Exploring Bias in Sclera Segmentation Models: A Group Evaluation Approach

Matej Vitek, Abhijit Das, Diego Rafael Lucio, Luiz Antonio Zanlorensi Jr., David Menotti, Jalil Nourmohammadi Khiarak, Mohsen Akbari Shahpar, Meysam Asgari-Chenaghlu, Farhang Jaryani, Juan E. Tapia, *Member, IEEE*, Andres Valenzuela, Caiyong Wang, Yunlong Wang, Zhaofeng He, Zhenan Sun, *Senior Member, IEEE*, Fadi Boutros, Naser Damer, *Member, IEEE*, Jonas Henry Grebe, Arjan Kuijper, Kiran Raja, *Senior Member, IEEE*, Gourav Gupta, Georgios Zampoukis, Lazaros Tsochatzidis, Ioannis Pratikakis, *Senior Member, IEEE*, S. V. Aruna Kumar, B. S. Harish, Umapada Pal, *Senior Member, IEEE*, Peter Peer, *Senior Member, IEEE*, and Vitomir Štruc, *Senior Member, IEEE*

***Abstract*—Bias and fairness of biometric algorithms have been key topics of research in recent years, mainly due to the societal, legal and ethical implications of potentially unfair decisions made by automated decision-making models. A considerable amount of work has been done on this topic across different biometric modalities, aiming at better understanding the main sources of algorithmic bias or devising mitigation measures. In this work, we contribute to these efforts and present the first study investigating bias and fairness of sclera segmentation models. Although sclera segmentation techniques represent a key component of sclera-based biometric systems with a considerable impact on the overall recognition performance, the presence of different types of biases in sclera segmentation methods is still underexplored. To address this limitation, we describe the results of a group evaluation effort (involving seven research groups), organized to explore the performance of recent sclera segmentation models within a common experimental framework and study performance differences (and bias), originating from various demographic as well as environmental factors. Using five diverse datasets, we analyze seven independently developed sclera segmentation models in different experimental configurations. The results of our experiments suggest that there are significant differences in the overall segmentation performance across the seven models and that among the considered factors, ethnicity appears to be the biggest cause of bias. Additionally, we observe that training with representative and balanced data does not necessarily lead to less biased results. Finally, we find that in general there appears to be a negative correlation between the amount of bias observed (due to eye color, ethnicity and acquisition device) and the overall segmentation performance, suggesting that advances in the field of semantic segmentation may also help with mitigating bias.**

***Index Terms*—Biometrics, sclera segmentation, ocular biometrics, bias, fairness.**

I. Introduction

OCULAR biometrics represents a branch of biometric recognition technology that exploits various characteristics of the eye for automatic identity inference [1]. Recognition techniques based on ocular traits have been successfully applied for access control applications, user-friendly verification schemes on mobile devices, as well as large scale identity-management programs, e.g., Aadhaar [2]. Research on ocular biometrics has long been focused on iris recognition technology, but more recently also expanded into other (visible) ocular modalities, such as the periocular region [3] and the vasculature of the sclera [4]. The sclera in particular has seen considerable interest, mainly due to its appealing characteristics, i.e.: (i) unlike iris recognition, sclera recognition performs best in the visible spectrum [5], and, hence, does not require any specialized acquisition hardware; and (ii) the vasculature of the sclera is considered to be highly discriminative and stable over time, (iii) while the presence of contact lenses can (purposely/inadvertently) degrade the performance of recognition techniques based on the iris or the periocular region, it has only a limited effect on sclera recognition models [5], [6].

A typical sclera recognition procedure consists of four main steps: sclera segmentation, vessel enhancement, feature extraction and matching. Each of these steps is critical for the overall *accuracy and trustworthiness* of the recognition procedure and has to ensure consistent performance across diverse data characteristics, e.g., gender, ethnicity, acquisition device, gaze direction. The recent interest in sclera biometrics has led to considerable advances with all four steps and among others resulted in powerful segmentation models [7], [8], [9], novel recognition techniques [4], [5], [10], but also multi-biometric systems with impressive performance characteristics [11], [12]. However, to the best of our knowledge, the

Manuscript received 21 March 2022; revised 13 July 2022 and 31 August 2022; accepted 4 October 2022. Date of publication 21 October 2022; date of current version 7 December 2022. This work was supported in part by the ARRS Research Program P2-0250(B) through Metrology and Biometric Systems and the ARRS Research Program P2-0214 through Computer Vision, in part by the National Natural Science Foundation of China under Grant 62106015, in part by the BUCEA Research Capacity Promotion Program for Young Scholars under Grant X21079, and in part by the Pyramid Talent Training Project of BUCEA under Grant JDYC20220819. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Emanuele Maiorana. *(Corresponding author: Matej Vitek.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Data Handling Committee (KRRP) at the Faculty of Computer and Information Science, University of Ljubljana.

Please see the Acknowledgment section of this article for the author affiliations.

This article has supplementary downloadable material available at https://doi.org/10.1109/TIFS.2022.3216468, provided by the authors.

Digital Object Identifier 10.1109/TIFS.2022.3216468

literature fails to address an important issue in this field: the bias and fairness of sclera-oriented biometric algorithms [13].

To address this gap, we describe in this paper a group evaluation effort, organized as a follow-up event to the 2020 edition of the annual Sclera Segmentation Benchmarking Competition (SSBC) [1], which focuses on the assessment of one of the key components of sclera-based recognition systems, i.e., sclera segmentation models. While SSBC 2020 studied the performance of modern sclera-segmentation models in mobile environments, the goal of the group evaluation was to benchmark segmentation performance across more diverse image characteristics, but more importantly to explore in a comprehensive manner the bias and fairness of contemporary sclera-segmentation techniques. This is a vital venue of research as questions of unfair treatment and bias in automated decision-making models have recently been a highly controversial and heavily researched topic in academia, industry, as well as society in general. Bias in machine learning algorithms has also been highlighted as one of the key topics of future research in various national and international strategies and acts [14]. It is, therefore, paramount to understand what kind of performance differentials can be expected from current state-of-the-art segmentation models, as this may impact the bias and fairness of all downstream tasks, including the final decisions made. Motivated by the importance of this topic, the group evaluation aimed at investigating the effect of different demographic and environmental characteristics, as well as the impact of training data on segmentation performance. Multiple segmentation models were developed for the evaluation, including extensions of some models that took part in SSBC 2020, but also novel approaches, developed specifically for the group evaluation. The models were benchmarked under a common experimental framework to provide answers to the following research questions:

- **Q1:** How well do contemporary models perform in the task of sclera segmentation with diverse input images?
- **Q2:** Which subject/data characteristics represent the most critical source of bias for sclera segmentation models?
- **Q3:** What impact do training data characteristics have on the bias exhibited by the segmentation models?
- **Q4:** Can we mitigate the bias exhibited by the segmentation models without losing segmentation accuracy?

The combined research efforts of multiple research groups helped provide answers to these questions and led to the following contributions that are presented in this work:

- A report on the current state-of-the-art in sclera segmentation, with a rigorous (independent) analysis of the main factors affecting segmentation performance of a representative sample of current segmentation models over five datasets with diverse image characteristics.
- A comprehensive evaluation of (algorithmic and representation) bias (and fairness) of sclera segmentation models across two environmental and two demographic factors. This includes novel performance measures for quantifying bias and a novel (public) dataset of ocular images.
- The introduction of several new (sclera) segmentation models developed exclusively for the group evaluation.

II. Background and Related Work

A. Bias and Fairness in Biometrics

Automated biometric recognition techniques can today be found in a variety of application areas that have an immediate impact on people's lives, including online banking, healthcare, access control, or surveillance and security [13], [15]. Because the (automated) decisions made by biometric systems have potentially critical consequences for individuals, it is paramount that the recognition techniques be free of biases and render fair decisions for all. While there is no (single) established definition of **bias** and **fairness** in the literature, we provide here the formulation of Drozdowski et al. [13], who defines an algorithm as being biased if it leads to significant performance differences for different subsets of data, where the subsets can be based on subject-specific (e.g., pose, expression), demographic (e.g., age, gender, ethnicity), or environmental (e.g., illumination, capture device) factors. The concept of fairness, on the other hand, can be viewed as an algorithmic property related specifically to demographic bias and is defined by Mehrabi et al. [15] as "*the absence of prejudice or favoritism toward an individual or group based on their innate or acquired characteristics*". Studies on bias and fairness in biometrics have been a central research topic in recent years [13], largely due to societal, legal and ethical implications of potentially *unfair* decisions made by automated machine learning models [16].

A considerable amount of work has been done to investigate (demographic) bias and fairness in face recognition systems, e.g., [17], [18], and [19], and potentially sensitive face-related tasks, such as age estimation [20], face image quality assessment [21], privacy protection [22], and face-morph detection [23] to name a few examples. Similar studies were also presented for fingerprints [24], [25], finger vein [26], and palm print [27] recognition systems among others. While much of this work aimed at identifying the presence of bias in various (learning-based) biometric systems and algorithms (e.g., [17], [20], [24], and [26]), a small number of works also tried to investigate causes of the observed performance differentials for different data groups, e.g., [19] and [28]. The insight and observations made by these studies provided critical understanding of the bias-related behavior of existing biometric algorithms and contributed towards various bias mitigation measures, e.g., [29], [30], and [31].

There has also been work exploring bias in the context of ocular biometrics. Krishnan et al., for example, investigated the presence of age and gender bias in recognition systems relying on the periocular regions in [32] and [33], respectively. Fang et al. [34] aimed at quantifying demographic bias in presentation attack detection (PAD) aimed at iris recognition systems, and Gorodnichy and Chumakov [35] explored age-induced performance differentials in biometric systems based on the iris. While these works presented empirical studies on the bias and fairness of different algorithms related to ocular biometrics, they have been limited to the iris and the periocular region only. Studies related to emerging ocular modalities, such as the sclera, on the other hand, are still largely missing from the literature. Given this limitation, we present a comprehensive analysis in this paper, focused on the overall performance but most of all bias of sclera segmentation models w.r.t. different demographic and environmental factors. Such segmentation models represent key components of sclera-based recognition systems and are, therefore, expected to have a considerable impact on their recognition performance.

B. Sclera Segmentation

The goal of sclera segmentation is to identify the region-of-interest (ROI) in the input image as accurately as possible, and, consequently, to ensure that all downstream tasks are applied only to relevant parts of the image that contain (discriminative) vascular patterns needed for identity inference. Several specific challenges make sclera segmentation a difficult task, including: (i) the low contrast between the foreground (i.e., the sclera) and the background (i.e., the surrounding region), which makes using traditional binarization techniques infeasible, (ii) the wide range of appearance variations caused by subject-specific and demographic factors such as eye color, ethnicity, sex/gender, or health; and (iii) the effects of external factors, e.g., the imaging device or ambient lighting.

