tvst

Artificial Intelligence

# Generalizability of Deep Neural Networks for Vertical Cup-to-Disc Ratio Estimation in Ultra-Widefield and Smartphone-Based Fundus Images

**Boon Peng Yap[1,*], Li Zhenghao Kelvin[2–4,*], En Qi Toh[3], Kok Yao Low[2,4], Sumaya Khan Rani[2,4], Eunice Jin Hui Goh[2,4], Vivien Yip Cherng Hui[2,4], Beng Koon Ng[1], and Tock Han Lim[2–4]**

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
[2] Department of Ophthalmology, Tan Tock Seng Hospital, Singapore, Singapore
[3] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[4] National Healthcare Group Eye Institute, Singapore, Singapore

**Correspondence:** Li Zhenghao Kelvin, Department of Ophthalmology, Tan Tock Seng Hospital, 11 Jln Tan Tock Seng, Singapore 308433, Singapore. e-mail: kelvin_li@ttsh.com.sg

**Received:** October 12, 2023
**Accepted:** February 19, 2024
**Published:** April 3, 2024

**Keywords:** optic disc; cup/disc ratio; deep learning

**Citation:** Yap BP, Kelvin LZ, Toh EQ, Low KY, Rani SK, Goh EJH, Hui VYC, Ng BK, Lim TH. Generalizability of deep neural networks for vertical cup-to-disc ratio estimation in ultra-widefield and smartphone-based fundus images. Transl Vis Sci Technol. 2024;13(4):6, https://doi.org/10.1167/tvst.13.4.6

**Purpose:** To develop and validate a deep learning system (DLS) for estimation of vertical cup-to-disc ratio (vCDR) in ultra-widefield (UWF) and smartphone-based fundus images.

**Methods:** A DLS consisting of two sequential convolutional neural networks (CNNs) to delineate optic disc (OD) and optic cup (OC) boundaries was developed using 800 standard fundus images from the public REFUGE data set. The CNNs were tested on 400 test images from the REFUGE data set and 296 UWF and 300 smartphone-based images from a teleophthalmology clinic. vCDRs derived from the delineated OD/OC boundaries were compared with optometrists' annotations using mean absolute error (MAE). Subgroup analysis was conducted to study the impact of peripapillary atrophy (PPA), and correlation study was performed to investigate potential correlations between sectoral CDR (sCDR) and retinal nerve fiber layer (RNFL) thickness.

**Results:** The system achieved MAEs of 0.040 (95% CI, 0.037–0.043) in the REFUGE test images, 0.068 (95% CI, 0.061–0.075) in the UWF images, and 0.084 (95% CI, 0.075–0.092) in the smartphone-based images. There was no statistical significance in differences between PPA and non-PPA images. Weak correlation ($r = -0.4046$, $P < 0.05$) between sCDR and RNFL thickness was found only in the superior sector.

**Conclusions:** We developed a deep learning system that estimates vCDR from standard, UWF, and smartphone-based images. We also described anatomic peripapillary adversarial lesion and its potential impact on OD/OC delineation.

**Translational Relevance:** Artificial intelligence can estimate vCDR from different types of fundus images and may be used as a general and interpretable screening tool to improve community reach for diagnosis and management of glaucoma.

## Introduction

Glaucoma is the second leading cause of blindness[1] and is projected to affect more than 111.8 million people in 2040.[2] Diagnosing glaucoma requires manual assessment of the optic disc regions to identify abnormal cupping of the optic nerve, a key characteristic of glaucoma. The vertical cup-to-disc ratio (vCDR), defined as the ratio between the vertical diameter of the optic cup (OC) and the vertical diameter of the optic disc (OD), has been shown to be useful for assessing the risk of glaucoma.[3] vCDR can be inferred from images acquired from different modalities, such as optical coherence tomography[4,5] and fundus photography.[6–8] The latter offers a low-cost solution for glaucoma screening, and portable fundus imaging devices based on smartphones have the potential to be used in large






translational vision science & technology

population-based glaucoma screening.[9] However, a precise measurement of vCDR would require detailed delineations of the OD and OC, which can be a time-consuming process, and hinder its use in large-scale screening.

Automated systems based on deep learning have been proven to be effective in classifying glaucoma in standard fundus images (with an angle of view within the range of 30–60 degrees),[10–13] ultra-widefield fundus images (200 degrees angle of view),[14] and smartphone-based images.[9] However, classification-based convolutional neural networks (CNNs) are often criticized for their black-box nature and lack of interpretability. While methods such as saliency map[15] attempts to explain the decision made by a CNN in a post hoc manner, the resulting explanations might not always be reliable.[16] One possible solution for an interpretable glaucoma diagnosis system is to design a modular pipeline consisting of multiple detection/segmentation-based CNNs, where each CNN focuses on detecting/delineating key structures such as OD/OC. Once the key structures have been delineated, quantitative measurements such as vCDR can be derived automatically as inputs for inherently interpretable machine learning models.

While prior work has considered deep learning for automated vCDR estimation from fundus images,[7,8] they were limited to standard fundus images and have not been validated across other types of fundus images, including ultra-widefield (UWF) and smartphone-based images. Having generic models that can interoperate between different image types would be beneficial in terms of scalability and maintainability of the system. The same system can be used when patients underwent screening via UWF imaging at a hospital setting, or it can be used in a telemedicine setting through portable/smartphone-based imaging devices operated by nonmedically trained personnel. The purpose of this study is to validate the effectiveness and generalizability of our deep learning system for vCDR estimation across standard, UWF, and smartphone-based images.

# Methods

## Standard Fundus Images

The REFUGE data set[17] is a publicly available data set consisting of 1,200 color fundus images with reference OD/OC delineations and glaucoma labels annotated by seven glaucoma specialists. The images were divided into train, validation, and test sets with a ratio of 1:1:1, each consisting of 40 glaucomatous and 360 nonglaucomatous cases. The images from the train set were acquired using a Zeiss Visucam 500 camera, Germany (with a resolution of 2124 × 2056 pixels) while the images from the validation and test set were acquired using a Canon CR-2 camera, Japan (with a resolution of 1634 × 1634 pixels). All images correspond to patients in a Chinese population and were retrieved retrospectively from multiple hospital and clinical studies. In this study, the REFUGE data set was used as the source of standard fundus images to investigate the feasibility of developing an automated vCDR estimation system that can generalize to a more diverse population and imaging devices.

**Table 1.** Demographics of the Patients in the AMK Study ($n = 151$; 300 Eyes)

| Characteristic | Value |
|---|---|
| Age, y | 61.54 ± 10.87 (27–87) |
| Sex | |
| Male | 87 (57.6) |
| Female | 64 (42.4) |
| Ethnicity | |
| Chinese | 109 (72.2) |
| Indian | 30 (19.9) |
| Malay | 10 (6.6) |
| Others | 2 (1.3) |

Values are presented as *n* (%) or mean ± SD (range).

## UWF and Smartphone-Based Images

In total, 151 patients aged 27 to 87 years attending regular teleophthalmology follow-ups at the Ang Mo Kio (AMK) Specialist Centre in Singapore were recruited for this project. Details of the subject demographics in the AMK study are shown in Table 1. These patients have a range of ophthalmologic problems, including diabetic retinopathy, cataracts, glaucoma, or age-related macular degeneration. All patients had their pupils dilated prior to having their fundus photos taken by the Zeiss CLARUS 500 camera and the oDocs Nun IR portable camera, oDocs Eye Care, New Zealand (as demonstrated in Fig. 1) in a dark consultation room. The study was performed in accordance with the ethical standards of the Declaration of Helsinki and was approved by the National Healthcare Group Domain Specific Review Board (2022/00556). Written informed consent was obtained from all participants.

