IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 17, 2021, accepted July 5, 2021, date of publication July 30, 2021, date of current version August 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3101521

A Study of Bilateral Symmetry in Color Fundus Photographs

SANGEETA BISWAS[1,2], **JOHAN ROHDIN**[1], **ANGKAN BISWAS**[1,3], **AND MARTIN DRAHANSKY**[1], **(Senior Member, IEEE)**
[1]Faculty of Information Technology, Brno University of Technology (BUT), 61200 Brno, Czech Republic
[2]Faculty of Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
[3]CAPM Company Ltd., Dhaka 1213, Bangladesh

Corresponding author: Sangeeta Biswas (biswas@fit.vutbr.cz)

This work was supported by the Project "Reliable, Secure, and Efficient Computer Systems," Brno University of Technology (BUT), Czech Republic, under Grant FIT-S-20-6427.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT We have previously shown that there is a high degree of bilateral similarity in the central retinal blood vessels (CRBVs), which are responsible for supplying blood to retinas and can be used as a strong biometric. We have also shown that a side-independent retina verification system can be developed based on the bilateral similarity in CRBVs. In this paper, we perform a similar investigation for color fundus photographs since color fundus photographs are much richer representations of retinas than CRBVs. We investigate whether the color fundus photographs of the left and right retinas possess strong enough bilateral symmetry so that we reliably tell whether a pair of the left and right retinas belong to a single subject. We evaluate and analyse the performance of both human- and deep neural network-based bilateral verification by experimenting on color fundus photographs of two publicly available data sets.

INDEX TERMS Retina, symmetry, color fundus photograph, deep neural network, biometric system.

I. INTRODUCTION

Bilateral symmetry can be defined as uniformity, equivalence, or exact similarity of two parts arranged on opposite sides of a median axis so that one part looks like a mirrored part of the other. It is the most common type of symmetry we can see in nature (e.g., flatworms, ribbon worms, clams, snails, octopuses, crustaceans, insects, spiders, brachiopods, sea stars, sea urchins, vertebrates, flowers from the orchid, pea, and figwort families, and the leaves of most of the plants). Paired organs, such as eyes, ears, hands, and legs, give a bilateral symmetrical look to the exterior of our human body by dividing it into two parts through an imaginary left-right axis from head to leg. We can easily see outward symmetry in our left and right eyes (as shown in the 1st row of Fig. 1). However, seeing symmetry in the left and right retinas, which are the internal parts of our two eyes, is not easy. As shown in the 2nd row of Fig. 1, the left and right retinas have an asymmetrical look at first glance in the 2D color fundus photographs (i.e., retinal images).

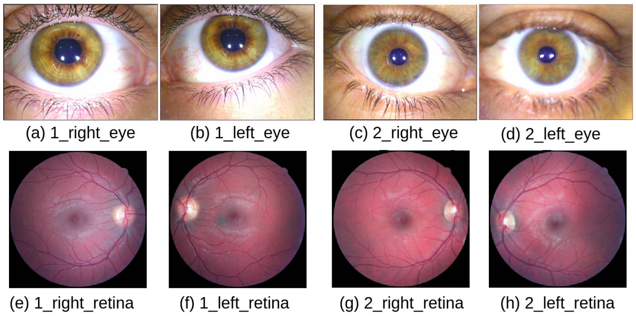


FIGURE 1. Bilateral symmetrical look of the eyes versus asymmetrical look of the retinas. Our pair of left and right eyes look symmetric from the outside, whereas the left and right retinas, the inner parts of our eyes, do not show easily visible bilateral symmetry. [1st row: two pairs of eyes belonging to two European subjects and 2nd row: two pairs of retinas corresponding to the pairs of eyes at the 1st row. The title of a sub-figure x_y_z indicates SubjectID_Side_Eye/Retina. Source of image: the private repository of the Security Technology Research and Development (STRaDe) group, Faculty of Information Technology, the Brno University of Technology, Czech Republic. We name it STRaDe data set.]

Even though the left and right retinas of a subject may look asymmetric at first glance, in this paper, we study to what extent there is a possibility that humans or deep neural

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

networks (DNNs) can detect bilateral symmetry in the color fundus photographs. Our work in this paper is an extension of our previous work reported in [1] and a complementary work reported in [2].

Similar to [1], [2], in this paper, we turn the problem of finding bilateral symmetry in the retina into a verification problem. If a system (either human or a DNN) can decide that a pair of fundus photographs of the left and right retinas belong to a single subject, we can assume that the system finds substantial bilateral symmetry in that pair. On the other hand, if the system decides that a pair of left and right retinas belong to two different subjects, we can assume that the system did not see any symmetrical properties in that pair. By manual and automatic verification, we investigate to what extent our assumptions are correct.

Similar to [1], in this work, we investigate bilateral symmetry in 2D color fundus photographs. However, contrary to [1], in this work, we involve the trained volunteers besides the untrained volunteers in the manual verification. We also study different setups of a Y-shaped convolutional neural network (which we name YNN) for automatic verification, while we reported only one setup of YNN in [1]. We also analyze here which parts of the retinas have a significant contribution to the decision of the YNN. We did not report this kind of analysis in [1].

In this work, we do experiments similar to the experiments done for central retinal blood vessels (CRBVs) in [2]. However, contrary to [2], in this work, we put light on whether human and neural networks are benefited by the color information and anatomical structures which are visible in color fundus photographs but not in the black-white representation of CRBVs. Further, we compare 2D color fundus photographs-based systems with our previous CRBVs-based systems. Moreover, we investigate here whether domain adaptation can improve the performance of a side-independent retina-based biometric system. We did not report domain adaptation-based results for CRBVs in [2].

Since 2D fundus photographs are widely used for automatically detecting pathology in the retina (e.g., [3]–[11]), our investigation on bilateral symmetry in 2D color fundus photographs can benefit medical science. It raises the possibility of matching both side retinal images of a subject (e.g., patient), captured at different sessions with different patient IDs. Fundus photographs captured over a long period can give valuable insight into the pathology progression in the retina. For example, based on fundus photographs captured at 5-year follow-up visits of patients, we can determine:

- the probability of progressing diabetic retinopathy in both retinas within 5 years.
- the probability of progressing to severe nonproliferative diabetic retinopathy (NPDR) or proliferative diabetic retinopathy (PDR) within five years of diagnosis for patients with mild NPDR and moderate NPDR.
- the probability of developing diabetic macular edema (DME) within five years for patients diagnosed with mild NPDR, moderate NPDR, severe NPDR, and PDR.

However, especially in developing countries, ophthalmologists do not often systematically store patients' retinal images in a single place for future analysis. The reasons behind this act can be their lack of awareness or lack of infrastructural support for storing retinal images. Another scenario is that patients change ophthalmologists frequently, especially in countries where patients are not registered to a specific hospital/clinic/ophthalmologist. Therefore, valuable retinal images are either lost or stored in a scattered way with non-uniform or wrong patient ID. Bilateral symmetry in the retina can help image processing-based approaches gather images of both side retinas of a specific patient in one place.

A retina-based biometric system would also be benefited from our investigation on bilateral symmetry in 2D color fundus photographs since it raises the possibility of developing a side-independent retina-based biometric system using color fundus photographs. Previously studied retina-based biometric systems (e.g., [12]–[20]) are mainly based on CRBVs and are side-dependent. In a side-dependent system, the same side retina as the one registered in the system must be used for authentication. A side-independent retina-based biometric system, in which any side retina can be used, increases user flexibility, especially when the registered side is affected by severe pathology. However, it has not been properly explored except our previous work reported in [2]. Regarding the use of color information, a few works can be found in the literature where the green channel of color fundus photographs was used for authentication, such as [21], [22]. Moreover, the systems proposed in those works were side-dependent.

We organize this paper as follows. In Section II, we briefly describe retina, retinal imaging, and previous works on bilateral symmetry in the retina. We explain our approaches for manual and automatic verification in Section III. In Section IV, we describe our experimental setup. In Section V we analyse our results. Finally, we draw our conclusions in Section VI.

II. BACKGROUND

A. RETINA AND RETINAL IMAGING

The retina is a thin, semi-transparent, multilayered neural tissue that covers two-thirds of the interior of each eye. It is mainly responsible for converting incoming electromagnetic signals from the world outside of our eye into neural signals and then handing them to the optic nerves. The neural signals, relaying through optic nerves, form images into the visual cortex of our brain, and therefore, we can have a sense of vision.

Different kinds of imaging (such as fundus photography, color fundus photography, adaptive optics scanning laser ophthalmoscopy (AOSLO), and optical coherence tomography (OCT)) have been developed for the clinical care and management of patients with retinal as well as systemic diseases. As shown in Fig. 2, in a color fundus photograph, we can see the optic disc, macula, and central retinal blood vessels (CRBVs) on circular and colored foreground displayed on a dark background. Depending on the fundus

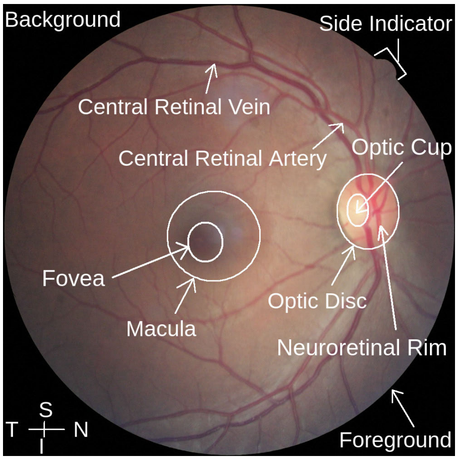


FIGURE 2. A color fundus photograph of a healthy right retina. We can see the optic disc, macula, and central retinal blood vessels (composed of central retinal arteries and central retinal veins) on a circular and colored foreground displayed on dark background in a 2D color fundus photograph. [T: Temporal, N: Nasal, S: Superior, I: Inferior side of the retina. The image was captured from a South-Asian subject. Source of image: STRaDe data set.]

camera, we may see a side indicator (i.e., a triangular- or oval-shaped bump) always at the right side, which helps us determine whether it is a left or right-side retina. As shown in Fig. 3, we can see different layers of the retina in tomograms obtained by tomography, such as OCT. By AOSLO we can see rods and cones, which contribute to our night-time and day-time vision, respectively. See [23]–[25] for details about retina and retinal imaging.

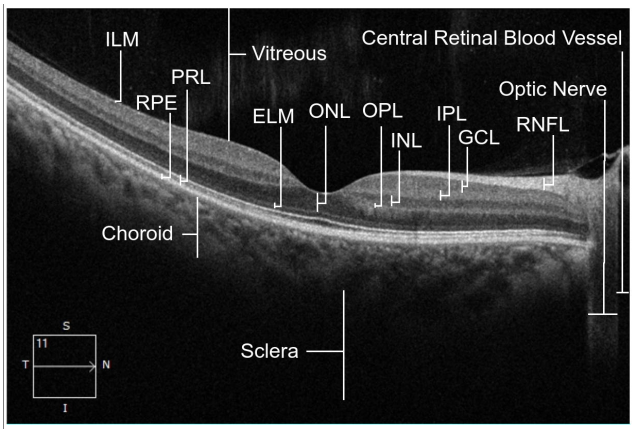


FIGURE 3. An optical coherence tomography of the retina shown in Fig. 2. Different layers of the multilayered retina, such as ILM, RNFL, GCL, IPL, INL, OPL, ONL, PRL, and RPE, are visible in the middle of the avascular vitreous and vascular choroid in the OCT image. [Source of image: STRaDe data set. Notations are ILM: inner limiting membrane, RNFL: retinal nerve fiber layer, GCL: ganglion cell layer, IPL: inner plexiform layer, INL: inner nuclear layer, OPL: outer plexiform layer, ONL: outer nuclear layer, OLM: outer limiting membrane and PRL: photoreceptor layer, RPE: retinal pigment epithelium.]

B. PREVIOUS WORKS ON BILATERAL SYMMETRY IN RETINA

The retina plays a vital role in our vision. Any disturbance in the retina can hurt our vision. Severe pathology in the retina can even cause irreversible partial or complete vision loss. Therefore, patients' retina is one of the focusing points for clinicians, ophthalmologists, medical researchers, and computer-aided diagnostic devices.

Bilateral symmetry in the retina can benefit medical science in two ways. First, it allows ophthalmologists to use one retina as a proxy for the other retina, especially for post-surgery analysis. When ophthalmologists do not have previous measurements or face difficulties in getting measurements about any sick retina, they can use measurements of the opposite retina for comparison purposes after doing any major surgery or using medicines to treat pathology in the retina.

Second, it helps to track pathology progress in a retina compared to the other retina. If any anatomic structure of the retina has bilateral symmetry, then a violation of that symmetry is a sign of developing pathology in the retina. For example, asymmetry of the physiologic cups in the two eyes is a sign of early glaucoma ([33]). If clinicians observe bilateral asymmetry in the retina, they can recommend a patient to have a further thorough examination by expert ophthalmologists, and, therefore, pathology in the retina can be detected at an earlier stage.

In previous works, bilateral symmetry in retina was reported: for CRBVs in [26], [34]–[37] using fundus photographs; for different anatomical structures and layers in [27]–[32], [38] using different kinds of tomography, such as OCT, Heidelberg retina tomography (HRT), and Spectral-domain OCT (SD-OCT); and for cones in [39], [40] using AOSLO. In these works, a specific attribute of the left and right retinas (such as length, area, or thickness of different anatomical structures and layers) of a group of patients was estimated at first. Then two types of analysis were performed.

The first type of analysis aims to answer whether a set of left retinas collected from different patients follows the same distribution as a set of right retinas collected from different patients (except for being mirrored). This analysis was performed by applying statistical significance tests such as paired t-test, Wilcoxon paired test, and Mann-Whitney U test, showing that the hypothesis that the distributions of the left and right attributes are equal cannot be rejected. Note that, strictly speaking, such analysis does not prove that the distributions are equal; it only shows that there is no strong evidence against this hypothesis.

The second type of analysis aims to answer the same question as our work, i.e., whether (on average) the left and the right retina from one patient are more similar than a left and a right retina from two different patients. This type of analysis was performed by measuring concordance (i.e., the degree of agreement) between the right and left retinas by estimating a correlation coefficient (such as the Pearson or Spearman correlation coefficient or the Li concordance correlation coefficient) or by plotting a Bland-Altman diagram. In the first case, the higher the value of the correlation coefficient was, the more symmetric were the left and right retinas considered to be. In the second case, if there was no bias in the

TABLE 1. Some of the previous works done by others on bilateral symmetry in the retina. [Notations are N: Number of subjects, OD: optic disc, OC: optic cup, NR: neuroretinal rim, FAZ: foveal avascular zon, CRAE: central retinal arteriolar equivalent, CRVE: central retinal venular equivalent, AVR: arteriole-to-venule ratio, TS: temporal-superior, NI: nasal-inferior, TI: temporal-inferior, CMT: central macular thickness, mGCC: macular ganglion cell complex, ACHM: achromatopsia, r: correlation coefficient, p: significance level, −: information is not provided, BAD: Bland-Altman diagram.]

Referred Work	N	Attributes	Mean ± SD		Significance Test		Concordance Analysis	
			Right	Left	Type	p	r	BAD
Leung et al. [26]	1546	CRAE	194.2 ± 21.2	194.2 ± 20.3	-	-	0.70	✓
		CRVE	225.0 ± 20.8	226.6 ± 20.2			0.77	
		AVR	0.865 ± 0.081	0.859 ± 0.077			0.54	
Budenz [27]	108	Vertical OC/OD	0.31 ± 0.15	0.30 ± 0.15	-	0.85	-	✗
		Horizontal OC/OD	0.29 ± 0.14	0.29 ± 0.14		0.20		
Li et al. [28]	1276	OD area (mm^2)	1.91 ± 0.40	1.92 ± 0.41	Paired t-test	0.372	0.737	✗
		Vertical OC/OD	0.34 ± 0.24	0.34 ± 0.24		0.387	0.665	
		OC/OD	0.23 ± 0.14	0.23 ± 0.15		0.487	0.735	
		NR/OD	0.77 ± 0.14	0.77 ± 0.15		0.649	0.645	
		OC area (mm^2)	0.46 ± 0.33	0.47 ± 0.35		0.232	0.763	
		NR area (mm^2)	1.45 ± 0.33	1.45 ± 0.33		0.919	0.641	
		OC shape	−0.17 ± 0.07	−0.17 ± 0.07		0.380	0.397	
Yang et al. [29]	86	RNFL thickness			Paired t-test or Wilcoxon paired test depending on normality of data			✗
		* TS (um)	156.6 ± 19	154.1 ± 20.5		0.131	0.71	
		* NI (um)	122.9 ± 22.7	122.2 ± 23.4		0.737	0.70	
		* TI (um)	160.7 ± 20.7	162.3 ± 20.7		0.283	0.82	
Zhou et al. [30]	158	Superior mGCC thickness (um)	98.54 ± 5.64	98.79 ± 5.79	Paired t-test	0.385	0.849	✓
		Inferior mGCC thickness (um)	98.13 ± 5.54	98.87 ± 5.67		0.343	0.835	
		Average mGCC thickness (um)	98.26 ± 5.54	98.07 ± 5.54		0.381	0.882	
Liu et al. [31]	87	FAZ area (mm^2)	0.33 ± 0.11	0.33 ± 0.12	Mann–Whitney U test	0.9	0.93	✗
		CMT (um)	241.5 ± 21.8	241.4 ± 22.0		0.68	0.93	
Mastey et al. [32]		Foveal ONL thickness (um)			Paired t-test			✓
	42	* Controls	112.9 ± 15.2	112.1 ± 13.9		0.434	0.911	
	76	* ACHM	79.7 ± 18.3	79.2 ± 18.7		0.636	0.899	

Bland-Altman diagram, then the left and right retinas were considered symmetric. See Table 1 for a summary of some of these works and see [41] for a review on retinal symmetry.

III. OUR APPROACH FOR FINDING BILATERAL SYMMETRY IN COLOR FUNDUS PHOTOGRAPHS

A. OUR TARGET

None of our paired body organs have identical left and right forms. That means our human body shows *approximate-bilateral* or *pseudo-bilateral* symmetry instead of *perfect-bilateral* symmetry. This approximate-bilateral symmetry is generally less obvious in the inner side of our body. The retina is an example of our inner body parts, which shows bilateral asymmetry at first glance. In the 2D color fundus photographs, the left and right retina look asymmetric mainly because of the unique tree-like structure of CRBVs spreading over the retina. Poor image quality increases the asymmetrical look by displaying different colors on the foreground (see Fig. 4 (b), (c), (e) & (f)) as well as overexposing (see Fig. 4 (d)) and underexposing (see Fig. 4 (f) & (g)) different parts of the retina. Besides, some pathologies affect only one eye which can cause bilateral asymmetry (see Fig. 4 (a), (g), & (i)). Moreover, which direction the subject looks at when the images are captured can affect the symmetric look of the retinas in 2D fundus photographs. Suppose the subject looks at two different directions when fundus photographs of the left and right retinas are captured. In that case, the alignment of different anatomical parts (e.g., the optic disc, macula, and CRBVs) can be different (see Fig. 4 (c)) and sometimes some anatomical parts can even be missing (see Fig. 4 (h)) in the left and right fundus photographs.

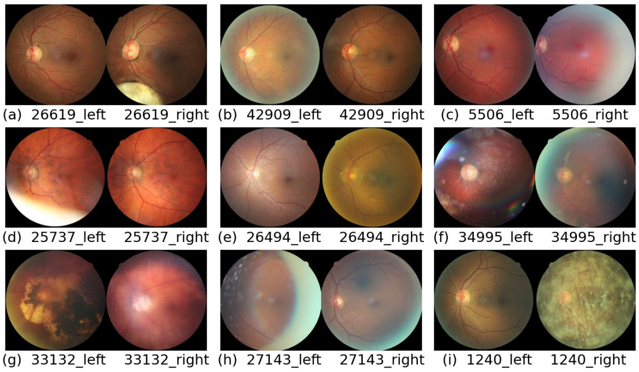

(a) 26619_left 26619_right (b) 42909_left 42909_right (c) 5506_left 5506_right
(d) 25737_left 25737_right (e) 26494_left 26494_right (f) 34995_left 34995_right
(g) 33132_left 33132_right (h) 27143_left 27143_right (i) 1240_left 1240_right

FIGURE 4. Nine pairs of fundus photographs where the left and right retinas look asymmetric for different reasons such as low image quality, mismatched alignment of anatomical structures (such as optic disc, macula, and central retinal blood vessels), pathology, and non-visible anatomical structures such as the optic disc and macula. [Source of images: Kaggle_SetB data set. Right side retinas are flipped horizontally.]

Even though, at first glance, a pair of the left and right retinas look asymmetric in 2D color fundus photographs, our first target is to show that there is an underlying symmetry in the left and right retinas from the same subject, which can be detected by both human and deep neural networks (DNNs). Second, we want to show how this bilateral symmetry in retinas is sufficient for building a side-independent retina biometric system.