Initial studies on sclera biometrics were mainly based on manual segmentation procedures [38], [39], but later evolved into automatic techniques designed around various clustering algorithms [40], [41], region-growing procedures [42], convex-hull based algorithms [12] and similar ad-hoc approaches [43]. While these techniques provided the basis for early sclera-based biometric systems, recent work is looking increasingly at deep learning models that have been shown to provide excellent segmentation performance for highly diverse input images. Examples of techniques from this group include convolutional neural networks (CNNs) with an encoder-decoder design [5], [9], [44], fully convolutional models [45], densely connected convolutional networks [46], generative networks [1], and other (custom) deep learning approaches [47], [48].

The evolution of sclera segmentation models has been documented and largely driven by a series of *Sclera Segmentation Benchmarking Competitions* (SSBC), held as part of major biometrics-oriented meetings and conferences [1], [48], [49], [50], [51], [52]. These competitions introduced segmentation benchmarks for the community [49], [50], examined segmentation performance under changes in gaze direction [51], in cross-sensor and cross-resolution settings [48], [52], and in mobile environments [1]. Segmentation models for the sclera (as well as for the pupil and iris) were also studied in the scope of Facebook's OpenEDS challenge [47], which aimed to compare existing models with data collected using head-mounted displays. In this paper, we further contribute to these efforts through a group assessment organized during 2021 as a follow-up to the 2020 edition of SSBC. As sclera segmentation models have matured and are now used in real-world applications, the goal of the evaluation was not only to investigate the performance of state-of-the-art models in various settings but also to better understand their behavior in terms of bias and fairness.

III. Benchmarking Methodology

A comprehensive experimental framework was designed to facilitate the group evaluation. This included the selection/collection of suitable datasets and the definition of common experimental protocols and appropriate performance measures. In this section we describe this experimental framework and provide details on the benchmarking methodology used throughout the group evaluation.

A. Datasets

Five dedicated datasets were utilized for the group evaluation. The datasets contain ocular images that differ in terms of acquisition device, gaze direction, ambient conditions, image quality, and demographics, and, hence, allow investigating various aspects of the developed segmentation models. Details on the datasets are given below.

- **The Multi-Angle Sclera Dataset** (MASD) [36] contains 2624 RGB images of 164 eyes from 82 different subjects, captured using a DSLR camera (specifically, NIKON D800 with 28-300 lenses). The images were manually cropped to a resolution of 7500 × 5000 pixels to extract the relevant region of interest (ROI). The images in MASD were acquired under 4 different gaze directions (left, right, straight, and up) and with 4 distinct images per gaze direction for each subject. The dataset contains images of male and female subjects captured in different lighting conditions and at different times of the day. It is annotated with high-quality manually generated sclera masks and is publicly available.[1]
- **The Sclera Mobile Dataset** (SMD) [37] contains 500 RGB images of 50 eyes from 25 subjects (10 images per eye) acquired with an 8MP (3264 × 2448) mobile phone (Micromax Canvas Knight A350) rear camera. The dataset is approximately gender-balanced with 12 male and 13 female subjects and comes with variations in the age and skin color of the subjects. Images in the dataset were captured in different lighting conditions and with image noise to more accurately represent realistic scenarios in which sclera segmentation methods need to operate. SMD ships with manually generated sclera annotations and is also publicly available[1].
- **The Sclera Liveness Dataset** (SLD) represents a novel dataset, captured specifically for the group evaluation, and consists of 108 genuine RGB images from both eyes of 27 individuals (in other words 54 different eyes). For each eye 2 sample images were captured. The dataset contains blurred images and images with blinking eyes. It includes both male and female subjects, of different ages and different skin tones. The images in SLD were taken at different times of the day to model natural environment-induced variations. Differences in image quality (blur, lighting condition, etc.) and acquisition conditions were included intentionally in the dataset to facilitate investigations into the performance of the segmentation models in non-ideal scenarios. High-resolution images (3264 × 2448) are included in the dataset. All images were captured using a mobile phone (Lenovo K3 Note) with an 8MP rear camera and are stored in JPEG format. SLD is publicly available[1].
- **The Sclera Blood Vessels, Periocular, and Iris** (SBVPI) [5], [9] dataset consists of 1858 RGB images of 110 eyes

[1] MASD, SMD and SLD are publicly available on request. Please contact abhijitdas2048@gmail.com for more information.

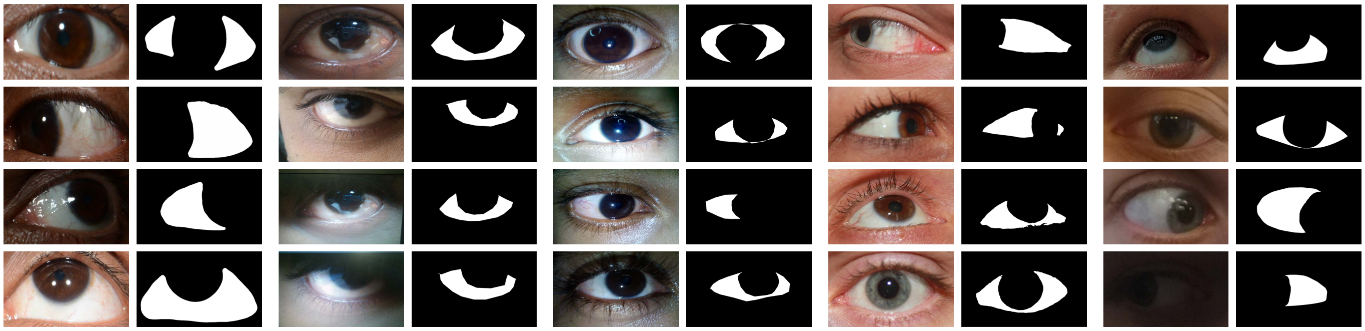

(a) MASD [36] (b) SMD [37] (c) SLD (newly collected) (d) SBVPI [5] (e) MOBIUS [1]

Fig. 1. Example images and corresponding ground truth annotations from the five datasets used in the group evaluation. The datasets contain images with diverse characteristics in terms of image quality, capture device, demographics, blur, gaze direction, eye color and others. The datasets are publicly available for research purposes[1,2].

(i.e., 55 subjects) captured with a DSLR camera (specifically, Canon EOS 60D with macro lenses). The images were manually cropped to extract the desired ROI while maintaining their aspect ratio, then rescaled to 3000 × 1700 pixels to maintain a consistent image size across the entire dataset. Images in the dataset were captured at the highest resolution and quality settings available in the camera and in a laboratory environment. Similarly to MASD, the dataset contains images taken under 4 different gaze directions, with a minimum of 4 images per direction for each subject. The appearance variability in SBVPI is due to identity, eye color, gender, and age. Manually generated markups of the sclera and periocular regions are present for all images. SBVPI is publicly available for research purposes.[2]

- **The Mobile Ocular Biometrics In Unconstrained Settings** (MOBIUS) [1] dataset comprises 16717 RGB images of 200 eyes from 100 subjects. The images were manually cropped to obtain the relevant ROI and resized to a resolution of 3000 × 1700 pixels to keep a consistent image size across the dataset. A subset of 3542 images from 35 subjects (70 eyes) is designated for segmentation research and contains high-quality manually generated (and later cleaned with a semi-automatic correction procedure [53]) annotations of the sclera, iris, and pupil regions. The dataset again contains 4 gaze directions for each eye, but exhibits a significantly higher degree of variability than other datasets due to the use of 3 different mobile phone cameras (Sony Xperia Z5 Compact, Apple iPhone 6s, and Xiaomi Pocophone F1) for image capture, and 3 ambient settings (i.e., sunny outside; inside with good illumination; and inside with poor illumination). Additionally, data about the subjects (e.g., identity, gender, eye color, age, eyewear, eye conditions and allergies) is also available to facilitate research into various data characteristics and their impact on segmentation performance[2].

We note that all datasets were collected with *consenting subjects*. A few illustrative example images from the experimental datasets and the corresponding sclera masks are presented in Fig. 1. A high-level comparison is given in Table I.

B. Evaluation Setup

1) Experimental Protocols: The research groups participating in the evaluation were given access to images from all five datasets. For the MASD, SMD, and SBVPI datasets both the raw images and the ground truth segmentation masks were made available, whereas only the raw images were made public for SLD and MOBIUS, while the ground truth remained sequestered. Based on this data, the participants were asked to develop sclera segmentation models under two distinct experimental protocols, i.e.:

- *The Complete Training Data (CTD)* protocol, where the segmentation models were trained on the full MASD, SMD, and SBVPI datasets (for a total of 4982 images from 162 subjects). The results for the group evaluation under this protocol were generated on the SLD and MOBIUS datasets. Since different datasets were used for training and testing in all experiments conducted under this protocol, there was no overlap in subjects between the training and testing data.
- *The Limited Training Data (LTD)* protocol, where it was only allowed to use specific training data to learn the models and results needed to be generated on predefined test datasets. This protocol resulted in multiple models with different train-test data configurations, depending on the bias aspect being explored in a given experiment. Details about the specific training and testing data used under various configurations of this protocol are provided in Section V.

The above protocols were designed for the analysis of different aspects of the developed segmentation models, as detailed in the experimental section.

2) Result Generation: Two types of results were requested for the analysis: (i) binarized (black-and-white) segmentation masks, with white pixels corresponding to the sclera region and black pixels to other image areas, and (ii) probabilistic segmentation maps, with the pixel intensities corresponding to the "probability" that the pixels belongs to the sclera region. Both types of results were submitted for all models trained under the CTD and LTD experimental protocols. A sample submission is shown in Fig. 2. These results were ultimately compared to the (sequestered) ground truth information for scoring purposes. In all experiments, the scoring was done with fixed-size images and ground truth masked, rescaled to 480 × 360 pixels, to ensure a common evaluation setting.

[2] The SBVPI and MOBIUS datasets are publicly available on request for research purposes. For more information visit sclera.fri.uni-lj.si.

TABLE I

HIGH-LEVEL COMPARISON OF THE DATASETS AND EXPERIMENTAL PROTOCOL USED IN THE GROUP EVALUATION

Dataset	#Images	#IDs	#Eyes	Capture Hardware	Capturing Resolution [px]	Processed Resolution [px]	Sources of Variability[†]	Role[‡]
MASD [36]	2624	82	164	DSLR camera	7360 × 4912	7500 × 5000	GZ, BL	TR
SMD [37]	500	25	50	Mobile camera	3264 × 2448	3264 × 2448	BL, CN	TR/TS*
SLD (new)	108	27	54	Mobile camera	3264 × 2448	3264 × 2448	BL, CN	TS
SBVPI [5], [9]	1858	55	110	DSLR camera	5184 × 3456	3000 × 1700	GZ, BL	TR
MOBIUS [1]	3542	35	70	Mobile cameras	4032 × 3024 or 5520 × 4140	3000 × 1700	MD, CN, GZ, BL	TS

[†] GZ - gaze, BL - blur, CN - acquisition condition, MD - mobile device; [‡] TR - training; TS - testing.

