLS achieved a good performance in both primary and external validations, which could potentially increase the efficiency and accuracy of SDOCT volumetric scans quality control by filtering out ungradable ones automatically. [DOI: 10.1117/1.NPh.6.4.041110]

Keywords: artificial intelligence; deep learning; optical coherence tomography; image quality control.

Paper 19056SSR received Jun. 8, 2019; accepted for publication Sep. 9, 2019; published online Nov. 1, 2019.

1 Introduction

Optical coherence tomography (OCT) is a noncontact and noninvasive imaging technology offering objective and quantitative assessment of human eye structures, including the cornea, macula, and optic nerve head (ONH) *in vivo*. The introduction of spectral-domain optical coherence tomography (SDOCT) in recent years has improved scanning speed and axial resolution, enabling high-resolution, three-dimensional (3-D) volumetric imaging that has made a great contribution to the wide application in clinics.[1] However, poor scan quality due to patients' poor cooperation, operators' skills, or device-dependent factors (e.g., inaccurate optic disc margin delineation) can affect the metrics generated from the SDOCT.[2,3] Specifically, insufficient image quality potentially leads to inaccurate measurements of retinal nerve fiber layer (RNFL) thickness, which is an important metric for detection of optic neuropathy such as glaucoma, a leading cause of irreversible blindness.[4] Other morphologies from ONH, such as neuroretinal rim and lamina cribrosa,[5] are also used to assess glaucoma, which also require sufficient quality of SDOCT volumetric data for such assessment. Thus, it is necessary to filter out ungradable scans and reoperate on patients with subpar images before any clinical assessment.

Conventionally, signal strength (SS) is the main parameter to include or exclude SDOCT scans for further quantitative analysis.[6] For the Cirrus high-definition SDOCT, image quality is indicated by SS ranging from 0 (worst quality) to 10 (best quality), representing the average of signal intensity of SDOCT volumetric scans, and scans with SS of 6 or above are usually defined as sufficient for further analysis.[7–9] However, even with acceptable SS, it is still hard to assess other SDOCT image quality issues, such as off-centration, out of registration, signal loss, motion artifacts, mirror artifacts, or blurriness of SDOCT volumetric data.[3] Such image quality assessment indeed requires highly trained operators and interpreters with specialized knowledge in SDOCT, which is a big challenge due to the lack of manpower and insufficient training time in clinics. In

*Address all correspondence to Carol Y. Cheung, E-mail: carolcheung@cuhk.edu.hk; Clement C. Tham, E-mail: clemtham@cuhk.edu.hk

†These authors contributed equally to the work as senior authors.

addition, it is impractical for human assessors to grade every SDOCT volumetric scan, which could be a time-consuming and tedious process in busy clinics.

Previous studies have proposed traditional computer-aided systems using hand-crafted features for automated image quality control in natural images.[10] However, the hand-crafted features were based on either geometric or structural quality parameters such as signal-to-noise ratio, which do not generalize well to new datasets. Moreover, unlike natural images, the gradability of medical images is not simply related to pixels, signals, noises, or distortion of an image itself. Human assessors' judgment on whether the quality of the entire image is sufficient for disease detection or further analysis is essential for discriminating the gradability of medical images.

Machine learning, under the broad name of artificial intelligence (AI), adopts a class of techniques called deep learning (DL).[11] In terms of image processing, convolutional neural networks (CNNs) are proven to be useful in image-related tasks. It is more efficient to extract and weigh features automatically rather than in a hand-crafted manner. Currently, CNN has been used for image quality control in various medical imaging, such as magnetic resonance imaging (MRI),[12] ultrasound imaging,[13] and fundus photography.[14] Generally, using DL for image quality control can be achieved with either unsupervised or supervised methods. Unsupervised anomaly detection is mainly used in highly imbalanced datasets to detect rare cases. It learns features from only one kind of input, then computes the similarity between the future input and the learned one. Generative-based works, such as variational autoencoder-based methods[15] and generative adversarial networks-based methods,[16] are commonly applied on more than one neural network. Nongenerative models, such as one-class neural networks,[17] require a pretrained deep autoencoder as well. Hence, a higher computational cost and a larger graphics processing unit (GPU) memory are needed for applying anomaly detection, especially on 3-D image tasks, which would be impractical in our study. The second method is binary classification, a supervised anomaly detection method to train a CNN model to recognize input images as binary labels. With residual connections proposed from ResNet,[18] deeper CNNs can be trained without degradation by reducing gradient vanishing or exploding problems. Other variants such as ResNeXt further improved the performance on classification benchmarks such as ImageNet data.[19] Apart from those, SENet proposed squeeze-and-excitation (SE) blocks, which introduced a channel-wise attention mechanism in a simple plug-in manner that could be applied in any DL models, and it surpassed other architectures in the competition ImageNet 2017.[20] Since the ground-truth label of each SDOCT volumetric scan is from highly trained human assessors, a model trained in a supervised manner would be better in our study. As far as we know, though CNN has been applied in medical imaging quality control, there is still a lack of DL-based method for quality control of SDOCT volumetric scans.

In this study, we aim to develop and validate a 3-D deep learning system (DLS) using SDOCT volumetric scans as input for filtering out ungradable volumes. We hypothesize that the 3-D DLS for filtering out ungradable SDOCT volumetric scans without hand-crafted features would perform well in both primary and external validations. The DLS would eventually increase the accuracy and efficiency of SDOCT volumetric data quality control and further make a contribution on accurate quantitative analysis and detection of diseases.

2 Materials and Methods

2.1 *Data Acquisition and Data Pre-Processing*

The dataset for training and validation was collected from the existing database of electronic medical and research records at the Chinese University of Hong Kong (CUHK) Eye Center and the Hong Kong Eye Hospital (HKEH) dated from March 2015 to March 2019. The inclusion criteria were any subjects who have undergone ONH SDOCT imaging by Cirrus SDOCT (Carl Zeiss Meditec, Dublin, California). A total of 5599 SDOCT volumetric scans from 1479 eyes were included for the development of the DLS. These data were from normal subjects or patients with any pathologies, and most of the patients had glaucoma. Two nonoverlapping datasets collected from Prince of Wales Hospital (PWH) and Tuen Mun Eye Center (TMEC) in Hong Kong were used as two external validation datasets, including 711 SDOCT scans from 509 eyes and 298 scans from 296 eyes, respectively. (Table 1)

An SDOCT volume was labeled as ungradable when SS was <5 or when any artifacts influenced the measurement circle or >25% of the peripheral area. Artifacts included (1) off-centration, (2) out of registration, (3) missing signal, (4) motion artifacts, (5) mirror artifacts, and (6) blurriness. An SDOCT volume was labeled as gradable when SS was ≥5 and absence of any artifacts or artifacts only influenced <25% of the peripheral area but not the RNFL calculation circle. The RNFL calculation circle was a circle of 3.46 mm in diameter evenly around its center based on the location of the optic disc, and it was automatically placed by Cirrus SDOCT machine (Cirrus User Manual). Before starting to grade, two highly trained human assessors were tested. A separated set of images containing 200 SDOCT volumetric scans were reviewed by the two assessors and kappa value of 0.96 was achieved, which indicated an almost perfect agreement.[21] Disagreed cases were further discussed with the senior assessor, a trained doctor with more than 5 years of clinical research experience in glaucoma imaging. After training and testing, two assessors worked separately to label each SDOCT volumetric scan from all the datasets as ungradable or gradable. Disagreements between the two assessors were resolved by consensus, and the cases without consensus were further reviewed by the senior assessor to make the final decision (examples are shown in Fig. 1).