B. OUR APPROACH

Finding bilateral symmetry in 2D color fundus photographs of the left and right retinas is a complex problem. We turn this complex problem into a binary classification problem. Instead of measuring specific attributes such as length and the number of blood vessels, or area and thickness of a specific layer, as is typical in the research of medical science (e.g., [34]–[38]), we investigate whether a system (human or a DNN) can tell whether a pair of left and right retinas belong to a single subject or two different subjects. If a system can decide that a pair of left and right retinas belong to a single subject, we can assume that the system finds substantial bilateral symmetry in that pair. On the contrary, if the system decides that a pair of left and right retinas belong to two different subjects, we can assume that the system did not find any symmetrical properties in that pair. From the biometric point of view, this is a side-independent verification task. Usually, in a retina-based biometric system, features extracted from the same side retina are compared for verification, whereas, in our task (see Subsection V-G), features extracted from different sides and the same side retinas are used for verification.

C. MANUAL VERIFICATION

We take opinions from two groups of human volunteers for manual verification: (1) untrained human volunteers and (2) trained human volunteers. We do not provide any information regarding symmetry in the retinas to untrained human volunteers. We ask them to decide by themselves how to do the verification. On the other hand, we provide the summary of the untrained volunteers' observations and 100 pairs of retinas with actual labels to the 2nd group of volunteers to train themselves as long as they want.

In principle, without observing any examples of positive pairs, PPs (i.e., the pairs from the same subjects) and negative pairs, NPs (i.e., the pairs from two different subjects), it is difficult to guess in which sense retinas from the same subject are symmetric. For example, without any prior knowledge, it is difficult to guess how much the foreground color or the optic disc can vary between two retinas from one subject and how much they can vary from subject to subject. By observing labeled training data, trained volunteers and DNNs may learn what similarities indicate that two retinas are from the same subject. Untrained volunteers do not have this opportunity. However, untrained volunteers may have a chance to do better than guessing randomly by training themselves gradually if they know that there are both PPs and NPs in the test. They can gradually learn where symmetry exists and assume that the pairs with the largest bilateral symmetry are positive. Keeping all these points in mind, we use a computer interface to take human volunteers' opinions. In the computer interface, we show four pairs of the left and right retina in a single frame at once, as shown in Fig. 7, and in total, we show 100 pairs in a random order, among which 50 pairs are PPs, and 50 pairs are NPs. We ask volunteers to make a binary decision about each pair in a frame. We describe the details about the experimental setup in Subsection IV-C.

D. AUTOMATIC VERIFICATION

In order to do the verification automatically, we need a DNN which can perform three tasks: (1) high-level feature extraction from retinal images, (2) merging of the extracted features, and (3) binary classification. Since convolutional neural networks (CNN) are widely used for extracting features from image data, we decide to use CNNs as a feature extractor.

Features extracted from two fundus photographs can be merged in many ways such as *concatenation*, *absolute subtraction*, *averaging*, *addition*, *multiplication*, *dot product*, *maximization*, and *minimization*. Two properties of the merging technique are important to consider. First, does it *preserve all information* in the two input features? That is, is it possible to reconstruct the input features from the merged feature? Second, is it *symmetric* with respect to the two input features? That is if we swap the two input features, will the output remain the same? We decide to explore only four merging/combining techniques:

1) concatenation: preserves all information, not symmetric
2) absolute subtraction: does not preserve all information, symmetric
3) averaging: does not preserve all information, symmetric
4) concatenation of *averaging* and *absolute subtraction*: preserves all information, symmetric

As a binary classifier, we decide to use a simple network having convolutional and dense layers as hidden layers and a neuron with *sigmoid* activity function in the output layer. The output of the sigmoid function is in the range $0 - 1$, and it is the standard choice of output activation function when one wants an NN with two output probabilities in a two-class problem. By deciding a threshold value, δ, we can easily get a binary decision.

Based on our preliminary decisions, we develop a Y-shaped DNN by connecting three neural networks (NNs) as shown in Fig. 5. We name it YNN. In the YNN, there are two convolutional neural networks (CNNs), each of which acts as a high-level feature extractor, and one neural network for binary classification that acts as a similarity score generator. Each retinal image in the pair to be verified is passed through a high-level feature extractor. The extracted high-level features are merged/combined and then processed further by the score generator, which generates a similarity score in the range $0 - 1$ for a pair. By setting $\delta = 0.5$, we get a binary decision about a pair of the left and right fundus photographs from the YNN.

We decide to train the YNN in an end-to-end manner, i.e., the feature extractor networks and the binary classification network are trained jointly for the binary classification task. During training, the YNN does not learn a model for

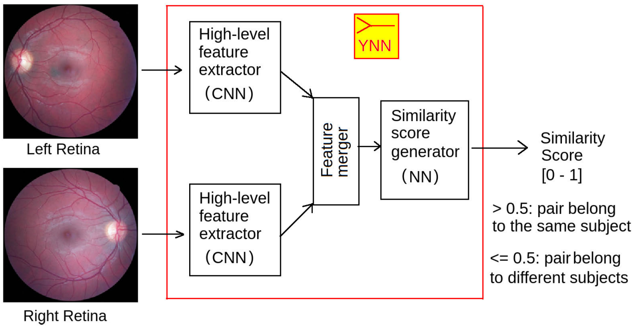

YNN
Left Retina
Right Retina
High-level feature extractor (CNN)
High-level feature extractor (CNN)
Feature merger
Similarity score generator (NN)
Similarity Score [0 - 1]
> 0.5: pair belong to the same subject
<= 0.5: pair belong to different subjects

FIGURE 5. YNN used for automatic verification of a pair of left and right retinas. YNN is a combination of three neural networks. The first two neural networks extract high-level features from the pair to be verified, and the third network generates a similarity score. If the score is below or equal to 0.5 for a pair, the fundus photographs of the pair are considered to belong to two different subjects. If the score is above 0.5 for a pair, the fundus photographs of the pair are considered to belong to a single subject. The architecture of YNN is depicted in more detail in Fig. 6 (f).

a specific subject. Instead, it learns where to look for symmetry in a pair of retinas. Therefore, after the training phase, it can verify pairs from subjects unseen in the training phase. It means the YNN can handle open-set verification.

We decide to explore whether:

- it is easier for the YNN to compare two retinas if we flip horizontally one image (e.g., the right retinal image) than to compare retinal images without flipping any image.
- sharing/tying parameters or using different parameters for the two feature extractors is optimal. Even when images are from the same subject in a PP, they are from retinas of two different sides. It is unclear whether they (even after one is flipped) follow the same distribution or two different distributions. If images from the left and right retinas follow the same distribution, it should be beneficial to share the parameters of two feature extractors. On the contrary, if they follow different distributions, then sharing parameters by the feature extractors may not be beneficial.
- a well-known pre-trained model (e.g., VGG16 [42]) would be a better feature extractor of YNN than its own feature extractor.
- a YNN can also be used as *side independent* verification system, by which we mean a system that can compare two retinas no matter whether they are from the same side or two different sides. In order to achieve this, we need to use a slightly different approach during training a YNN. We need to include not only *left-right* pairs after flipping a specific side image, but also *right-right* and *left-left* pairs while training the YNN.

We also decide to investigate the impact of intersession variability. Ideally, we should use only pairs of fundus photographs from different sessions, but among the databases, we have access to, this is only possible in one database, which is too small for a detailed experimental analysis. We, therefore, use that database to investigate the impact of the intersession variability on the results, whereas we use a larger database for the main experiments. We describe the architecture of YNN and its training procedure in detail in Subsection IV-D.

If two feature extractors of the YNN share parameters, the YNN turns to a siamese network proposed independently for fingerprint recognition in [43] and signature verification in [44], and successfully used for face verification (e.g., [45]), one-shot image recognition (e.g., [46]) and depth information extraction (e.g., [47]). Note that the NN proposed in [43] was not named a siamese network. The term *siamese network* was coined from [44] to indicate all networks which have two identical sub-networks to extract features from the inputs. As shown in Fig. 6, the differences among the different siamese networks are mainly the different arrangements of the convolutional layers in the CNNs, different kinds of dimension reduction approaches (e.g., max-pooling, subsampling, striding), and different similarity score generation approaches.

IV. EXPERIMENTAL SETUP

In this section, we briefly describe our hardware and software setup in Subsection IV-A, data sets in Subsection IV-B, setup for manual verification in Subsection IV-C, setup for automatic verification in Subsection IV-D, and image pre-processing steps in Subsection IV-E.

A. HARDWARE & SOFTWARE TOOLS

We did all experiments using a machine with TensorFlow's Keras API 2.0.0, OpenCV 4.2.0, and Python 3.6.9. The machine is a standard PC with 32 GB memory, AMD Ryzen Threadripper 2950X CPU with 16 cores per socket, and one GeForce RTX 2080 Super GPU with 8 GB memory.

B. DATA SETS

The publicly available retina data sets are mainly prepared for automatic pathology detection (e.g., [48]–[55]), segmentation of anatomical structures such as optic disc, macula, CRBVs (e.g., [56]–[59]), assessing quality of retinal image (e.g., [60], [61]), retinal image registration (e.g., [62], [63]) and so on. Most of these data sets have images from only one side and one session. Therefore, most of these data sets were not appropriate for our purpose. Few data sets have images from both side retinas, among which we chose three data sets to do the verification: the Kaggle data set ([48]), the Longitudinal diabetic retinopathy screening data set ([63]), and Messidor-2 data set ([52])

1) Kaggle data set: This data set is prepared for the competition of diabetic retinopathy detection. It is provided by EyePACS and publicly available via the Kaggle online community of data scientists and machine learners. It has 44, 351 pairs of images. There are left and right retinal images belonging to a single patient ID number in each pair. Therefore, there are in total 88, 702 RGB retinal images belonging to 44, 351 subject IDs. Images were captured under a variety of conditions. There is no information whether images from

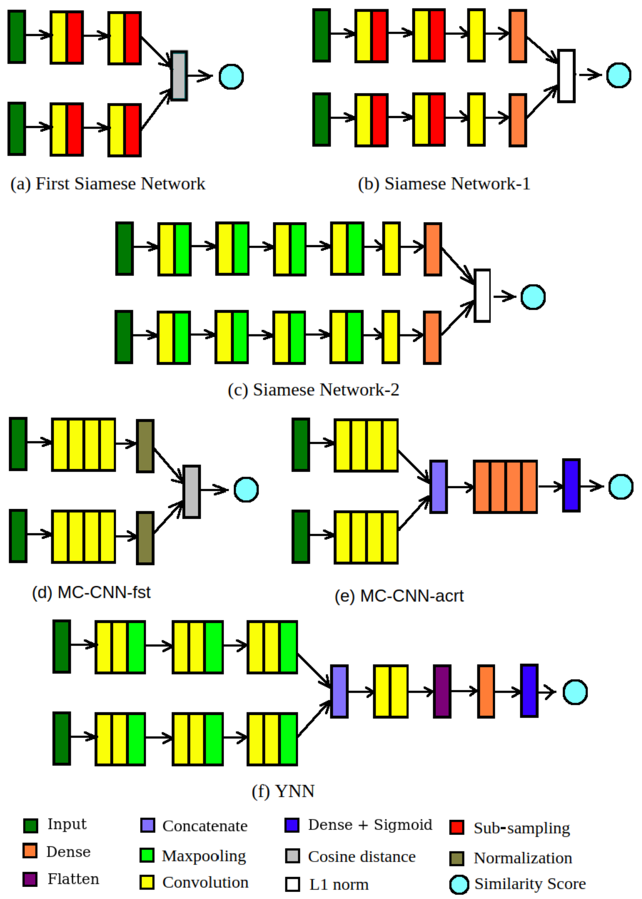


FIGURE 6. **Varieties of siamese networks, i.e., the networks which have two identical sub-networks to extract high-level features from the inputs. (a) A siamese network proposed in [44] for verifying whether two signatures are signed by the same subject. (b) A siamese network proposed in [45] for verifying whether faces belong to the same subject. (c) A siamese network proposed in [46] for one-shot image recognition. (d) & (e) Siamese networks, named MC-CNN-fst and MC-CNN-acrt, proposed in [47] for extracting depth information from a rectified image pair. (f) Our YNN proposed for verifying whether two fundus photographs (which could be from the left side, right side, or from both left and right side) belong to the same subject. Notice that our YNN is more similar to MC-CNN-acrt than other siamese networks because it does not use any predefined distance metric to estimate the similarity score.**

the same patient were captured at the same session, but this is likely the case. Therefore, this database cannot be used to study the effect of session variability. We prepared three sets, i.e., Kaggle_A, Kaggle_B, and Kaggle_C, from the images having resolutions 3264 × 4928 and 3168 × 4752 for three purposes (see Table 2 for details).

We prepared two test sets (i.e., Kaggle_SetA.1 and Kaggle_SetA.2) using the images of Kaggle_SetA and one test set (i.e., Kaggle_SetC) using the images of Kaggle_SetC. In principle, it is possible to build 150 positive pairs (i.e., the left and right retinal images of a pair belonging to a single subject ID) and $150 \times 149 = 22,350$ negative pairs (i.e., the left and right retinal images of a pair belonged to two different subject IDs) using the 150 pairs of Kaggle_SetA. However, for human volunteers, it is difficult and time-consuming to decide about $150 + 22,350 = 22,500$ pairs. Therefore, we decided to reduce the number of pairs while keeping the variability among pairs as much as possible. For fulfilling that, we divided 150 subjects into three groups: the first 50 subjects were for the positive pairs (PPs), the second 50 were for the left side of negative pairs (NPs), and the third 50 were for the right side of NPs. In this way, we kept only 50 PPs and $50 \times 50 = 2,500$ NPs in Kaggle_SetA.1, and 50 PPs and 50 NPs in Kaggle_SetA.2. The PPs were the same in both test sets, and the NPs of Kaggle_SetA.2 were a subset of the NPs of Kaggle_SetA.1. In Kaggle_SetC, there were 1,752 PPs and 1,752 NPs. Even though it was possible to make $1,752 \times 1,751 = 3,067,752$ NPs from 1,752 pairs, we chose only 1,752 NPs in order to keep a balance between the PPs and NPs. Contrary to Kaggle_SetA.1 and Kaggle_SetA.2, there was a subject overlap between the PPs and NPs and the left and right sets of NPs in Kaggle_SetC.

2) Longitudinal diabetic retinopathy screening data set: This data set was prepared for fundus image registration methods. It has 1120 in total color fundus photographs of 70 patients in the diabetic retinopathy screening program of the Rotterdam Eye Hospital (Rotterdam, The Netherlands). For each patient, there are four types of color fundus photos for both left and right retinas: macula-centered, optic nerve-centered, superior, and temporal fundus images. The images were captured in two sessions, and there was a 1-week gap between the two sessions. We prepared one set named RODREP_SetA by taking only macula-centered images (i.e., two images for each patient's per retina) from this data. Since there is only one session with a macula-centered retinal image for the right side retina for the subject having Patient ID 62, we excluded images of this subject. Therefore, we had 276 images (138 images from the left side and 138 images from the right side) from 69 subjects. We were able to prepare 138 PPs of left-right retinas from the same session (i.e., RODREP_SetA.1), 138 PPs of left-right retinas from different sessions (i.e., RODREP_SetA.2), 69 PPs of left-left retinas from different sessions (i.e., RODREP_SetA.3), and 69 PPs of right-right retinas from different sessions (i.e., RODREP_SetA.4). For all cases, we prepared NPs equal to the PPs. We chose macula-centered images mainly because YNN's training set, i.e., Kaggle_SetB, has macula-centered images.

3) Messidor-2 data set: This data set is mainly prepared for computer-assisted diagnoses of diabetic retinopathy. It contains two macula-centered eye fundus images (one per eye) per subject. There are 874 subjects (i.e., patients); therefore, in total, 1748 images. There is no information about whether images from the same patient were captured at the same session, but this is likely the case. We used all images of this data set.

C. SETUP FOR MANUAL VERIFICATION

Twenty-three human volunteers participated in manual verification. The majority of them were not familiar with the fundus photographs, as shown in Table 3. Even if any volunteer was familiar with the color fundus photographs, looking for

TABLE 2. Data Sets used in our experiments. [# Subj: Number of subjects, # Pairs: Number of pairs, PPs: Positive pairs, NPs: Negative pairs, LR-SS: pairs of left-right retinas from the same session, LR-DS: pairs of left-right retinas from different sessions, LL-DS: pairs of left-left retinas from different sessions, RR-DS: pairs of right-right retinas from different sessions. The Kaggle sets and Messidor-2 have only pairs of left-right retinas from the same session.]

Data Set	Height × Width	# Subjs.	# Pairs		Purpose
			PPs	NPs	
Kaggle_SetA.1	3264 × 4928	150	50	50 (Different 50 NPs randomly chosen from 2,500 NPs for different volunteers)	Test set for manual verification
Kaggle_SetA.2				50 (Same NPs for all volunteers)	Test set for manual and automatic verification
Kaggle_SetB.1	3168 × 4752	6834	6834	41,229,522	Training set of YNN
Kaggle_SetB.2		200	200	200	Validation set of YNN
Messidor-2	1536 × 2304	151	874	763,002	Adaptation set of YNN
	960 × 1440	132			
	1488 × 2240	154			
Kaggle_SetC	3264 × 4928	1752	1752	1752	Test set for automatic verification
RODREP_SetA.1	1312 × 2000	69	138 (LR-SS)	138	
RODREP_SetA.2			138 (LR-DS)	138	
RODREP_SetA.3			69 (LL-DS)	69	
RODREP_SetA.4			69 (RR-DS)	69	

TABLE 3. Volunteers' familiarity level with fundus photographs. [Level-0: Not familiar with fundus photographs, Level-1: Have some basic knowledge about fundus photographs, Level-2: Work on fundus photographs.]

Level of familiarity with fundus photographs	# Untrained Volunteers	# Trained Volunteers	Total
Level-0	13	1	14
Level-1	6	0	6
Level-2	1	2	3
Total	20	3	23

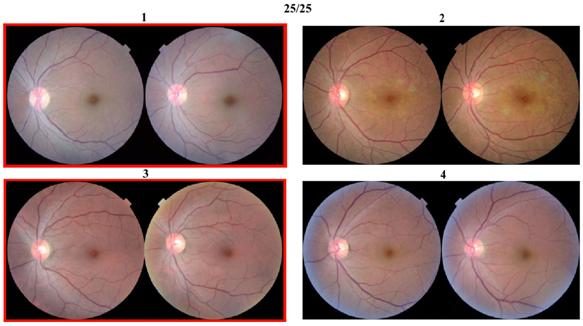


FIGURE 7. An example frame for collecting volunteers' opinions. When a volunteer clicked on a pair, its boundary turned into red color and it meant that the volunteer thought this pair belonged to a single subject. Numbers 1, 2, 3, 4 were the pair numbers and 25/25 was the frame number.

symmetry in the left and right retinas and making a decision about the subjects was entirely new for him/her. None of the volunteers thought about it before participating in the test, i.e., they did not have any training. Among 23 volunteers, we did not provide any information regarding symmetry in retinas to the first 20 volunteers. They directly participated in the test, knowing only some basic rules about the test. On the other hand, the last three volunteers were provided the summary of the information noted down by the first 20 volunteers. They got time to train themselves to figure out symmetry. Therefore, we considered the first 20 volunteers as untrained volunteers while the last three volunteers as trained volunteers.

In the test, 25 frames were shown to each volunteer, where each frame contained four pairs of retinas side-by-side (as shown in Fig. 7). The right side retinas were flipped to make the comparison task easier for human volunteers. The task of the human volunteers was to click on a pair when they thought that the pair belongs to a single person. Volunteers were allowed to select/deselect any pair as many times as they wanted and could spend as much time on the verification task as they wanted. However, after closing any frame, they were not allowed to see it again. After closing the last frame, each volunteer was asked to write about the factors they considered to make a decision. Twenty-three volunteers participated in 23 separate sessions. None of them were aware of the actual answers. All volunteers were requested not to share their assumptions with other volunteers. When writing their points, untrained volunteers were informed of retina-related terms to make their writing easier. After individual participation in the test, three volunteers participated together. This time, they were allowed to consult with each other and decide about a pair based on the majority opinion.