*SMD was used for training and testing in different experiments, but never for both in the same experiment.

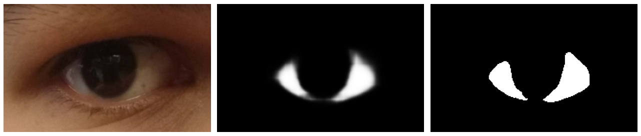


Fig. 2. Illustration of the results generated for the group evaluation. For each input image (left), a probabilistic (middle) and binary segmentation mask (right) had to be generated and submitted for scoring.

C. Scoring Criteria

The main goal of the group evaluation is to analyze two key aspects of recent (sclera) segmentation models: (i) the overall segmentation performance, and (ii) the exhibited biases. Two sets of performance indicators are, therefore, used to report results of the group evaluation.

1) Overall Segmentation Performance: In accordance with standard evaluation methodology [1], [48], we use the following indicators to score segmentation performance:

- **Precision**, i.e., the proportion of correctly identified sclera pixels in relation to all pixels determined as belonging to the sclera by a given model: $(\frac{TP}{TP+FP})$ [54].
- **Recall**, i.e., the number of correctly identified sclera pixels in relation to all pixels marked as belonging to the sclera region in the ground truth: $(\frac{TP}{TP+FN})$ [54].
- $\boldsymbol{F_1}$**- score**, i.e., the harmonic mean between precision and recall: $(2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}})$ [1].
- **Intersection over Union (IoU)** or Jaccard index, i.e., the quotient between the size of the intersection of the predicted and actual sclera regions, and the size of the union of the two, computed as: $(\frac{TP}{TP+FP+FN})$ [1].

Here, TP, FP, and FN stand for the number of *true positives*, *false positives* and *false negatives* generated by the models with respect to the ground truth. Additionally, we report complete *precision-recall (PR) curves* [55], [56] based on the computed probabilistic predictions and the corresponding *Area Under the precision-recall Curve* (AUC) [57] as another aggregate performance indicator that provides a more holistic view on the performance of the evaluated models.

2) Bias Evaluation: Bias is commonly quantified through a measure of performance (or error) dispersion across different subgroups of the evaluation data [18], [29]. Following this established practice, we report the standard deviation (STD) and mean absolute deviation (MAD) of the computed performance indicators as two measures of bias in our experiments [30], i.e.:

- **Standard Deviation (STD)**, defined as the square root of the average squared deviation between the performance of specific subgroups p_g and the mean performance across all groups $\overline{p}$:
$$\text{STD} = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(p_g - \overline{p}\right)^2}. \tag{1}$$
- **Mean Absolute Deviation (MAD)**, defined as the average absolute deviation between the performance on specific subgroups p_g and the mean performance across all groups $\overline{p}$:
$$\text{MAD} = \frac{1}{G}\sum_{g=1}^{G}\left|p_g - \overline{p}\right|. \tag{2}$$

In the above equations G refers to the number of different subgroups in the data, p_g denotes the group-specific performance (in our case the F_1 score), and $\overline{p} = \frac{1}{G}\sum_{g=1}^{G} p_g$. While the two measures capture similar aspects of the bias, STD gives larger importance to outliers (worst case scenario), whereas MAD is influenced more by the majority of subgroups.

In general, STD and MAD quantify the performance variations across different data subgroups (i.e., bias) but ignore the *innate* variations of the data that also cause performance differences. Based on this observation and the insights from [58] we, therefore, propose and introduce *two disparity measures* that weigh the computed group-specific dispersion against the observed dispersion on some reference data:

- **Control Group Disparity (CGD)**, which we define as the ratio between the standard deviation of the performance scores between different data subgroups and the corresponding standard deviation computed on control groups:
$$\text{CGD} = \frac{\text{STD}}{\sqrt{\frac{1}{G}\sum_{c=1}^{G}\left(p_c - \overline{p_C}\right)^2}}, \tag{3}$$
where there are G control groups in total, and each control group c matches the size of one of the original attribute-specific data subgroups, but contains randomly chosen samples. Additionally, $\overline{p_C} = \frac{1}{G}\sum_{c=1}^{G} p_c$.
- **Fisher Disparity (FSD)**, which we define as the ratio between the standard deviation of the performance scores across different data subgroups and the mean standard deviation within the subgroups, i.e.:
$$\text{FSD} = \frac{\text{STD}}{\frac{1}{G}\sum_{g=1}^{G}\sqrt{\frac{1}{n_g}\sum_{i=1}^{n_g}\left(p_i - p_g\right)^2}}, \tag{4}$$
where p_i is the performance on the i-th data instance (i.e., a single image) and n_g is the number of images in the g-th subgroup.

Both FSD and CGD consider reference variations when quantifying bias, but do so based on different assumptions,

TABLE II

HIGH-LEVEL COMPARISON OF THE SEGMENTATION MODELS DEVELOPED FOR THE GROUP EVALUATION. THE MODELS EXHIBIT DIVERSITY ACROSS THE BASE ARCHITECTURE USED, THE FORMAT OF THE INPUT DATA, THE USE OF AUGMENTATION STRATEGIES, NORMALIZATION PROCEDURE, PROBLEM FORMULATION, LEARNING OBJECTIVE AND COMPLEXITY

Segmentation Model	**Contributor**	**Architecture Base**[†]	**Color Space**	**Augmentation**	**Normalization**	**Problem Formulation**	**Objective**	**#Params.**	**FLOPs**
ScleraSegNet	CAS	U-Net	RGB	Yes (Heavy)	No	Semantic Segmentation	Single-task	24M	217G
ScleraU-Net2	DUTH	U-Net	RGB	Yes	No	Semantic Segmentation	Single-task	3M	180G
MU-Net	WUT	U-Net	RGB	Yes	No	Semantic Segmentation	Single-task	409K	180G
FCN8	UFPR	FCN (VGG16$^+$)	RGB	No	No	Semantic Segmentation	Single-task	138M	15G
CGANs2020CL	HDA	cGAN	Y Channel	Yes (Heavy)	No	Image Translation	Dual-task	11.5M	N/A
RGB-SS-Eye-MS	IGD	CRN	RGB	Yes (Heavy)	Standard $0-1$	Semantic Segmentation	Single-task	22.7M	N/A
ScleraMaskRCNN	NTNU	Mask R-CNN	RGB	No	No	Instance Segmentation	Multi-task	69M	N/A

[†]FCN – Fully Convolutional Network; VGG16$^+$ – Pretrained VGG16 (ImageNet); cGAN – Conditional GAN; CRN – Convolutional Refinement Network (custom)

and, therefore, provide complementary information w.r.t. the observed performance differences across data subgroups.

IV. SUMMARY OF DEVELOPED MODELS

Seven research teams developed segmentation models for the group evaluation, i.e., the Chinese Academy of Sciences (CAS), the Democritus University of Thrace (DUTH), the Warsaw University of Technology (WUT), the Federal University of Parana (UFPR), the Hochschule Darmstadt (HDA), Fraunhofer IGD (IGD), and the Norwegian University of Science and Technology (NTNU). More details on the submitted models are provided in the following section, along with links to the corresponding source code repositories, to ensure reproducibility of the results and to provide additional implementation details on all developed segmentation approaches.

A. Model Descriptions

ScleraSegNet[3] (CAS). The CAS group designed an attention-assisted U-Net-based [59] model for sclera segmentation [60], called ScleraSegNet. The model incorporates modules for channel- and spatial-attention into both the central bottleneck, as well as the skip-connection part of the base U-Net architecture. This helps improve the sensitivity of the model to foreground/background pixels and also alleviates the interference of noise factors. ScleraSegNet is trained with images resized to a width of 600 pixels (regardless of the original size of the images in the training data), while maintaining the original aspect ratio. Heavy data augmentation is performed, including random resizing, blurring, translating, flipping, rotating and cropping (to 321 × 321 pixels) to avoid overfitting. Binary cross-entropy is utilized as the loss function in training. When generating the binary masks for the evaluation, the binarization threshold is set to 0.5.

ScleraU-Net2[4] (DUTH). The DUTH group developed a novel U-Net-inspired model based on the ScleraU-Net architecture designed initially for the SSBC 2020 competition [1]. Compared to the original U-Net, ScleraU-Net2 has a reduced number of convolutional layers and, therefore, exhibits decreased network complexity. This leads to a more light-weight architecture that is better tailored towards the sclera segmentation problem. Specifically, ScleraU-Net2 comprises 8 filter kernels in its first convolutional layer, compared to the 64 kernels of the original U-Net. For the subsequent layers the number of filters is doubled after every pooling operation. Another key improvement in ScleraU-Net2 is the use of Group Normalization (GN) after each convolutional layer. Group normalization [61] is used as a replacement for Batch Normalization (BN) and is paramount for models trained with relatively small batch sizes, where BN layers may fail to properly capture the distribution parameters, resulting in poor normalization with adverse effects on generalization. Finally, the activations of all convolutional layers are replaced with GELUs [62], for which improved performance across many vision tasks has been reported in the literature. Training is performed with a fixed learning rate of 10^{-4} and a batch size of 6. Data augmentation is applied in an online fashion, by random horizontal flipping and a limited amount of rotation and shear. The probability maps are converted to binary maps by a fixed-value thresholding of 0.5.

MU-Net[5] (WUT). The main idea behind the approach of the WUT group is to utilize a light-weight architecture designed for mobile-computing that allows for efficient learning of segmentation models with limited training data. Along these lines, the WUT group designed a U-Net-like encoder-decoder model, named MU-Net, with a MobileNetV2 [63] encoder pretrained on ImageNet. The model is fine-tuned for sclera segmentation using the provided training data, augmented with horizontal flips, and the standard binary cross-entropy learning objective. Because the encoder model, MobileNetV2, has fewer parameters than the encoder of the original U-Net, the entire model converges quickly, while also ensuring high segmentation accuracy and good generalization. The model produces a probabilistic segmentation prediction for each pixel location. A binarisation threshold is, therefore, chosen by iterating over the validation images and fixing the threshold at the value that achieves the highest F1-score to produce the binary segmentation results required for the evaluation.

FCN8[6] (UFPR). Inspired by the work of Long et al. [64], the UFPR group designed a segmentation model, FCN8, based on a Fully Convolutional Network (FCN). Due to the fully convolutional structure, the model is applicable with input images of arbitrary size and produces corresponding segmentation results. For FCN8, an architecture similar to the one proposed by Teichmann et al. [65] is utilized and relies on a VGG-16 model (without the FC layers) in the encoder and a three-layer decoder for upsampling. A unique aspect of FCN8 is the design of the bottleneck module in the encoder-decoder architecture, which retains the spatial dimension instead of compressing all of the information of the input image into a vectorized latent representation. This design choice allows

[3]ScleraSegNet is available from github.com/xiamenwcy/ScleraSegNet.