All UWF and smartphone fundus photos were then manually graded according to a modified version of the Singapore Integrated Diabetic Retinopathy Program

Figure 1. Demonstration of the image acquisition process using the smartphone-based oDocs Nun IR camera.

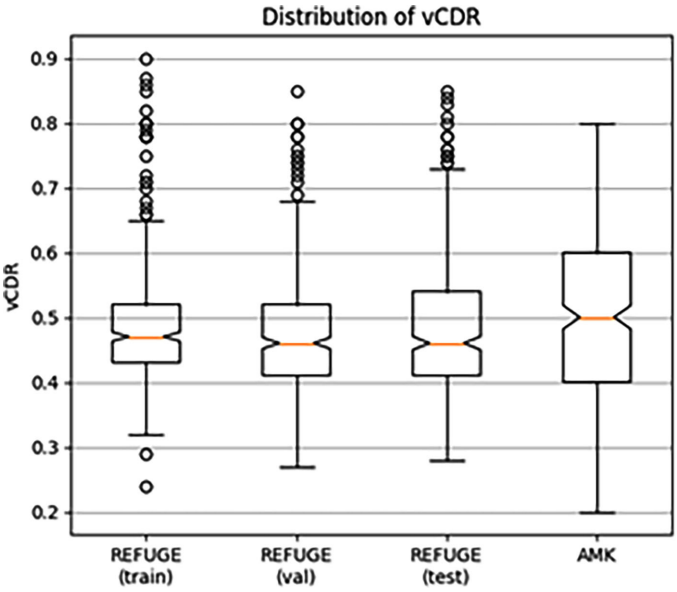


Figure 2. Distribution of vCDR in each data set.

Table 2. Gradeability of Smartphone-Based Fundus Photos Using a Modified Version of the SiDRP Criteria

| Characteristic | Left Eye ($n = 149$), $n$ (%) | Right Eye ($n = 151$), $n$ (%) |
|---|---|---|
| Gradable | 100 (67.1) | 97 (64.2) |
| Ungradable | 49 (32.9) | 54 (35.8) |
| Ungradable causes | | |
| Uncorrectable underexposure | 2 (1.3) | 6 (4.0) |
| Uncorrectable overexposure | 14 (9.4) | 14 (9.3) |
| Severe obscuration | 30 (20.1) | 38 (25.2) |
| Insufficient focus | 18 (12.1) | 14 (9.3) |
| Incorrect positioning | 1 (0.7) | 1 (0.7) |

Each ungradable image may correspond to more than one ungradable cause.

(SiDRP) criteria for fundus image quality assessment (see Supplementary Table S1 for the details of the criteria). ImageJ version 1.53t (National Institutes of Health, Bethesda, MD, USA) was used to optimize the brightness and contrast of the images using an in-built slider function. All UWF images were deemed gradable while the causes for each smartphone-based photo to be ungradable were also recorded and tabulated in Table 2. The original SiDRP criteria require visibility of the fundus up to 1 disc diameter beyond the vascular arcade, which converted all smartphone-based photos to ungradable. As such, this particular criterion was not taken into consideration in the grading of the photos. The limited view of the smartphone-based photos is likely an inherent limitation, as the images retain most of the key features (e.g., fovea, optic disc, vascular arcades) despite not fulfilling that criterion.

The reference vCDR values for the AMK UWF images were manually annotated by optometrists from the AMK Specialist Centre via visual estimation on the captured UWF images. The annotations were then vetted by the primary eye care team at Tan Tock Seng Hospital. The distributions of the reference vCDR values for each data set are shown in Figure 2.

## Automated vCDR Estimation System

An automated system based on deep learning was developed for vCDR estimation. The pipeline of the vCDR estimation system is shown in Figure 3. The automated system consists of a detection CNN and a segmentation CNN, both based on the U-Net architecture[18] with a ResNet-18 encoder.[19] The detection CNN was trained to detect ODs from whole fundus images by fitting square bounding boxes around the ODs to generate region of interest (ROI) images for the segmentation CNN, which in turn was trained to delineate detailed boundaries for OD and OC. The ROI image was dynamically extracted based on the diameter of the detected OD to account for the heterogenic variability in the optic nerve head size. Specifically, a square bounding box with sides of 1.1 times the diameter of the detected OD was extracted as the ROI

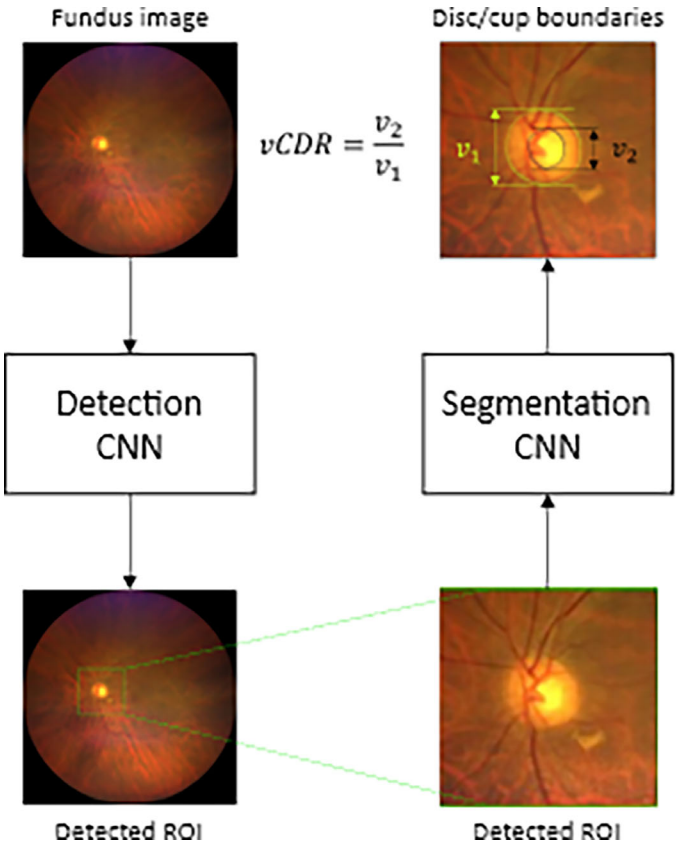


**Figure 3.** Pipeline of the automated vCDR estimation system. $v_1$ and $v_2$ represent the vertical optic disc diameter and vertical optic cup diameter, respectively.

image. The vCDR value was calculated by measuring and dividing the vertical diameter of the OC boundary with the vertical diameter of the OD boundary.

The outputs of the CNNs were normalized to a range of 0 to 1 using the sigmoid function to obtain the probability maps. A threshold value of 0.5 was then used to convert the probability maps into binary masks of OD/OC. If the binary masks were empty, a fallback mechanism based on Otsu's method[20] was used to determine the threshold value automatically. Finally, a postprocessing step based on the binary closing operation was performed to fill in small holes in the binary masks to generate the OD/OC delineations.