After summarizing the features reported by the first 20 volunteers, we gave them to the last three volunteers. We prepared a similar interface for them to train themselves. The only difference between this interface and the one used in the test phase was that after clicking on an NP, they saw a blue-colored boundary, whereas after clicking on a PP, they saw a red-colored boundary so that they could know which

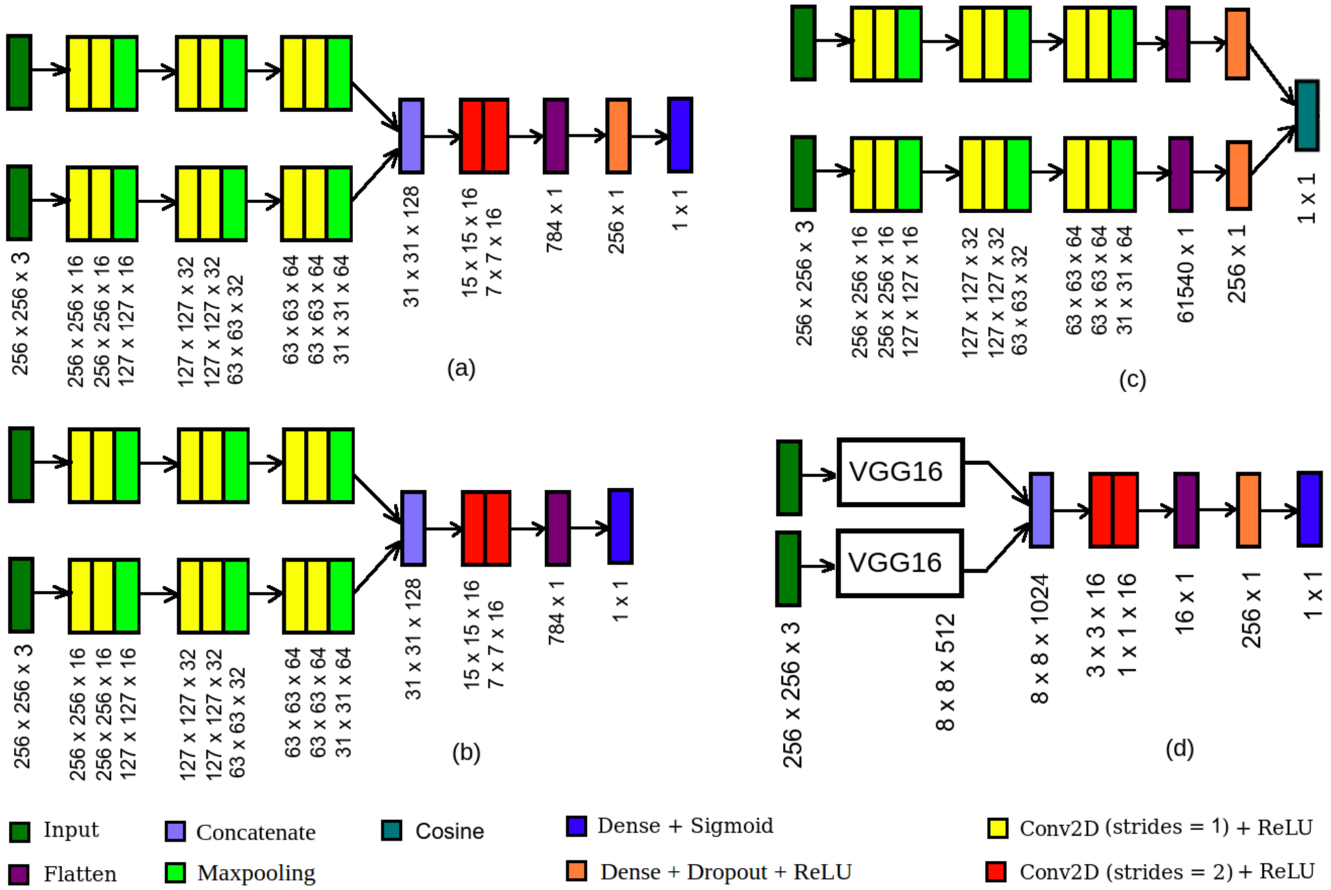


FIGURE 8. **Architectures of deep neural networks used in our experiments: (a) YNN, (b) Model1: NN having one dense layer less than the YNN, (c) Model2: NN having the same layers as the YNN but with an extra dense layer in each feature extractor and a Cosine distance layer as feature merger and similarity score generator, and (d) Model3: NN using the pre-trained VGG16 provided by Keras, without the last three dense layers, as a feature extractor. [Note that only the concatenation approach is depicted as the feature merger for the YNN even though four merging approaches were explored. The vertical text shows the output shape of the corresponding layer.]**

pairs are PPs and which pairs are NPs and analyze their decisions. They had to decide about a pair before seeing the correct answer. They were given three weeks to train themselves. In total, 100 pairs (50 PPs and 50 NPs) from the Kaggle_SetB were seen by them in random order during training. Only when they felt the confidence to participate in the final test, they were shown the pairs of the Kaggle_SetA.2.

All volunteers (i.e., ID 1-23) saw the same 50 PPs but in random orders. Volunteers with ID 1-10 saw 50 different NPs, which were randomly chosen from the 2, 500 NPs of Kaggle_SetA.1 so that they were not exhausted by seeing too many NPs. Volunteers with ID 11-23 saw the same 50 NPs but in random orders.

There is a similar work for irises. Bilateral texture similarity in pairs of left and right irises was investigated based on human opinions in [64]. The setup used to get humans' opinions had one pair of irises per frame. Twenty-seven humans had to choose an option from the list of five degrees of certainty after seeing a pair only for three seconds. There were 105 PPs and 105 NPs. At the beginning of the test, all humans got a chance to train themselves a bit by three PPs and two NPs with labels.

D. SETUP FOR DEEP NEURAL NETWORKS

We trained four kinds of deep neural networks (NNs): YNN, Model1, Model2, and Model3. Each of these models can be split into three parts: (1) two identical feature extractors, (2) feature merger and (3) similarity score generator. As shown in Fig. 8 (a), each feature extractor of the YNN consisted of three blocks of convolutional-maxpooling layers, where each block had two consecutive convolutional layers followed by one max-pooling layer. In all convolutional layers, $stride = 1$, $padding = same$ and $kernel\_size = 3 \times 3$ and in all max-pooling layers, $stride = 2$ and $pool\_size = 3 \times 3$. In each block, the number of channels was increased twice by increasing the number of filters, whereas the height and width of the feature map (i.e., the output of each block) were reduced to half. Four kinds of merging approaches were applied as the feature merger. The similarity score generator had two consecutive convolutional layers with $filters\_no = 16$, $kernel\_sz = 3\times3$, $stride = 2$ and $padding = valid$. Because of the settings $stride = 2$ and $padding = valid$, the size of feature maps was reduced to half by each convolutional layer. Therefore, a max-pooling layer was not used for shrinking the size of the feature maps. Except for the neuron of the output layer, the rectified linear units (ReLU) was used as the activation function for all neurons in all other layers. For the neuron of the output layer (i.e., the last dense layer), *sigmoid* function was used as the activation function.

As shown in Fig. 8 (b), Model1 had one dense layer less than the YNN. Model2 had the same layers as the YNN but with an extra dense layer in each feature extractor and a Cosine distance layer as the feature merger and the similarity score generator as shown in Fig. 8 (c). In Model3, the pre-trained VGG16 provided by Keras was used as a feature extractor as shown in Fig. 8 (d). The VGG16 model has 13 convolutional layers, five pooling layers, and three dense layers. The pre-trained VGG16 model provided by Keras was trained mainly for classifying images of 1000 classes. To use this model for extracting high-level features from fundus photographs, we had to exclude the last three dense layers. We trained these models in four different ways:

- without tying parameters of the feature extractors and without flipping the right-side retina.
- without tying parameters of the feature extractors, but flipping the right-side retina.
- by tying parameters of the feature extractors and without flipping the right-side retina.
- by tying parameters of the feature extractors but flipping the right-side retina.

For the YNN, we tried four different operations to merge the features extracted from the fundus photographs of the left and right retinas: concatenation, average, absolute subtraction, and concatenation of absolute subtraction and average. For Model1, Model2, and Model3, we merged features only by concatenation.

For training the YNN, Model1, Model2, and Model3, we used left-right paired images. We also trained a variation of the YNN (we name it Model4) using the left-right, right-left, left-left, and right-right pairs. Since in the Kaggle_SetB, there is only one image per side for a subject, we used data augmentation to create extra left-left and right-right pairs for Model4. We did data augmentation by rotating an image in the range $0 - 90$ degrees, shifting it at most 10% both along the horizontal and vertical direction, changing brightness in the range $0.2 - 1.0$, zooming in the range $0.8 - 1.2$ and shearing, i.e., displacing image at most 18 degrees in a counter-clockwise direction. For adapting Model4 to reduce domain mismatch between the training and test data sets, we used pairs from the Messidor-2 data set. Three training and adaptation pairs from two subjects are shown in Fig. 9 as examples.

We trained YNN using grayscale, red channel, green channel, and blue channel images along with RGB colored images to understand which color channel is the best. To avoid the effect of randomness caused by different factors, including weight initialization and dropout, on the estimation of performance, we trained all models five times. That means, in total, we trained the YNN using RGB color images 80 (i.e., $4 \times 4 \times 5$) times, whereas we trained color channel-based YNN, Model1, Model2, and Model3 five times and Model4 twice: for the first time without adaptation and for the second time with an adaptation technique.

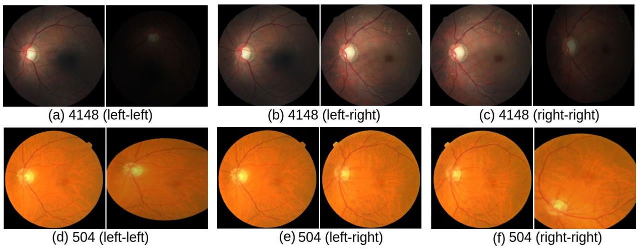


FIGURE 9. Example pairs used for training and adapting Model4. 1st row: example pairs from the training set, Kaggle_SetB, after using data augmentation. 2nd row: example pairs from the adaptation set, Messidor-2, after using data augmentation. [Note that images of the right retinas were flipped horizontally.]

TABLE 4. Performance (mean ± standard deviation) of human volunteers to verify 100 pairs of color fundus photographs where 50 pairs were positive pairs and 50 pairs were negative pairs.

	Kaggle_SetA.1	Kaggle_SetA.2	
	Untrained Volunteers	Untrained Volunteers	Trained Volunteers
Accuracy	0.79 ± 0.07	0.79 ± 0.05	0.86 ± 0.02
Sensitivity	0.74 ± 0.15	0.72 ± 0.09	0.82 ± 0.04
Specificity	0.84 ± 0.17	0.86 ± 0.07	0.91 ± 0.05
F1	0.77 ± 0.09	0.77 ± 0.07	0.86 ± 0.02

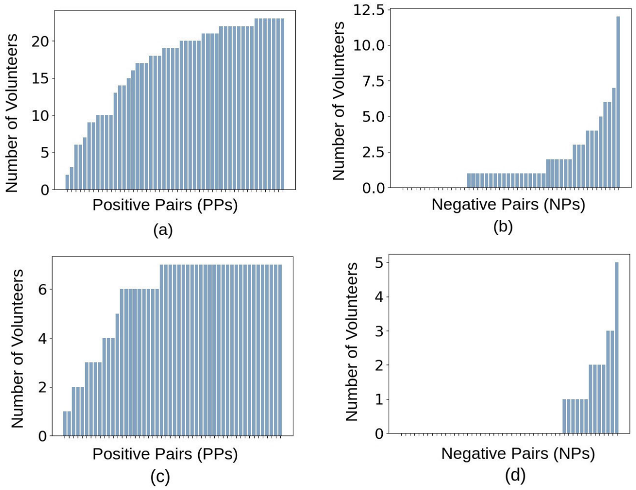
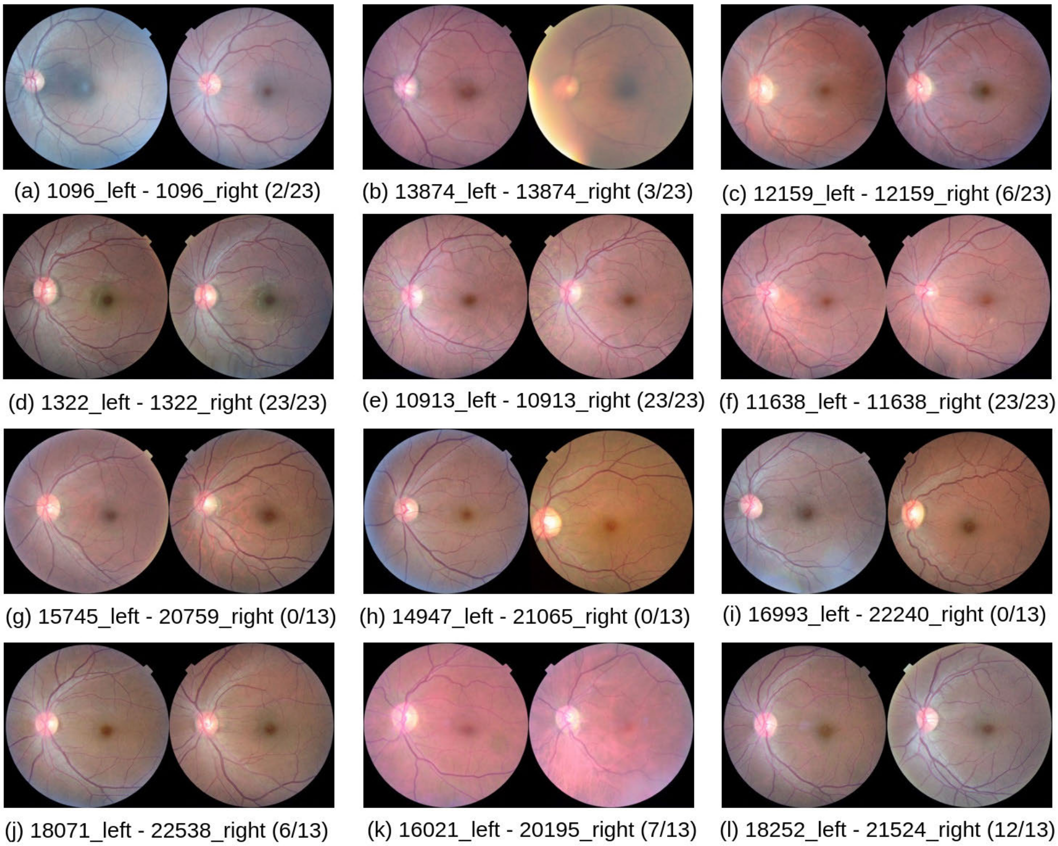


FIGURE 10. The number of volunteers who thought the pairs belonged to the same subject, i.e., who classified each pair as a positive pair (PP): (a) the number of volunteers among 23 volunteers for each PP, (b) the number of volunteers among 13 volunteers for each negative pair (NP), (c) the number of volunteers among seven volunteers (i.e., top four untrained volunteers + three trained volunteers) for each PP and (d) the number of volunteers among five volunteers (i.e., two untrained volunteers among top four untrained volunteers + three trained volunteers) for each NP.

To train all deep NNs, we set *mean squared error* as the loss function; RMSProp with a learning rate of 0.0001 as the optimizer and *epoch_no* = 50. We reduced the learning rate if there was no change in the *validation_accuracy* for more than three consecutive epochs. We stopped training if *validation_accuracy* did not change in 15 consecutive

TABLE 5. Individual performance of human volunteers. [V. ID: Volunteer ID, Acc.: Accuracy, Sens.: Sensitivity, Spec.: Specificity, Tog.: Together.]

	Kaggle_SetA.1										Kaggle_SetA.2													
	Untrained Volunteers										Untrained Volunteers										Trained Volunteers			
V. ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Tog.
Acc.	0.75	0.81	0.71	0.89	0.84	0.86	0.65	0.73	0.83	0.82	0.71	0.83	0.75	0.88	0.86	0.81	0.79	0.79	0.74	0.74	0.84	0.87	0.87	0.88
Sens.	0.82	0.76	0.44	0.84	0.82	0.76	0.86	0.52	0.66	0.94	0.52	0.74	0.78	0.82	0.80	0.84	0.64	0.72	0.64	0.68	0.80	0.88	0.84	0.78
Spec.	0.68	0.86	0.98	0.94	0.86	0.96	0.44	0.94	1.00	0.70	0.90	0.92	0.72	0.94	0.92	0.78	0.94	0.86	0.84	0.80	0.88	0.86	0.90	0.98
F1	0.77	0.80	0.60	0.88	0.84	0.84	0.71	0.66	0.80	0.84	0.64	0.81	0.76	0.87	0.85	0.82	0.75	0.77	0.71	0.72	0.83	0.87	0.87	0.87



(a) 1096_left - 1096_right (2/23)

(b) 13874_left - 13874_right (3/23)

(c) 12159_left - 12159_right (6/23)

(d) 1322_left - 1322_right (23/23)

(e) 10913_left - 10913_right (23/23)

(f) 11638_left - 11638_right (23/23)

(g) 15745_left - 20759_right (0/13)

(h) 14947_left - 21065_right (0/13)

(i) 16993_left - 22240_right (0/13)

(j) 18071_left - 22538_right (6/13)

(k) 16021_left - 20195_right (7/13)

(l) 18252_left - 21524_right (12/13)

FIGURE 11. Examples of easy-to-recognize pairs and difficult-to-recognize pairs for human volunteers. Easy-to-recognize pairs were correctly recognized, whereas difficult-to-recognize pairs were incorrectly classified by the majority of volunteers. [1st row: three difficult-to-recognize PPs, 2nd row: three easy-to-recognize PPs, 3rd Row: three easy-to-recognize NPs, and 4th Row: two comparatively difficult-to-recognize NPs (j) & (k) + one difficult-to-recognize NP (l). The title of a sub-figure indicates SubjectID_left - SubjectID_right (the number of volunteers who thought the pairs belonged to the same subject / total number of volunteers). Source of image: Kaggle_SetA.]

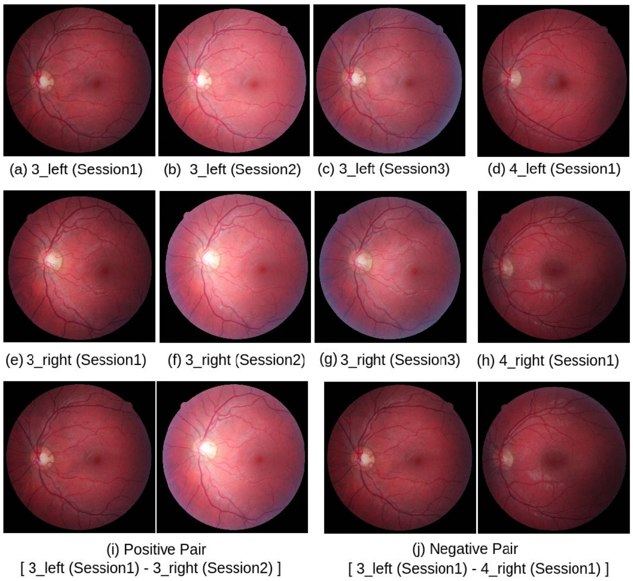
epochs. For adaptation, we trained Model4 for ten epochs using the Kaggle_SetB data set at first and then retrained Model4 using the Messidor-2 data set, keeping the same set up as we used for all other models. We set $batch\_size = 64$ when training a model except Model3 (i.e., the VGG16 based model). Due to memory limitations, we set $batch\_size = 32$ when training Model3. For all other settings, we used the default values of TensorFlow's Keras API 2.6.1-tf.

E. IMAGE PRE-PROCESSING

We cropped the dark pixels of the background, which do not provide any information about the retina, using a simple background cropping algorithm. At first, we converted an RGB-colored fundus photograph into a grayscale image. Then we blurred the grayscale image using a 5 × 5 Gaussian kernel. After that, we detected edges using *Canny's edge detection* algorithm ([65]). After that, we found the contour which had the maximum area. After that, we estimated the radius of the circle that minimally enclosed that contour. Using that radius, we cropped the background. We used the functions of the OpenCV library for this part. After cropping the background, since the different data sets have different resolutions, we re-sized all images to 256 × 256 by bicubic interpolation. Then we re-scaled pixel values to [0, 1] for each channel of each image independently. Except the above processes, no pre-processing was applied to any images.

TABLE 6. Frequency of symmetrical properties noted down by trained and untrained volunteers. [Notations are NUV: number of untrained volunteers who found symmetry, NTV: number of trained volunteers who found symmetry, TUV: total number of untrained volunteers, TTV: total number of trained volunteers.]

Symmetrical properties found in	NUV/TUV	NTV/TTV
CRBVs	17/20	3/3
· Pattern in CRBVs	4/20	2/3
· Pattern in thick CRBV	1/20	0/3
· Branching pattern in CRBVs	3/20	1/3
· Density of CRBVs	4/20	0/3
· Spreadness of CRBVs	5/20	2/3
· End of a CRBV	1/20	0/3
· Tortuosity of CRBVs	5/20	0/3
· Thickness of CRBVs	1/20	0/3
· Color intensity of CRBVs	3/20	1/3
Foreground	15/20	3/3
· Foreground color	15/20	2/3
· Color of the border of the foreground	0/20	1/3
· Artifacts in the border of the foreground	1/20	0/3
Optic disc (OD)	12/20	2/3
· Size & shape of OD	4/20	2/3
· Pattern and thickness of CRBVs in OD	3/20	0/3
· Exit of CRBVs from OD	3/20	1/3
· Number of branches coming out from OD	2/20	0/3
· Border of OD	1/20	0/3
· Visibility of optic cup (OC)	1/20	0/3
· Color of OC	2/20	0/3
· Orientation of OD	1/20	0/3
Macula	8/20	0/3
· Overall macula	1/20	0/3
· Color of macula	1/20	0/3
· Size of macula	1/20	0/3
· Shape of macula	1/20	0/3
· Alignment of OD with macula	5/20	0/3
· Position of macula	1/20	0/3
Choroidal blood vessels	5/20	0/3


(a) 3_left (Session1) (b) 3_left (Session2) (c) 3_left (Session3) (d) 4_left (Session1)

(e) 3_right (Session1) (f) 3_right (Session2) (g) 3_right (Session3) (h) 4_right (Session1)

(i) Positive Pair
[3_left (Session1) - 3_right (Session2)]

(j) Negative Pair
[3_left (Session1) - 4_right (Session1)]

FIGURE 12. Effect of camera settings and environment on the foreground color. (a), (b) & (c): images were captured from the left retina of Subject-3 in three different sessions, (e), (f) & (g): images were captured from the right retina of Subject-3 in three different sessions, (d) & (h): images were captured from the left and right retinas of Subject-4 in the same session, (i): a positive pair prepared by the left and right retinas of Subject-3 and (j): a negative pair prepared by the left retina of Subject-3 and the right retina of Subject-4. The foreground colors of the images captured from the same retina but in different sessions are different. On the contrary, the foreground colors of the images captured from different subjects but in the same session are same. Therefore, humans will make wrong decisions about pairs in (i) & (j), if their decisions are based on the foreground color. [Note that Subject-3 and Subject-4 were male and European. Camera settings and lighting conditions were different in different sessions, but the same for two subjects in the same session. There was one month gap between the 1st and 2nd sessions and 14 days gap between the 2nd and 3rd sessions. Right side images are horizontally flipped. Source of image: STRaDe data set.]