[4]ScleraU-Net2 link: github.com/georgezampoukis/ScleraU-Net2_SSBC.

[5]MU-Net is available from github.com/Jalilnkh/MU-Net.

[6]FCN8 is available from github.com/diegorafaellucio/FCN8.

for the use of a simple decoder that does not need to learn decoding spatial information from the latent representation and can, therefore, be trained efficiently with a limited amount of data. FCN8 is learned with a cross-entropy training objective.

CGANs2020CL[7] (HDA). The HDA group contributed an approach that framed the sclera segmentation task as a patch-based image-translation problem [66] and used a Conditional Generative Adversarial Network (cGAN) as the basis for the segmentation. The goal of the cGAN in this setting is to implement a mapping from the gray-scale real-valued (ocular image) domain to the binary sclera domain. The backbone of the model is a ResNet-101 [67] learned from scratch using only the provided training data. To avoid overfitting and ensure that a well performing model with good generalization capabilities is learned, aggressive data-augmentation is performed using the *Imgaug library*.[8] Here, various augmentation strategies were considered, including rotations, flipping, cropping, color manipulations and others. The model is trained using a weighted sum of the GAN and L_1-reconstruction losses.

RGB-SS-Eye-MS[9] (IGD). The IGD team developed a model that extends the multi-scale eye segmentation solutions (Eye-MS) from [68]. The model represents a convolutional neural network (CNN) that refines segmentation progressively using different input resolutions. Each of the refinement modules consists of two convolutional layers, followed by a normalization layer and a LReLU non-linearity. RGB-SS-Eye-MS is trained with the Intersection over Union (IoU) loss using the SGD optimizer with a learning rate of 0.1 and a batch size of 32. Heavy data augmentation in the form of random cropping, horizontal flipping brightness and contrast changes, blurring and noise infusions is employed to improve generalization. The predicted segmentation is rounded to the nearest integer values to generate binary segmentation masks.

ScleraMaskRCNN[10] (NTNU). The solution designed by the NTNU group, called ScleraMaskRCNN, follows the two-stage approach from Mask-RCNN originally proposed in [69]. In the first stage, ScleraMaskRCNN generates region proposals for the sclera in the input image and in the second stage then classifies these proposals into the most likely class (i.e., sclera/other). A pixel-level mask is also computed in the second stage to facilitate the (instance) segmentation procedure. The model uses a ResNet-101 [67] as the backbone for feature extraction and is trained using a joint objective that combines losses for classification/localization and segmentation-mask prediction. No augmentation is used during training.

B. Comparative Analysis

A high-level comparison of the developed segmentation models is presented in Table II. In accordance with recent trends in the segmentation literature [1], [9], [48], [70], all of the contributed models use deep learning to efficiently capture the complex appearance variations present in the ocular images. The majority of solutions rely on encoder-decoder architectures (e.g., U-Net-based, FCN) with an information bottleneck as the basis for segmentation, but custom designs (CRN) and Masked-RCNNs are also represented among the contributed models. Noteworthy, all models learn from color as well as texture information, except for the solution from the HDA group that relies exclusively on texture (i.e., processes gray-scale images). The developed models also differ in terms of problem statement (semantic vs. instance segmentation and segmentation vs. image translation) and corresponding learning objectives. Finally, we observe considerable differences in the number of trainable parameters, ranging from 409K for the most light-weight model to 138M for the heaviest one. Overall, the developed models represent a rich and diverse set of segmentation techniques for the group evaluation.

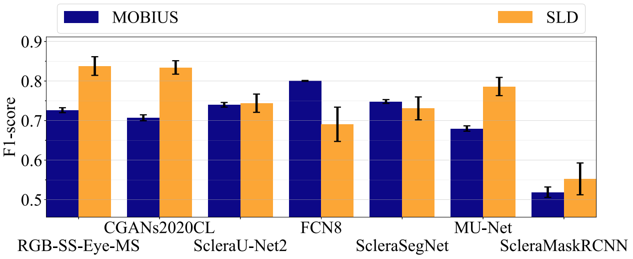


Fig. 3. Segmentation performance of the developed models on the MOBIUS (blue) and SLD (yellow) datasets with the CTD experimental protocol. Shown are the average F_1 scores (and corresponding standard deviations, $n = 5$) computed from the binary segmentation masks. Results are sorted w.r.t. the harmonic F_1 mean over the two datasets.

V. Experiments and Results

In this section, we present the results of the group evaluation that: (i) analyze the performance of different sclera segmentation models over multiple test datasets, (ii) investigate performance differences of the models across various data subgroups and training configurations, and (iii) study the correlations between bias and overall segmentation performance. We make our evaluation code publicly available to ensure the reproducibility of our results.[11]

A. Segmentation Performance

In the first series of experiments, we benchmark the developed models with respect to the overall segmentation performance. We consider the *complete-training-data* (CTD) protocol for these experiments and use the (unseen) MOBIUS and SLD test images for scoring. We separately analyze results based on: (i) the submitted binary segmentation masks, where the participating groups performed the binarization procedure on their own, and (ii) the probabilistic masks, processed independently by the organizers of the group evaluation.

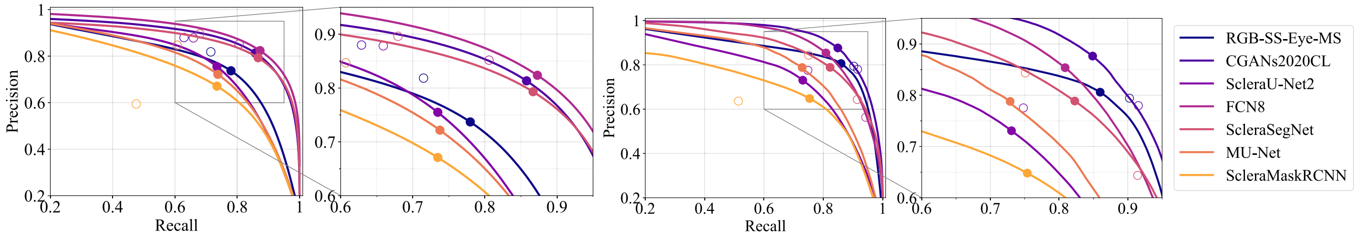
1) Results on Binary Masks: In Fig. 3 we show the average F_1 scores obtained by the developed models together with the corresponding standard deviations computed from $n = 5$ (disjoint) stratified subsets of images sampled from each of the two test datasets. These results provide insight into the performance of the segmentation models, but also the variability of the observed scores. More detailed results across the remaining

[7] CGANs2020CL is available from github.com/jedota/Sclera-Segmentation.
[8] Imgaug is available from: github.com/aleju/imgaug.
[9] RGB-SS-Eye-MS is available from github.com/fdbtrs/SS-Eye-MS.
[10] ScleraMaskRCNN link: github.com/NTNUGE/ScleraMaskRCNN.
[11] The evaluation code is available from github.com/MatejVitek/GE.

TABLE III

COMPARISON OF THE OVERALL SEGMENTATION PERFORMANCE. SHOWN ARE RESULTS FOR THE BINARISED MASKS AND PROBABILISTIC PREDICTIONS. F_1^{opt} DENOTES THE HIGHEST F_1 SCORE ON THE PRECISION-RECALL CURVE. THE SECOND COLUMN BELOW EACH MEASURE IS THE HARMONIC MEAN ACROSS THE TWO TEST DATASETS. RESULTS ARE SORTED BASED ON THE (BINARY) HARMONIC F_1 MEAN

Model	**Evaluation data**	**From binary masks**								**From probabilistic predictions**			
		F_1		**Precision**		**Recall**		**IoU**		F_1^{opt}		**AUC**	
RGB-SS-Eye-MS	MOBIUS SLD	0.726 0.838	0.778	0.818 0.794	0.806	0.715 0.903	0.798	0.618 0.730	0.669	0.748 0.852	0.797	0.804 0.874	0.837
CGANs2020CL	MOBIUS SLD	0.707 0.834	0.765	0.880 0.780	0.827	0.629 0.915	0.746	0.611 0.725	0.663	0.843 0.885	0.863	0.894 0.938	0.915
ScleraU-Net2	MOBIUS SLD	0.740 0.744	0.742	0.879 0.776	0.824	0.659 0.748	0.701	0.618 0.614	0.616	0.749 0.751	0.750	0.795 0.779	0.787
FCN8	MOBIUS SLD	0.800 0.691	0.741	0.852 0.563	0.678	0.806 0.943	0.869	0.702 0.544	0.613	0.857 0.854	0.856	0.918 0.914	0.916
ScleraSegNet	MOBIUS SLD	0.748 0.731	0.739	0.896 0.644	0.749	0.680 0.914	0.780	0.650 0.606	0.627	0.831 0.830	0.830	0.880 0.888	0.884
MU-Net	MOBIUS SLD	0.680 0.786	0.729	0.847 0.843	0.845	0.607 0.751	0.671	0.559 0.654	0.603	0.717 0.803	0.757	0.783 0.821	0.802
ScleraMaskRCNN	MOBIUS SLD	0.522 0.555	0.538	0.592 0.630	0.610	0.482 0.522	0.501	0.454 0.474	0.463	0.591 0.583	0.587	0.746 0.711	0.728



(a) MOBIUS

(b) SLD

Fig. 4. Precision-recall (PR) curves of the experiments – best viewed in color and zoomed in. The optimal F_1 scores on the PR curves (F_1^{opt}) are marked with full circles, whereas the F_1 points of the binary masks are marked with empty circles. Results are presented separately for the (a) MOBIUS and (b) SLD datasets.

performance indicators are given in Table III. Here, only the mean scores are reported to keep the results uncluttered. The models are sorted with respect to the harmonic F_1 mean calculated across the two evaluation datasets.

As can be seen, RGB-SS-Eye-MS and CGANs2020CL perform the best overall in this setting with harmonic F_1 means of 0.778 and 0.765, respectively, followed closely by ScleraU-Net2 with a score of 0.742, mostly due to the more consistent performance across both test datasets. FCN8 and ScleraSegNet exhibit a slightly weaker performance to ScleraU-Net2 in terms of the harmonic F_1 mean with scores of 0.741 and 0.739, respectively, but achieve the best and second best performance on MOBIUS. MU-Net and ScleraMaskCNN yield the sixth and seventh best results and rank behind the best performing models with corresponding harmonic F_1 means of 0.729 and 0.538. It is interesting to note that among the top performers, models using a fixed thresholding procedure for generating the binary masks (ScleraU-Net2 and ScleraSegNet) result in more consistent F_1 scores across the datasets than models using dynamic thresholding (RGB-SS-Eye-MS, CGANs2020CL and FCN8). Nonetheless, finding a good trade-off between precision and recall scores appears to be challenging for all models regardless of the thresholding strategy used, as evidenced by the difference between the two performance scores and their variability across MOBIUS and SLD in Table III. Overall, we observe that 6 of the 7 submitted models are within a performance difference of less than 0.05 in terms of the harmonic F_1 mean. However, larger performance variations are observed with other (individual) performance scores (recall, precision, IoU) over each of the two datasets.