Data augmentation strategies, including random flipping, random rotating, and random shifting, were used to enhance the training samples and alleviate overfitting. The original SDOCT volumes were with size of $200 \times 200 \times 1024$ in three axes, x axis, y axis, and z axis, respectively. To mimic the real SDOCT imaging in the clinics, some data augmentation methods were only applied on one or two axes for the whole volume. For instance, 20% chance random flipping and 15-deg random rotation were applied on only x (200) and y-axes (200), respectively. The color channel was set to 1 since all OCT images were grayscaled.

2.2 *Irrelevancy Reduction and Attention Mechanism*

Generally, for this specific task, i.e., discriminating the ungradability from an SDOCT scan, there is a high level of information that could disturb the ungradable features, such as the anatomic changes of ONH, the shadow of vessels, and the noise speckles in the choroid or vitreous. Hence, the features in ungradable SDOCT volumes do not follow any specific feature patterns,

Table 1 Summary of all the study subjects.

	Ungradable	Gradable	*P* value
Training and primary validation dataset			
No. of SDOCT volumes	1353	4246	—
No. of subjects	260	549	—
Gender (male/female)	139/121	229/320	**0.002**
Age, years (mean ± SD)	63.2 ± 15.8	57.0 ± 17.1	**0.003**
No. of eyes	402	1077	—
Eye (right/left)	222/180	533/544	**0.050**
External validation dataset 1 (PWH)			
No. of SDOCT volumes	181	530	—
No. of subjects	80	227	—
Gender (male/female)	50/30	102/125	**0.009**
Age, years (mean ± SD)	73.6 ± 10.5	68.3 ± 11.9	0.367
No. of eyes	122	387	—
Eye (right/left)	60/62	204/183	0.497
External validation dataset 2 (TMEC)			
No. of SDOCT volumes	60	238	—
No. of subjects	42	158	—
Age, years (mean ± SD)	61.7 ± 14.5	60.7 ± 12.4	0.177
Gender (male/female)	23/19	77/81	0.603
No. of eyes	60	234	—
Eye (right/left)	35/25	114/120	0.185

Note: OCT, optical coherence tomography; SD, standard deviation. Unpaired *t*-test for numerical data and chi-square test for categorical data were used for comparison between ungradable and gradable groups. All the hypotheses tested were two-sided, and *p*-value <0.05 were considered to be significant which were values in bold.

which may lead the neural networks to misinterpret the appearance of those aforementioned irrelevances as ungradable features. To address the problem, we trialed two methods—irrelevancy reduction and attention mechanism—for a better model performance.

Irrelevancy reduction omits the parts of irrelevant signals that should not be noticed by the signal receiver, which potentially improves the performance.[22] Intuitively, denoising was used as one of the strategies to reduce the irrelevancies of OCT scans since the noise of SDOCT scan impeded the medical analysis either visually or programmatically.[23] Thus, in experiment 1, we used a model based on ResNet blocks to compare the performance between the original and the irrelevancy reduced data. For denoising, we used nonlocal means[24] as the strategy, which performed both vertically (along x, z facets) and horizontally (along x, y facets) with different sets of parameters. Vertically, the template window size was set to 10, whereas the search window size was set to 5 with a filter strength of 5. Horizontally, the template window size was set to 5, and search window size was set to 5 with a filter strength of 5.

In experiment 2, we applied a self-attention mechanism by combing the SE block that introduced a channel-wise attention mechanism to the ResNet model. The self-attention mechanism could make the model pay attention to the more important areas and extract features automatically in the original SDOCT volumes. Furthermore, we experimented on the combination of data denoising and the attention mechanism by training the SE-ResNet model with denoised volumes, with the aim of achieving a better performance.

In experiment 3, we substituted the ResNet blocks with ResNeXt blocks with the consideration of the performance improvement. Then we fine-tuned the cardinality of transformation layers to reduce the GPU cost.

2.3 *Development and Validation of the Deep Learning System*

The model for the DLS was implemented with Keras and Tensorflow, on a workstation equipped with i9-7900X and Nvidia GeForce GTX 1080Ti. Figure 2(a) shows the building block of the ResNet model. First, there were 32 filters with $7 \times 7 \times 7$ kernel convolution layer with the stride of 2, along with a $3 \times 3 \times 3$ max pooling with the same stride setting. Second, the obtained feature maps went through 18 ResNet blocks. A pooling size 2 with stride 2 average pooling was performed every three blocks to aggregate the learned features. Channel-wise batch normalization and rectified linear unit activation were performed after all convolution operations. Finally, a global average pooling followed by a fully connected softmax layer was used to produce the binary output as gradable or ungradable. This ResNet-based model was taken as the benchmark model. Next, we further experimented with the SE-ResNet-block[20] [Fig. 2(b)] and SE-ResNeXt-block[19] [Fig. 2(c)], as the basic building block. In each SE-ResNet or SE-ResNeXt block, the SE reduction ratio was set to 4 and the cardinalities of the transformation layer were set to 8, with 32 filters. The constructed models are depicted in Fig. 2(d).

A total of 1353 ungradable and 4246 gradable SDOCT volumetric scans collected from CUHK Eye Center and HKEH were randomly divided for training (80%) and primary validation (20%). In each set, we kept the similar distribution of ungradable versus gradable scans and distributed the eyes from the same patient to the same set in order to prevent data leakage and biased estimation of the performance. Cross entropy and Adam were used as the loss function and the optimizer. During the training, 3000 volumetric scans were selected with data balancing. Batch size was set to 1 due to the limited GPU memory. The initial learning rate was set to 0.0001, and then reduced by multiplying 0.75 in every two epochs. In addition, to validate the generalizability of the proposed DLS, SDOCT scans from PWH (181 ungradable versus 530 gradable) and TMEC (60 ungradable versus 238 gradable) were utilized for external validation separately.

3 Experiments and Results

3.1 *Evaluation Metrics*

In the experiments, the area under the receiver operating characteristic (ROC) curve (AUC) with 95% confidence intervals (CIs), sensitivity, specificity, and accuracy were used to evaluate