V. RESULTS & ANALYSIS

In this section, we present and analyse the results of several experiments. In Subsection V-A and V-C, we present the results of manual and automatic *different-side* verification, respectively. We analyze what patterns human volunteers and deep neural networks are looking at in Subsection V-B and V-D, respectively. We compare the performance of whole color fundus photographs-based automatic verification with the CRBVs-based automatic verification in Subsection V-E. In Subsection V-F, we analyse the agreement of manual and automatic verification. In Subsection V-G, we compare results of different-side with same-side verification as well as different-session image verification. We compare results of different NNs having slightly different architectures comparing to YNN in Subsection V-H. Finally, we compare different channel based YNNs in Subsection V-I.

A. MANUAL VERIFICATION

As can be seen in Table 4, on average, the untrained volunteers had an accuracy of 79%, whereas the trained volunteers had an accuracy of 89%. As shown in Table 5, even the result of the volunteer with the lowest score was better than the result expected from random guesses. Therefore, we can conclude that human volunteers can catch substantial bilateral symmetry in color fundus photographs if they investigate carefully.

Most of the untrained volunteers were not familiar with fundus photographs, as shown in Table 3. It seems familiarity with color fundus photographs did not positively impact the performance of untrained volunteers. The worst performer (having volunteer ID 7) was familiar with color fundus photographs; on the other hand, the best performer (having volunteer ID 4) was not.

There was, on average, a relatively small difference between the performance of the top four untrained volunteers (having volunteer ID 4, 14, 6 & 15)) and three trained volunteers (having volunteer ID 21, 22, & 23). The first reason could be that four pairs of left and right retinas were in a frame. The untrained volunteers knew beforehand that there would be both PPs and NPs in the test, although not necessarily in every frame. This prior knowledge helped them to gradually make rules about symmetry, many of which were correct. The second reason could be that the trained

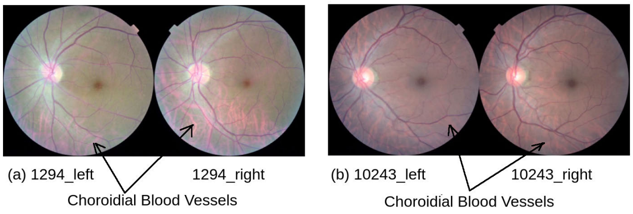


FIGURE 13. Visibility of choroidal blood vessels can make a positive pair easy to recognize by humans. [Source of image: Kaggle_SetA.]

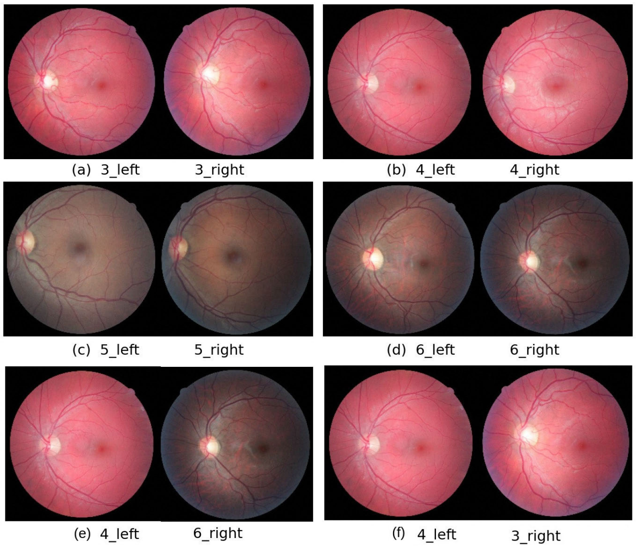


FIGURE 14. Variability in the foreground color caused by ethnicity. (a) & (b): two positive pairs (PPs) from two Europeans, (c) & (d): two PPs from two South-Asians, (e): one negative pair (NP) prepared by combining retinas from an European and a South-Asian and (f) one NP prepared by combining retinas of two Europeans. Notice that Europeans and South-Asians have obvious different foreground colors. By combining retinas from subjects belonging to these two ethnic groups, we can prepare easily recognizable NPs as shown in (e). By combining retinas from subjects belonging to the same ethnic group, we can prepare NPs which are difficult to recognize, as shown in (f). [Note that, the right side images are horizontally flipped. Source of image: STRaDe data set.]

volunteers were not as focused as some untrained volunteers during the test.

The pairs that the majority of volunteers correctly recognized can be considered easy-to-recognize pairs; contrary, the pairs that the majority of volunteers incorrectly classified can be considered difficult-to-recognize pairs. Some example pairs of these two categories are shown in Fig. 11. As shown in Fig. 10 (a), among 50 positive pairs (PPs), there were seven easy-to-recognize PPs which were recognized by all 23 volunteers. As shown in Fig. 10 (b), among 50 negative pairs (NPs), there were 15 easy-to-recognize NPs which were correctly recognized by all 13 volunteers. Note that among 23 volunteers, all volunteers saw the same 50 PPs, but only 13 volunteers (having volunteer ID 10-23) saw the same NPs.

Among top four untrained volunteers and three trained volunteers, all seven volunteers (having volunteer ID 4, 14, 6, 15, 21, 22, 23) saw the same 50 PPs, but only five volunteers (having volunteer ID 14, 15, 21, 22, 23) saw the same NPs. Interestingly, these volunteers had the same opinions about the majority of PPs and NPs (as shown in Fig. 10 (c) & (d)). Seven volunteers correctly recognized 28 out of 50 PPs (as shown in Fig. 10 (c)). Five volunteers correctly recognized 37 out of 50 NPs (as shown in Fig. 10 (d)).

B. WHAT HUMANS LOOK FOR

The human volunteers discovered symmetrical properties in the foreground, optic disc, macula, CRBVs, and choroidal blood vessels for verification. Every volunteer considered multiple properties while making a decision about a pair. Although the untrained volunteers did not influence each other, interestingly, the properties they discovered overlapped with each other. Some of the properties are summarized in Table 6.

The most frequently mentioned properties are related to the CRBVs. Even though CRBVs give an asymmetrical look in a pair of fundus photographs at first glance, the majority of volunteers found symmetry in CRBVs. Interestingly, the findings of the human volunteers match well with the findings in the previous studies [26], [34]–[36] (which were conducted much before our experiments). For example:

- By measuring diameters of all retinal arterioles and venules located 0.5 – 1.0 disc diameters from the OD margin of 1546 subjects, bilateral symmetry in the left and right retina was reported in [26].
- By measuring the fractal dimension of the retinal vascular network as a means of quantifying the branching pattern, bilateral symmetry in the retina was reported in [37].

By using masks of CRBVs (images where CRBVs are visible as white on a black background), we have analysed the symmetry in CRBVs in more detail with the help of 24 human volunteers in our previous work [2]. Among the volunteers in that work, 20 untrained and two trained volunteers also participated in this work, i.e., 2D color fundus photographs-based experiments.

After CRBVs, the foreground color was the property that influenced most volunteers in their decisions. However, our further analysis found that emphasizing on the foreground color to make a decision can sometimes be misleading. It is true that retinas from the same subject, in general, have the same foreground color and different colored retinas for a single subject are pretty rare to find (actually, it might happen for eyes having heterochromia iris). However, the images captured from the same retina of a subject may have different colors, whereas images captured from retinas of different subjects may have the same color. As shown in Fig. 12 (a), (b), & (c), the foreground colors of the images captured from the left retina of Subject-3 in three different sessions are different. Same phenomenon happened for the right side retina as shown in Fig. 12 (e), (f), & (g). On the contrary, images shown in Fig. 12 (a) & (d) have the same color even they are captured from the left retinas of two different subjects. Same phenomenon happened for the images shown in Fig. 12 (e) & (h). Different factors can be the

TABLE 7. Performance of different types of YNNs for verifying pairs of color fundus photographs of the left and right retinas. Randomization caused by random weight initialization and dropout had a large effect. [Note that each type of YNNs was trained five times. The results are in the mean ± standard deviation form. Notations are AUC: Area under the receiver operating characteristic (ROC) curve, EER: Equal error rate, NF: NoFlip, i.e., right side retinas were not flipped. RF: RightFlip, i.e., right side retinas were flipped. Ctt, Avg, Sub, and Ctt(Sub,Avg) are applied for merging the high-level features. Ctt: Concatenation, Avg: Average, Sub: Absolute subtraction, Ctt(Sub,Avg): Concatenation of absolute subtraction and average.]

	Kaggle_SetA.2							
	Untied							
	NF				RF			
	Ctt	Avg	Sub	Ctt(Sub,Avg)	Ctt	Avg.	Sub	Ctt(Sub,Avg)
Accuracy	0.890 ± 0.024	0.900 ± 0.011	0.880 ± 0.028	0.896 ± 0.017	0.824 ± 0.031	0.774 ± 0.083	0.830 ± 0.068	0.804 ± 0.076
Sensitivity	0.984 ± 0.008	0.980 ± 0.000	0.984 ± 0.008	0.984 ± 0.008	0.984 ± 0.008	0.980 ± 0.000	0.976 ± 0.008	0.984 ± 0.008
Specificity	0.796 ± 0.046	0.820 ± 0.022	0.776 ± 0.051	0.808 ± 0.041	0.664 ± 0.066	0.568 ± 0.166	0.684 ± 0.137	0.624 ± 0.159
F1	0.900 ± 0.020	0.908 ± 0.009	0.892 ± 0.023	0.905 ± 0.014	0.849 ± 0.022	0.816 ± 0.054	0.854 ± 0.047	0.837 ± 0.050
AUC	0.958 ± 0.015	0.962 ± 0.006	0.941 ± 0.016	0.960 ± 0.023	0.915 ± 0.019	0.916 ± 0.029	0.942 ± 0.009	0.900 ± 0.047
EER	0.068 ± 0.027	0.068 ± 0.010	0.080 ± 0.013	0.064 ± 0.027	0.132 ± 0.045	0.128 ± 0.037	0.104 ± 0.027	0.116 ± 0.034
	Tied							
	NF				RF			
	Ctt	Avg.	Sub	Ctt(Sub,Avg)	Ctt	Avg.	Sub	Ctt(Sub,Avg)
Accuracy	0.906 ± 0.024	0.810 ± 0.026	0.942 ± 0.012	0.896 ± 0.033	0.882 ± 0.018	0.748 ± 0.092	0.876 ± 0.026	0.870 ± 0.017
Sensitivity	0.980 ± 0.000	0.880 ± 0.123	0.984 ± 0.015	0.984 ± 0.008	0.988 ± 0.010	0.972 ± 0.027	0.996 ± 0.008	1.000 ± 0.000
Specificity	0.832 ± 0.048	0.740 ± 0.097	0.900 ± 0.018	0.808 ± 0.069	0.776 ± 0.041	0.524 ± 0.206	0.756 ± 0.059	0.740 ± 0.033
F1	0.913 ± 0.021	0.819 ± 0.042	0.944 ± 0.011	0.905 ± 0.028	0.894 ± 0.014	0.800 ± 0.058	0.890 ± 0.020	0.885 ± 0.013
AUC	0.956 ± 0.019	0.893 ± 0.025	0.987 ± 0.006	0.956 ± 0.011	0.963 ± 0.008	0.883 ± 0.038	0.951 ± 0.016	0.930 ± 0.037
EER	0.076 ± 0.027	0.168 ± 0.027	0.044 ± 0.020	0.044 ± 0.015	0.072 ± 0.020	0.164 ± 0.039	0.060 ± 0.025	0.080 ± 0.013

	Kaggle_SetC							
	Untied							
	NF				RF			
	Ctt	Avg.	Sub	Ctt(Sub, Avg)	Ctt	Avg.	Sub	Ctt(Sub, Avg)
Accuracy	0.864 ± 0.029	0.839 ± 0.011	0.837 ± 0.023	0.857 ± 0.033	0.795 ± 0.041	0.776 ± 0.062	0.826 ± 0.059	0.772 ± 0.064
Sensitivity	0.983 ± 0.006	0.985 ± 0.007	0.990 ± 0.002	0.983 ± 0.012	0.976 ± 0.011	0.972 ± 0.010	0.970 ± 0.009	0.989 ± 0.002
Specificity	0.745 ± 0.064	0.694 ± 0.027	0.684 ± 0.048	0.731 ± 0.078	0.613 ± 0.091	0.581 ± 0.126	0.683 ± 0.121	0.555 ± 0.129
F1	0.879 ± 0.022	0.860 ± 0.008	0.859 ± 0.017	0.874 ± 0.025	0.828 ± 0.028	0.815 ± 0.040	0.850 ± 0.041	0.815 ± 0.042
AUC	0.951 ± 0.016	0.944 ± 0.013	0.935 ± 0.017	0.935 ± 0.020	0.918 ± 0.014	0.917 ± 0.031	0.932 ± 0.033	0.906 ± 0.037
EER	0.105 ± 0.013	0.114 ± 0.018	0.115 ± 0.011	0.107 ± 0.012	0.128 ± 0.016	0.142 ± 0.028	0.120 ± 0.028	0.131 ± 0.034
	Tied							
	NF				RF			
	Ctt	Avg.	Sub	Ctt(Sub, Avg)	Ctt	Avg.	Sub	Ctt(Sub, Avg)
Accuracy	0.883 ± 0.032	0.800 ± 0.039	0.940 ± 0.008	0.891 ± 0.031	0.879 ± 0.014	0.754 ± 0.073	0.887 ± 0.022	0.868 ± 0.018
Sensitivity	0.983 ± 0.009	0.878 ± 0.158	0.974 ± 0.009	0.987 ± 0.007	0.981 ± 0.002	0.963 ± 0.036	0.983 ± 0.002	0.985 ± 0.007
Specificity	0.784 ± 0.072	0.721 ± 0.089	0.906 ± 0.021	0.794 ± 0.069	0.778 ± 0.027	0.545 ± 0.174	0.791 ± 0.046	0.752 ± 0.042
F1	0.895 ± 0.027	0.807 ± 0.066	0.942 ± 0.007	0.901 ± 0.026	0.891 ± 0.011	0.800 ± 0.043	0.897 ± 0.019	0.882 ± 0.014
AUC	0.961 ± 0.012	0.893 ± 0.027	0.986 ± 0.003	0.961 ± 0.013	0.961 ± 0.007	0.874 ± 0.032	0.968 ± 0.010	0.951 ± 0.012
EER	0.081 ± 0.020	0.183 ± 0.025	0.054 ± 0.007	0.066 ± 0.009	0.082 ± 0.011	0.193 ± 0.031	0.066 ± 0.006	0.071 ± 0.005

reasons behind these examples, such as different settings of cameras, lighting conditions where the fundus photographs are captured, and the size of the pupil when images are captured. As a result, we would make the wrong decisions for both PPs and NPs if we consider the foreground color as the main factor.

Many volunteers also found symmetric properties in the optic disc, macula, and choroidal blood vessels. The number of volunteers who found symmetry in the optic disc was larger than the number of volunteers who found symmetry in the macula. Interestingly, the top four untrained volunteers used the presence of the choroidal blood vessels while making decisions. Our further analysis found that the choroidal blood vessels are not visible in fundus photographs of all subjects. If they are visible, they are visible on both sides of the retinas and have a similar pattern in both retinas as shown in Fig. 13.

Some of the properties noted by the volunteers were incorrectly assumed to be related to the subject's identity, i.e., artifacts in the foreground border, the orientation of the optic disc, the position of the macula, and alignment of the optic disc with the macula. The first one depends on the camera lens, while the others depend on which direction the subject looks at while the retinal photograph is captured. Accordingly, these properties can be different for the same eye of a subject. Therefore, these properties should not be used for making decisions about any pairs.

Many factors, such as experience level of the operator, operator's finger movement or shaking, different settings of fundus cameras, subject's eye movement or blinking, different amounts of light reflected by different parts of the retina because of its natural curved structure, inadequate illumination, variation of pupil dilation, and poor focus, can result in poor-quality retinal images. Poor quality fundus photographs can mislead observers. Therefore, the quality of fundus photographs should be assessed, and poor-quality fundus photographs should be discarded before taking

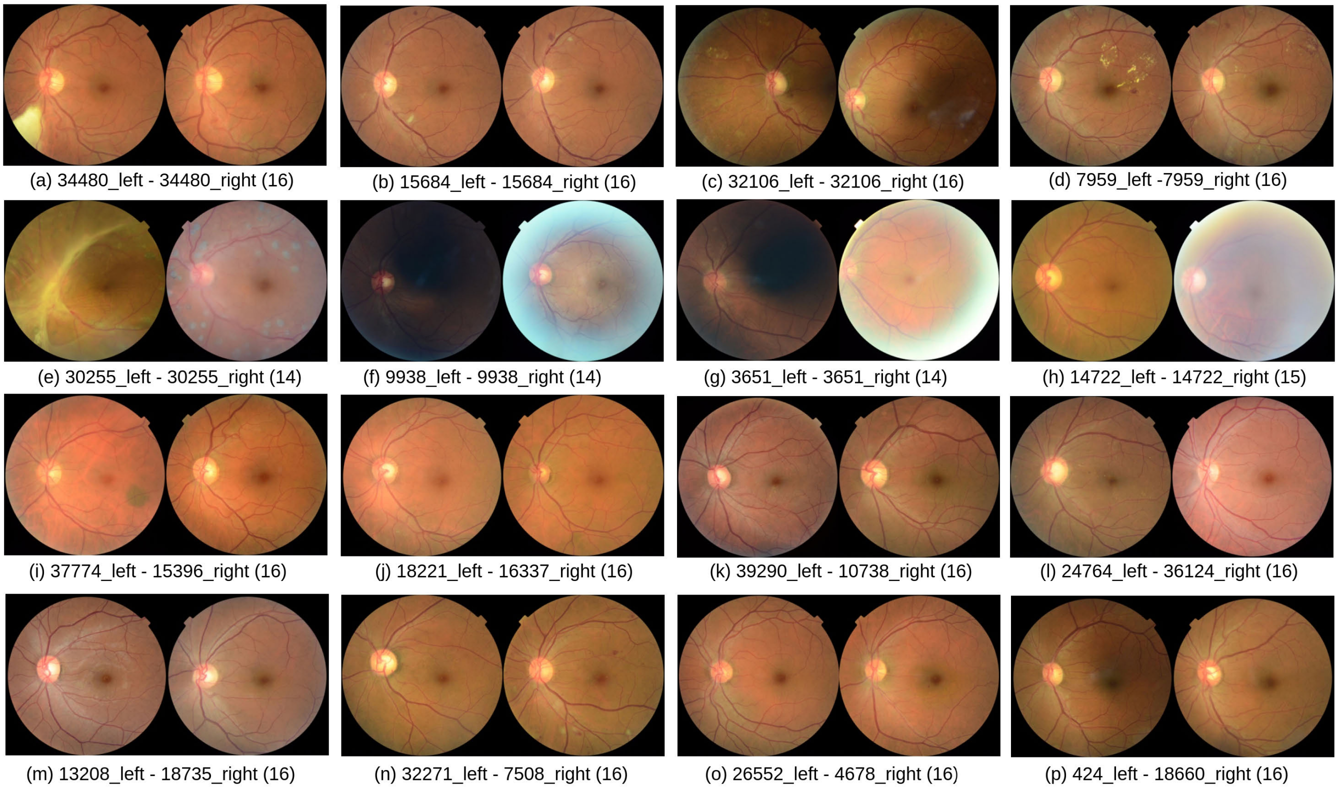

(a) 34480_left - 34480_right (16)
(b) 15684_left - 15684_right (16)
(c) 32106_left - 32106_right (16)
(d) 7959_left -7959_right (16)
(e) 30255_left - 30255_right (14)
(f) 9938_left - 9938_right (14)
(g) 3651_left - 3651_right (14)
(h) 14722_left - 14722_right (15)
(i) 37774_left - 15396_right (16)
(j) 18221_left - 16337_right (16)
(k) 39290_left - 10738_right (16)
(l) 24764_left - 36124_right (16)
(m) 13208_left - 18735_right (16)
(n) 32271_left - 7508_right (16)
(o) 26552_left - 4678_right (16)
(p) 424_left - 18660_right (16)

FIGURE 15. Examples of easy-to-recognize pairs and difficult-to-recognize pairs for YNNs. Easy-to-recognize pairs were recognized correctly and difficult-to-recognize pairs were recognized wrongly by the majority of YNNs. 1st row: positive pairs (PPs) recognized correctly by 16 varieties of YNN. 2nd row: PPs recognized wrongly as negative pairs (NPs) by 14 or 15 out of 16 varieties of YNNs. 3rd row: NPs recognized correctly as NPs by 16 varieties of YNN. 4th row: NPs recognized as PPs by 16 varieties of YNNs. [Note that the title of each sub-figure indicates SubjectID_left - SubjectID_right (Number of YNNs decided that the pair belonged to the same subject). Source of image: Kaggle_SetC.]

opinions from human volunteers. Different kinds of retinal image quality assessment algorithms (e.g. [66]–[74]) can be used to assess quality of fundus photographs automatically. Finding an appropriate algorithm is kept as future work.