2) Results on Probabilistic Masks: To get better insight into the performance of the segmentation models, we generate precision-recall curves from the probabilistic segmentation masks and show these together with the optimal operating point in terms of F_1 score in Fig. 4. Additionally, we also visualize the operating points that correspond to the binary masks in the same graph. Numerical results computed based on the curves are summarized in the right part of Table III. Several observations can be made based on these results:

- **Binarization.** Using an optimal threshold for generating segmentation masks from the probabilistic predictions in general improves results for all models in terms of the harmonic F_1 mean. Additionally, the binary operating points are often not located on the PR curves due to different strategies used for either producing the probabilistic predictions or determining the binarization threshold. This suggests that even with fixed segmentation models, efficient mechanisms for segmentation mask generation are critical for performance and suitable trade-offs between precision and recall scores.
- **Generalization and Calibration.** We observe that comparable performances can be achieved on both datasets with an optimal binarization threshold. Except for RGB-SS-Eye-MS and MU-Net, where the harmonic F_1^{opt} means still exhibit a larger difference, all other models generalize well across MOBIUS and SLD. The results

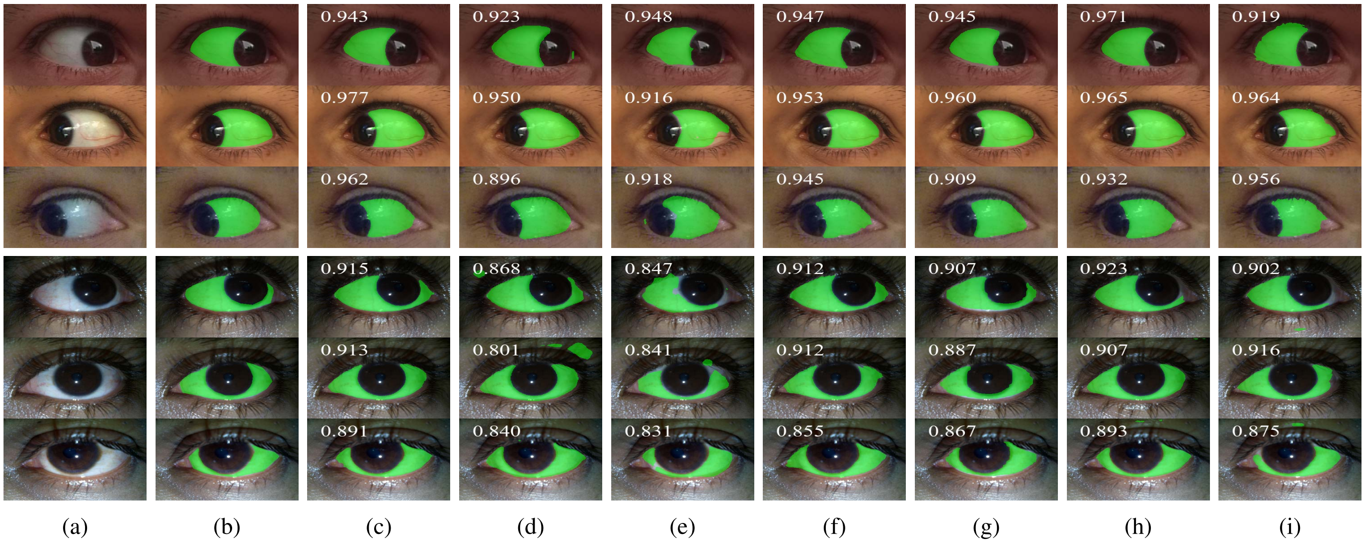


Fig. 5. Qualitative comparison of the (binary) segmentation results on the samples from the MOBIUS and SLD datasets that resulted in some of the best segmentation performance across the different models. The 3 rows at the top half of the figure show MOBIUS samples captured in different lighting conditions, specifically in top-to-bottom order: with natural lighting, indoor lighting, and poor lighting. The F_1 scores achieved are superimposed. The columns represent: (a) the original image; (b) the ground truth mask; and the predicted binary masks for: (c) CGANs2020CL, (d) FCN8, (d) MU-Net, (e) RGB-SS-Eye-MS, (f) ScleraMaskRCNN, (g) ScleraSegNet, and (h) ScleraU-Net2. Best viewed zoomed-in.

generated from the binary masks, on the other hand, vary wildly and in general correspond to higher precision scores with lower recalls compared to the optimum on MOBIUS and the other way around on SLD.

- **Ranking.** Compared to the binary results from Fig. 3, we notice a change in the ranking of the models, where CGANs2020CL and FCN8 are now the clear top performers with harmonic F_1^{opt} means of 0.863 and 0.856 across the two datasets, respectively. ScleraSegNet ranks third with a score of 0.830, whereas the rest perform weaker. These results suggest that mechanisms that allow for efficient training with limited training data (e.g., heavy augmentation, use of pretrained models) lead to the most competitive segmentation performance.

3) Qualitative Results: In Figs. 5 and 6 we visualize the (binary) segmentation masks generated by the evaluated models for a few example images from the two experimental datasets that produced the best and the worst segmentation results across the evaluated models. We can see that for the well-performing samples in Fig. 5, all evaluated models generate competitive results and generalize well across different gaze directions. For the more challenging samples from Fig. 6, on the other hand, the segmentation models result in very different errors. While some are able to reasonably well identify the sclera region in the presence of partially closed eyes and eyelash occlusions (e.g., see the results for CGANs2020C and FCN8), others struggle to locate parts of the sclera region or introduce visible artifacts.

B. Bias Analysis

As emphasized by Mehrabi et al. in [15], biases come in various shapes and forms and may raise issues related to the fairness of automated decisions made by machine learning algorithms. To better understand the behavior of sclera segmentation models in this regard, we explore two types of biases in this section, i.e.:

- **Algorithmic Bias:** The first type of bias originates from the machine learning algorithms and is typically associated with the design choices made, the optimization objective used, the regularizations considered and similar algorithm-specific characteristics [71]. We study algorithmic bias in the following sections by comparing the developed models on subgroups of the test samples with fixed and predefined training data.
- **Representation/Sampling Bias:** The second type of bias stems from the way the data is sampled from a population during the data collection process [15], [71]. Unrepresentative training data or potential biases in the data are typically inherited by the machine learning models and (may) eventually lead to unfair decisions. We investigate representation bias within the group evaluation by analyzing models learned with different sets of training data.

To ensure a comprehensive analysis, experiments are conducted with subgroups defined based on different data characteristics. Specifically, we consider subgroups generated based on *demographic* (eye color and ethnicity) as well as *environmental* factors (acquisition device and gaze direction), which represent two of the main groups of data characteristics most critical from a bias perspective according to [13]. The selection of characteristics is also motivated by the annotations available in the datasets utilized for the group evaluation. We note that all experiments are performed with stratified subgroups to mitigate issues related to different sample sizes.

1) Algorithmic Bias: When investigating algorithmic bias, we consider F_1 scores computed from the binary segmentation masks as the basis for the analysis. The binarization threshold is, thus, set automatically during training, similarly to a real-world operational scenario.

Eye-Color Bias. An important ocular characteristic, also often associated with race, is the eye color of the individuals. The vast majority of people of Asian and African origins, for example, have brown eyes, whereas people of Caucasian origin typically exhibit a wider spectrum of eye colors. To explore the impact of eye color on segmentation performance, we conduct an analysis on the test images from the MOBIUS dataset with (stratified) subgroups that correspond to subjects with brown,

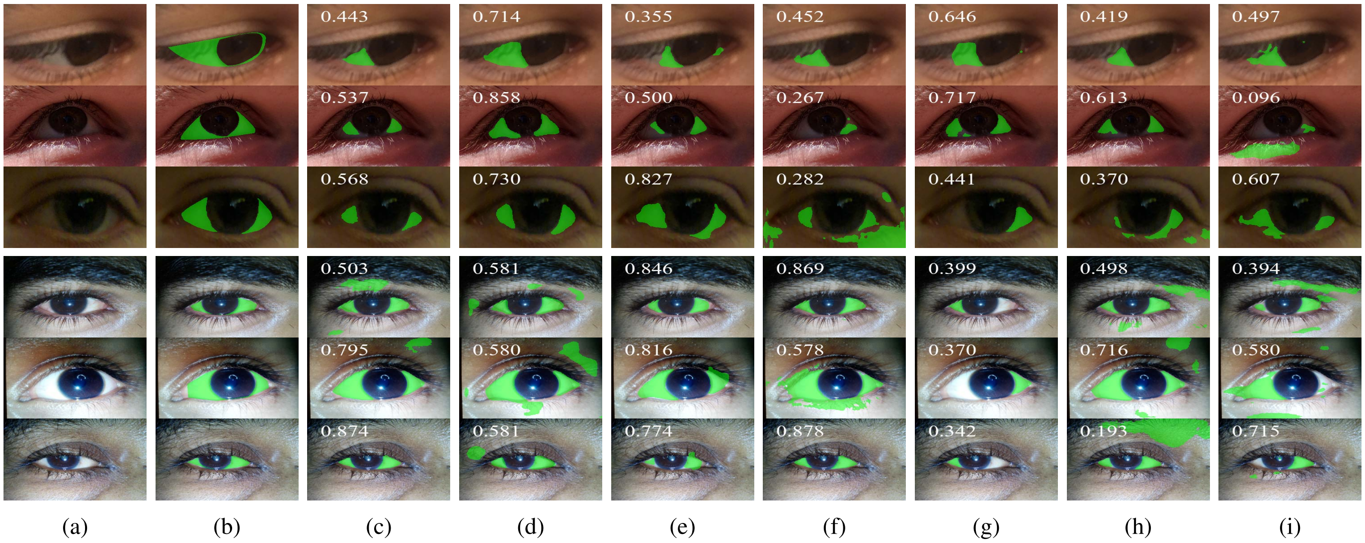
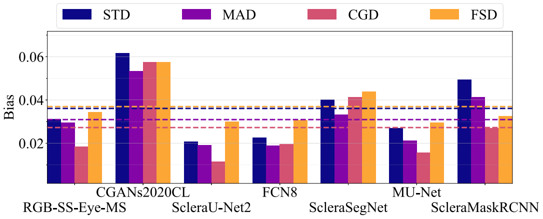


Fig. 6. Qualitative comparison of the (binary) segmentation results on samples from the MOBIUS and SLD datasets that resulted in the worst performance across the different models. The 3 rows in the top half of the figure show MOBIUS samples captured in different lighting conditions, specifically in top-to-bottom order: with natural lighting, indoor lighting, and poor lighting. The images additionally have their corresponding F_1 scores superimposed. The columns represent: (a) the original image; (b) the ground truth mask; and the predicted binary masks for: (c) CGANs2020CL, (d) FCN8, (e) MU-Net, (f) RGB-SS-Eye-MS, (g) ScleraMaskRCNN, (h) ScleraSegNet, and (i) ScleraU-Net2. Best viewed zoomed-in.