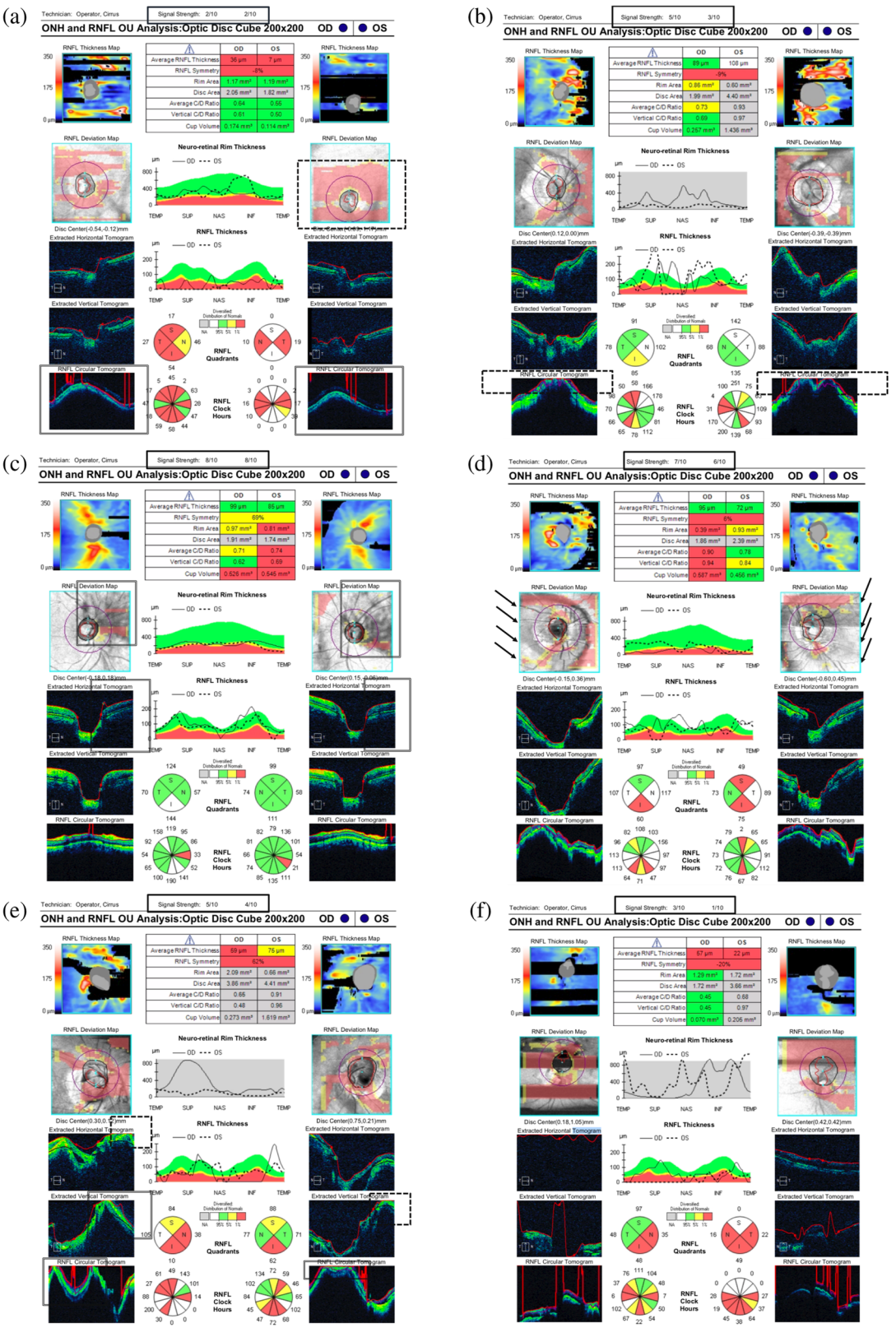

Fig. 1 Examples of ungradable SDOCT images due to different kinds of artifacts. (a) SDOCT signal loss on both eyes (double bordered boxes) and off-centration artifact on left eye (dashed box). (b) SDOCT signal out of register superiorly on both eyes (dashed boxes). (c) SDOCT signal loss in the measurement circle on both eyes (double bordered boxes) while the SS was above 6. (d) Motion artifacts on both eyes (black arrows) resulting in inaccurate segmentation while the SS was above or equal to 6 in both eyes. (e) Mirror artifacts in the peripheral area on both eyes (double bordered boxes) and the SDOCT signal out of register superiorly on both eyes (dashed boxes). (f) SDOCT signal loss on right eye and blurriness artifact on left eye. OD, right eye; OS, left eye.

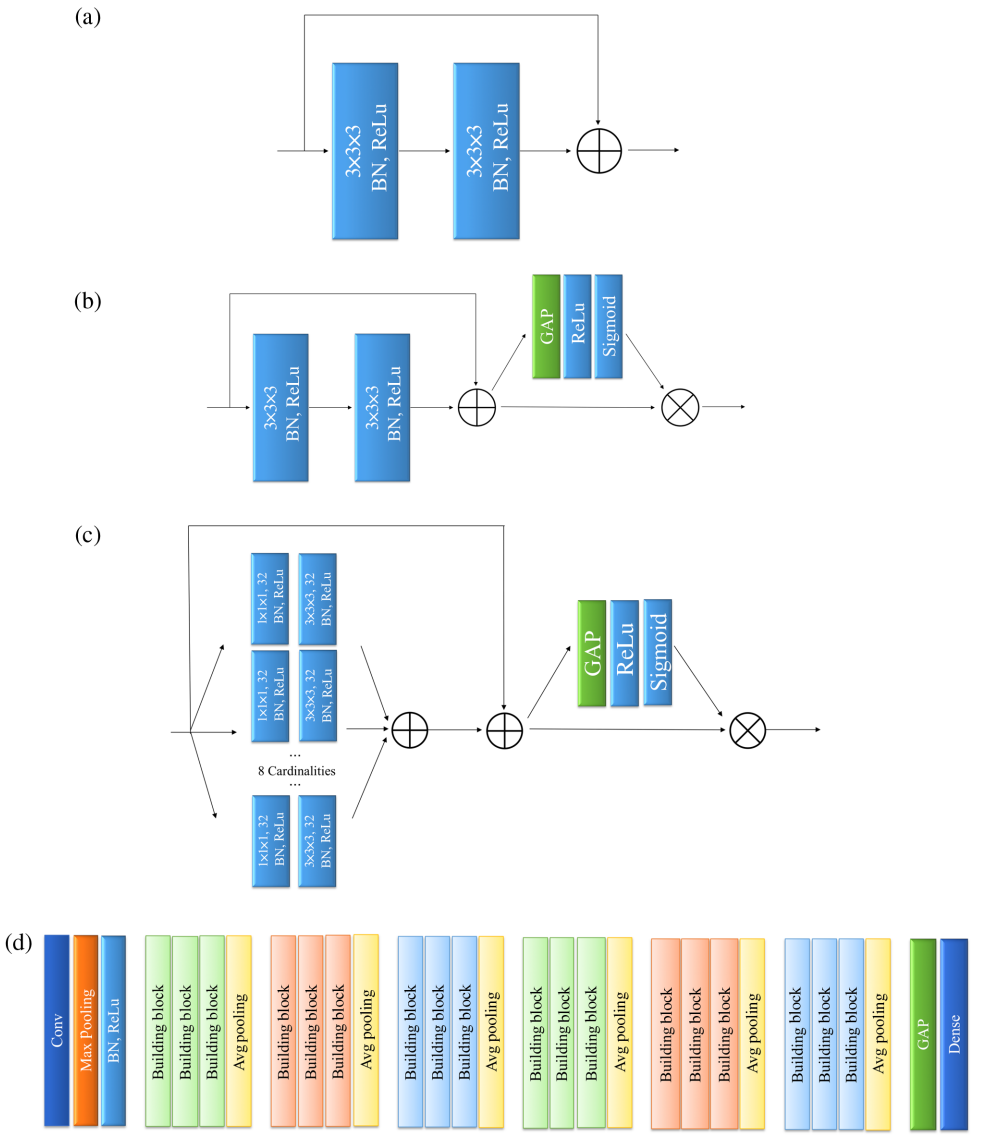

Fig. 2 A diagram illustrating the architecture of basic building blocks and the architecture of different models. (a) The architecture of ResNet building block. (b) The architecture of SE-ResNet building block. (c) The architecture of SE-ResNeXt building blocks. We used eight transformation layers along with 32 filters for each transformation layer. (d) The architecture of different models. Despite the difference in building blocks, the architectures of the ResNet, SE-ResNet, and SE-ResNeXt models were the same. BN = batch normalization, GAP = global average pooling, Conv = convolutional, Avg = average.

the diagnostic performance of the DLS discriminating ungradable or gradable scans. Training-validation loss curves were observed (Fig. 3). Heatmaps were generated by gradient-weighted class activation map (Grad-CAM)[25] to evaluate the performance qualitatively.