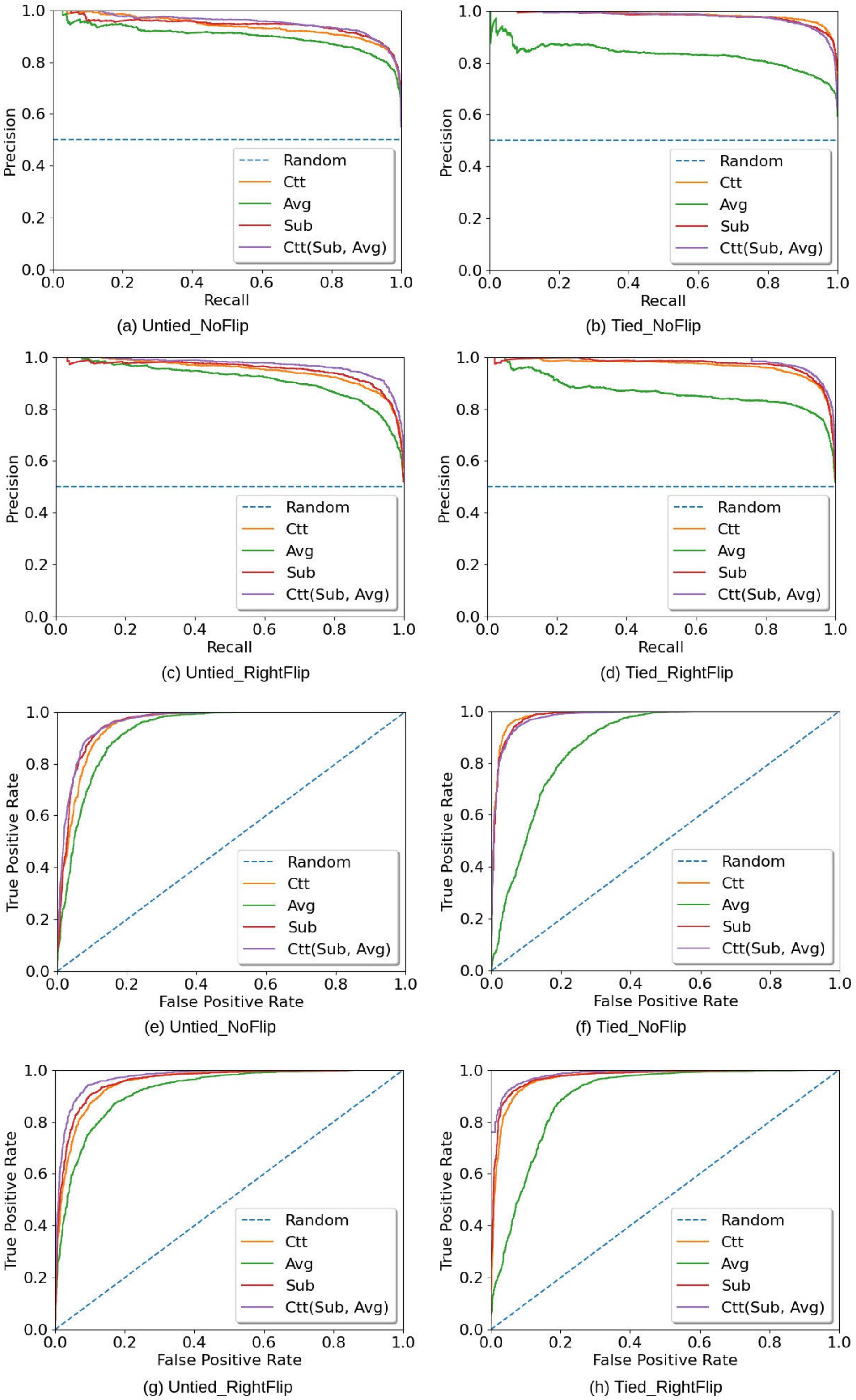
Our other observation is that ethnicity and gender differences in the retinas of a pair can make it an easily recognizable NP. There are potential ethnic and gender differences in retinal vascular geometric parameters and retinal structures, e.g.:

- Among Indians, Malays, and Chinese subjects, Indians have the largest retinal arteriolar and venular calibers, whereas Chinese have the largest retinal arteriolar and venular tortuosity and venular fractal dimension [75].
- There are larger retinal arteriolar and venular calibers in Blacks and Hispanics than Whites and Chinese [76]. Both retinal arteriolar and venular calibers are substantially wider in the East-Asian than in the Caucasian children [77].
- The fovea is significantly less thick in females than in males. It is less thick in African-Americans than in Caucasians [78], [79]. The mean foveal thickness of Hispanics is in the middle of Caucasians and African-Americans [79].
- There is a significant difference in the macular pigment density between white non-Hispanic and African subjects. A parafoveal ring is significantly more frequent in African subjects than in white non-Hispanic subjects [80].
- East Asian children have similar mean vertical OD diameters to European-Caucasians but 30-43% larger mean vertical OC diameters, resulting in larger mean OC/OD ratios [81].
- The foreground color of the retina depends mainly on the amount of melanin in the RPE layer. Different ethnic groups have different amounts of melanin. Therefore, a wide color spectrum can be seen for the retina. Caucasians have a strong red component, whereas African-Americans have a much stronger blue component in the retina [82].

When retinas in an NP are from two different ethnic groups (as shown in Fig. 14 (e)), it is easy for humans to decide that the retinas in the pair are from two different subjects by looking at only the foreground color. However, retinas in an NP from the same ethnic group (as shown in Fig. 14 (f)) are difficult to be recognized. In such a case, humans need to consider other properties.

C. AUTOMATIC VERIFICATION

Table 7 and Fig. 16 show the performance of YNN for different ways of: tying parameters (tied/untied),



(a) Untied_NoFlip

(b) Tied_NoFlip

(c) Untied_RightFlip

(d) Tied_RightFlip

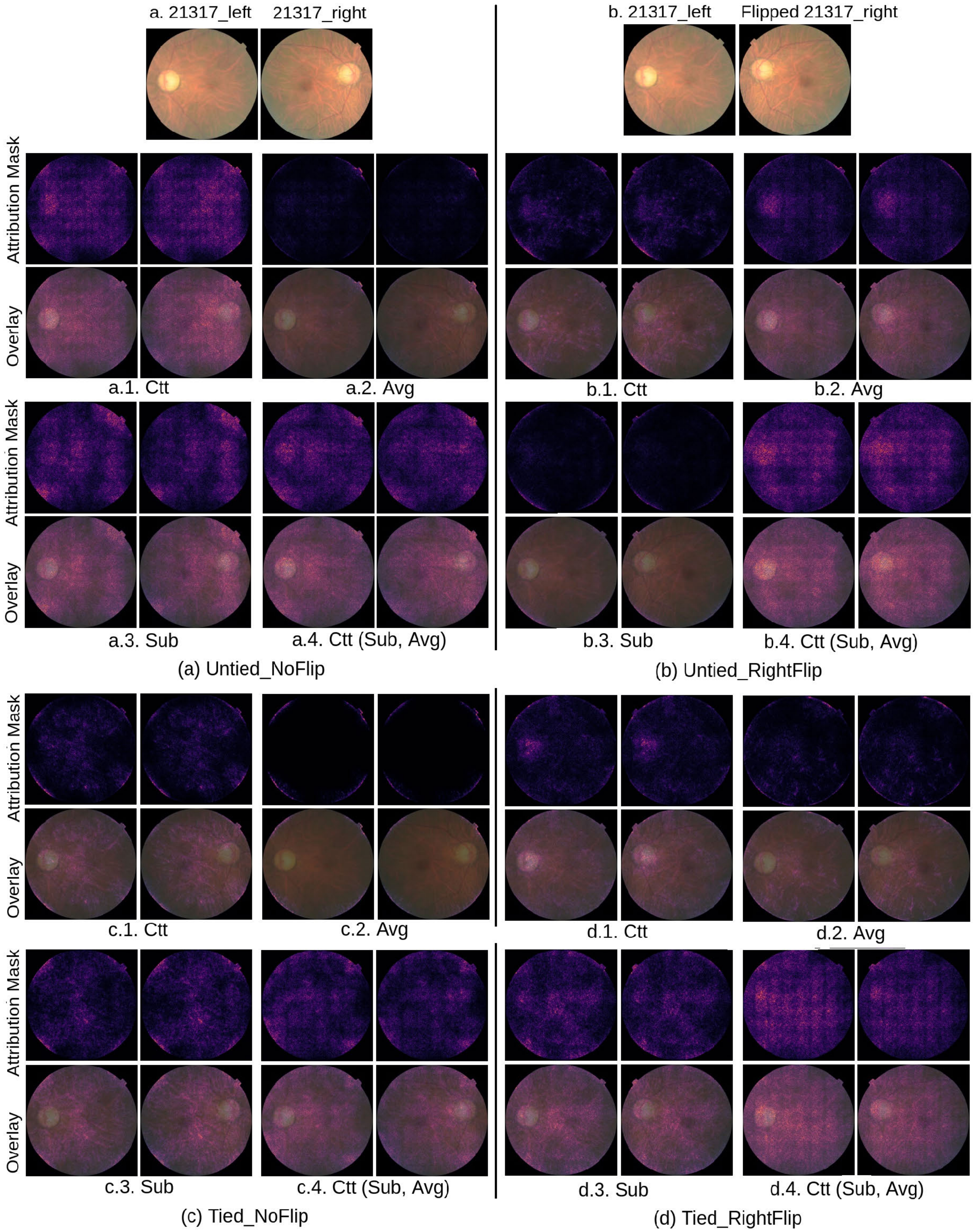
(e) Untied_NoFlip

(f) Tied_NoFlip

(g) Untied_RightFlip

(h) Tied_RightFlip

FIGURE 16. Precision-Recall curves [(a) - (d)] and Receiver Operating Characteristic (ROC) curves [(e)-(g)] for different types of YNNs for the Kaggle_SetC data set. For all operating points in both precision-recall and ROC curves, *Avg* is the worst for all cases and *Ctt(avg,sub)* is the best for most of the cases. It can also be noticed that for the precision-recall, the ranking of the system depends on the operating point, i.e., *Ctt* is better than *Sub* when high precision is important but *Sub* is better than *Ctt* when high recall is important.



(a) Untied_NoFlip

(b) Untied_RightFlip

(c) Tied_NoFlip

(d) Tied_RightFlip

FIGURE 17. What YNNs with different feature mergers looked at for verifying a positive pair of color fundus photographs. The integrated gradients attribution method [84] is used to highlight the contribution of important parts of the pair in the similarity score.

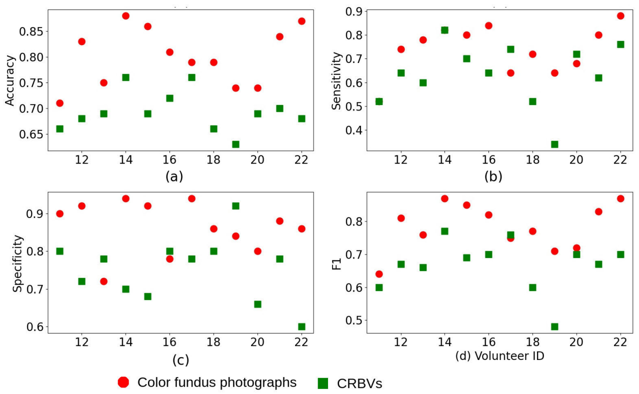
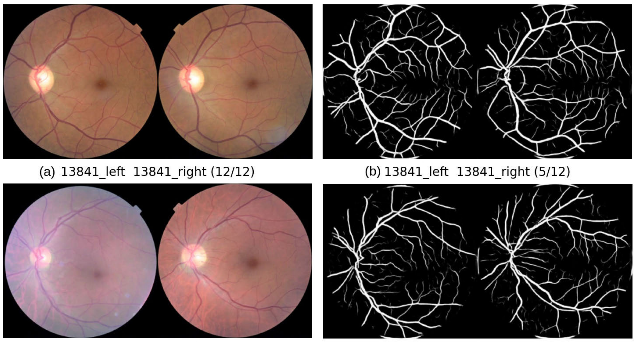


FIGURE 18. Comparison between manual verification based on color fundus photographs and manual verification based on masks of CRBVs. Most of the volunteers performed better in color fundus photograph based verification. [Note that verification was performed on the Kaggle_SetA data set.]



(a) 13841_left 13841_right (12/12)

(b) 13841_left 13841_right (5/12)

(c) 16496_left 20288_right (0/12)

(d) 16496_left 20288_right (6/12)

FIGURE 19. Examples of cases when color fundus photograph based verification is better than masks of CRBV based verification. (a) & (c): color fundus photographs of a positive pair (PP) and a negative pair (NP). (b) & (d): masks of CRBVs generated from the color fundus photographs in (a) & (c), respectively. Human volunteers took advantage of visibility of the optic disc, macula and foreground color in the color fundus photographs to make correct decisions. Therefore, more human volunteers correctly recognize a PP and an NP in color fundus photographs than in masks of CRBVs. [Note that (x/y) in each subtitle indicates that x out of y number of volunteers selected that pair as a PP. Source of image: Kaggle_SetA.]

merging features (concatenation/subtraction/average), and pre-processing inputs (no flipping / flipping the right side). From the results, it is clear that YNNs performed substantially better than human volunteers. The results for Kaggle_SetA.2 and Kaggle_SetC are almost the same. It is expected since these sets have the same properties. Since Kaggle_SetC is much larger than Kaggle_SetA.2 it provides more reliable results. Randomization caused by random weight initialization and dropout had a bigger effect than we thought. Our other observations are that:

- overall, non-flipping the right retina is better than flipping the right retina.
- except for *average* merging, tying parameters is better than non-tying parameters.

TABLE 8. Percentage of agreement between color fundus photograph based verification and masks of CRBV based verification. [Notations are PP: positive pairs, NP: negative pairs, C(S,A): Concatenation of absolute subtraction and average.]

Merger	Flip	Kaggle_SetA				Kaggle_SetC			
		Untied		Tied		Untied		Tied	
		PP	NP	PP	NP	PP	NP	PP	NP
Concat	NF	96	72	94	82	96	70	95	77
	RF	94	66	96	66	95	62	95	68
Avg	NF	94	78	52	58	95	64	42	42
	RF	94	58	92	62	91	52	89	43
Sub	NF	98	74	98	70	97	68	98	69
	RF	92	62	100	68	95	56	97	55
C(S,A)	NF	94	70	96	70	98	57	98	64
	RF	96	72	98	66	96	62	98	59

TABLE 9. Scores of the six positive pairs (PPs) and six negative pairs (NPs) shown in Fig. 11, decided by human volunteers and YNN. The human scores were decided by averaging the decision of 23 volunteers for PPs and 13 volunteers for NPs. In these 12 pairs, there were three easily recognized PPs, three easily recognized NPs, three difficult PPs, and three difficult NPs. Note that scores near 1.0 are better for PPs, while scores near 0.0 are better for NPs. [Notations are Vlt: Volunteers, C(S, A): Concatenation of absolute subtraction and average, and YNN_Untied_NoFlip: the parameters of the feature extractors of the YNN were untied, and the right side retinal images were not horizontally flipped.]

	Pair ID	Vlt.	YNN_Untied_NoFlip			
			Ctt	Avg	Sub	C(S,A)
PPs	1096_left-1096_right	0.09	1.00	1.00	1.00	1.00
	13874_left-13874_right	0.13	0.02	0.25	0.03	0.00
	12159_left-12159_right	0.26	0.97	1.00	1.00	1.00
	1322_left-1322_right	1.00	1.00	1.00	1.00	1.00
	10913_left-10913_right	1.00	1.00	1.00	1.00	1.00
	11638_left-11638_right	1.00	1.00	1.00	1.00	1.00
NPs	15745_left-20759_right	0.00	0.89	0.97	1.00	0.99
	14947_left-21065_right	0.00	0.00	0.00	0.01	0.00
	16993_left-22240_right	0.00	0.99	1.00	1.00	1.00
	18071_left-22538_right	0.46	0.00	0.00	0.00	0.00
	16021_left-20195_right	0.54	0.73	1.00	1.00	1.00
	18252_left-21524_right	0.92	0.00	0.00	0.00	0.00

TABLE 10. The number of times volunteers and YNN agreed with each other for 50 PPs and 50 NPs of the Kaggle_SetA.2. [Notations are Vlt: Volunteers and YNN_Untied_NoFlip: the parameters of the feature extractors of the YNN were untied and the right side retinal images were not horizontally flipped.]

		Vlt.	YNN_Untied_NoFlip			
			Ctt	Avg	Sub	Ctt(Sub,Avg)
PPs	Vlt	-	39	40	40	39
	Ctt	39	-	49	49	50
	Avg	40	49	-	50	49
	Sub	40	49	50	-	49
	Ctt(Sub,Avg)	39	50	49	49	-
NPs	Vlt	-	41	41	41	37
	Ctt	41	-	44	50	46
	Avg	41	44	-	44	40
	Sub	41	50	44	-	46
	Ctt(Sub,Avg)	37	46	40	46	-

- absolute subtraction performs consistently better than other merging approaches for both the Kaggle_SetA.2 and Kaggle_SetC data sets.

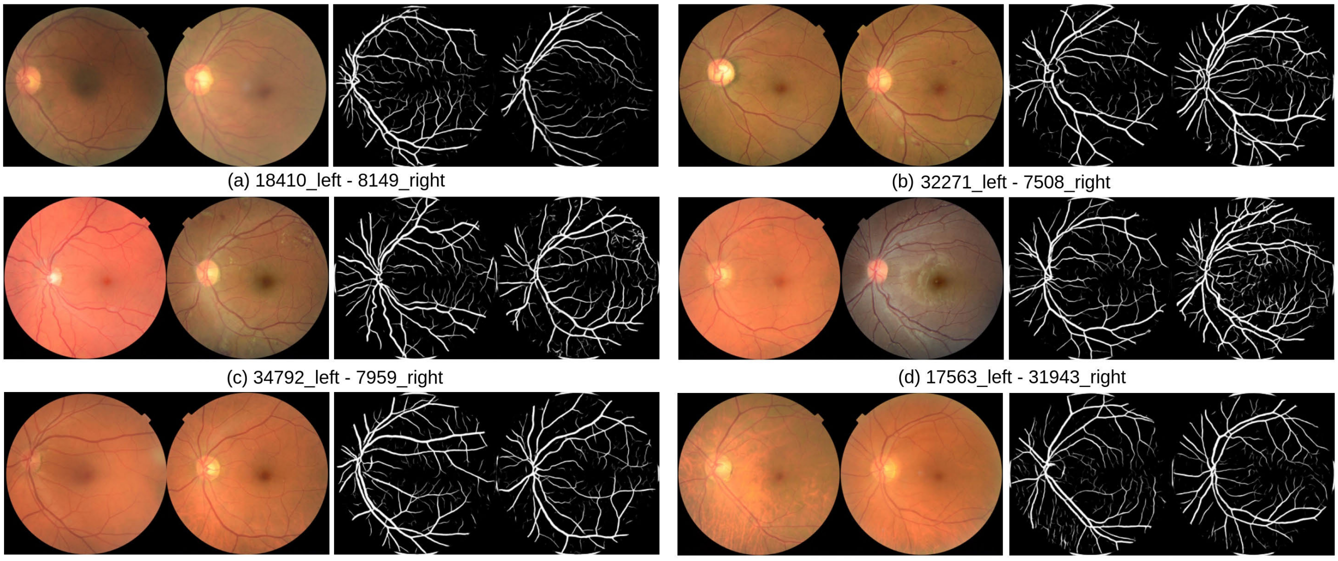

(a) 18410_left - 8149_right
(b) 32271_left - 7508_right
(c) 34792_left - 7959_right
(d) 17563_left - 31943_right
(e) 33156_left - 29105_right
(f) 3987_left - 35137_right

FIGURE 20. Examples of agreement and disagreement between color fundus photograph based verification and masks of CRBV based verification by 16 varieties of YNN. (a) & (b) negative pairs (NPs) that YNNs mistakenly considered as positive pairs (PPs) in both types of images based verification. (c) & (d) NPs that were correctly recognized by YNNs when color fundus photographs were used but mistakenly recognized as PPs when masks of CRBVs were used. (e) & (f) NPs that were correctly recognized by YNNs when masks of CRBVs were used but mistakenly recognized as PPs when color fundus photographs were used.

TABLE 11. The number of times volunteers and YNN were correct when disagreeing with each other about 50 PPs and 50 NPs of the Kaggle_SetA.2. [A cell value *x*/*y* means that the system in the row and column header disagreed *y* times and that the system in the row header was correct *x* times. Notation is YNN_Untied_NoFlip: the parameters of the feature extractors of the YNN were untied and the right side retinal images were not horizontally flipped.]