To improve the robustness and quality of the OD/OC delineations, test-time augmentation (TTA) was applied during the inference process. TTA is commonly used to remove noise by averaging the outputs obtained from multiple augmented versions of an input image. The augmentations used for TTA in the automated vCDR estimation system include image scaling (at 80%/90%/100% of the input resolution), rotation (90/180/270 degrees), and flipping (horizontal/vertical).

## Training of Deep Neural Networks

A summary of the data sets used in this study is given in Table 3. The 400 images from the validation set of the REFUGE data set were used to guide the selection of hyperparameters. In the model development phase, a set of models was trained on the 400 training images of the REFUGE data set to select the best hyperparameters (learning rate, weight decay) based on the performance on validation set. After the optimal hyperparameters were determined, a new model was trained on 800 images from both the training and validation sets to make use of all the images available in the development phase to improve model performance. The trained model was then used for evaluating the segmentation and vCDR estimation performance on independent test images. Both CNNs were initialized with parameters pretrained on the ImageNet data set.[21] The detection CNN was trained using an Adam optimizer with a learning rate of 0.1 and a batch size

**Table 3.** Details of the Data Sets Used in This Study

| Characteristic | REFUGE | AMK-UWF | AMK-Smartphone |
|---|---|---|---|
| Number of training images | 400 | — | — |
| Number of validation images | 400 | — | — |
| Number of testing images | 400 | Total: 296<br>Left: 147<br>Right: 149 | Total: 300<br>Left: 149<br>Right: 151 |
| Device | ZEISS VISUCAM 500/Canon CR-2 | ZEISS CLARUS 500 | oDocs Nun IR |
| Angle of view | 30°/45° | 200° | 45°–55° |
| Resolution (pixels) | 2124 × 2056/1632 × 1634 | 1110 × 1111 | 2880 × 2160 |

For REFUGE, the validation images were merged with the training images to train the final model after the hyperparameter selection phase.

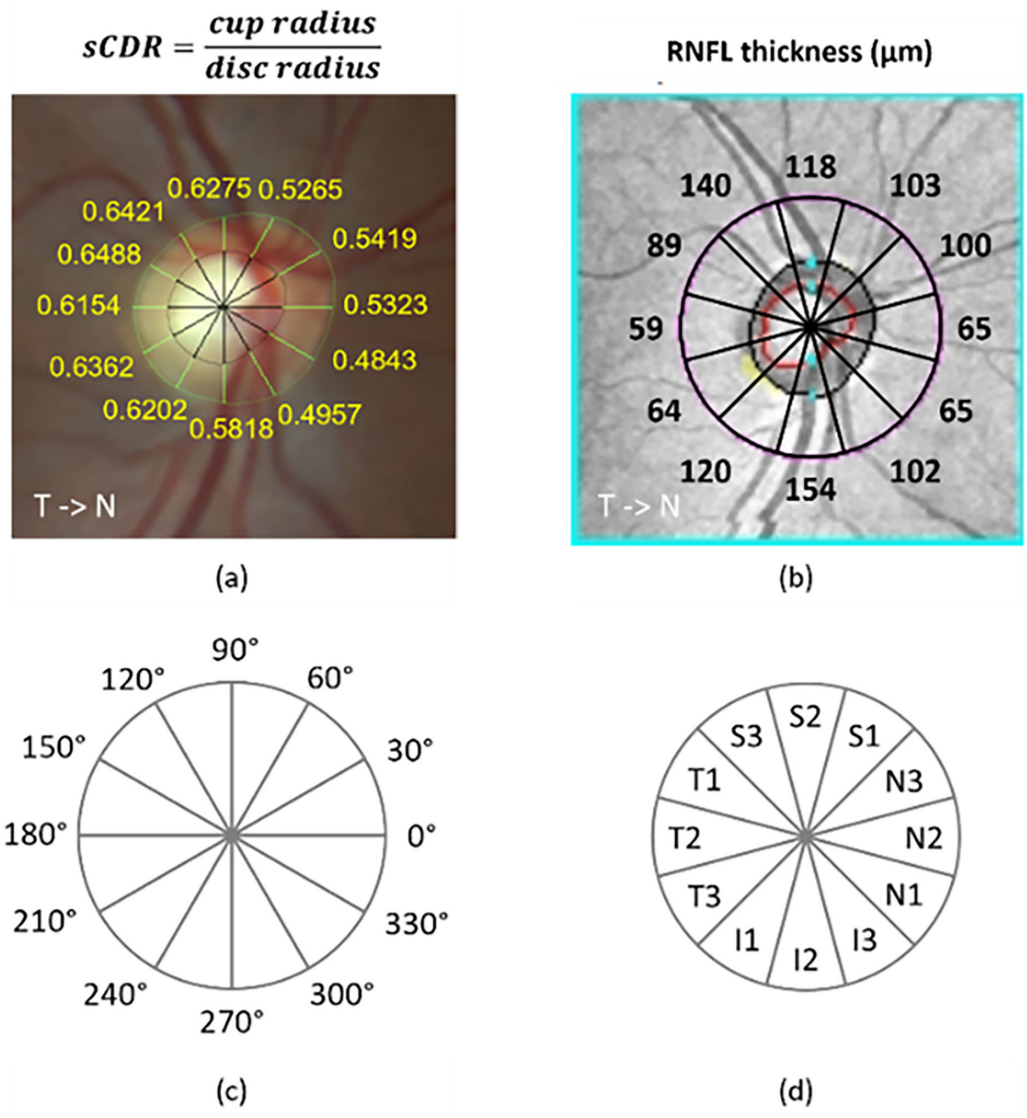


**Figure 4.** (**a**) Measurement of sCDR on the ROI of UWF image. (**b**) RNFL thickness measured on each sector of the OCT en face fundus image. (**c**) Labels of the angle correspond to each sector where sCDR was measured. (**d**) Labels of each sector where RNFL thickness was measured.

of 8 for 100 epochs, while the segmentation CNN was trained using an Adam optimizer with a learning rate of 0.001 and a batch size of 5 for 100 epochs. The learning rates were decreased gradually using a cosine scheduler. The input images were resized to 800 × 800 pixels and 256 × 256 pixels for the detection and segmentation CNN, respectively. Pixel values of each image were normalized to a range of 0 to 1, and standard data augmentations including random scaling (in the range of 0.6 to 1.0), random rotations (up to 45 degrees), random horizontal and vertical flips, random brightness/saturation/contrast shifts (in the range of 0.6 to 1.4), and random hue shifts (in the range of −0.1 to 0.1) were used to reduce the chance of overfitting during the training process. The CNNs were implemented using the PyTorch framework[22] on Python 3.9.16 and trained on a desktop computer with a 24-GB NVIDIA RTX 3090 GPU.

## Studying the Effect of Peripapillary Atrophy

Peripapillary atrophy (PPA) is characterized by the choroidal thinning of the retinal layers around the optic nerve. As it changes the appearance of regions surrounding the optic nerve, PPA may be a potential confounder that would result in a wrong OD segmentation and vCDR value. To study whether PPA would negatively impact the performance of vCDR

estimation, the UWF images collected from the AMK Specialist Centre were individually assessed for the presence of PPA by a trained ophthalmologist. Among the 296 UWF images, 57.8% of the images ($n = 171$) were found to contain PPA, and findings from UWF images were applied to the paired smartphone-based images.