3.2 *Performance Comparison*

We tested the feasibility of irrelevancy reduction and attention mechanism in experiments 1 and 2, respectively. We also explored whether the performance would improve by combining the two approaches. Experiment 3 was further performed by refining the model structures. The experimental results were shown in Table 2 and Figs. 3 and 4.

In experiment 1, we developed ResNet models fed with original volumes and denoised volumes, respectively. We can observe from the training-validation loss curves in Figs. 3(a) and 3(b) that the ResNet model trained with original volumes was highly overfitted, whereas the model trained with irrelevancy reduced volumes fitted better with some level of oscillation. As a result, ResNet trained with denoised volumes reached much better AUCs than the one trained with original volumes (primary validation: 0.806 versus 0.640, external validation 1: 0.645 versus 0.535, external validation 2: 0.792 versus 0.697).

In experiment 2, the SE block was implemented to introduce the channel-wise attention to the benchmark model, which could help the method suppress the noisy features for the more essential features to discriminate ungradable patterns. As illustrated in Figs. 3(c) and 3(d), with the introduced attention mechanism, the training-validation loss was well converged without significant oscillations. As shown in Table 2 and Fig. 4, the SE-ResNet model fed with original volumes performed much better than

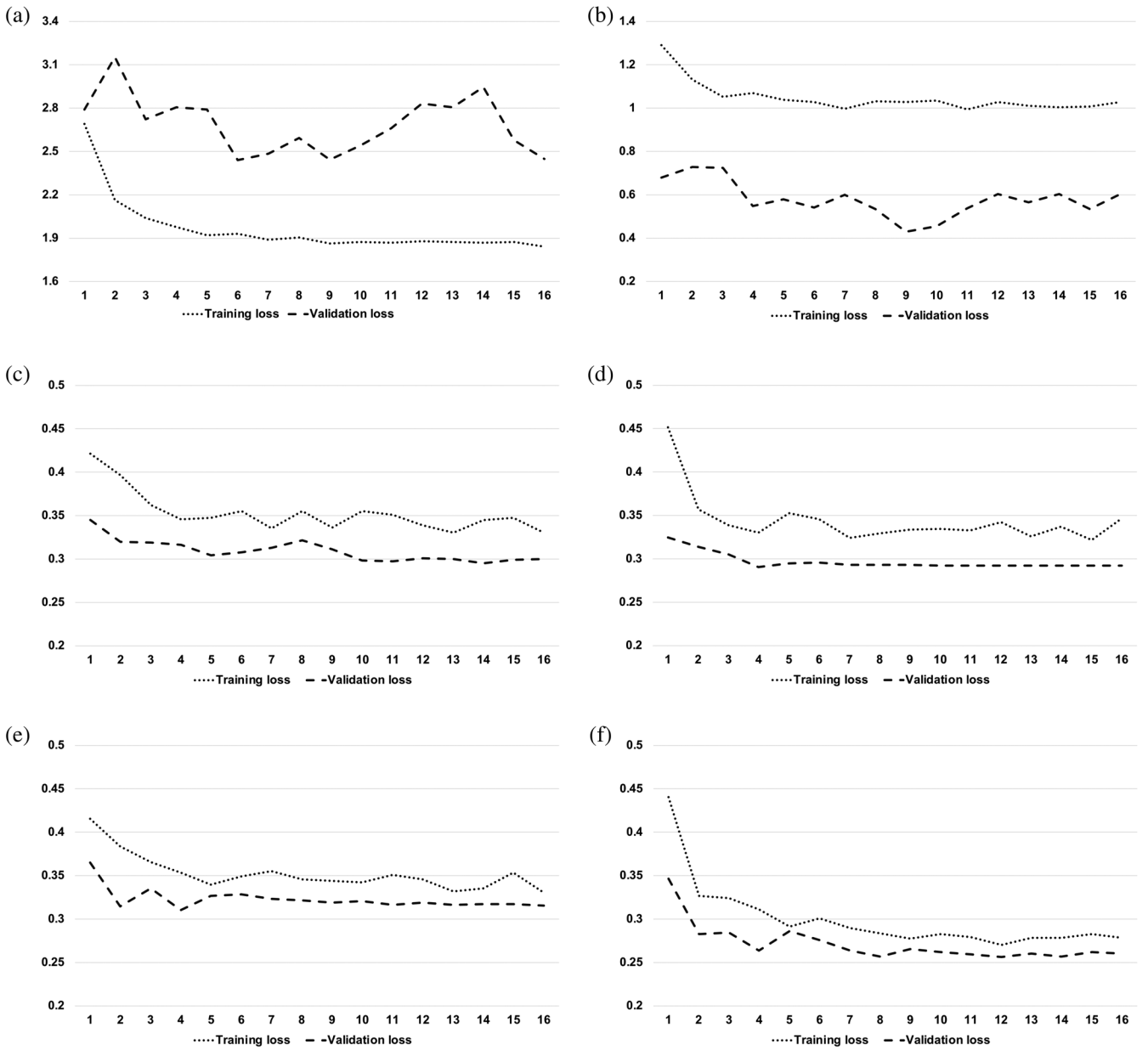

Fig. 3 Training-validation loss curves of different models along training epochs. During the training, we observed the cross-entropy loss to measure the model training effect. Theoretically, a good model shall keep the lowest loss and should have the smallest difference between the training-validation losses without significant oscillation. (a) The highly overfitted ResNet model trained with original SDOCT volumes. (b) The highly overfitted ResNet model trained with denoised SDOCT volumes. (c) The stable and well-fitted SE-ResNet model trained with original SDOCT volumes. (d) The stable and well-fitted SE-ResNet model trained with denoised SDOCT volumes. (e) The stable and better-fitted SE-ResNeXt model with denoised SDOCT volumes. (f) The stable and best-fitted SE-ResNeXt model with denoised SDOCT volumes.

the ResNet with original volumes in both primary and external validations (primary validation: 0.905 versus 0.640, external validation 1: 0.815 versus 0.535, external validation 2: 0.858 versus 0.697). We further evaluated the performance of SE-ResNet model fed with the denoised volumes, which also led to much higher AUCs compared to the ResNet model with the same input (primary validation: 0.943 versus 0.806, external validation 1: 0.811 versus 0.645, external validation 2: 0.844 versus 0.793).