		Vlt.	YNN_Untied_NoFlip			
			Ctt	Avg	Sub	Ctt(Sub,Avg)
PPs	Vlt	-	0/11	0/10	0/10	0/11
	Ctt	11/11	-	1/1	1/1	0/0
	Avg	10/10	0/1	-	0/0	0/1
	Sub	10/10	0/1	0/0	-	0/1
	Ctt(Sub,Avg)	11/11	0/0	1/1	1/1	-
NPs	Vlt	-	8/9	7/9	8/9	12/13
	Ctt	1/9	-	2/6	0/0	4/4
	Avg	2/9	4/6	-	4/6	8/10
	Sub	1/9	0/0	2/6	-	4/4
	Ctt(Sub,Avg)	1/13	0/4	2/10	0/4	-

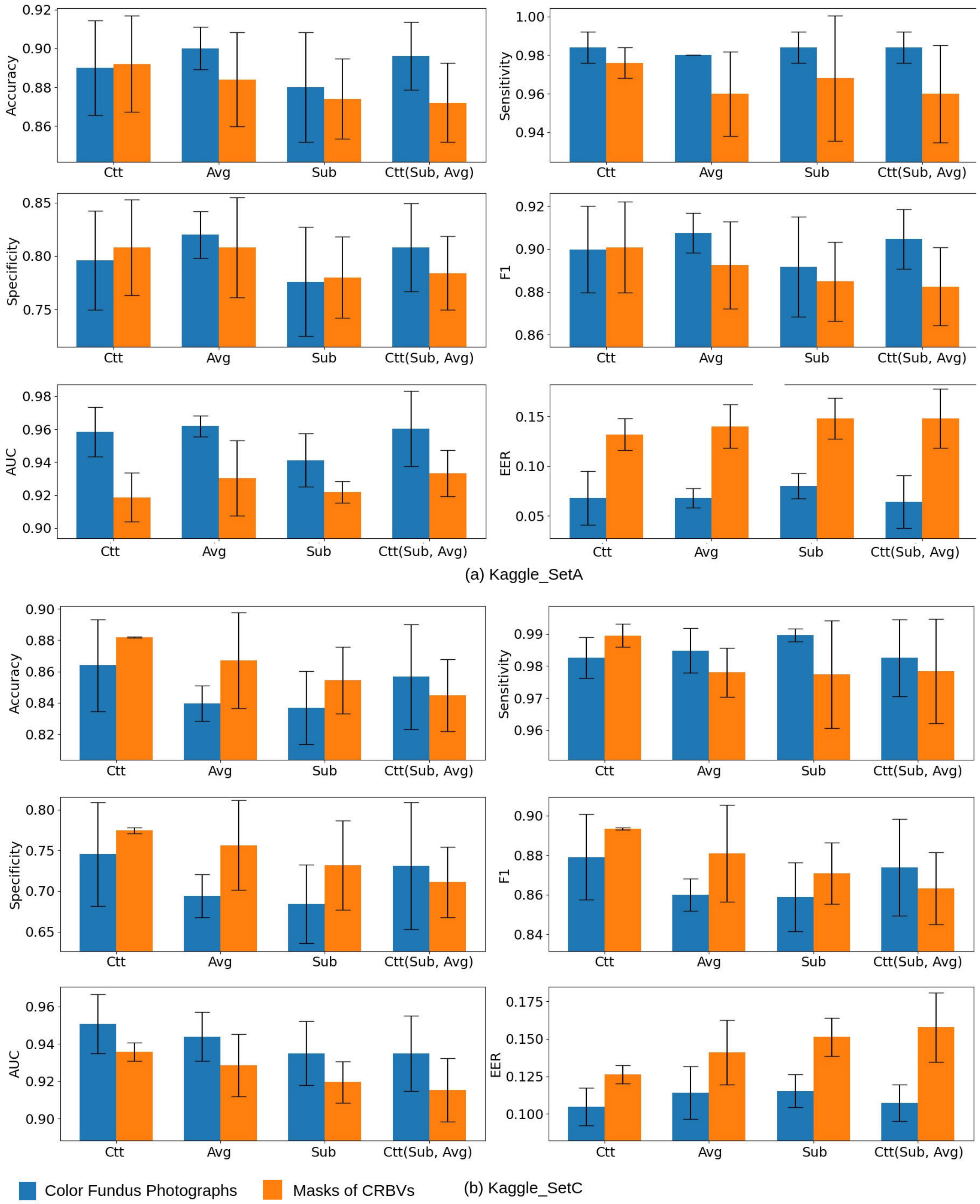
- merging features by the *average* operation after flipping the right retina performs worse than the non-flipping case.
- most of the varieties of YNN perform badly when poor quality image pairs are compared. For example, poor quality PPs are recognized as NPs, as shown in the 2nd row of Fig. 15.
- all varieties of YNN recognize NPs correctly when there is a clear foreground color difference as shown in the 3rd row of Fig. 15.
- all varieties of YNN cannot recognize NPs having almost the same foreground color for both sides as shown in the 4th row of Fig. 15. Increasing the number of NPs having the same colored retinas in the training set of YNN or reducing domain mismatch issue can mitigate this problem to some extent.

D. WHAT NEURAL NETWORKS LOOK AT

To find answers of questions such as "which parts of retinas contribute significantly to the output of the YNN", "Does our YNNs find similarity in the same places of retinas as human volunteers do", or "Do the different YNN architectures look at different things?", we applied an *attribution method*. The task of an attribution method is to estimate how much each input feature contributes to the decisions of the network [83]. We used the integrated gradients method [84] that generates a heatmap representing how important a region of the input image was for the decision. According to this method, for each pair of retinas, we generated a path of 64 image pairs by interpolating between a pair of black baseline images and the original pair of retinal images. We calculated the gradient of the YNN output with respect to each pixel of the pairs of retinas, summed the gradients over each interpolated image, and took the absolute value. As shown in Fig. 17, similar to different human volunteers, different YNNs used different features such as illumination of the foreground, boundary area of the foreground, choroidal blood vessels, optic disc, and the area near the optic disc, to generate a similarity score for a pair of fundus photographs.

E. COLOR FUNDUS PHOTOGRAPHS VS MASKS OF CRBVs

Using a U-shaped convolutional neural network, we generated masks of CRBVs from the color fundus photographs and did the same kind of tests as we did for color fundus



(a) Kaggle_SetA

(b) Kaggle_SetC

FIGURE 21. Comparison between automatic verification based on color fundus photographs and automatic verification based on masks of CRBVs. The difference in performance between the two types of verification was marginal. [Note that verification was performed on (a) the Kaggle_SetA data set and (b) the Kaggle_SetC data set.]

photographs (see [2] for the details of the tests). Overall, the majority of human volunteers recognized more pairs of color fundus photographs correctly than the pairs of CRBVs. Fig. 18 shows the performance of 12 volunteers (ten untrained + two trained volunteers), who saw the same PPs and NPs for both types of images. It seems most of the time, the visibility of optic disc, macula, and foreground color in the color fundus photographs made it easier for human volunteers to make a correct decision about a pair of the left and right retina (see Fig. 19 for examples).

Contrary to the human volunteers, YNN had a marginal difference in performance between the two types of images based verification. Interestingly, most of the time, color fundus photograph based YNN and masks of CRBV based YNN agreed about PPs. On the other hand, both types of YNNs disagreed with NPs quite often (see Table 8). Some examples

TABLE 12. Results of side-independent verification. Domain adaptation by fine-tuning parameters improved the performance of the YNN. [Note that left-right, left-left, and right-right pairs of color fundus photographs were used to train the YNN. The two feature extractors of the YNN shared parameters. Concatenation was used as a merging approach. The right side retinal images were horizontally flipped. Notations are LR-SS: pairs of the left-right retinas from the same session, LR-DS: pairs of left-right retinas from different sessions, LL-DS: pairs of left-left retinas from different sessions, RR-DS: pairs of right-right retinas from different sessions.]

	YNN trained by Kaggle_SetB only					YNN trained by Kaggle_SetB and adapted by Messidor2				
	Kaggle_SetC	RODREP_SetA				Kaggle_SetC	RODREP_SetA			
	LR-SS	LR-SS	LR-DS	LL-DS	RR-DS	LR-SS	LR-SS	LR-DS	LL-DS	RR-DS
Accuracy	0.8773	0.5725	0.5797	0.5942	0.5870	0.8610	0.8406	0.8116	0.8333	0.9058
Sensitivity	0.9720	0.9783	0.9783	1.0000	1.0000	0.8288	0.8768	0.8551	0.8406	0.9275
Specificity	0.7825	0.1667	0.1812	0.1884	0.1739	0.8933	0.8043	0.7681	0.8261	0.8841
F1	0.8879	0.6959	0.6996	0.7113	0.7077	0.8564	0.8458	0.8193	0.8345	0.9078
AUC	0.9424	0.6002	0.6048	0.6232	0.6087	0.9265	0.8817	0.8753	0.8806	0.9468
EER	0.0959	0.6232	0.6594	0.6087	0.7391	0.1358	0.1812	0.1812	0.1739	0.1014

TABLE 13. Performance of varieties of YNN. [NF: NoFlip, i.e., right side retinas were not flipped. RF: RightFlip, i.e., right side retinas were flipped horizontally. Model1: Neural network (NN) having one dense layer less than the YNN, (c) Model2: NN having the same layers as the YNN but with an extra dense layer in each feature extractor and a cosine distance layer as the feature merger and the similarity score generator, and (d) Model3: NN using the pre-trained VGG16 provided by Keras, without the last three dense layers, as feature extractors.]

	Kaggle_SetA.2							
	Untied							
	NF				RF			
	YNN	Model1	Model2	Model3	YNN	Model1	Model2	Model3
Accuracy	0.890 ± 0.024	0.880 ± 0.014	0.594 ± 0.043	0.738 ± 0.048	0.824 ± 0.031	0.770 ± 0.074	0.618 ± 0.055	0.780 ± 0.033
Sensitivity	0.984 ± 0.008	0.980 ± 0.013	0.900 ± 0.151	0.924 ± 0.054	0.984 ± 0.008	0.964 ± 0.043	0.784 ± 0.174	0.976 ± 0.008
Specificity	0.796 ± 0.046	0.780 ± 0.028	0.288 ± 0.232	0.552 ± 0.135	0.664 ± 0.066	0.576 ± 0.173	0.452 ± 0.187	0.584 ± 0.073
F1	0.900 ± 0.020	0.891 ± 0.012	0.686 ± 0.023	0.780 ± 0.026	0.849 ± 0.022	0.811 ± 0.045	0.666 ± 0.066	0.817 ± 0.022
AUC	0.958 ± 0.015	0.936 ± 0.026	0.500 ± 0.000	0.837 ± 0.022	0.915 ± 0.019	0.883 ± 0.067	0.500 ± 0.000	0.882 ± 0.023
EER	0.068 ± 0.027	0.104 ± 0.023	0.348 ± 0.032	0.208 ± 0.048	0.132 ± 0.045	0.140 ± 0.028	0.356 ± 0.050	0.180 ± 0.028
	Tied							
	NF				RF			
	YNN	Model1	Model2	Model3	YNN	Model1	Model2	Model3
Accuracy	0.906 ± 0.024	0.898 ± 0.017	0.714 ± 0.051	0.786 ± 0.040	0.882 ± 0.018	0.824 ± 0.043	0.582 ± 0.064	0.852 ± 0.025
Sensitivity	0.980 ± 0.000	0.980 ± 0.000	0.900 ± 0.074	0.964 ± 0.027	0.988 ± 0.010	0.984 ± 0.008	0.816 ± 0.159	0.956 ± 0.015
Specificity	0.832 ± 0.048	0.816 ± 0.034	0.528 ± 0.165	0.608 ± 0.093	0.776 ± 0.041	0.664 ± 0.089	0.348 ± 0.143	0.748 ± 0.048
F1	0.913 ± 0.021	0.906 ± 0.015	0.760 ± 0.023	0.819 ± 0.026	0.894 ± 0.014	0.849 ± 0.031	0.657 ± 0.069	0.866 ± 0.020
AUC	0.956 ± 0.019	0.956 ± 0.013	0.566 ± 0.070	0.863 ± 0.025	0.963 ± 0.008	0.923 ± 0.034	0.506 ± 0.012	0.917 ± 0.005
EER	0.076 ± 0.027	0.068 ± 0.020	0.232 ± 0.032	0.196 ± 0.029	0.072 ± 0.020	0.100 ± 0.018	0.340 ± 0.099	0.124 ± 0.015

	Kaggle_SetC							
	Untied							
	NF				RF			
	YNN	Model1	Model2	Model3	YNN	Model1	Model2	Model3
Accuracy	0.864 ± 0.029	0.844 ± 0.028	0.611 ± 0.054	0.742 ± 0.038	0.795 ± 0.041	0.732 ± 0.057	0.621 ± 0.044	0.777 ± 0.009
Sensitivity	0.983 ± 0.006	0.983 ± 0.011	0.928 ± 0.073	0.912 ± 0.054	0.976 ± 0.011	0.958 ± 0.055	0.851 ± 0.125	0.939 ± 0.018
Specificity	0.745 ± 0.064	0.705 ± 0.065	0.294 ± 0.181	0.573 ± 0.116	0.613 ± 0.091	0.506 ± 0.161	0.392 ± 0.182	0.615 ± 0.036
F1	0.879 ± 0.022	0.864 ± 0.019	0.706 ± 0.015	0.781 ± 0.020	0.828 ± 0.028	0.784 ± 0.030	0.690 ± 0.032	0.808 ± 0.004
AUC	0.951 ± 0.016	0.916 ± 0.034	0.500 ± 0.000	0.836 ± 0.020	0.918 ± 0.014	0.866 ± 0.045	0.500 ± 0.000	0.882 ± 0.004
EER	0.105 ± 0.013	0.131 ± 0.011	0.304 ± 0.040	0.236 ± 0.018	0.128 ± 0.016	0.154 ± 0.020	0.344 ± 0.042	0.189 ± 0.009
	Tied							
	NF				RF			
	YNN	Model1	Model2	Model3	YNN	Model1	Model2	Model3
Accuracy	0.883 ± 0.032	0.871 ± 0.017	0.690 ± 0.051	0.809 ± 0.019	0.879 ± 0.014	0.810 ± 0.051	0.589 ± 0.046	0.872 ± 0.010
Sensitivity	0.983 ± 0.009	0.980 ± 0.012	0.885 ± 0.084	0.958 ± 0.015	0.981 ± 0.002	0.983 ± 0.007	0.828 ± 0.170	0.950 ± 0.008
Specificity	0.784 ± 0.072	0.761 ± 0.045	0.495 ± 0.179	0.659 ± 0.052	0.778 ± 0.027	0.636 ± 0.106	0.351 ± 0.169	0.794 ± 0.027
F1	0.895 ± 0.027	0.884 ± 0.013	0.742 ± 0.018	0.834 ± 0.012	0.891 ± 0.011	0.839 ± 0.037	0.662 ± 0.066	0.881 ± 0.008
AUC	0.961 ± 0.012	0.948 ± 0.009	0.557 ± 0.061	0.891 ± 0.010	0.961 ± 0.007	0.933 ± 0.023	0.508 ± 0.011	0.940 ± 0.007
EER	0.081 ± 0.020	0.097 ± 0.013	0.246 ± 0.038	0.159 ± 0.017	0.082 ± 0.011	0.101 ± 0.014	0.347 ± 0.110	0.113 ± 0.007

of NPs, about which both types of YNNs agreed or disagreed, are shown in Fig. 20.

As shown in Fig. 21, based on two threshold independent metrics, *AUC* and *EER*, we can say that YNN was slightly better for recognizing pairs of color fundus photographs than the pairs of CRBVs. However, based on four threshold dependent metrics, *Accuracy*, *Sensitivity*, *Specificity* and *F1*, we cannot declare anyone winner.

F. HUMAN VS. YNN

To estimate the agreement among human volunteers, we applied majority voting to the decisions of the human

TABLE 14. The number of epochs, number of parameters, and training time of different models. [# Parm: number of trainable parameters, Epoch: Average number of epochs, and Time: Average training time.]

	Flip	YNN			Model1			Model2			Model3		
		# Parm.	Epoch	Time	# Parm.	Epoch	Time	# Parm.	Epoch	Time	# Parm.	Epoch	Time
Untied	NF	366,145	50	37.83435	165,713	50	34.73527	15,889,442	19	15.74792	4,874,017	24	59.48771
	RF		50	45.12250		50	42.03457		19	20.56558		25	61.35203
Tied	NF	294,065	50	35.76289	93,633	50	35.82669	15,817,362	21	28.37987	2,514,209	50	69.09019
	RF		47	44.31338		40	35.22006		48	31.31578		43	76.16682

TABLE 15. Performance of YNNs for different types of images. Notice that YNN's accuracy is slightly better for the red channel than for all other channels, including the RGB images. [Note that, for each type of image, the YNN was trained independently. Feature extractors of each YNN were untied. Concatenation was used for merging features. Right side retinal images were not horizontally flipped. CRBVs-based results reported here were taken from [2] for the purpose of comparison.]

	Kaggle_SetA					
	CRBV	RGB	Gray	Red	Green	Blue
Accuracy	0.892 ± 0.025	0.890 ± 0.024	0.882 ± 0.013	0.920 ± 0.017	0.886 ± 0.022	0.870 ± 0.032
Sensitivity	0.976 ± 0.008	0.984 ± 0.008	0.988 ± 0.010	0.996 ± 0.008	0.980 ± 0.022	0.984 ± 0.008
Specificity	0.808 ± 0.045	0.796 ± 0.046	0.776 ± 0.034	0.844 ± 0.034	0.792 ± 0.055	0.756 ± 0.065
F1	0.901 ± 0.021	0.900 ± 0.020	0.893 ± 0.010	0.926 ± 0.015	0.896 ± 0.018	0.884 ± 0.026
AUC	0.919 ± 0.015	0.958 ± 0.015	0.935 ± 0.026	0.973 ± 0.010	0.946 ± 0.010	0.934 ± 0.017
EER	0.132 ± 0.016	0.068 ± 0.027	0.100 ± 0.028	0.076 ± 0.023	0.088 ± 0.024	0.128 ± 0.027
	Kaggle_SetC					
	CRBV	RGB	Gray	Red	Green	Blue
Accuracy	0.882 ± 0.000	0.864 ± 0.029	0.844 ± 0.026	0.877 ± 0.009	0.855 ± 0.027	0.855 ± 0.024
Sensitivity	0.989 ± 0.004	0.983 ± 0.006	0.986 ± 0.011	0.982 ± 0.007	0.984 ± 0.015	0.975 ± 0.015
Specificity	0.774 ± 0.003	0.745 ± 0.064	0.703 ± 0.059	0.773 ± 0.022	0.726 ± 0.066	0.736 ± 0.061
F1	0.893 ± 0.000	0.879 ± 0.022	0.864 ± 0.018	0.889 ± 0.007	0.872 ± 0.020	0.872 ± 0.017
AUC	0.936 ± 0.005	0.951 ± 0.016	0.934 ± 0.011	0.950 ± 0.007	0.944 ± 0.006	0.943 ± 0.011
EER	0.126 ± 0.006	0.105 ± 0.013	0.130 ± 0.013	0.116 ± 0.013	0.106 ± 0.009	0.122 ± 0.022

volunteers (i.e., 23 volunteers for PPs, 13 volunteers for NPs) for the Kaggle_SetA.2 (see Table 9 for the scores of the 12 pairs shown in Fig. 11 as examples). We then compared how well the majority of human volunteers and different YNNs agreed. The result is shown in Table 10. Interestingly, there is a high agreement between manual and automatic verification and a very high agreement between different YNN architectures for automatic verification. Whenever there was a disagreement about PPs between the human majority vote and the YNNs, the YNNs were correct most of the time, while for NPs, human volunteers were correct most of the time, as shown in Table 11. For example, the human majority vote and the YNN, which merged extracted features by *concatenation*, disagreed for 11 out of the 50 PPs. For these 11 PPs, the human vote was not correct for any PPs, but the YNN was correct for all 11 PPs. On the contrary, they disagreed with nine out of the 50 NPs. For these nine NPs, the human vote was correct for eight NPs, and the YNN was correct for only one NP.

G. SIDE INDEPENDENT VERIFICATION

In the previously proposed automatic verification system (e.g., [19], [20], [85]), the same side retinal images were used to verify a subject. Our experiments in Section V-C revealed that two different side retinas could also be used for verifying a subject by an automatic verification system. Now the first question is, can a single system be used for both cases, i.e., can we verify a subject using a single system no matter whether image pairs are from the same side or two different sides of retinas? The second question is, how much does the performance reduce if the two images in a pair are from different sessions (e.g., the images are captured on different days)? We cannot answer these questions using the Kaggle data set used in the previous experiments because this data set has only one left and one right retinal image for each subject, captured at the same session. To answer these questions, we instead used the RODREP_SetA data set. This data set has two sessions per subject, and in each session, there is a left and a right retinal image. Accordingly, we can compare

- same-session, left-right verification,
- different-session, left-right verification,
- different-session, left-left verification,
- different-session, right-right verification.