Segment. Model	F_1^{all} ↑	**Eye-color specific** F_1^{eye} ($F_1^{all} - F_1^{eye}$)				**Disparities**	
		Blue	**Green**	**Gray**	**Brown**	**CGD** ↓	**FSD** ↓
RGB-SS-Eye-MS	0.726	0.762 (−0.036)	0.690 (0.037)	0.768 (−0.042)	0.721 (0.005)	2.63	0.129
CGANs2020CL	0.707	0.748 (−0.041)	0.617 (0.090)	0.753 (−0.046)	0.726 (−0.019)	8.23	0.215
ScleraU-Net2	0.740	0.765 (−0.025)	0.714 (0.025)	0.779 (−0.039)	0.733 (0.007)	**1.64**	0.113
FCN8	0.800	0.827 (−0.026)	0.763 (0.038)	0.823 (−0.023)	0.804 (−0.004)	2.79	0.115
ScleraSegNet	0.748	0.783 (−0.035)	0.682 (0.066)	0.801 (−0.053)	0.754 (−0.006)	5.90	0.164
MU-Net	0.680	0.707 (−0.027)	0.640 (0.040)	0.710 (−0.030)	0.683 (−0.003)	2.23	**0.110**
ScleraMaskRCNN	0.519	0.522 (−0.004)	0.442 (0.077)	0.570 (−0.051)	0.540 (−0.021)	3.85	0.122

(a) Differential performance and bias disparities (eye color)



(b) Bias scores (eye color)

Fig. 7. Differential performance and bias scores with respect to eye color evaluated on the MOBIUS dataset. For the disparities in (a) the best score is presented in bold, the second best is underlined. In (b), the disparities (CGD and FSD) are normalized to the range of STD and MAD for visualization purposes. The mean value of each bias score is shown as a dashed line and lower values imply better performance/behavior, i.e., ↓. The figure is best viewed electronically and in color.

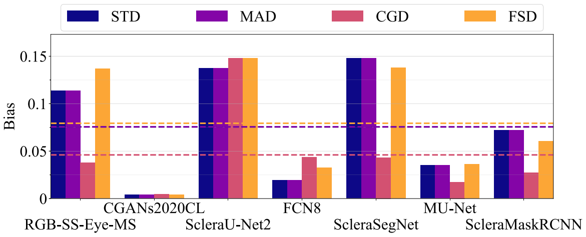
green, blue and gray eyes. The models scored in the analysis are trained using the CTD protocol, so all eye colors are well represented in the training data.

Fig. 7(a) shows that all models underperform with green eyes, where green-color-specific F_1^{eye} scores between 3.4% (ScleraU-Net2) and 14.8% (ScleraMaskRCNN) below the average performance across all images F_1^{all} are observed, and F_1^{eye} scores between 8.4% (ScleraU-Net2) and 22.0% (ScleraMaskRCNN) below the best performing eye-color are seen. The (differential) results for the remaining eye-colors are closer in general, with blue and gray eyes consistently yielding the highest scores for all tested models and brown eyes resulting in somewhat lower but still above average F_1^{eye} values. These systematic performance differentials are unexpected, especially given the fact that the blue- and gray-colored eyes are less represented in the provided training data than the gray and brown eyes, suggesting that *eye-color represents a critical image characteristics* with considerable impact on the segmentation performance and fairness of the evaluated models, regardless of their design. While the presented F_1 scores provide an initial idea about the performance differentials due to eye color, they are based on selected subgroup samples that may contain additional sources of variability that affect performance [19]. Since these sources cannot be easily accounted for (as they are in general unknown), we report the proposed bias disparities, CGD and FSD, in the right part of Fig. 7(a). As can be seen, all models exhibit a CGD score above 1, suggesting that the variability in segmentation performance due to color variations is larger (even though moderately so) than the variability seen in randomly sampled subgroups. The lowest performance differentials are seen with the ScleraU-Net2 model with a CGD score of 1.64 and the largest for the CGANs2020CL approach with a CGD value of 8.23. Interestingly, this model is also the only one trained on gray-scale images. The fact that the only model working with gray-scale images exhibited by far the highest degree of eye-color bias suggests that using color information (for training and at run-time) is beneficial for stable results across different eye colors. A similar ranking can also be observed when normalizing the bias scores using within-group variations in FSD. Here, MU-Net and ScleraU-Net2 exhibit the most stable performance across the eye-color subgroups, whereas CGANs2020CL again results in the largest performance differences. In Fig. 7(b) we compare the

Segment. Model	$F_1^{all} \uparrow$	Ethnicity specific F_1^{etn} $(F_1^{all} - F_1^{etn})$		Disparities	
		Caucasian	Indian	CGD ↓	FSD ↓
RGB-SS-Eye-MS	0.741	0.768 (−0.027)	0.538 (0.202)	13.0	0.526
CGANs2020CL	0.666	0.667 (−0.001)	0.656 (0.010)	**1.61**	**0.015**
ScleraU-Net2	0.701	0.734 (−0.033)	0.461 (0.240)	50.6	0.569
FCN8	0.770	0.774 (−0.004)	0.740 (0.030)	14.8	0.124
ScleraSegNet	0.620	0.653 (−0.033)	0.380 (0.240)	14.8	0.530
MU-Net†	0.253	0.262 (−0.009)	0.185 (0.067)	5.98	0.139
ScleraMaskRCNN	0.597	0.583 (0.014)	0.703 (−0.106)	9.34	0.232

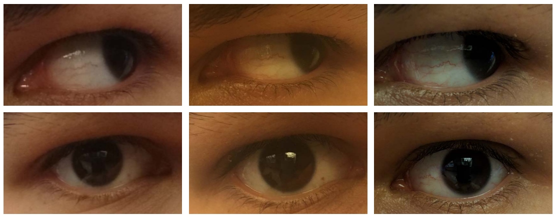
†MU-Net apparently did not converge given the training data available in this experiment

(a) Differential performance and bias disparities (ethnicity)



(b) Bias scores (ethnicity)

Fig. 8. Differential performance and bias scores with respect to ethnicity on a chimeric test dataset (MOBIUS+SMD+SLD). The best disparity score in (a) is presented in bold, the second best is underlined. In (b), the disparities (CGD and FSD) are normalized to the range of STD and MAD for visualization purposes. The mean value of each bias score is shown as a dashed line and lower values imply better performance/behavior, i.e., ↓. The figure is best viewed electronically and in color.



(a) Xperia (b) iPhone (c) Xiaomi

Fig. 9. Illustration of MOBIUS images captured with three different acquisition devices in an indoor setting. Note that the devices produce images of different characteristics in terms of color tone, sharpness, and focus.

segmentation models with respect to all four bias scores and relative to the average performance across models (dashed line). In this relative comparison, MU-Net and FCN8 are the only two models that yield below-average bias scores across all performance indicators. On the other end of the spectrum are the CGANs2020CL and ScleraSegNet models, which exhibit above-average bias scores with all considered measures, exceeding the bias scores of the best performing models by a factor of more than 2×.

Ethnicity Bias. The datasets used for the group evaluation contain individuals of different ethnicities. MOBIUS predominantly consists of Caucasian (white) subjects, whereas SMD and SLD contain subjects of Indian descent. To explore ethnicity-related bias, we construct a *chimeric dataset* from the test images in the MOBIUS, SMD and SLD datasets. We note that the three datasets also differ to some extent in image characteristics other than ethnicity (e.g., due to the capturing equipment and lighting), so cross-talk from other attributes may be present in the reported results. This cross-talk needs to be taken into account when interpreting results, but is accounted for (partially) by the disparity measures. To ensure that there is no overlap between the training and testing images, we use the *limited-training-data* (LTD) protocol for the analysis with MASD and SBVPI (having a total of 4482 images from 137 distinct subjects) serving as the training data.

From the results in Fig. 8(a)[12] we observe that several of the tested models produce considerable performance differences for subjects of different ethnicities. RGB-SS-Eye-MS, ScleraU-Net2, and ScleraSegNet, for example, show a difference of 29.9%, 37.2% and 41.8% in the ethnicity-specific F_1^{etn} scores, respectively. The most stable models in terms of performance differentials, CGANs2020CL and FCN8, on the other hand, generate F_1^{etn} differences between the two ethnicities of below 5%. When looking at the disparity measures, we notice that ethnicity induces significantly larger performance variations than eye color on average, with CGD scores reaching values above 10 for most models and FSD scores above 0.5. While the performance of the fairest (most unbiased) model, CGANs2020CL, is comparable to the best model from Fig. 7, we still observe performance differentials that are larger than in the control group. As illustrated in Fig. 8(b) three (valid) models achieve below-average performance differences across the ethnicities when considering all four bias scores, i.e., CGANs2020CL, FCN8 and ScleraMaskRCNN. It needs to be noted, though, that the overall segmentation performance is quite different for the three models, with F_1^{all} scores of 0.666, 0.770 and 0.597 for CGANs2020CL, FCN8 and ScleraMaskRCNN, respectively.

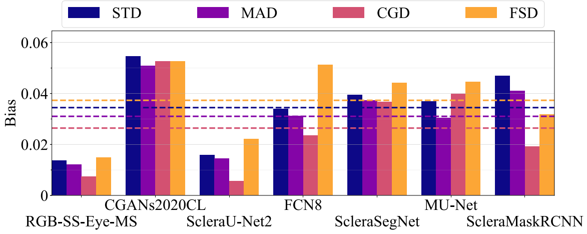
Acquisition-Hardware Bias. Next, we focus on performance differentials induced by the acquisition devices. For this part of the analysis, we again consider the test images from the MOBIUS dataset, which come from three different mobile phones. We use the CTD protocol to make sure examples from all capture devices are present in equal amounts in the training data. We note that, in general, all acquisition devices generate images of reasonable quality but with differences in color tone, sharpness and focus, as shown in Fig. 9.