In experiment 3, we replaced all the ResNet blocks with ResNeXt blocks with the aim of achieving a better performance. We then fine-tuned the cardinality of ResNeXt transformations to 8. As illustrated in Figs. 3(e) and 3(f), SE-ResNeXt model obtained more stable and better-fitted training-validation curves, compared to experiment 2. Furthermore, the performance of SE-ResNeXt model fed with original and denoised volumes were increased with the AUCs of 0.938 (95% CI: 0.919 to 0.957) and 0.954 (95% CI: 0.938 to 0.970), respectively, as shown in Fig. 4. More importantly, the overall diagnostic of SE-ResNeXt fed with denoised volumes was the best with sensitivity of 86.2% (95% CI: 80.0% to 92.4%), specificity of 92.6% (95% CI: 86.8% to 96.9%), and accuracy of 91.0% (95% CI: 87.3% to 93.5%) in primary validation and sensitivities of 69.1% (95% CI: 58.0%

Table 2 The performance comparisons of the 3-D DLS using different training strategies in both primary and external validations.

Strategies	AUC (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Accuracy, % (95% CI)
ResNet with original volumes				
Primary validation	0.640 (0.594 to 0.686)	53.3% (46.7 to 60.0)	67.5% (63.9 to 71.3)	63.9% (60.9 to 66.9)
External validation 1	0.535 (0.484 to 0.587)	45.3% (1.0 to 66.3)	67.4% (45.9 to 98.1)	62.2% (50.9 to 76.9)
External validation 2	0.697 (0.616 to 0.779)	60.0% (45.0 to 80.0)	80.7% (57.6 to 89.1)	76.2% (61.4 to 82.2)
ResNet with denoised volumes				
Primary validation	0.806 (0.772 to 0.841)	71.9% (65.7 to 77.6)	74.9% (71.6 to 78.2)	74.1% (71.2 to 77.1)
External validation 1	0.645 (0.598 to 0.693)	51.4% (41.4 to 62.4)	74.9% (66.8 to 82.1)	68.9% (63.6 to 73.1)
External validation 2	0.793 (0.732 to 0.853)	75.0% (58.3 to 91.7)	74.8% (53.8 to 87.4)	75.2% (60.7 to 83.2)
SE-ResNet with original volumes				
Primary validation	0.905 (0.875 to 0.934)	82.9% (77.1 to 89.1)	93.3% (88.2 to 96.2)	90.6% (87.3 to 92.6)
External validation 1	0.815 (0.779 to 0.852)	68.0% (59.7 to 80.7)	83.2% (68.1 to 87.9)	79.2% (70.6 to 82.3)
External validation 2	0.858 (0.801 to 0.915)	78.3% (63.3 to 90.0)	81.1% (73.5 to 92.9)	81.4% (75.2 to 88.7)
SE-ResNet with denoised volumes				
Primary validation	0.943 (0.925 to 0.961)	83.1% (77.1 to 88.6)	93.1% (87.6 to 96.4)	90.6% (87.1 to 92.7)
External validation 1	0.811 (0.774 to 0.847)	68.0% (59.1 to 80.1)	83.2% (69.1 to 88.3)	79.2% (71.3 to 82.7)
External validation 2	0.844 (0.784 to 0.847)	78.3% (63.3 to 91.7)	81.5% (73.5 to 93.7)	81.5% (75.2 to 89.3)
SE-ResNeXt with original volumes				
Primary validation	0.938 (0.919 to 0.957)	84.8% (78.6 to 91.0)	91.6% (86.2 to 95.6)	89.7% (86.4 to 92.4)
External validation 1	0.801 (0.764 to 0.838)	68.0% (53.6 to 86.2)	78.9% (58.7 to 90.6)	76.1% (66.1 to 82.4)
External validation 2	0.854 (0.796 to 0.912)	76.7% (61.7 to 90.0)	**84.9% (74.0 to 95.4)**	**83.2% (75.8 to 90.3)**
SE-ResNeXt with denoised volumes				
Primary validation	**0.954 (0.938 to 0.970)**	**86.2% (80.0 to 92.4)**	92.6% (86.8 to 96.9)	**91.0% (87.3 to 93.5)**
External validation 1	**0.816 (0.780 to 0.852)**	**69.1% (58.0 to 84.0)**	81.3% (64.0 to 89.4)	**78.2% (68.8 to 82.7)**
External validation 2	**0.857 (0.800 to 0.914)**	**78.3% (61.7 to 91.7)**	**82.8% (71.9 to 94.6)**	**82.6% (74.2 to 89.9)**

Note: AUC, area under the receiver operator characteristic curve; CI, confidence interval; DLS, deep learning system.
Primary validation: CUHK Eye Center and HKEH;
External validation 1: PWH, Hong Kong;
External validation 2: TMEC, Hong Kong.
The bold values were the highest values in respective categories.

to 84.5%) and 78.3% (95% CI: 61.7% to 91.7%), specificities of 81.3% (95% CI: 64.0% to 89.4%) and 82.8% (95% CI: 71.9% to 94.6%), and accuracies of 78.2% (95% CI: 68.8% to 82.7%) and 82.6% (95% CI, 74.2% to 89.9%) in the external validations, respectively, as shown in Table 2. The results showed that the SE-ResNeXt-based model with denoised SDOCT volumes stood out of all other models fed with either original or denoised volumes.

In general, the model with denoised scans was better than the one with original scans with a significant improvement on the primary dataset and similar performance on external validations. The introduced SE blocks achieved a comparable result even on original volumes. It proved that either irrelevancy reduction or attention mechanism could significantly improve the performance, compared to the benchmark model—ResNet—fed with original volumes. Moreover, our proposed method combining both irrelevancy reduction and attention mechanism has achieved the highest AUCs in our experiments in both primary and external validations (primary validation: 0.954 versus 0.640 to 0.943, external validation 1: 0.816 versus 0.535 to 0.815, external validation 2: 0.857 versus 0.697 to 0.857). Referring to other diagnostic metrics, such as sensitivity, specificity, and accuracy, this model also outperformed the other models in both primary and external validations in general.