Developing a general model which performs well on unseen data is a fundamental problem of deep learning like any other machine learning algorithm. If the unseen data have a different distribution from the training set, i.e., a *domain shift* exists, the problem of generalization becomes significantly difficult ([86]–[88]). Different approaches have been proposed for adapting a deep model trained in the source domain to the target domain. The necessity of domain adaptation arose while experimenting with the RODREP_SetA data set. The Kaggle_SetB, which was used for training YNNs had a domain mismatch with the RODREP_SetA data set. Seeing the foreground color, we assume that subjects of the RODREP_SetA data set belong to a single ethnic group. On the other hand, the Kaggle_SetB data set has subjects from

TABLE 16. Number of pairs (mean ± standard deviation) in a specific range of similarity scores. Most of the positive pairs (PPs) were in the range 0.9 – 1.0, while most of the negative pairs (NPs) were in the range 0.0 – 0.1 for all types of images. [Note that, for each type of image, the YNN was trained independently. Feature extractors of the YNN were untied. Concatenation was used for merging features. Right side retinal images were not horizontally flipped. CRBVs-based results reported here are taken from [2] only for the comparison purpose.]

Score Range	Kaggle_SetA.2					
	PPs					
	CRBV	RGB	Gray	Red	Green	Blue
0.0-0.1	0.400 ± 0.490	0.600 ± 0.490	0.200 ± 0.400	0.000 ± 0.000	0.200 ± 0.400	0.400 ± 0.490
0.1-0.2	0.200 ± 0.400	0.000 ± 0.000	0.400 ± 0.490	0.200 ± 0.400	0.000 ± 0.000	0.200 ± 0.400
0.2-0.3	0.200 ± 0.400	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.200 ± 0.400	0.200 ± 0.400
0.3-0.4	0.200 ± 0.400	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.200 ± 0.400	0.000 ± 0.000
0.4-0.5	0.200 ± 0.400	0.200 ± 0.400	0.000 ± 0.000	0.000 ± 0.000	0.400 ± 0.800	0.000 ± 0.000
0.5-0.6	0.400 ± 0.490	0.000 ± 0.000	0.200 ± 0.400	0.000 ± 0.000	0.200 ± 0.400	0.400 ± 0.490
0.6-0.7	0.200 ± 0.400	0.000 ± 0.000	0.200 ± 0.400	0.000 ± 0.000	0.600 ± 1.200	0.800 ± 1.600
0.7-0.8	0.600 ± 0.490	0.200 ± 0.400	0.000 ± 0.000	0.400 ± 0.490	0.400 ± 0.800	2.000 ± 3.098
0.8-0.9	2.000 ± 1.414	0.600 ± 0.800	2.600 ± 4.716	2.200 ± 0.748	1.600 ± 3.200	2.800 ± 3.763
0.9-1.0	45.600 ± 1.744	48.400 ± 0.800	46.400 ± 5.200	47.200 ± 0.980	46.200 ± 6.112	43.200 ± 8.795
Score Range	NPs					
	CRBV	RGB	Gray	Red	Green	Blue
0.0-0.1	36.600 ± 1.200	37.400 ± 2.417	35.400 ± 1.625	36.800 ± 1.470	35.800 ± 1.470	34.400 ± 1.497
0.1-0.2	1.600 ± 1.356	0.600 ± 0.490	1.800 ± 1.939	1.400 ± 0.800	1.400 ± 1.356	0.600 ± 0.800
0.2-0.3	0.600 ± 0.800	0.600 ± 0.800	0.400 ± 0.490	1.000 ± 0.632	0.800 ± 0.748	1.200 ± 0.980
0.3-0.4	1.000 ± 1.095	0.400 ± 0.490	1.000 ± 0.894	2.000 ± 0.632	1.000 ± 0.632	0.800 ± 1.166
0.4-0.5	0.600 ± 0.800	0.800 ± 0.748	0.200 ± 0.400	1.000 ± 0.894	0.600 ± 0.800	0.800 ± 0.748
0.5-0.6	1.000 ± 0.894	0.800 ± 1.166	0.400 ± 0.800	0.000 ± 0.000	0.400 ± 0.490	0.600 ± 0.490
0.6-0.7	0.800 ± 0.980	0.800 ± 1.166	0.600 ± 0.800	1.200 ± 0.748	1.400 ± 0.800	0.600 ± 1.200
0.7-0.8	0.600 ± 0.800	1.000 ± 0.632	1.400 ± 1.020	1.000 ± 0.632	0.600 ± 0.490	1.600 ± 0.800
0.8-0.9	0.600 ± 0.800	2.200 ± 1.600	1.200 ± 1.166	1.600 ± 0.800	1.000 ± 1.095	2.000 ± 1.095
0.9-1.0	6.600 ± 0.800	5.400 ± 3.137	7.600 ± 1.020	4.000 ± 1.265	7.000 ± 2.280	7.400 ± 2.871

Score Range	Kaggle_SetC					
	PPs					
	CRBV	RGB	Gray	Red	Green	Blue
0.0-0.1	4.000 ± 0.894	9.800 ± 4.167	7.400 ± 4.409	7.000 ± 3.742	7.600 ± 2.653	11.000 ± 4.858
0.1-0.2	3.000 ± 2.449	5.800 ± 2.561	3.200 ± 2.135	4.400 ± 3.007	4.800 ± 5.706	6.200 ± 7.467
0.2-0.3	3.200 ± 1.470	3.200 ± 1.720	4.400 ± 3.441	5.800 ± 2.315	3.200 ± 2.638	5.400 ± 4.030
0.3-0.4	3.200 ± 1.470	5.000 ± 2.449	4.400 ± 5.571	8.200 ± 2.561	4.200 ± 5.036	9.800 ± 6.013
0.4-0.5	5.000 ± 3.688	6.800 ± 3.919	5.800 ± 4.400	6.400 ± 3.499	8.200 ± 10.477	11.000 ± 8.414
0.5-0.6	8.000 ± 4.775	7.200 ± 3.544	8.600 ± 9.091	13.000 ± 5.477	10.800 ± 13.790	19.000 ± 14.792
0.6-0.7	10.400 ± 3.774	12.600 ± 5.571	16.400 ± 16.978	20.200 ± 6.969	21.000 ± 26.283	35.200 ± 29.158
0.7-0.8	31.400 ± 9.708	18.400 ± 11.056	32.800 ± 36.251	40.000 ± 16.697	38.600 ± 56.333	84.600 ± 79.636
0.8-0.9	83.400 ± 34.696	44.000 ± 30.906	103.200 ± 115.977	100.200 ± 33.187	89.400 ± 139.872	187.000 ± 139.529
0.9-1.0	1600.400 ± 55.413	1639.200 ± 55.765	1565.800 ± 195.921	1546.800 ± 64.384	1564.200 ± 261.504	1382.800 ± 286.521
Score Range	NPs					
	CRBV	RGB	Gray	Red	Green	Blue
0.0-0.1	1241.600 ± 34.857	1125.200 ± 113.500	1073.400 ± 95.569	1170.800 ± 34.799	1106.600 ± 96.257	1076.800 ± 56.947
0.1-0.2	41.200 ± 15.549	67.400 ± 12.706	56.200 ± 6.306	66.600 ± 5.122	56.400 ± 14.935	65.600 ± 17.783
0.2-0.3	27.800 ± 8.818	44.800 ± 14.372	39.200 ± 8.232	48.600 ± 2.871	39.600 ± 9.265	51.400 ± 17.001
0.3-0.4	21.200 ± 6.969	37.600 ± 8.089	31.400 ± 5.571	36.400 ± 8.065	37.200 ± 3.544	46.000 ± 9.612
0.4-0.5	24.800 ± 7.082	30.800 ± 6.493	31.600 ± 4.800	31.600 ± 4.271	31.800 ± 4.750	49.200 ± 13.182
0.5-0.6	25.800 ± 5.528	33.000 ± 11.009	34.000 ± 3.033	33.400 ± 7.761	36.400 ± 9.562	53.000 ± 5.831
0.6-0.7	25.400 ± 5.607	37.000 ± 5.967	41.200 ± 7.359	40.800 ± 9.704	41.800 ± 8.280	58.200 ± 8.158
0.7-0.8	36.400 ± 12.435	40.400 ± 9.478	54.200 ± 5.036	47.600 ± 9.972	44.200 ± 7.756	69.000 ± 10.431
0.8-0.9	51.000 ± 8.222	72.200 ± 11.652	88.200 ± 14.077	65.200 ± 2.561	70.800 ± 10.008	88.800 ± 24.095
0.9-1.0	256.800 ± 26.210	263.600 ± 103.905	302.600 ± 96.344	211.000 ± 16.334	287.200 ± 104.505	194.000 ± 82.712

multiple ethnic groups. However, there are not enough images in the Kaggle_SetB data set to capture representative features of the ethnic group of the RODREP_A data set in the YNN. The possibility of seeing many NPs from the same ethnic group during training is very low for the YNN trained on the Kaggle_SetB. Therefore, YNN cannot learn distinguishable features for the retinas in NPs from the same ethnic group and cannot make correct decisions. Fine-tuning parameters of YNN (i.e., retraining YNN trained by using the Kaggle_SetB data set for few epochs) by the Messidor2 data set was able to mitigate this problem to some extent. More investigation needs to be done in the future regarding this issue to improve the performance of YNN. Note that, since there is no publicly available metadata about the ethnic group/race for any of these data sets, we had to depend on the foreground color to guess about the similarity of the ethnicity among the subjects, which is sometimes misleading, as we discussed in Subsection V-A.

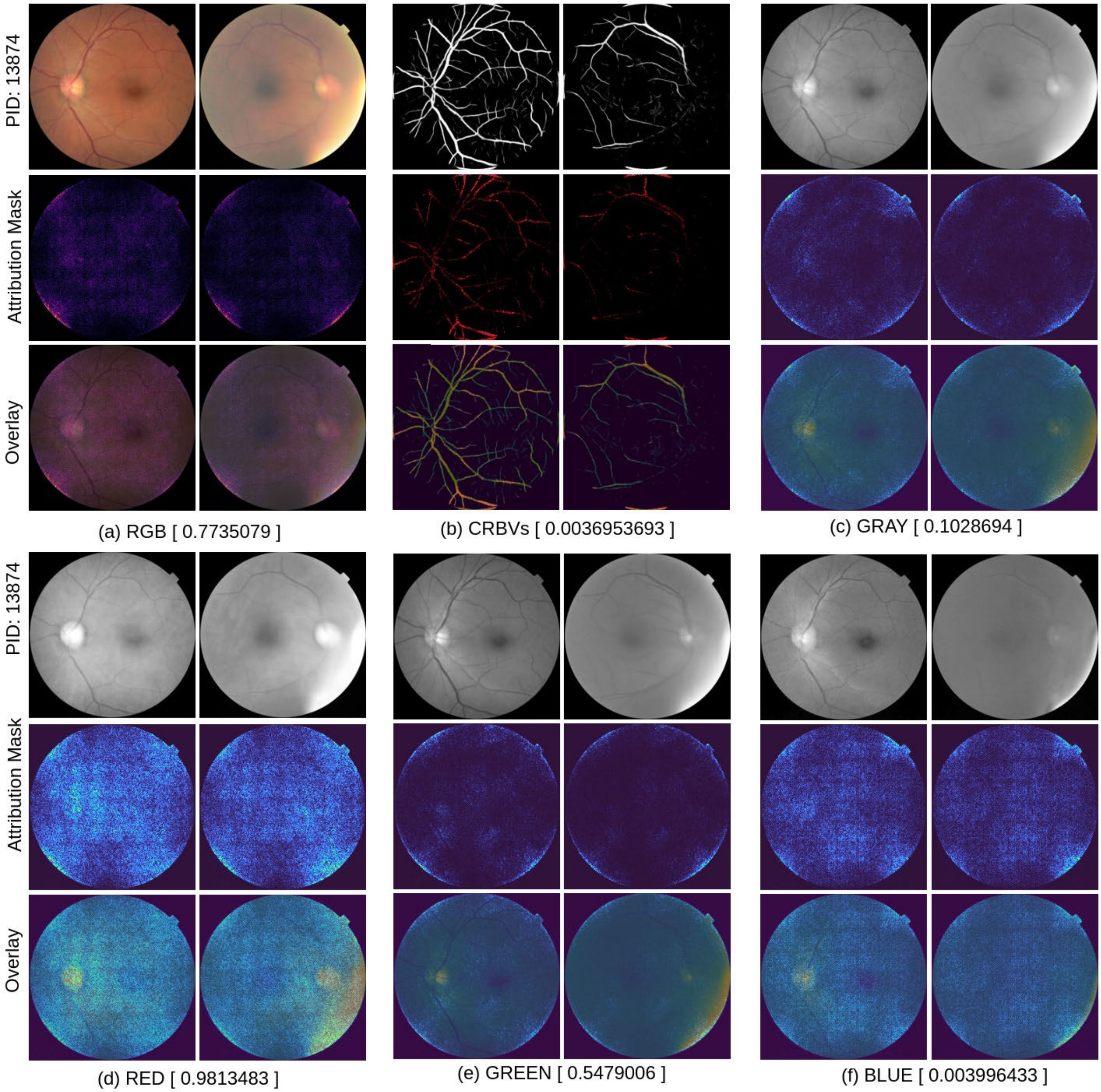


FIGURE 22. What a YNN looked at in different types of retinal images for verifying a positive pair. An attribution method based on integrated gradients [84] was used to highlight the contribution of important parts of the pair in the similarity score. Notice that the YNNs focused on different areas of the pair in different types of retinal images. [Note that, for each image type, a YNN was trained independently. Feature extractors of each YNN were untied. Concatenation was used for merging features. Right side retinal images were not horizontally flipped. Each subtitle indicates Image_Type [Similarity_Score]. CRBVs-based results reported here were taken from [2] only for comparison purposes. Source of image: Kaggle_SetA.]

We used the YNN trained by using left-left, left-right, and right-right pairs of color fundus photographs. From the results given in Table 12, our observations are that:

- domain adaptation by fine-tuning parameters helped to mitigate the gap between the domain of the RODREP_SetA data set and the Kaggle_SetB data set and improved the performance of YNN.
- domain adaptation degraded performance of YNN for the Kaggle_SetA.
- one single YNN is almost equally good for handling pairs from the same side, both sides, and different sessions. That means it is clear that a single retina-based system can be used for side-independent verification. However, verification accuracy (81%-90%) is not good

enough for practical use. More investigation is necessary regarding this issue.

As far as we know, side-independent verification has not been studied in the retina-based biometric field before. Therefore, our findings will advance retina-based biometric systems in the future.

H. VARIETIES OF YNN

To figure out the appropriateness of the YNN architecture, we explored three varieties of YNN: Model1, Model2, and Model3. As shown in Table 13, these varieties did not perform better than the YNN. Our other observations are:

- Model1 performed slightly worse than the YNN. However, it had 54.74% and 68.16% less trainable parameters for untied and tied cases, respectively, comparing to the YNN (see Table 14). In our current hardware support, we did not achieve a noticeable advantage by reducing parameters. However, reducing trainable parameters of the YNN would reduce training time if we had less powerful hardware support. It indicates that there could be other design of the YNN which could perform better and faster than the current design of the YNN. However, we will have numerous design choices if we alter the number of convolution, pooling, and fully connected layers, change the ordering of layers or the hyper-parameters for each type of layer (e.g., the kernel size, stride, and the number of kernels for a convolution layer). Therefore, a manual search for a suitable architecture of the YNN is infeasible. An automatic algorithm such as MetaQNN [89] can be beneficial for finding an appropriate YNN architecture. We kept this issue as future work.
- Model2, i.e., cosine distance model, performed worse than the other models. However, they still performed better than random guesses.
- Transfer learning by the pre-trained VGG16 model was not very beneficial.

I. SYMMETRY IN DIFFERENT COLOR CHANNELS

Since foreground color played an important role for both human volunteers and YNNs, we checked YNN's performance separately in each type of images (i.e., RGB colored image, grayscale image, red channel, green channel, and blue channel image) to figure out which type of image played the most important part. As shown in Table 15, YNN's accuracy is slightly better in the *red* channel than all other channels, even than RGB colored images for both Kaggle_SetA.2 and Kaggle_SetC.

As shown in Table 16, with few exceptions, we saw a similar pattern in the similarity scores for all kinds of images. For all models, most of the similarity/symmetry scores of positive pairs (PPs) are in the range $0.9 - 1.0$, whereas most of the scores of negative pairs (NPs) are in the range $0.0 - 0.1$. This trend reveals that all YNNs were quite confident about their decisions. They were able to draw a clear, distinguishable line between the PPs and NPs. Using an integrated gradient-based attribution approach, we investigated which area of the retina played important parts. As shown in Fig. 22, different YNNs focused on different areas of a pair.

VI. CONCLUSION

The target of our study was to confirm within-subject bilateral symmetry in color fundus photographs. To this end, we investigated if it is possible to decide whether a pair of left and right retinas belong to a single subject or two different subjects. Twenty-three human volunteers participated in a manual verification experiment. We designed a deep neural network for automatic verification. We named it YNN. Both humans and the YNN could verify subjects based on two side retinas. For human volunteers, the accuracy of the verification task was typically in the range 65%-89%, whereas for 16 varieties of YNN, the accuracy was in the range 74%-90%. In both cases, the accuracy was well above the result of random guesses. Therefore, we concluded that there is a high degree of bilateral symmetry in color fundus photographs of a subject (e.g., patient). We also found that:

- both human volunteers and YNNs found symmetry in foreground color, central retinal blood vessels (CRBVs), area of the optic disc, and choroidal blood vessels,
- randomness, caused by the parameter initialization and dropout layer, created a noticeable fluctuation in the performance of the YNN,
- there was a high agreement between human volunteers and YNNs and between different YNNs, which reveals the fact that the majority of retina pairs possess strong bilateral symmetry,
- poor quality images made both human volunteers and YNNs confused; therefore, false verification occurred.
- by minimizing domain mismatch between the training set and target set, we can improve the YNN's performance,
- one single YNN can verify pairs from the same side and two different sides. It had slightly better accuracy for the *same-side* and *same-session* verification than the *different-side* and *different session* verification, respectively,
- YNN's accuracy was slightly better for the red channel than all other channels, even than RGB images, and
- the RGB based system was better than CRBV based system for the calibration insensitive metrics.

Even though our investigation found much exciting information about the bilateral symmetry in color fundus photographs, this topic needs to be investigated from many other directions so that we can develop a side-independent retina-based biometric system for practical use. For that, we need to prepare a big data set with good quality color fundus photographs captured at multiple sessions from both sides retinas of many subjects belonging to different ethnic groups. We also need a better approach to detect poor quality retinal images, tackle the domain mismatch issue, and decide the YNN's architecture.

REFERENCES

[1] S. Biswas, J. Rohdin, and M. Drahansky, "Interretinal symmetry in color fundus photographs," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1980–1983.

[2] S. Biswas, J. Rohdin, A. Kavetskyi, G. Saraiva, A. Biswas, and M. Drahansky, "Investigation of bilateral similarity in central retinal blood vessels," *IEEE Access*, vol. 9, pp. 63012–63028, 2021.

[3] A. Sopharak, U. Bunyarit, and S. Barman, "Automatic exudate detection from non-dilated diabetic retinopathy retinal images using fuzzy C-means clustering," *Sensors*, vol. 9, no. 3, pp. 2148–2161, 2009.

[4] M. U. Akram, S. Khalid, A. Tariq, and M. Y. Javed, "Detection of neovascularization in retinal images using multivariate m-Mediods based classifier," *Comput. Med. Imag. Graph.*, vol. 37, nos. 5–6, pp. 346–357, Jul. 2013.

[5] T. Jaya, J. Dheeba, and N. A. Singh, "Detection of hard exudates in colour fundus images using fuzzy support vector machine-based expert system," *J. Digit. Imag.*, vol. 28, no. 6, pp. 761–768, Dec. 2015.

[6] V. Gulshan, L. Peng, and M. Coram, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[7] C. Lam, C. Yu, L. Huang, and D. Rubin, "Retinal lesion detection with deep learning using image patches," *Invest. Ophthalmol. Vis. Sci.*, vol. 59, no. 1, pp. 590–596, 2018.

[8] Z. Wang and J. Yang, "Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 514–520.

[9] J. Sahlsten, J. Jaskari, J. Kivinen, L. Turunen, E. Jaanio, K. Hietala, and K. Kaski, "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Sci. Rep.*, vol. 9, no. 1, p. 10750, Dec. 2019.

[10] G. Quellec, M. Lamard, P.-H. Conze, P. Massin, and B. Cochener, "Automatic detection of rare pathologies in fundus photographs using few-shot learning," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101660.

[11] R. Kinouchi, S. Ishiko, K. Hanada, H. Hayashi, D. Mikami, and A. Yoshida, "Identification of risk factors for retinal vascular events in a population-based cross-sectional study in Rumoi, Japan," *Sci. Rep.*, vol. 11, no. 1, p. 6340, Dec. 2021.

[12] Z.-W. Xu, X.-X. Guo, X.-Y. Hu, and X. Cheng, "The blood vessel recognition of ocular fundus," in *Proc. Int. Conf. Mach. Learn. Cybern.*, 2005, pp. 4493–4498.

[13] C. Mariño, M. G. Penedo, M. Penas, M. J. Carreira, and F. Gonzalez, "Personal authentication using digital retinal images," *Pattern Anal. Appl.*, vol. 9, no. 1, pp. 21–33, May 2006.