### Comparing Retinal Nerve Fiber Layer Thickness With Sectoral CDR

Among the 151 patients enrolled in this study, 17 patients underwent additional optical coherence tomography (OCT) screening using the Zeiss CIRRUS HD-OCT system, resulting in 34 OCT scans with corresponding retinal nerve fiber layer (RNFL) thickness maps. The sectoral RNFL thickness values extracted from 32 (excluding two scans with signal strength <6) RNFL thickness maps were compared against the sectoral cup-to-disc ratio (sCDR) inferred from delineated UWF images. Figure 4 shows how the sCDR was measured and how it is compared with the RNFL thickness. A circle centered at the OC boundary was divided into 12 equal sectors to match the sectors of the RNFL thickness map. The sCDR of each sector was then calculated as the ratio of the cup-to-disc radius measured at the center line of each sector. Before calculating the sCDRs, the images of the left eyes were flipped horizontally to align the nasal and temporal regions with the right-eyed images. Partial Pearson correlation test was conducted to study the correlation between sCDR and RNFL thickness at each sector, with the confounding effect of the optic disc area[23,24] (extracted from the OCT scan report) removed.

### Statistical Analysis

The performance of the automated system in delineating OD/OC was evaluated on the test set of the REFUGE data set using the intersection over union (IoU) metrics. Mean absolute error (MAE) was used to compare the estimated vCDR to the reference standard. vCDRs estimated from the REFUGE test set were rounded to the nearest two decimal places while the estimations from the AMK UWF and smartphone-based images were rounded to the nearest 0.05 to match the precision of the optometrists' annotations. The 95% confidence intervals (CIs) were estimated for all performance metrics. The agreement between the automated system and references was examined using Bland–Altman plot analysis. All statistical analyses were performed using the SciPy software package[25] on Python 3.9.16.

## Results

### Performance of vCDR Estimations

The best IoUs achieved on the 400 validation images from the REFUGE data set during the model development phase were 0.901 (95% CI, 0.893–0.910) and 0.774 (95% CI, 0.762–0.786) for optic disc and optic cup segmentation, respectively. The final segmentation CNN (trained on the combination of 400 training + 400 validation images from the REFUGE data set) were evaluated on 400 test images from the REFUGE data set and achieve an IoU of 0.909 (95% CI, 0.905–0.914) on OD delineation and an IoU of 0.796 (95% CI, 0.787–0.805) on OC delineation. For vCDR estimation, the automated system achieved MAEs of 0.040 (95% CI, 0.037–0.043), 0.068 (95% CI, 0.061–0.075), and 0.084 (95% CI, 0.075–0.092) in the test set of the REFUGE data set ($n = 400$), AMK UWF images ($n = 296$), and AMK smartphone-based images ($n = 300$), respectively. Comparisons of the MAE achieved on the REFUGE test set with the top submissions in the REFUGE 2018 challenge[17] are shown in Table 4. When considering the gradeability of the smartphone-based images under the SiDRP guidelines, the MAEs on the gradable images (0.076; 95% CI, 0.067–0.085; $n = 197$) were lower ($P = 0.0228$) compared to ungradable images (0.098; 95% CI, 0.079–0.116; $n = 103$). Scatterplots comparing the estimated vCDRs with the reference values are provided in Figure 5.

Examples of the qualitative results from the automated system are shown in Figures 6 and 7. The Bland–Altman plots in Figure 8 visualize the differences of the estimated vCDRs with the reference standards in each testing image. On the test set of the

**Table 4.** Comparisons of vCDR Estimation Performance in Terms of MAE with the Top 5 Submissions in the REFUGE 2018 Challenge

| Team | Rank | MAE |
|---|---|---|
| Masker | 1 | 0.0414 |
| CUHKMED | 2 | 0.0450 |
| BUCT | 3 | 0.0456 |
| NKSG | 4 | 0.0465 |
| AIML | 5 | 0.0469 |
| This work | — | 0.040 |

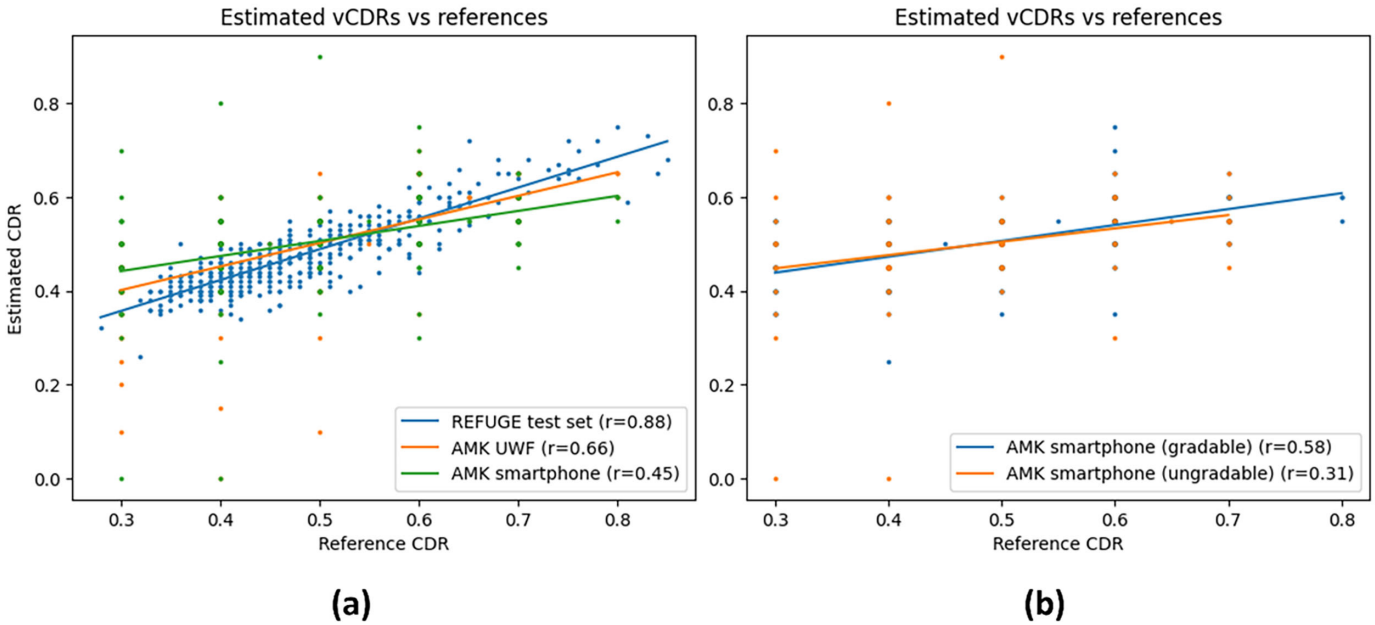


**Figure 5.** (**a**, **b**) Scatterplots of the estimated vCDRs versus the reference values in each data set. Pearson correlation coefficient of the predicted vCDR versus reference vCDR is provided for each testing set.

REFUGE data set ($n = 400$, Fig. 8a), the mean difference was 0.005 with the 95% limit of agreement of −0.095 to 0.105. The mean differences in the AMK UWF images ($n = 296$, Fig. 8b) were −0.005 with a 95% limit of agreement of −0.182 to 0.173. For the gradable ($n = 197$, Fig. 8c) and ungradable ($n = 103$, Fig. 8d) smartphone-based images, the mean differences were −0.004 with a 95% limit of agreement of −0.196 to 0.188 and −0.021 with a 95% limit of agreement of −0.284 to 0.242, respectively.