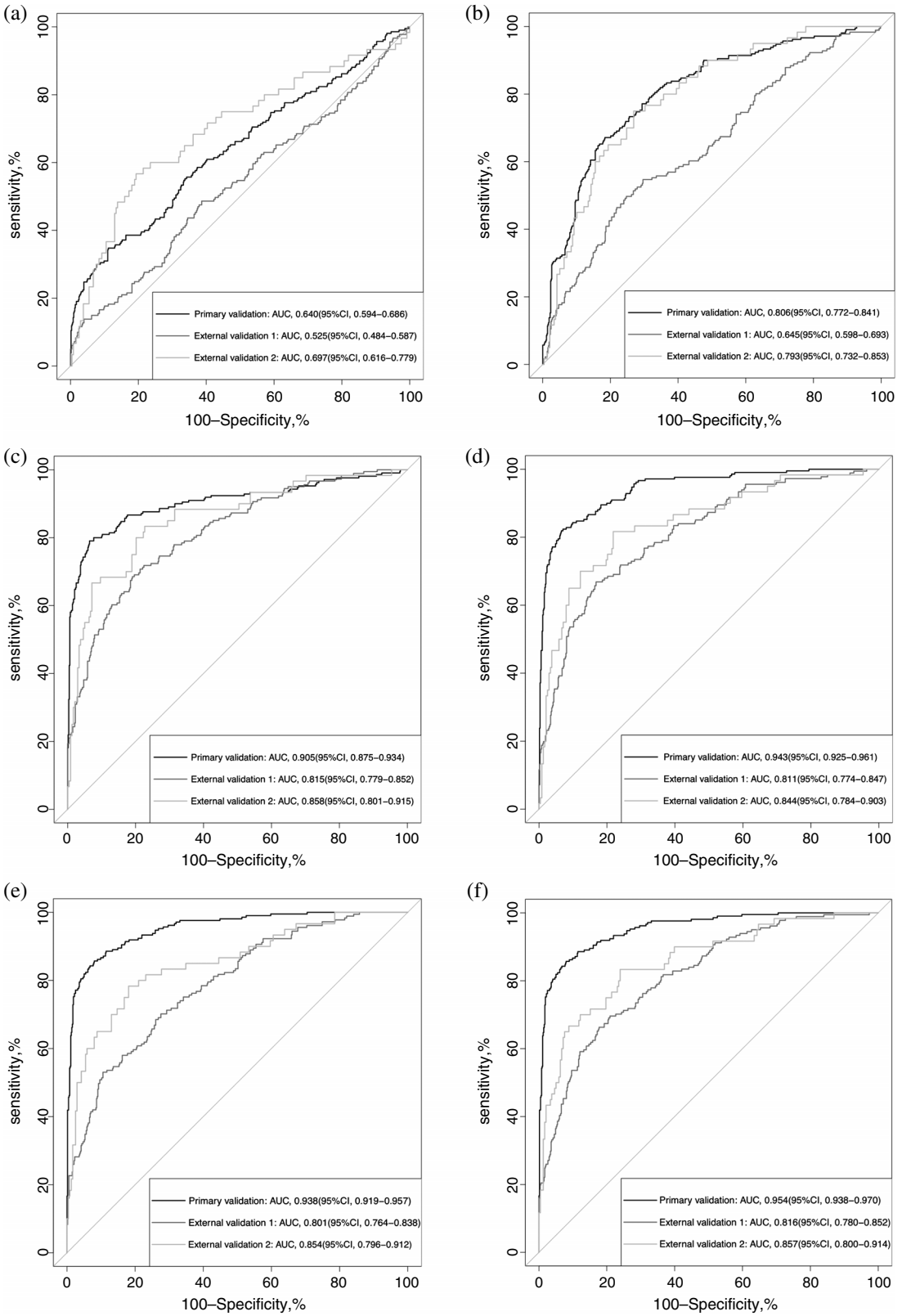

Fig. 4 The AUC of different models in the primary validation dataset and two external validation datasets. (a) AUCs of the ResNet model trained with original SDOCT volumes. (b) AUCs of the ResNet model trained with denoised SDOCT volumes. (c) AUCs of the SE-RexNet model trained with original SDOCT volumes. (d) AUCs of the SE-ResNet model trained with denoised SDOCT volumes. (e) AUCs of the SE-ResNeXt model trained with original SDOCT volumes. (f) AUCs of the SE-ResNeXt model trained with denoised SDOCT volumes. In general, the AUCs of the SE-ResNeXt model outperformed both ResNet and SE-ResNet models. Particularly, the SE-ResNeXt model trained with denoised data achieved the highest AUCs in both primary and external validations and an overall better diagnostic performance with regard to other metrics, such as sensitivity, specificity, and accuracy.

3.3 *Qualitative Evaluation*

We generated heatmaps (Fig. 5) based on the best performing model—SE-ResNeXt fed with denoised volumes—where the red-orange color represented more discriminative areas for ungradable information. We observed that for the truly discriminated ungradable volumes, there was no regular pattern in the area highlighted by the DLS due to the variances from different artifacts. However, in general, we still found that the DLS could detect ungradable features well, especially signal loss, mirror