[14] H. Farzin, H. Abrishami-Moghaddam, and M.-S. Moin, "A novel retinal identification system," *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, Dec. 2008, Art. no. 280635.

[15] A. Arakala, J. S. Culpepper, J. Jeffers, A. Turpin, S. Boztaş, K. J. Horadam, and A. M. McKendrick, "Entropy of the retina template," in *Advances in Biometrics*, M. Tistarelli and M. S. Nixon, Eds. Berlin, Germany: Springer, 2009, pp. 1250–1259.

[16] M. Ortega, M. G. Penedo, J. Rouco, N. Barreira, and M. J. Carreira, "Personal verification based on extraction and characterisation of retinal feature points," *J. Vis. Lang. Comput.*, vol. 20, no. 2, pp. 80–90, Apr. 2009.

[17] M. Agopov, "Retinal identification," in *IntechOpen Biometrics*, J. Yang, Ed. IntechOpen, 2011, ch. 5.

[18] W. Barkhoda, F. Akhlaqian, M. D. Amiri, and M. S. Nouroozzadeh, "Retina identification based on the pattern of blood vessels using fuzzy logic," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, p. 113, Dec. 2011.

[19] S. M. Lajevardi, A. Arakala, S. A. Davis, and K. J. Horadam, "Retina verification system based on biometric graph matching," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3625–3635, Sep. 2013.

[20] M. Frucci, D. Riccio, G. S. di Baja, and L. Serino, "Using direction and score information for retina based person verification," *Expert Syst. Appl.*, vol. 94, pp. 1–10, Mar. 2018.

[21] U. Suripon, "Retina recognition using compression-based joint transform correlator," *Opt. Eng.*, vol. 50, no. 9, Sep. 2011, Art. no. 098201.

[22] M. Sabaghi, S. R. Hadianamrei, M. Fattahi, M. R. Kouchaki, and A. Zahedi, "Retinal identification system based on the combination of Fourier and wavelet transform," *J. Signal Inf. Process.*, vol. 3, no. 1, pp. 35–38, 2012.

[23] H. Kolb, "Simple anatomy of the retina," in *The Organization of the Retina and Visual System*, H. K. H, E. Fernandez, and R. Nelson, Eds. Salt Lake City, UT, USA: Webvision, 1995.

[24] S. C. Nemeth, C. Shea, M. DiSclafani, and M. Schluter, "The posterior segment," in *Ocular Anatomy and Physiology*, 2nd ed. Thorofare, NJ, USA: Slack Incorporated, 2008, ch. 9, pp. 88–99.

[25] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 169–208, 2010.

[26] H. Leung, J. J. Wang, E. Rochtchina, A. G. Tan, T. Y. Wong, L. D. Hubbard, R. Klein, and P. Mitchell, "Computer-assisted retinal vessel measurement in an older population: Correlation between right and left eyes," *Clin. Experim. Ophthalmol.*, vol. 31, no. 4, pp. 326–330, Aug. 2003.

[27] D. L. Budnez, "Symmetry between the right and left eyes of the normal retinal nerve fiber layer measured with optical coherence tomography (an AOS thesis)," *Trans. Amer. Ophthalmol. Soc.*, vol. 106, pp. 252–275, Dec. 2008.

[28] H. Li, P. R. Healey, Y. M. Tariq, E. Teber, and P. Mitchell, "Symmetry of optic nerve head parameters measured by the Heidelberg retina tomograph 3 in healthy eyes: The Blue Mountains Eye study," *Amer. J. Ophthalmol.*, vol. 155, no. 3, pp. 518–523, 2013.

[29] M. Yang, W. Wang, Q. Xu, S. Tan, and S. Wei, "Interocular symmetry of the peripapillary choroidal thickness and retinal nerve fibre layer thickness in healthy adults with isometropia," *BMC Ophthalmol.*, vol. 16, no. 1, p. 182, Dec. 2016.

[30] M. Zhou, B. Lu, J. Zhao, Q. Wang, P. Zhang, and X. Sun, "Interocular symmetry of macular ganglion cell complex thickness in young Chinese subjects," *PLoS ONE*, vol. 11, no. 7, Jul. 2016, Art. no. e0159583.

[31] G. Liu, K. Keyal, and F. Wang, "Interocular symmetry of vascular density and association with central macular thickness of healthy adults by optical coherence tomography angiography," *Sci. Rep.*, vol. 7, no. 1, p. 16297, Dec. 2017.

[32] R. R. Mastey, M. Gaffney, K. M. Litts, C. S. Langlo, E. J. Patterson, M. R. Strampe, A. Kalitzeos, M. Michaelides, and J. Carroll, "Assessing the interocular symmetry of foveal outer nuclear layer thickness in achromatopsia," *Transl. Vis. Sci. Technol.*, vol. 8, no. 5, p. 21, Sep. 2019.

[33] I. P. Conner, J. S. Schuman, and D. L. Epstein, *Examination of the Optic Nerve*, 5th ed. Thorofare, NJ, USA: Slack Incorporated, 2013, ch. 8, pp. 81–94.

[34] D. J. Couper, R. Klein, L. D. Hubbard, T. Y. Wong, P. D. Sorlie, L. S. Cooper, R. J. Brothers, and F. J. Nieto, "Reliability of retinal photography in the assessment of retinal microvascular characteristics: The atherosclerosis risk in communities study," *Amer. J. Ophthalmol.*, vol. 133, no. 1, pp. 78–88, Jan. 2002.

[35] T. Y. Wong, R. Klein, F. J. Nieto, B. E. K. Klein, A. R. Sharrett, S. M. Meuer, L. D. Hubbard, and J. M. Tielsch, "Retinal microvascular abnormalities and 10-year cardiovascular mortality: A population-based case-control study," *Ophthalmology*, vol. 110, no. 5, pp. 933–940, 2003.

[36] T. Y. Wong, M. D. Knudtson, R. Klein, B. E. K. Klein, S. M. Meuer, and L. D. Hubbard, "Computer-assisted measurement of retinal vessel diameters in the Beaver Dam Eye Study: Methodology, correlation between eyes, and effect of refractive errors," *Ophthalmology*, vol. 111, no. 6, pp. 1183–1190, 2004.

[37] A. M. Taylor, T. J. MacGillivray, R. D. Henderson, L. Ilzina, B. Dhillon, J. M. Starr, and I. J. Deary, "Retinal vascular fractal dimension, childhood IQ, and cognitive ability in old age: The Lothian birth cohort study 1936," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0121119.

[38] A. Caramoy, K. M. Droege, B. Kirchhof, and S. Fauser, "Retinal layers measurements in healthy eyes and in eyes receiving silicone oil-based endotamponade," *Acta Ophthalmol.*, vol. 92, no. 4, pp. e292–e297, Jun. 2014.

[39] K. M. Litts, M. Georgiou, C. S. Langlo, E. J. Patterson, R. R. Mastey, A. Kalitzeos, R. E. Linderman, B. L. Lam, G. A. Fishman, M. E. Pennesi, C. N. Kay, W. W. Hauswirth, M. Michaelides, and J. Carroll, "Interocular symmetry of foveal cone topography in congenital achromatopsia," *Current Eye Res.*, vol. 45, no. 10, pp. 1257–1264, Oct. 2020.

[40] J. A. Cava, M. T. Allphin, R. R. Mastey, M. Gaffney, R. E. Linderman, R. F. Cooper, and J. Carroll, "Assessing interocular symmetry of the foveal cone mosaic," *Investigative Opthalmol. Vis. Sci.*, vol. 61, no. 14, p. 23, Dec. 2020.

[41] J. R. Cameron, R. D. Megaw, A. J. Tatham, S. McGrory, T. J. MacGillivray, F. N. Doubal, J. M. Wardlaw, E. Trucco, S. Chandran, and B. Dhillon, "Lateral thinking–interocular symmetry and asymmetry in neurovascular patterning, in health and disease," *Prog. Retinal Eye Res.*, vol. 59, pp. 131–157, Jul. 2017.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[43] P. Baldi and Y. Chauvin, "Neural networks for fingerprint recognition," *Neural Comput.*, vol. 5, no. 3, pp. 402–418, May 1993.

[44] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 737–744, 1993.

[45] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.

[46] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015.

[47] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 65, pp. 1–32, Apr. 2016.

[48] J. Cuadros and G. Bresnick, "EyePACS: An adaptable telemedicine system for diabetic retinopathy screening," *J. Diabetes Sci. Technol.*, vol. 3, no. 3, pp. 509–516, May 2009.

[49] W. Abdulla and R. J. Chalakkal. (2018). *University of Auckland Diabetic Retinopathy (UoA-DR) Database-END USER LICENCE AGREEMENT*. [Online]. Available: https://auckland.figshare.com/articles/UoA-DR_Database_Info/5985208/5

[50] M. Niemeijer, B. Van Ginneken, M. J. Cree, A. Mizutani, G. Quellec, C. I. Sánchez, B. Zhang, R. Hornero, M. Lamard, C. Muramatsu, and X. Wu, "Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 185–195, Jan. 2010.

[51] H. Fu, F. Li, J. I. Orlando, H. Bogunovic, X. Sun, J. Liao, Y. Xu, S. Zhang, and X. Zhang, "PALM: Pathologic myopia challenge," in *Proc. IEEE Dataport*, 2019, doi: 10.21227/55pk-8z03.

[52] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, M. Lamard, D. C. Moga, G. Quellec, and M. Niemeijer, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA Ophthalmol.*, vol. 131, no. 3, pp. 351–357, 2013, doi: 10.1001/jamaophthalmol.2013.1743.

[53] A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan, "Retinal fundus images for glaucoma analysis: The RIGA dataset," in *Proc. SPIE, Med. Imag., Imag. Informat. Healthcare, Res., Appl.*, vol. 10579, 2018, pp. 55–62, doi: 10.1117/12.2293584.

[54] E. Decencière, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis, "TeleOphta: Machine learning and image processing methods for teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, Apr. 2013, doi: 10.1016/j.irbm.2013.01.010.

[55] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, "DIARETDB1 diabetic retinopathy database and evaluation protocol," in *Proc. Brit. Mach. Vis. Conf.*, 2007, pp. 15.1–15.10, doi: 10.5244/C.21.15.

[56] C. G. Owen, A. R. Rudnicka, R. Mullen, S. A. Barman, D. Monekosso, P. H. Whincup, J. Ng, and C. Paterson, "Measuring retinal vessel tortuosity in 10-year-old children: Validation of the computer-assisted image analysis of the retina (CAIAR) program," *Investigative Ophthalmol. Vis. Sci.*, vol. 50, no. 5, pp. 2004–2010, 2009. [Online]. Available: https://staffnet.kingston.ac.uk/ ku15565/CHASE_DB1/assets/CHASEDB1.zip

[57] E. Carmona, M. Rincón, J. García-Feijoó, and J. M. Martínez-de-la-Casa, "Identification of the optic nerve head with genetic algorithms," *Artif. Intell. Med.*, vol. 43, no. 3, pp. 59–243, 2008.

[58] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. V. Ginneken, "Ridge based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004. [Online]. Available: https://www.isi.uu.nl/Research/Databases/DRIVE/download.php

[59] A. Budai, R. Bock, A. Maier, J. Hornegger, and G. Michelson, "Robust vessel segmentation in fundus images," *Int. J. Biomed. Imag.*, vol. 2013, pp. 1–11, Dec. 2013. [Online]. Available: https://www5.cs.fau.de/fileadmin/research/datasets/fundus-images/all.zip

[60] R. Pires, H. F. Jelinek, J. Wainer, E. Valle, and A. Rocha, "Advancing bag-of-visual-words representations for lesion classification in retinal images," *PLoS ONE*, vol. 9, no. 6, p. 96814, 2014. [Online]. Available: https://www.isi.uu.nl/Research/Databases/DRIVE/download.php

[61] T. Köhler, A. Budai, M. F. Kraus, J. Odstrčilik, G. Michelson, and J. Hornegger, "Automatic no-reference quality assessment for retinal fundus images using vessel segmentation," in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2013, pp. 95–100. [Online]. Available: https://www5.cs.fau.de/fileadmin/research/datasets/fundus-images/allQuality.zip

[62] C. Hernandez-Matas, X. Zabulis, A. Triantafyllou, P. Anyfanti, S. Douma, and A. A. Argyros, "FIRE: Fundus image registration dataset," *J. Model. Ophthalmol.*, vol. 1, no. 4, pp. 16–28, 2017. [Online]. Available: https://projects.ics.forth.gr/cvrl/fire/

[63] K. M. Adal, P. G. van Etten, J. P. Martinez, L. J. van Vliet, and K. A. Vermeer, "Accuracy assessment of intra- and intervisit fundus image registration for diabetic retinopathy screening," *Investigative Ophthalmol. Vis. Sci.*, vol. 56, no. 3, pp. 1805–1812, Mar. 2015.

[64] K. W. Bowyer, S. Lagree, and S. Fenker, "Human versus biometric detection of texture similarity in left and right irises," in *Proc. 44th Annu. IEEE Int. Carnahan Conf. Secur. Technol.*, Oct. 2010, pp. 323–329.

[65] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[66] M. Niemeijer, M. Abramoff, and B. Vanginneken, "Image structure clustering for image quality verification of color retina images in diabetic retinopathy screening," *Med. Image Anal.*, vol. 10, no. 6, pp. 888–898, Dec. 2006.

[67] J. M. P. Dias, C. M. Oliveira, and L. A. da Silva Cruz, "Retinal image quality assessment using generic image quality indicators," *Inf. Fusion*, vol. 19, pp. 73–90, Sep. 2014.

[68] U. Sevik, C. Köse, T. Berber, and H. Erdöl, "Identification of suitable fundus images using automated quality assessment methods," *J. Biomed. Opt.*, vol. 19, no. 4, Apr. 2014, Art. no. 046006.

[69] D. Veiga, C. Pereira, M. Ferreira, L. Gonçalves, and J. Monteiro, "Quality evaluation of digital fundus images through combined measures," *J. Med. Imag.*, vol. 1, no. 1, Apr. 2014, Art. no. 014001.

[70] L. Abdel-Hamid, A. El-Rafei, S. El-Ramly, G. Michelson, and J. Hornegger, "Retinal image quality assessment based on image clarity and content," *SPIE J. Biomed. Opt.*, vol. 21, no. 9, 2016, Art. no. 096007.

[71] P. Costa, A. Campilho, B. Hooi, A. Smailagic, K. Kitani, S. Liu, C. Faloutsos, and A. Galdran, "EyeQual: Accurate, explainable, retinal image quality assessment," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 323–330.

[72] F. Shao, Y. Yang, Q. Jiang, G. Jiang, and Y.-S. Ho, "Automated quality assessment of fundus images via analysis of illumination, naturalness and structure," *IEEE Access*, vol. 6, pp. 869–878, 2018.

[73] G. T. Zago, R. V. Andreão, B. Dorizzi, and E. O. T. Salles, "Retinal image quality assessment using deep learning," *Comput. Biol. Med.*, vol. 103, pp. 64–70, Dec. 2018.

[74] M. A. Zapata, D. Royo-Fibla, O. Font, J. I. Vela, I. Marcantonio, E. U. Moya-Sanchez, A. Sanchez-Perez, D. Garcia-Gasulla, U. Cortés, E. Ayguade, and J. Labarta, "Artificial intelligence to identify retinal fundus images, quality validation, laterality evaluation, macular degeneration, and suspected glaucoma," *Clin. Ophthalmol.*, vol. 14, pp. 419–429, Feb. 2020.

[75] X. Li, W. L. Wong, C. Y.-L. Cheung, C.-Y. Cheng, M. K. Ikram, J. Li, K. S. Chia, and T. Y. Wong, "Racial differences in retinal vessel geometric characteristics: A multiethnic study in healthy Asians," *ARVO Investigative Ophthalmol. Vis. Sci.*, vol. 49, pp. 3650–3656, 2013.

[76] T. Y. Wong, F. M. A. Islam, R. Klein, B. E. K. Klein, M. FrancesCotch, C. Castro, A. R. Sharrett, and E. Shahar, "Retinal vascular caliber, cardiovascular risk factors, and inflammation: The multi-ethnic study of atherosclerosis (MESA)," *ARVO Investigative Ophthalmol. Vis. Sci.*, vol. 47, no. 6, pp. 2341–2350, 2006.

[77] E. Rochtchina, J. J. Wang, B. Taylor, T. Y. Wong, and P. Mitchell, "Ethnic variability in retinal vessel caliber: A potential source of measurement error from ocular pigmentation?—The Sydney Childhood Eye Study," *ARVO Investigative Ophthalmol. Vis. Sci.*, vol. 49, pp. 1362–1366, Apr. 2008.

[78] P. J. Kelty, J. F. Payne, R. H. Trivedi, J. Kelty, E. M. Bowie, and B. Burger, "Macular thickness assessment in healthy eyes based on ethnicity using Stratus OCT optical coherence tomography," *ARVO Investigative Ophthalmol. Vis. Sci.*, vol. 49, no. 6, pp. 2668–2672, 2008.

[79] A. H. Kashani, I. E. Zimmer-Galler, S. M. Shah, L. Dustin, D. V. Do, D. Eliott, J. A. Haller, and Q. D. Nguyen, "Retinal thickness analysis by race, gender, and age using stratus OCT," *Amer. J. Ophthalmol.*, vol. 149, no. 3, pp. 496–502, Mar. 2010.

[80] U. E. K. Wolf-Schnurrbusch, N. Röösli, E. Weyermann, M. R. Heldner, K. Höhne, and S. Wolf, "Ethnic differences in macular pigment density and distribution," *ARVO Investigative Ophthalmol. Vis. Sci.*, vol. 48, pp. 3783–3787, Aug. 2007.

[81] C. Samarawickrama, J. J. Wang, S. C. Huynh, A. Pai, G. Burlutsky, K. A. Rose, and P. Mitchell, "Ethnic differences in optic nerve head and retinal nerve fibre layer thickness parameters in children," *Brit. J. Ophthalmol.*, vol. 94, no. 7, pp. 871–876, Jul. 2010.

[82] L. Giancardo, F. Meriaudeau, T. P. Karnowski, E. Chaum, and K. Tobin, "Quality assessment of retinal fundus images using elliptical local vessel density," in *New Developments in Biomedical Engineering*, D. Campolo, Ed. Rijeka, Croatia: IntechOpen, 2010, ch. 11.

[83] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.

[84] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[85] M. Ortega, M. G. Penedo, J. Rouco, N. Barreira, and M. J. Carreira, "Retinal verification using a feature points-based biometric pattern," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, Dec. 2009, Art. no. 235746.

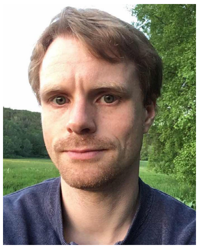
[86] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.

[87] M. Kim, P. Sahu, B. Gholami, and V. Pavlovic, "Unsupervised visual domain adaptation: A deep max-margin Gaussian process approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4380–4390.

[88] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "A closer look at domain shift for deep learning in histopathology," in *Proc. MICCAI Workshop Comput. Pathol. (COMPAY)*, 2019, pp. 1–8.

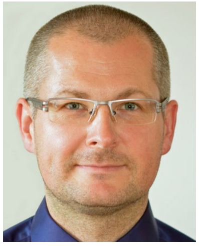
[89] B. Baker, N. N. O. Gupta, and R. Raska, "Designing neural network architectures using reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–18.

SANGEETA BISWAS received the master's degree in computer science and engineering from the University of Rajshahi (RU), Bangladesh (BD), in 2007, and the master's degree in computer science and the Ph.D. degree from Tokyo Institute of Technology (TokyoTech), Japan, in 2011 and 2016, respectively. From 2018 to 2020, she worked as a Postdoctoral Researcher with the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Czech Republic (CZ). She is currently working as an Assistant Professor with the Faculty of Engineering, RU. She is also collaborating with the STRaDe Team of the FIT, BUT, as a Remote Researcher. Her current research interests include biometrics, medical image processing, and crowd counting.



JOHAN ROHDIN received the M.Sc. degree in engineering physics and mathematics from Chalmers University of Technology, in 2008, and the Ph.D. degree in computer science from Tokyo Institute of Technology, in 2015. In 2016, he joined the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Czech Republic, as a Postdoctoral Researcher, sponsored by the South Moravian Program for Distinguished Researchers (SoMoPro) for working on neural network approaches to speaker recognition. He currently works at the FIT, BUT, as an Assistant Professor. His primary research interests include statistical pattern recognition and speech processing.

ANGKAN BISWAS received the M.B.S. degree in accounting from the National University at Bangladesh, Bangladesh, in 2011. Since January 2020, he has been collaborating with the STRaDe Team of the Faculty of Information Technology (FIT), Brno University of Technology (BUT), Czech Republic, as a Remote Researcher. His research interests include image processing and artificial intelligence.



MARTIN DRAHANSKY (Senior Member, IEEE) graduated from the Faculty of Electrotechnics and Computer Science, Brno University of Technology (BUT), Czech Republic, and the Faculty of Electrotechnics, FernUniversität in Hagen, Germany, in 2001. He received the Ph.D. degree from the Faculty of Information Technology (FIT), BUT, in 2005. In 2009, he became an Associate Professor with the FIT, BUT. Since 2017, he has been working as a Full Professor with the FIT, BUT. His research interests include biometrics, security and cryptography, artificial intelligence, and sensory systems.