In terms of the agreement of the estimated vCDRs within the paired UWF and smartphone-based images ($n = 296$), the Bland–Altman plots in Figure 9a show a high degree of agreement on the gradable images ($n = 194$, excluding three smartphone-based images without a paired UWF image) with zero mean difference and a 95% limit of agreement of −0.143 to 0.142. For the ungradable images ($n = 102$, excluding one smartphone-based image without a paired UWF image), the mean difference was −0.015, with a 95% limit of agreement of −0.235 to 0.205 (Fig. 9b).

## Effect of Peripapillary Atrophy

Table 5 shows the MAEs of the estimated vCDRs and the results of independent $t$-tests comparing the groups with PPA and the group without PPA. The qualitative results for images with PPA are given in Figure 10.

## Association Between OCT RNFL Thickness and Fundus sCDR

Table 6 shows the results of the correlation study of OCT RNFL thickness versus fundus CDR at each sector. Without correcting for optic disc size, there was no correlation between OCT RNFL thickness and fundus CDR in all sectors. After correcting for disc area, a weak negative correlation ($r = -0.4046$, $P < 0.05$) was observed at the 90°/S2 sector.

# Discussion

In this study, we have demonstrated that CNNs trained for vCDR estimation using standard images can generalize to UWF and smartphone-based imaging devices, even though the angle of view and image quality were drastically different from the training images. This is partly achieved through postprocessing techniques, including the hole-filling operation and TTA, which remove noise by averaging over multiple outputs. The developed models perform comparably with the top submissions in the REFUGE 2018 challenge. Among the top three submissions for vCDR estimation, Team BUCT (rank 3) leveraged two different U-Net models to separately segment the OD and OC, Team CUHKMED (rank 2) utilized an adversar-

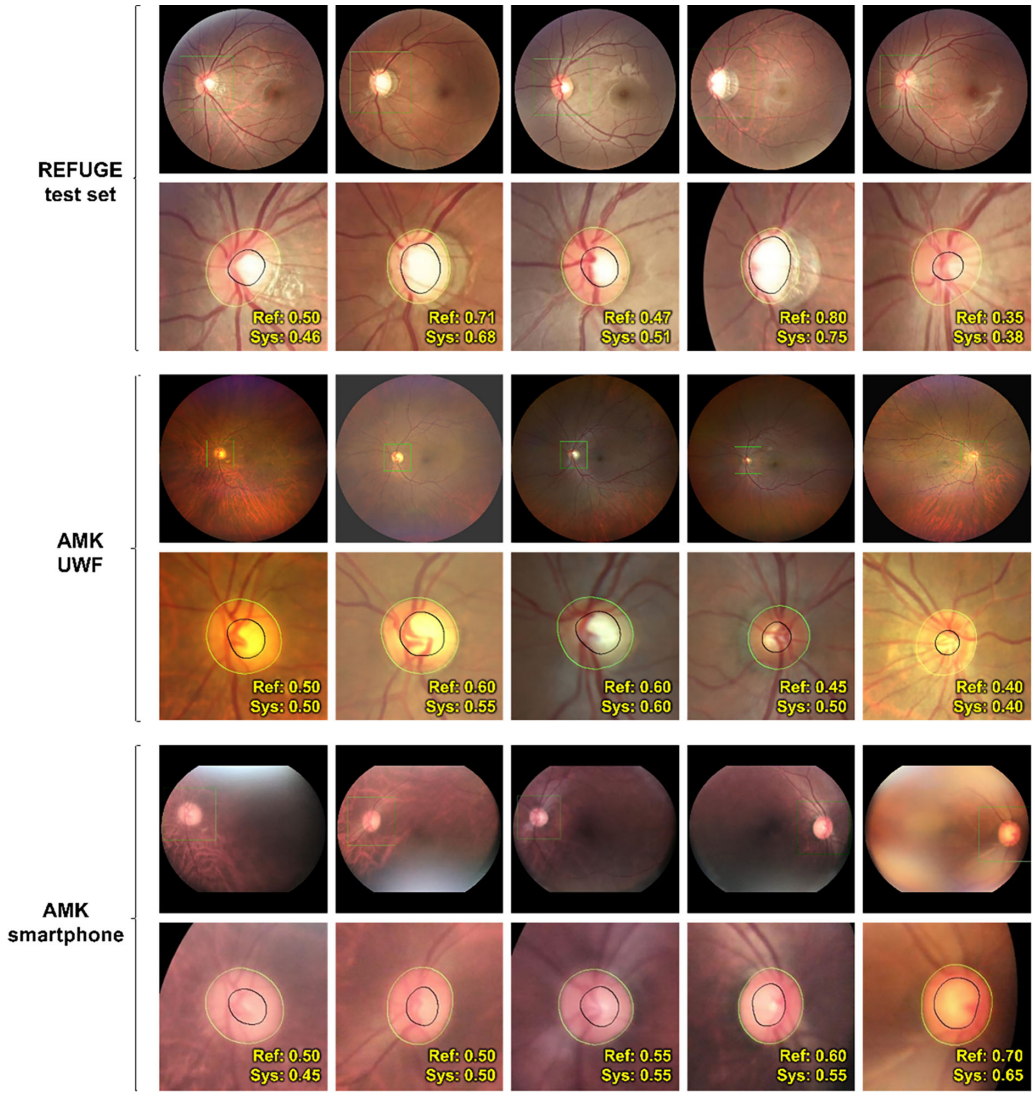


**Figure 6.** Qualitative results from the detection and segmentation CNNs in each data set. *Green bounding box* indicates the detected ROI while the *green* and *black outlines* delineate the OD and OC, respectively. The reference vCDR (Ref) and estimated vCDR (Sys) are shown at the *bottom right* of each ROI image.

ial learning framework to reduce performance degradation on the testing data set and the final prediction was obtained by an ensemble of five models, while Team Masker (rank 1) adopted a multiarchitecture approach to train multiple segmentation models on different subsets of the training data set. Our automated vCDR estimation system was based on two U-Net models, one for ROI extraction and another one for OD/OC segmentation. The main advantage of this two-stage design is that the ODs are roughly centered in the ROIs extracted by the first model, which allows the second model to focus only on delineating the boundaries of OD and OC instead of learning both ROI extraction and delineation tasks at the same time. By decoupling ROI extraction and OD/OC delineation tasks, it also provides some degree of future upgradability where either of the two models may be swapped out and replaced with an improved model. It was shown to be robust to different image qualities and was able to delineate reasonable boundaries for OD and OC from images that were deemed ungradable under the modified SiDRP guidelines for fundus image quality assessment. The interoperability between multiple imaging types allows the same vCDR estimation system to be deployed to different clinical setting with different imaging capability. Since the CNNs were based on relatively lightweight ResNet-18 encoders, they can potentially be embedded directly into a smartphone, allowing real-time vCDR estimation in an offline fashion. This will be useful