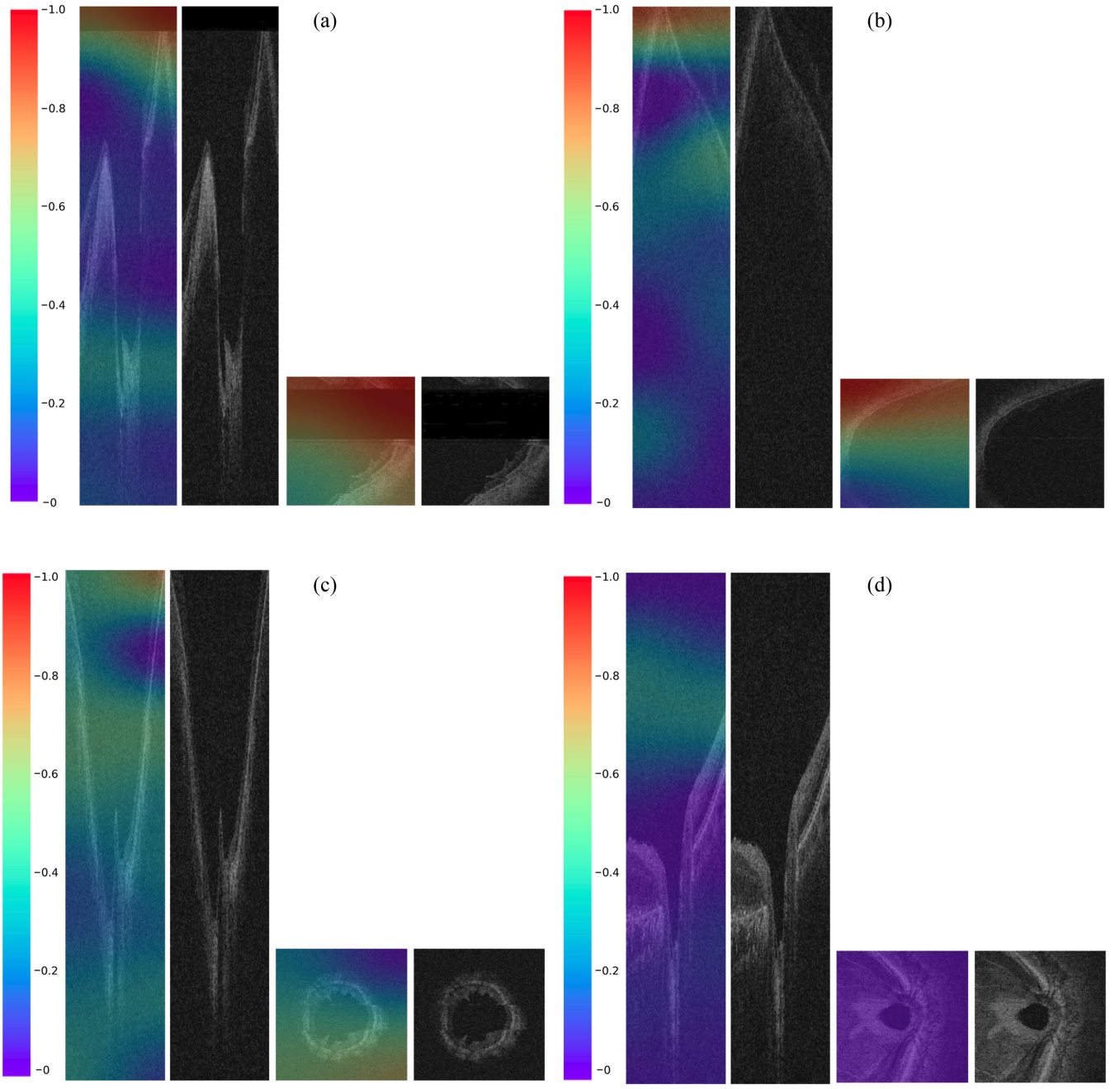

Fig. 5 Examples of truly classified cases and the corresponding heatmaps generated by Grad-CAM. Referring to the color distribution of the color bar, the red-orange color represented more discriminative areas for ungradable information, which were of value 1 to 0.8, whereas the blue-purple color represented the nondiscriminative areas that were of value 0.2 to 0: (a) cross-sectional SDOCT B-scan and *en-face* SDOCT C-scan of an ungradable volume with signal loss; (b) cross-sectional SDOCT B-scan and *en-face* SDOCT C-scan of an ungradable volume with mirror artifact; (c) cross-sectional SDOCT B-scan and *en-face* SDOCT C-scan of an ungradable volume due to blurring; and (d) cross-sectional SDOCT B-scan and *en-face* SDOCT C-scan of a gradable volume. In general, for the truly discriminated ungradable volumes, the DLS could well detect ungradable features, especially signal loss, mirror artifacts, or blurriness, as illustrated in (a), (b), and (c). On the other hand, there was seldom warmer color in the truly discriminated gradable volumes, but relatively highlighted regions were mainly distributed in the vitreous or choroid, as illustrated in (d).

artifacts, or blurriness. Area with the appearance of these artifacts was exactly covered by warmer color.

4 Discussion

In this study, we developed and validated a 3-D DLS to discriminate ungradable SDOCT ONH volumetric scans automatically. The proposed method, SE-ResNeXt fed with denoised volumes, achieved best performance in both primary and external validations. Experimental results show that ungradable SDOCT volumetric scans can be discriminated without any human interventions. It may potentially increase the efficiency of SDOCT image quality control and further help with disease detection, which could be a novel application in clinics.

Our proposed DLS offers a powerful tool for filtering out ungradable scans in clinics. Currently, one of the main challenges for SDOCT image quality control is the irregular ungradable feature patterns due to varying artifacts. Using traditional index for the quality assessment, such as SS, is insufficient to assess different kinds of artifacts such as off-centration, out of register, motion artifacts, and mirror artifacts. Scans with acceptable SS could also be ungradable for disease detection, as illustrated by the examples in Figs. 1(c) and 1(d). Nevertheless, manually assessing all the scans would be tedious and impractical in clinics. To address this problem, a DLS with the proposed SE-ResNeXt model was developed and trained for the auto-assessment. From experiment 1, we found that the overall diagnostic performance improved significantly by denoising, which indicated that the irrelevant information in SDOCT scans could strongly affect the model training. Better performance and better generalizability were obtained by reducing the irrelevant information. In experiment 2, we introduced the attention mechanism to further extract the important features out of the whole SDOCT volume automatically. The AUCs of SE-ResNet models were significantly increased, compared to ResNet models. In addition, the results proved that the combination of attention mechanism and irrelevancy reduction, the nonlocal means denoising, showed a more stable training-validation curve and outperformed either one of the previous two strategies. It proved that the channel-wise attention could help the model learn from noisy data with a much stable loss and a better generalizability.

In experiment 3, we replaced all the ResNet blocks with ResNeXt blocks for a better performance and a lower GPU cost. Our final proposed model was developed by SE-ResNeXt structure and trained with denoised full-size SDOCT volumes. According to the activation heatmaps in Fig. 5, the proposed model learned ungradable features similar to what human assessors would observe. Referring to the color distribution of the color bar, the red-orange color represented more discriminative areas for ungradable information, which were of value 0.8 to 1, whereas the blue-purple color represented the nondiscriminative areas that were of value 0 to 0.2. In general, we found that the DLS could well detect the ungradable features such as signal loss, mirror artifacts, and severe motion artifacts. Meanwhile, some ungradable features, i.e., blurriness or optic disc dislocation, were highlighted on the whole retina. There was seldom warmer color in the truly discriminated gradable volumes, but the relatively highlighted regions were mainly distributed in the vitreous and choroid for almost every truly detected gradable volumes. It might be caused by the appearance of more noise speckles in vitreous and choroid, compared to retina. The results from all the experiments illustrated that our proposed model trained with the denoised full-sized volumes was the best-fitted model that also achieved the optimum diagnostic performance among all the models in both primary and external validations.

It is hard to apply traditional computer-aided image quality control to a new dataset since the hand-crafted features are usually based on the objective features, either geometric or structural quality parameters, while some features are subjective in the real-world cases. A previous study on MRI image quality control also proved that deep neural networks got an overall better performance compared with the traditional machine learning method.[26] In our work, the proposed DLS achieves good performance on two totally unseen datasets from different clinics, which means the model has a good generalizability that may be applied to other clinics directly.

At present, multiple DLSs have been developed based on OCT in ophthalmology, such as referable retina diseases detection,[27,28] glaucoma quantification and classification,[29–31] and antivascular endothelial growth factor treatment.[32] These studies perfectly underscored the promise of DL to lower the cost of disease interpretation from OCT images. Hence, it would be necessary to filter out ungradable images beforehand for a better precision. However, at present, most of the ungradable images were filtered out manually before ground-truth labeling for abnormalities. Thus, our DLS could potentially be incorporated with other DLSs for further disease detection. Another important future application of our DLS is to be installed in SDOCT machines so that operators could be informed to repeat image acquisitions immediately if the DLS classifies the acquired image as ungradable. It would largely alleviate the burden of image quality control manually and efficiently provide images with better quality for further analysis.

There are several points to strengthen the training and evaluation of our proposed model. First, highly trained SDOCT human assessors reviewed both volumetric scans and reports rather than reviewing printout reports only for the precise labeling. Second, the external validation datasets were collected from different eye clinics, which enlarged the distribution variances of the dataset. Third, we generated activation heatmaps to visualize the discriminative regions for the model output reasoning. However, in our study, only optic disc scans from one type of SDOCT device were used, which might limit the applicability to other devices. In the next version, we shall develop a DLS trained with more kinds of scans, such as macular scan, from various types of devices. In addition, 3-D CNNs consume higher GPU memories, which might cause great extra cost for clinic usage. In the future, a model compression shall be applied to save the GPU memory cost.

5 Conclusions

Image quality control for the SDOCT volumetric scans is vital for accurate disease detection. Since it is time-consuming and requires the expertise of human graders, manual assessment for every volumetric scan would be tedious and even unfeasible, especially in a clinical center without experienced graders. To improve the efficiency and accuracy of image quality control, a computer-aided system based on DL was developed in our study.

The proposed DLS utilized irrelevancy reduction methods and an attention mechanism for the best diagnostic performance with the highest AUCs, better sensitivity, specificity, and accuracy in both primary and external validations, compared with other experimented models. Combining the observation from

the heatmaps, it proved that the proposed DLS learns similar features as human assessors do. Hence, as an automated filtering system, our proposed DLS could give more accurate and reasonable predictions. It would further advance the research on SDOCT image quality control as well as make SDOCT more feasible and reliable for disease detection.

Disclosures

No conflict of interest exists for any of the authors.

Acknowledgments

The work described in this paper was supported by the Research Grants Council – General Research Fund, Hong Kong (Reference No. 14102418), and the Bright Focus Foundation (Reference No. A2018093S).