The results in Fig. 10(a) show that the performance differences due to the capture device are overall larger than those originating from eye-color, but are below the differentials observed for ethnicities. In general, most models (except RGB-SS-Eye-MS) perform strongest with images from the Xiaomi phone, whereas the other two acquisition devices produce mixed rankings across the models. We see performance differences in the range of 5.1% (RGB-SS-Eye-MS) to 15.4% (CGANs2020CL) between the best and worst device-specific F_1^{hdw} scores and observe that even when normalized against reference data variations (in CGD and FSD), the acquisition device still has a considerable impact on segmentation performance. When comparing the models in terms of all four bias scores in Fig. 10(b), we notice below-average performance differentials in terms of all four scores for the RGB-SS-Eye-MS and ScleraU-Net2 models and above-average differentials for the CGANs2020CL and ScleraSegNet models. Both of the strongest models in this experiment frame the sclera segmentation task as a semantic segmentation problem and are among

[12] The MU-Net model appears to have not been trained well using the limited amount of data available in this configuration and is, therefore excluded from the analysis.

Segment. Model	F_1^{all} ↑	Hardware specific F_1^{hdw} ↑ ($F_1^{all} - F_1^{hdw}$)			Disparities	
		Xiaomi	Xperia	iPhone	CGD ↓	FSD ↓
RGB-SS-Eye-MS	0.726	0.717 (0.009)	0.744 (−0.018)	0.706 (0.021)	3.34	**0.055**
CGANs2020CL	0.707	0.793 (−0.086)	0.671 (0.036)	0.688 (0.019)	24.0	0.193
ScleraU-Net2	0.740	0.757 (−0.017)	0.744 (−0.004)	0.718 (0.022)	**2.55**	0.081
FCN8	0.800	0.856 (−0.055)	0.773 (0.028)	0.796 (0.005)	10.7	0.188
ScleraSegNet	0.748	0.810 (−0.062)	0.730 (0.018)	0.722 (0.026)	16.7	0.163
MU-Net	0.680	0.733 (−0.053)	0.646 (0.034)	0.687 (−0.007)	18.1	0.163
ScleraMaskRCNN	0.519	0.589 (−0.070)	0.479 (0.039)	0.519 (−0.000)	8.76	0.116

(a) Differential performance and bias disparities (acquisition device)

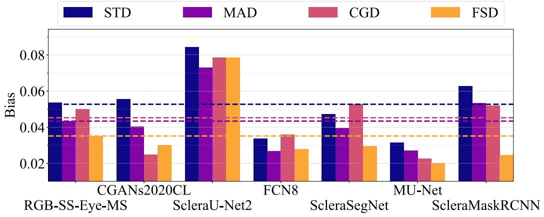


(b) Bias scores (acquisition device)

Fig. 10. Differential performance and bias scores w.r.t. acquisition hardware on the MOBIUS dataset. The best disparity CGD and FSD scores in (a) are presented in bold, the second best are underlined. In (b), the disparities (CGD and FSD) are normalized to the range of STD and MAD for visualization purposes. The mean value of each bias score is shown as a dashed line and lower values imply better performance/behavior, i.e., ↓. The figure is best viewed electronically and in color.

Segment. Model	F_1^{all} ↑	Gaze specific F_1^{gaze} ($F_1^{all} - F_1^{gaze}$)				Disparities	
		Up	Left	Straight	Right	CGD ↓	FSD ↓
RGB-SS-Eye-MS	0.726	0.815 (−0.089)	0.712 (0.014)	0.664 (0.062)	0.714 (0.013)	10.7	0.230
CGANs2020CL	0.707	0.782 (−0.075)	0.711 (−0.004)	0.626 (0.081)	0.710 (−0.003)	10.8	0.189
ScleraU-Net2	0.740	0.785 (−0.045)	0.791 (−0.051)	0.596 (0.144)	0.790 (−0.050)	18.9	0.485
FCN8	0.800	0.855 (−0.054)	0.796 (0.004)	0.761 (0.039)	0.790 (0.010)	**4.42**	0.177
ScleraSegNet	0.748	0.830 (−0.082)	0.737 (0.011)	0.704 (0.044)	0.722 (0.026)	10.3	0.197
MU-Net	0.680	0.735 (−0.055)	0.655 (0.024)	0.663 (0.017)	0.666 (0.014)	12.5	**0.131**
ScleraMaskRCNN	0.519	0.630 (−0.111)	0.468 (0.050)	0.498 (0.021)	0.479 (0.040)	6.08	0.168

(a) Differential performance and bias disparities (eye gaze)



(b) Bias scores (eye gaze)

Fig. 11. Differential performance and bias scores with respect to eye gaze evaluated on the MOBIUS dataset. For the disparities in (a) the best score is presented in bold, the second best is underlined. In (b), the disparities (CGD and FSD) are normalized to the range of STD and MAD for visualization purposes. The mean value of each bias score is shown as a dashed line and lower values imply better performance/behavior, i.e., ↓. The figure is best viewed electronically and in color.

the lighter models in terms of trainable parameters, which helps to generate stable segmentation results with limited performance variations across different capture devices.

Gaze-Direction Bias. Another potential source of differences in segmentation performance is the gaze direction, in which the eye was imaged. To explore the impact of gaze on segmentation performance, we conduct an analysis on the test images from the MOBIUS dataset with (stratified) subgroups that correspond to images captured with subjects looking up, left, straight, and right. The models scored in the analysis are trained using the CTD protocol, so all gaze directions are well represented in the training data, as both the MASD and SBVPI datasets (which form the majority of the training data in the CTD protocol) contain images with varying gaze directions – refer to Table I for details on the dataset characteristics.

The results in Fig. 11(a) show that the worst performance is fairly consistently achieved with the *straight* gaze direction, where straight-gaze-specific F_1^{gaze} scores between 2.5% (MU-Net) and 19.5% (ScleraU-Net2) below the average performance across all images F_1^{all} are observed. Note that this result appears despite the fact that the straight gaze direction is slightly overrepresented in the provided training data (since the 500 images in SMD are captured in the straight gaze direction only). The upwards gaze direction consistently results in the best segmentation performance, with upward-gaze-specific F_1^{gaze} scores between 6.1% (ScleraU-Net2) and 21.4% (ScleraMaskRCNN) above F_1^{all}. The left and right directions result in roughly equivalent performances across the board, mostly falling between the performances on the upwards- and the straight-gaze direction.

The (poor) results with the straight-gaze direction can be attributed to the fact that under this direction, the sclera commonly appears in the form of two distinct areas of roughly the same size with possibly different brightness values – due to the external illumination conditions. This also explains why the worst individual performances were observed with the sunny and well-lit samples in Fig. 6, even though the images captured in sunny weather and well-lit rooms achieve better segmentation performance than images captured in poorly-lit rooms on average [1]. Conversely, with the upwards-gaze direction the sclera typically takes the form of a single contigous area with potential gradual changes in illumination and contrast, leading to above-average segmentation results. The weaker results with the straight-direction images are somewhat in conflict with the feature extraction stage, where matching with the straight-gaze direction images was shown to lead to better recognition results than matching with other gaze direction images [5]. This implies that in real-world scenarios, where sclera segmentation is still a relatively difficult problem (unlike in the laboratory conditions explored in [5]), a balance in the performance of the methods addressing these two steps has to be achieved for a successful overall recognition pipeline.

As all models exhibit a CGD score significantly above 1, we can conclude that the variability in segmentation performance due to gaze variations is larger than the variability seen in randomly sampled subgroups. The lowest performance differentials are seen with the FCN8 model with a CGD score of 4.42 and the largest for the ScleraU-Net2 approach with a CGD value of 18.9. A similar ranking can also be observed when normalizing the bias scores using within-group variations in FSD. Here, MU-Net and ScleraMaskRCNN exhibit the most stable performance across the gaze-direction subgroups, whereas ScleraU-Net2 again results in the largest performance differences. In Fig. 11(b) we compare the segmentation models with respect to all four bias scores and relative to the average performance across models (dashed line). In this relative comparison, FCN8 and ScleraSegNet are the only two models that yield below-average bias scores across all performance

TABLE IV

DIFFERENTIAL PERFORMANCE WITH RESPECT TO CHANGES IN THE TRAINING DATA (SBVPI VS. MASD+SBVPI). RESULTS ARE COMPUTED FROM THE BINARY SEGMENTATION MASKS AND PRESENTED IN THE FORM $F_1^{etn}(F_1 - F_1^{etn})$. RESULTS ARE REPORTED WITH THE IMAGES FROM MOBIUS, SMD, AND SLD SERVING AS THE TEST DATA

Model	**Trained on SBVPI**		**Trained on MASD+SBVPI**	
	Caucasian	**Indian**	**Caucasian**	**Indian**
RGB-SS-Eye-MS	0.675 (−0.024)	0.477 (0.175)	0.768 (−0.027)	0.538 (0.202)
CGANs2020CL	0.560 (−0.011)	0.466 (0.082)	0.667 (−0.001)	0.656 (0.010)
ScleraU-Net2	0.732 (−0.028)	0.496 (0.208)	0.734 (−0.033)	0.461 (0.240)
FCN8	0.798 (−0.007)	0.738 (0.053)	0.774 (−0.004)	0.740 (0.030)
ScleraSegNet	0.503 (−0.016)	0.366 (0.121)	0.653 (−0.033)	0.380 (0.240)
MU-Net[†]	0.390 (−0.017)	0.251 (0.123)	0.262 (−0.009)	0.185 (0.067)
ScleraMaskRCNN	0.631 (−0.013)	0.519 (0.099)	0.583 (0.014)	0.703 (−0.106)

[†] MU-Net apparently did not converge given the training data available in this experiment

indicators. On the other end of the spectrum is ScleraU-Net2, which exhibits above-average bias scores with all considered measures, exceeding the bias scores of the best performing models by a factor of 4×.

2) Representation Bias: Performance differentials across different data subgroups are often ascribed to biased (or unbalanced) training data [13], [15]. To provide insight into this issue, we study the representation (or sampling) bias in the context of performance differentials induced by ethnicities in the next series of experiments. For the analysis we consider two configurations of the limited-training-data (LTD) experimental protocol: (i) in the first configuration, the SBVPI data, with exclusively Caucasian subjects, (i.e., 1858 images from 55 subjects) is used as the training data, and (ii) in the second configuration, MASD (having exclusively Indian subjects) as well as SBVPI (for a combined 4482 images from 137 subjects) are utilized for training. The test set consists of images from the MOBIUS, SMD and SLD datasets, which were not seen during training. We note that the MU-Net model did not converge properly using the limited amount of training data available in this experiment and is, therefore, excluded from the following analysis.