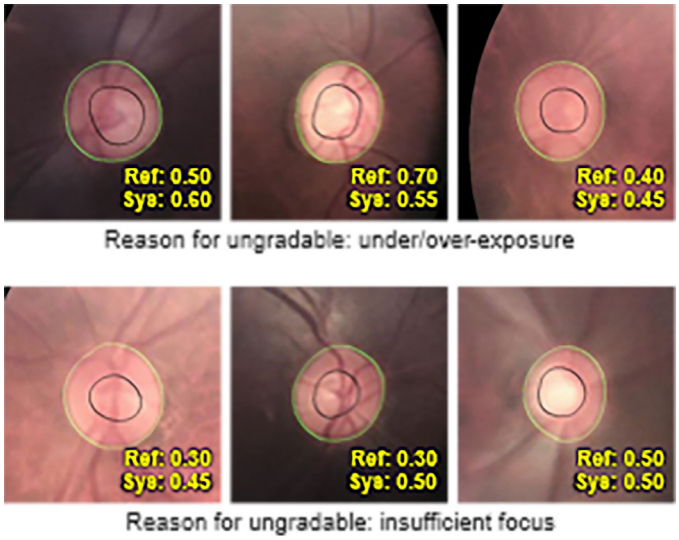


**Figure 7.** Qualitative results of the OD/OC boundaries on ungradable smartphone-based images. *Green* and *black outlines* delineate the OD and OC, respectively. The reference vCDR (Ref) and estimated vCDR (Sys) are shown at the *bottom right* of each image.

in a large-scale community screening setting, where the portable imaging devices might be operated by nonspecialists.

As a relatively new fundus imaging device, the smartphone-based camera offers greater flexibility and portability compared to its tabletop counterpart, which makes it an ideal choice for community screening in remote locations. As part of the study, image quality control was performed against a widely accepted national standard. However, there is still a gap in the image quality between tabletop and portable devices. Image quality is an important factor for artificial intelligence (AI)–based diagnostic systems, and some AI-based diagnostic software for fundus images[26,27] include an image quality assessment algorithm to detect images with insufficient quality. In terms of vCDR estimation, the quality of the captured smartphone-based images matters as the MAE is higher in ungradable images compared to gradable images. Thus, further effort is required to improve the quality of smartphone-based images, such as improved camera sensor and stabilization to minimize unwanted image artifacts. Additionally, an automated image quality assessment algorithm may be developed and embedded directly into smartphone's display to guide the image acquisition process in real time.

The Bland–Altman plot analysis showed that the developed system achieves a good agreement with the optometrist's annotation in both UWF and gradable smartphone-based images. The results are similar to previous work on other standard fundus data sets, which reported a mean difference of 0.0034 and 95% limit of agreement of −0.2931 to 0.2998.[8] In addition, vCDRs estimated for paired UWF and smartphone-based images show a high degree of agreement in gradable images, suggesting that portable devices may be a viable option for vCDR estimation in locations without access to UWF or a standard fundus imaging device. For the ungradable images, there was a moderate degree of agreement with the UWF counterpart (Fig. 9b) and a larger deviation with the reference values toward lower vCDRs (Fig. 5b). This further proves that low image quality will degrade the performance of vCDR estimation. In addition, the results in Table 5 suggest that the differences in vCDR estimation performance between the group with PPA and the group without PPA are statistically insignificant. Furthermore, as shown in the qualitative results, the developed models are also reasonably robust against other confounders such as cotton wool spots (Fig. 10,

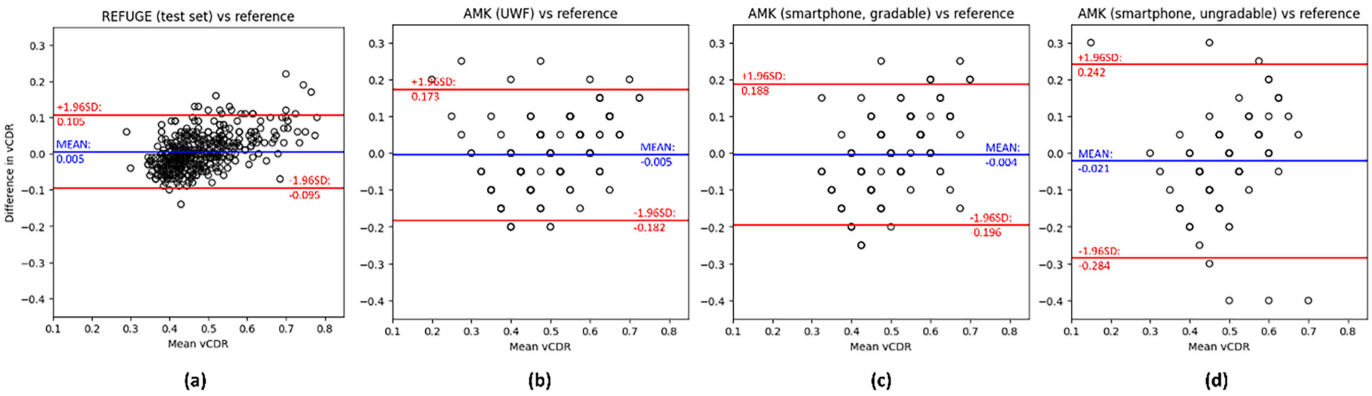


**Figure 8.** (**a–d**) Bland–Altman plots comparing the estimated vCDRs of the automated system with optometrists' annotations in standard, UWF, and smartphone-based (gradable versus ungradable) images.

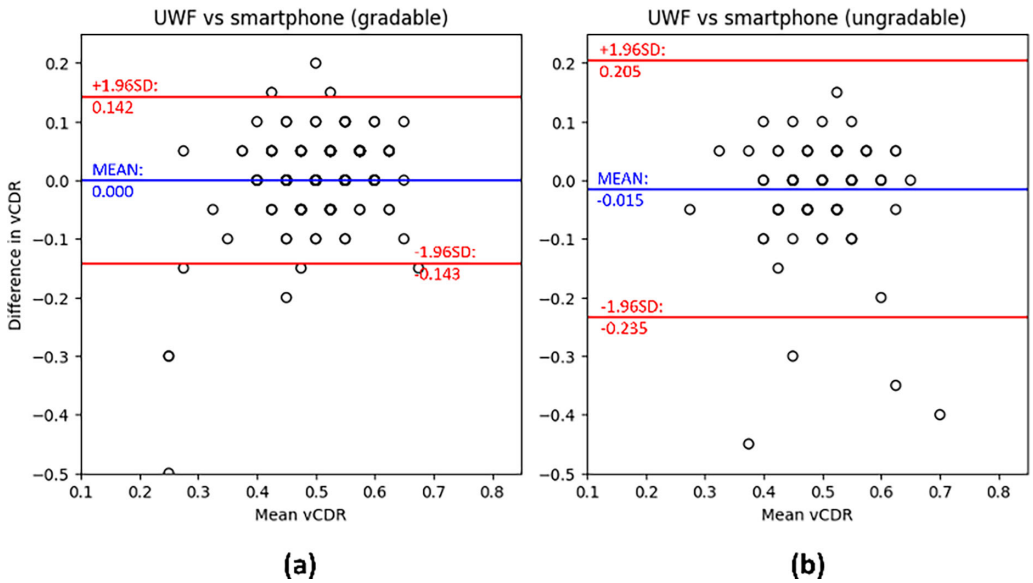


**Figure 9.** (**a**, **b**) Bland–Altman plots comparing the vCDRs estimated for the UWF and smartphone-based images using the same automated system.