References

1. J. F. de Boer et al., "Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography," *Opt. Lett.* **28**(21), 2067–2069 (2003).
2. J. S. Hardin et al., "Factors affecting cirrus-HD OCT optic disc scan quality: a review with case examples," *J. Ophthalmol.* **2015**, 1–16 (2015).
3. J. Chhablani et al., "Artifacts in optical coherence tomography," *Saudi J. Ophthalmol.* **28**(2), 81–87 (2014).
4. S. Asrani et al., "Artifacts in spectral-domain optical coherence tomography measurements in glaucoma," *JAMA Ophthalmol.* **132**(4), 396–402 (2014).
5. J. C. Downs and C. A. Girkin, "Lamina cribrosa in glaucoma," *Curr. Opin. Ophthalmol.* **28**(2), 113–119 (2017).
6. S. Liu et al., "Quality assessment for spectral domain optical coherence tomography (OCT) images," *Proc. SPIE* **7171**, 71710X (2009).
7. R. Lee et al., "Factors affecting signal strength in spectral-domain optical coherence tomography," *Acta Ophthalmol.* **96**(1), e54–e58 (2018).
8. C. Y. L. Cheung et al., "Relationship between retinal nerve fiber layer measurement and signal strength in optical coherence tomography," *Ophthalmology* **115**(8), 1347–1351.e2 (2008).
9. C. Y. Cheung, N. Chan, and C. K. Leung, "Retinal nerve fiber layer imaging with spectral-domain optical coherence tomography: impact of signal strength on analysis of the RNFL map," *Asia Pac. J. Ophthalmol.* **1**(1), 19–23 (2012).
10. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.* **15**(11), 3440–3451 (2006).
11. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
12. T. Kustner et al., "Automated reference-free detection of motion artifacts in magnetic resonance images," *Magn. Reson. Mater. Phys. Biol. Med.* **31**(2), 243–256 (2018).
13. L. Wu et al., "FUIQA: fetal ultrasound image quality assessment with deep convolutional networks," *IEEE Trans. Cybern.* **47**(5), 1336–1349 (2017).
14. F. Yu et al., "Image quality classification for DR screening using deep learning," in *39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Vol. 2017, pp. 664–667 (2017).
15. S. K. Kumaran et al., "Video trajectory classification and anomaly detection using hybrid CNN-VAE," arXiv: 1812.07203v1 (2018).
16. D. Li et al., "Anomaly detection with generative adversarial networks for multivariate time series," arXiv: 1809.04578v3 (2019).
17. R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," arXiv: 1802.06360v2 (2019).
18. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 770–778 (2016).
19. S. Xie et al., "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.* (2017).
20. J. Hu et al., "Squeeze-and-excitation networks," arXiv:1709.01507 (2018).
21. M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. Med.* **22**(3), 276–282 (2012).
22. N. Pandey and H. Agrawal, "Hybrid image compression based on fuzzy logic technology," *Int. J. Sci. Eng. Technol. Res.* **3**(9), 2438–2441 (2014).
23. Y. Ma et al., "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," *Biomed. Opt. Express* **9**(11), 5129–5146 (2018).
24. A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Process. On Line* **1**, 208–212 (2011).
25. R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *IEEE Int. Conf. Comput. Vision*, pp. 618–626 (2017).
26. T. Kustner et al., "A machine-learning framework for automatic reference-free quality assessment in MRI," *Magn. Reson. Imaging* **53**, 134–147 (2018).
27. J. De Fauw et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.* **24**(9), 1342–1350 (2018).
28. D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell* **172**(5), 1122–1131.e9 (2018).
29. F. A. Medeiros, A. A. Jammal, and A. C. Thompson, "From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs," *Ophthalmology* **126**, 513–521 (2019).
30. A. C. Thompson, A. A. Jammal, and F. A. Medeiros, "A deep learning algorithm to quantify neuroretinal rim loss from optic disc photographs," *Am. J. Ophthalmol.* **201**, 9–18 (2019).
31. A. R. Ran et al., "Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis," *Lancet Digital Health* **1**(4), e172–e182 (2019).
32. P. Prahs et al., "OCT-based deep learning algorithm for the evaluation of treatment indication with anti-vascular endothelial growth factor medications," *Graefes Arch. Clin. Exp. Ophthalmol.* **256**(1), 91–98 (2018).

An Ran Ran is a PhD student in the Department of Ophthalmology and Visual Science, Chinese University of Hong Kong (CUHK). She obtained her master's degree in ophthalmology from the Capital Medical University and her medical degree from the Shanghai Jiao Tong University. Her research interests are glaucoma, ocular imaging, and artificial intelligence (AI). She has recently published a paper titled "Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography—a retrospective training and validation deep-learning analysis" as a first author and other seven papers as co-author.

Jian Shi is a research assistant (computer vision) at CUHK. He was a former GE Power employee and was awarded an Impact Award. He received his MSc degree in computer science from the University of Leicester in 2018.

Amanda K. Ngai is a fourth year medical student at CUHK. She spent 8 years studying abroad in the United Kingdom. Her current research focuses on using AI to aid the processing and grading of optical coherence tomography scans.

Wai-Yin Chan received her BSc degree in biochemistry with first class honors from CUHK, in 2017. She then continued studying medicine at CUHK. She has been involved in laboratory and clinical research in the past years, with projects in apoptosis, cancer drug toxicology, and epidemiology of gastric cancer.

Poemen P. Chan is a glaucoma specialist and an assistant professor in the Department of Ophthalmology and Visual Sciences, CUHK, and honorary associate consultant at Hong Kong Eye Hospital (HKEH).

Alvin L. Young is the cluster coordinator in the Department of Ophthalmology and Visual Sciences, Prince of Wales Hospital (PWH) and Alice Ho Miu Ling Nethersole Hospital and deputy hospital chief executive at PWH. He was the chairman of the Hong Kong Hospital Authority Coordinating Committee in Ophthalmology (2013 to 2017). He serves as clinical professor (honorary) of DOVS at CUHK and is a visiting professor at STU-CUHK Joint Shantou International Eye Center.

Hon-Wah Yung is a consultant at Tuen Mun Eye Center, Hong Kong, an honorary clinical associate professor in the Department of Ophthalmology and Visual Sciences, CUHK, and a council member of the College of Ophthalmologists of Hong Kong (COHK).

Clement C. Tham is the chairman of the Department of Ophthalmology and Visual Sciences, CUHK; S.H. Ho Professor of ophthalmology and visual sciences, CUHK; honorary chief-of-service, HKEH; director, CUHK Eye Center, CUHK; vice president (general affairs), COHK; secretary general and CEO, Asia-Pacific Academy of Ophthalmology; treasurer, International Council of Ophthalmology; vice president, Asia-Pacific Glaucoma Society; and chair, Academia Ophthalmologica Internationalis.

Carol Y. Cheung is an assistant professor at Department of Ophthalmology and Visual Sciences, CUHK, council board member and treasurer of Asia Pacific Tele-Ophthalmology Society, and secretary general of Asia Pacific Ocular Imaging Society. She has been working in the field of ocular imaging for more than 10 years, focusing on development and application of image analysis techniques and AI for studying eye diseases, including diabetic retinopathy, glaucoma, and relationship between retinal imaging markers and Alzheimer's disease.