**Table 5.** MAE of the Estimated vCDRs and Independent *t*-test Results Comparing the Difference Between Images With and Without PPA

| Characteristic | Without PPA ($n = 125$) | With PPA ($n = 171$) | $t(294)$ | *P* Value |
|---|---|---|---|---|
| AMK UWF | $0.070 \pm 0.010$ | $0.066 \pm 0.009$ | 0.4536 | 0.6504 |
| AMK smartphone | $0.078 \pm 0.014$ | $0.085 \pm 0.012$ | −0.8568 | 0.3923 |

second image) and nerve fiber layer reflection (Fig. 10, fourth image). However, as illustrated in Figure 11, the model may fail to delineate accurate OD boundaries when the images contain unusual abnormalities or are underexposed. We termed these abnormalities found around the optic disc as an anatomic peripapillary adversarial lesion (APAL) given that they act as adversarial signals to the deep learning system while being naturally occurring anatomic features. While the issue of underexposure can be resolved by recapturing the photos, the APAL requires further investigation (e.g., through better model design or training on more images that exhibit such patterns). We will leave this investigation for future work.

The system developed in this study can be used to delineate the boundaries of OD and OC automatically, which enables automatic quantification and extraction of CDR at different sectors of the fundus image (Fig. 4a), allowing us to study any potential association between the sectoral RNFL thickness and sCDR. A previous study reported that the mean global peripapillary RNFL thickness does not correlate with the CDR derived from fundus images ($r = 0.029$; $P = 0.858$) in patients with suspected pediatric glaucoma,[28] while

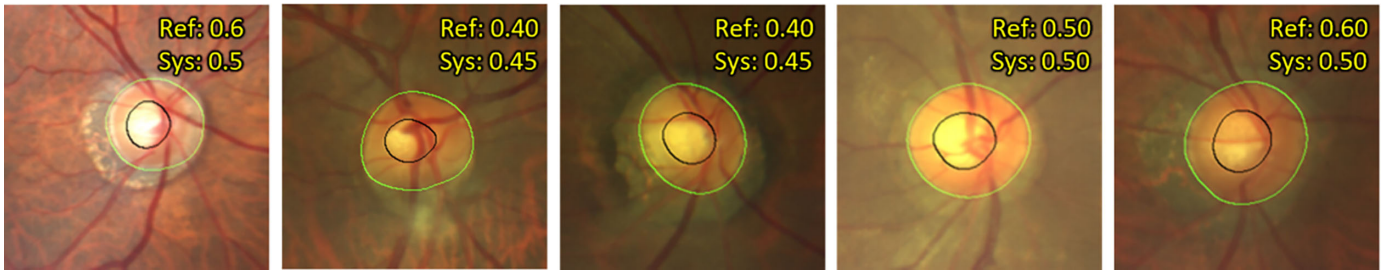


**Figure 10.** Qualitative results on UWF images with PPA.

**Table 6.** Pearson Correlation Coefficient of sCDR Versus RNFL Thickness at Each Sector

| sCDR Angle | RNFL Sector | *r* | *P* Value | *r*[a] | *P* Value[a] |
|---|---|---|---|---|---|
| 0° | N2 | −0.2041 | 0.2624 | 0.0698 | 0.7091 |
| 30° | N3 | 0.0083 | 0.9640 | 0.0205 | 0.9128 |
| 60° | S1 | −0.1550 | 0.3970 | −0.3092 | 0.0906 |
| 90° | S2 | −0.2302 | 0.2050 | −0.4046 | **0.024** |
| 120° | S3 | −0.0206 | 0.9110 | 0.1384 | 0.4578 |
| 150° | T1 | −0.1149 | 0.5312 | 0.1016 | 0.5867 |
| 180° | T2 | −0.0077 | 0.9667 | 0.0032 | 0.9862 |
| 210° | T3 | −0.0739 | 0.6878 | −0.0089 | 0.962 |
| 240° | I1 | 0.1766 | 0.3336 | 0.2756 | 0.1334 |
| 270° | I2 | 0.0574 | 0.7549 | 0.1032 | 0.5807 |
| 300° | I3 | −0.1973 | 0.2791 | −0.2022 | 0.2754 |
| 330° | N1 | 0.0177 | 0.9232 | 0.0686 | 0.7137 |

Bold type indicates statistical significance.
[a]Results of partial Pearson correlation corrected for disc area.

another study reported a weak correlation between average RNFL thickness in Stratus OCT with CDR derived from slit-lamp examination ($r = -0.58918$).[29] In contrast, our sectoral-based study reveals no correlations in all sectors except the 90° sector. To the best of our knowledge, this study is the first to investigate the association between the sectoral fundus CDR and OCT RNFL thickness.

Besides enabling association study on a sectoral basis, vCDR or sCDR derived from the automated system can be combined with other variables such as intraocular pressure and visual acuity score to aid clinicians in the diagnostic decision-making process, or it can be feed into an inherently interpretable models such as a decision tree[30] to make decisions in a fully automated fashion. Under this design, the derivation of each quantitative measurement that contributes to a diagnostic decision is traceable and verifiable through visual inspection of the model outputs. For example, the OD/OC delineations responsible for vCDR derivation can be visually inspected to identify/troubleshoot any failure in the delineation process and

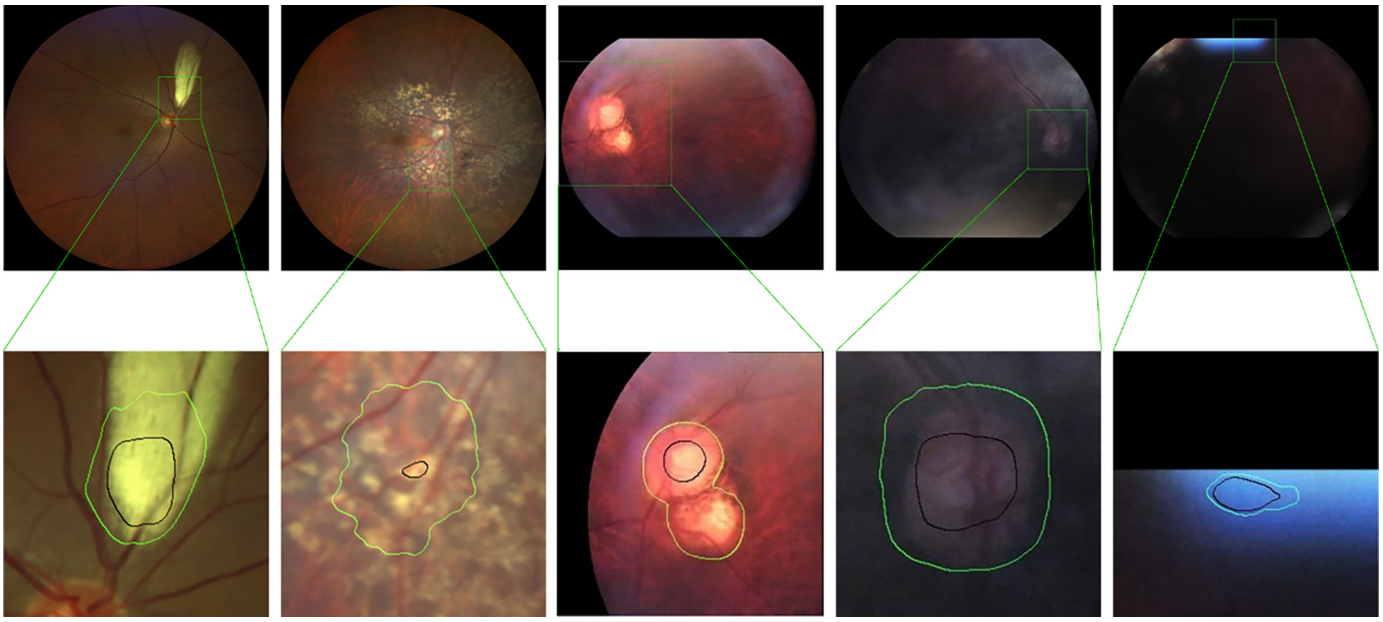

**Figure 11.** Examples of the failure case. The first two columns correspond to UWF images while the last three columns correspond to smartphone-based images.

rectified before it is used in the diagnostic process. This is in contrast to regression-based approaches[31,32] where the vCDR is directly estimated in an end-to-end manner. While the regression-based model alleviates the requirement of pixel-level segmentation masks during the training phase, its inherent black-box nature limits the interpretability of the predicted outputs.

This study has a few limitations. First, due to TTA, 18 augmented images were constructed from each image (four rotations plus two flips for each of the three resolutions). This resulted in an average processing time of 0.7 seconds per image, which is 2.7 times longer compared to 0.26 seconds per image without TTA. Nevertheless, a subsecond performance of 0.7 seconds per image is acceptable for a near real-time use case and could be further improved with better optimization and hardware. Second, the system was validated on a relatively small population that was predominantly Chinese. Lastly, the correlation study on fundus sCDR versus OCT RNFL thickness was limited to a small sample size and nonglaucomatous cases. We believe that the automated OD/OC detection and boundary segmentation models will be a valuable tool to encourage further studies in this direction.

## Acknowledgments

The authors thank the optometrists involved in the annotation of vCDR: Lim Yunchong, Aw Lingli, Tan Huiling Sheryl, Yau Siew Lian, Leanne Lee Huixian, Vivian Ng Voon Li, Jasmine Chua Ling Ling, Marilyn Puah Geok Ling, Kang Kok Kai, Tay Yuan Fang, Quek Zuoling, Joannabell Tan Mei Geok, Sarah Koh Sian Keng, Zhuo Jialong, Ang Jian Xiong, and Yang Lijun.

Supported by the Ng Teng Fong Healthcare Innovation Programme (Innovation Track), project code NTF_DEC2021_1_C1_D_03.

Disclosure: **B.P. Yap**, None; **L.Z. Kelvin**, None; **E.Q. Toh**, None; **K.Y. Low**, None; **S.K. Rani**, None; **E.J. Hui Goh**, None; **V.Y.C. Hui**, None; **B.K. Ng**, None; **T.H. Lim**, None

* BPY and LZK are co-first authors.

## References

1. Quigley HA. Number of people with glaucoma worldwide. *Br J Ophthalmol*. 1996;80:389–393.
2. Tham Y-C, Li X, Wong TY, et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 2014;121:2081–2090.
3. Jonas JB, Bergua A, Schmitz-Valckenberg P, Papastathopoulos KI, Budde WM. Ranking of optic disc variables for detection of glaucomatous optic nerve damage. *Invest Ophthalmol Vis Sci*. 2000;41(7):1764–1773.
4. Paunescu LA, Schuman JS, Price LL, et al. Reproducibility of nerve fiber thickness, macular thickness, and optic nerve head measurements using StratusOCT. *Invest Ophthalmol Vis Sci*. 2004;45:1716–1724.
5. Arnalich-Montiel F, Muñoz-Negrete F, Rebolleda G, et al. Cup-to-disc ratio: agreement between slit-lamp indirect ophthalmoscopic estimation and stratus optical coherence tomography measurement. *Eye*. 2007;21:1041–1049.
6. Joshi GD, Sivaswamy J, Krishnadas SR. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. *IEEE Trans Med Imaging*. 2011;30(6):1192–1205.
7. Fu H, Cheng J, Xu Y, Wong DW, Liu J, Cao X. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Trans Med Imaging*. 2018;37(7):1597–1605.
8. Guo J, Azzopardi G, Shi C, Jansonius NM, Petkov N. Automatic determination of vertical cup-to-disc ratio in retinal fundus images for glaucoma screening. *IEEE Access*. 2019;7:8527–8541.
9. Nakahara K, Asaoka R, Tanito M, et al. Deep learning-assisted (automatic) diagnosis of glaucoma using a smartphone. *Br J Ophthalmol*. 2022;106:587–592.
10. Li Z, He Y, Keel S, et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*. 2018;125:1199–1206.
11. Fu H, Cheng J, Xu Y, et al. Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Trans Med Imaging*. 2018;37:2493–2501.
12. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. 2019;137:1353.
13. Asaoka R, Tanito M, Shibata N, et al. Validation of a deep learning model to screen for glaucoma using images from different fundus cameras and data augmentation. *Ophthalmol Glaucoma*. 2019;2(4):224–231.

14. Li Z, Guo C, Lin D, et al. Deep learning for automated glaucomatous optic neuropathy detection from ultra-widefield fundus images. *Br J Ophthalmol*. 2021;105(11):1548–1554.
15. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. In: *6th International Conference on Learning Representations*. Vancouver, BC, Canada: OpenReview.net; 2018.
16. Saporta A, Gui X, Agrawal A, et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat Mach Intell*. 2022;4:867–878.
17. Orlando JI, Fu H, Breda JB, et al. Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal*. 2020;59:101570.
18. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi A, eds. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*. Munich, Germany: Springer Part III 18; 2015:234–241.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA: IEEE; 2016:770–778.
20. Otsu N. A threshold selection method from gray level histograms. *IEEE Trans Syst Man Cybern*. 1979;SMC-9:62–66.
21. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Miami, USA: IEEE; 2009:248–255.
22. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. *Adv Neural Info Process Syst*. 2019;32:8024–8035.
23. Bengtsson B. The variation and covariation of cup and disc diameters. *Acta Ophthalmol*. 1976;54(6):804–818.
24. Jonas JB, Gusek GC, Naumann GO. Optic disc, cup and neurorefinal rim size, configuration and correlations in normal eyes. *Invest Ophthalmol Vis Sci*. 1988l;29(8):1151–1158.
25. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–272.
26. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1(1):39.
27. van der Heijden AA, Abràmoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96(1):63–68.
28. Mocan MC, Machen L, Jang I, Cao D. The relationship between optic nerve cup-to-disc ratio and retinal nerve fiber layer thickness in suspected pediatric glaucoma. *J Pediatr Ophthalmol Strabismus*. 2020;57(2):90–96.
29. Chandra A, Bandyopadhyay AK, Bhaduri G. A comparative study of two methods of optic disc evaluation in patients of glaucoma. *Oman J Ophthalmol*. 2013;6(2):103.
30. Murthy SK. Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Mining Knowledge Discov*. 1998;2(4):345–389.
31. Hemelings R, Elen B, Barbosa-Breda J, et al. Deep learning on fundus images detects glaucoma beyond the optic disc. *Sci Rep.* 2021;11:20313.
32. Hemelings R, Elen B, Schuster AK, et al. A generalizable deep learning regression model for automated glaucoma screening from fundus images. *npj Digit Med*. 2023;6:112.