Several observations can be made from the results in Table IV: (i) The segmentation performance increases for both ethnicities in terms of F_1^{etn} scores when adding more training data for the majority of models, i.e., SBVPI → MASD+SBVPI, suggesting that the added training samples contribute towards better segmentation results. (ii) The performance with Caucasian subjects is consistently higher for all models regardless of the training data used. The only notable exception here is ScleraMaskRCNN, which performs better with Indian subjects when the mixed-ethnicity data is used for training. (iii) While the performance differentials between the two ethnicities range between 7.5% (FCN8) and 32.2% (ScleraU-Net2) in terms of F_1^{etn} scores when the models are trained on the SBVPI data, the range of performance differentials changes to between 1.65% (CGANs2020CL) and 41.8% (ScleraSegNet), when both MASD and SBVPI are utilized for the training procedure. As also seen from Fig. 12, where differences in the bias scores due to the training data are presented, i.e., Bias$\Delta = \psi_{masd+sbvpi} - \psi_{sbvpi}$; $\psi \in \{\text{STD}, \text{MAD}, \text{CGD}, \text{FSD}\}$, several models (CGANs2020CL and FCN8) are able to significantly reduce the performance differences with more representative training examples in addition to improving their overall F_1 scores, whereas others (e.g., RGB-SS-Eye-MS, ScleraU-Net2, and ScleraSegNet) improve segmentation performance but also increase the differences in the ethnicity-specific F_1^{etn} values. This observation is consistent with prior work studying representation bias in other problem domains, e.g., [72] – informative training data may help to reduce performance differentials with well-designed and trained models, but this is by no means guaranteed.

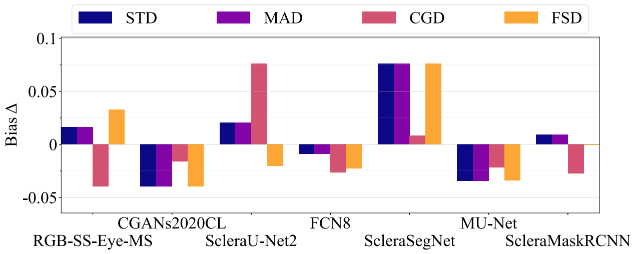


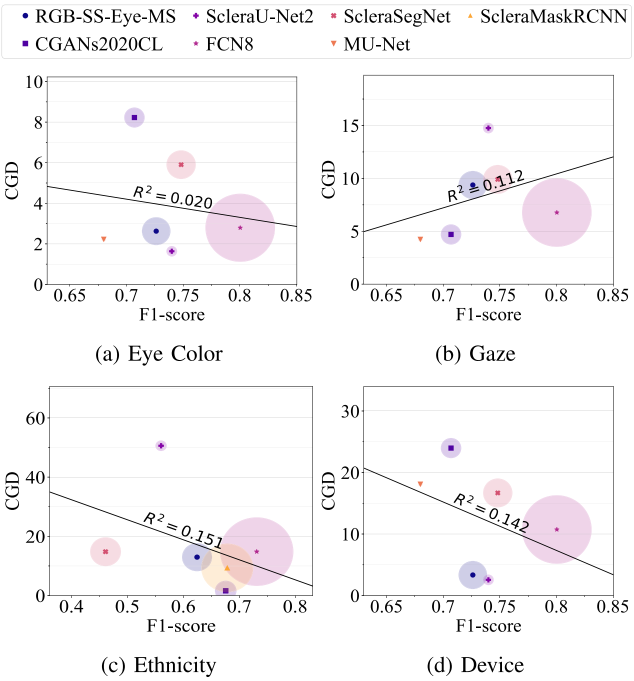
Fig. 12. Differences in the ethnicity-induced bias scores generated with two different training configurations, one containing only Caucasian subjects, and the other containing Caucasian as well as Indian subjects. Positive values imply larger bias scores were observed with the mixed-ethnicity training configuration, while negative values mean that larger bias scores were observed with the Caucasian-only training configuration.

C. Bias vs. Segmentation Performance

In the final analysis we investigate the relationship between algorithmic bias across eye color, gaze direction, ethnicity, and capture device and the overall segmentation performance. The analysis for each of the four factors is conducted with the same experimental setup in terms of training and testing data as in the corresponding experiments from Section V-B. Thus, all models are trained on the same data to ensure a fair evaluation.

In Fig. 13 we plot the calculated CGD disparities against the F_1 scores for each experiment and as a function of the model size, i.e., the number of model parameters. The figure, thus, captures the trade-off the developed models offer in terms of bias, segmentation performance and model footprint. In an ideal setting, the models would have low bias (CGD scores on the y-axis), high performance (F_1 scores on the x-axis) and a low parameter count (circle areas), and would as such be located at the lower right in the presented graphs. To capture the relationships between the bias and performance scores, we fit a line to the data points in a least-squares manner. Since certain models performed much worse on certain training configurations, possibly due to insufficient training or errors in training data handling (see for instance Fig. 8(a) and Table IV), we eliminate outliers with an F_1 z-score above 2 (i.e., models with an F_1 score that is more than 2 standard deviations from the mean F_1 score) and fit the line to the remaining data.

As can be seen, there is a weak but consistent negative correlation between the performance differentials and overall segmentation performance for all considered factors. This suggests that better performing models tend to produce smaller performance differences over data subgroups. With improvements in visual segmentation techniques, reductions in the performance differentials may, therefore, also be expected. Active research on reducing bias with existing models is nevertheless a key concern going forward. If we look at the performance-bias trade-off from the perspective of model size, we observe that the largest model, FCN8, is consistently among the best models located at the bottom right of Fig. 13, while the smaller models (MU-Net, CGANs2020CL, and ScleraU-Net2) trend more toward the left (low perfor-



(a) Eye Color (b) Gaze

(c) Ethnicity (d) Device

Fig. 13. Scatter plots of the CGD disparities relative to the F_1 values achieved by the models. Lines fitted to the points in the graphs are also shown, along with their corresponding R^2 scores. The areas of the circles around the points represent the model sizes in terms of the number of parameters.

mance) and top (high bias) of the graphs. This observation may suggest that model scaling can also have a beneficial impact on the segmentation models, similarly to what has been observed recently in other areas, where larger models were found to have a significant edge over their smaller counterparts [73], [74], [75].

VI. Discussion

The following observations were made based on the available sample of models with respect to the research questions laid out in the introductory section of the paper.

Q1: How well do contemporary sclera-segmentation models perform with diverse input images?

Significant performance differences were observed across the evaluated models. While many of the best performing models (RGB-SS-Eye-MS, CGANs2020CL, FCN8, ScleraU-Net2 or ScleraSegNet) achieved F_1 scores above 0.7 on the challenging MOBIUS dataset (mobile setting, different devices, gaze directions and environments), some of the weaker models yielded F_1 scores closer to 0.5. Similarly, on the newly collected SLD dataset, F_1 scores varied from above 0.8 for the strongest models to 0.55 for the weakest one. Nonetheless, the results suggest that given the current state of technology, it is possible to train segmentation models that generalize well across data characteristics and produce usable segmentation results even with challenging input images.

Q2: What are the most critical sources of bias?

Different characteristics were taken into account when exploring algorithmic bias with the developed segmentation models, including eye color, ethnicity, acquisition hardware and gaze direction. The largest performance differences were observed across ethnicities, where 6 out of 7 tested models exhibited a clear preference for Caucasian subjects, despite the fact that the ethnicity groups were equally represented in the training data (with a slight under-representation for Caucasian subjects). The bias due to eye color was overall the lowest in our experiments. Nonetheless, all 7 models performed worst with green eyes and 6 out of the 7 models performed best with gray eyes, suggesting that eye color represents a systematic (yet limited) source of algorithmic bias in sclera segmentation models. The bias scores observed with different acquisition devices were overall higher than what was observed due to eye color in the experiments, but the ranking w.r.t. devices was not consistent across the segmentation models. While all 7 models performed best with images captured by the Xiaomi phone, the ranking on the other two phones was mixed, implying that, while the acquisition hardware is still a significant source of bias, various segmentation methods respond differently to the image characteristics introduced by the capturing hardware. The bias scores observed for gaze directions were comparable to the scores observed for the acquisition devices, with 6 out of 7 models exhibiting the worst performance with the straight-gaze directions, and similarly 6 out of 7 performing best for the upwards-gaze direction, again pointing to the presence of systematic bias with respect to gaze directions.

Q3: What impact do training data characteristics have on the bias exhibited by the segmentation models?

To study the impact of training data characteristics on the segmentation accuracy and ethnicity bias, two different training configurations were explored: (i) one that contained only Caucasian subjects, and (ii) another one that contained an approximately balanced number of images of Indian and Caucasian subjects. Two main observation were made. The overall segmentation performance of 6 of the 7 models improved with the larger, more representative training dataset. However, only 3 out of the 7 models managed to also reduce the performance differences between the Caucasian and Indian subjects, for 2 models the results were mixed, whereas for the last 2, the bias in fact increased with the balanced dataset. This confirms prior observations [29], [72] that balanced datasets do not automatically lead to unbiased performance, as algorithmic bias is not necessarily related only to unbalanced training data.

Q4: Can we mitigate algorithmic bias without degrading segmentation performance?

There appeared to be a consistent (albeit weak) negative correlation between segmentation accuracy and the CGD bias score across all of our bias experiments. This implies that improving the overall segmentation performance of the models also simultaneously reduces its inherent bias on average. Advances in semantic segmentation can therefore be expected to also address bias and fairness issues to a certain degree.

VII. Conclusion

In this paper, we presented the results of a group evaluation, organized to benchmark the performance and bias of sclera segmentation models under a common experimental setting. Seven research groups participated in the effort and contributed seven distinct models to the evaluation for scoring.

The results of the group evaluation suggest that contemporary models are able to ensure useful segmentation performance with diverse input images and that more accurate models consistently also achieve lower bias scores with respect to different factors. Increasing the model complexity was also observed to lead to better performance and lower bias. Given such results, recent advances in modern model architectures (such as transformers) may help provide better performance-bias trade-offs in the future. However, note that the improvement in this case may come at the cost of higher memory usage and computational intensity, which could be problematic for applications running on less-capable hardware.

As part of our future work, we plan to explore correlation-based measures for quantifying bias, applicable to groups of (machine learning) models. Such measures are expected to ensure additional insights into the behavior of the models and help identify important trends and model/data characteristics affecting performance and performance differentials across subgroups of data.

Acknowledgment

Matej Vitek and Peter Peer are with the Computer Vision Laboratory, Faculty of Computer and Information Science, University of Ljubljana (UL), 1000 Ljubljana, Slovenia (e-mail: matej.vitek@fri.uni-lj.si).

Abhijit Das is with the Department of Computer Science and Information Science, Birla Institute of Technology and Science Pilani (BITS), Pilani 333031, India.

Diego Rafael Lucio, Luiz Antonio Zanlorensi Jr., and David Menotti are with the Department of Informatics, Federal University of Paraná (UFPR), Curitiba 81531-980, Brazil.

Jalil Nourmohammadi Khiarak is with the Institute of Control and Computation Engineering, Warsaw University of Technology (WUT), 00-661 Warszawa, Poland.

Mohsen Akbari Shahpar and Meysam Asgari-Chenaghlu are with the Department of Computer Engineering, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz 51368, Iran.

Farhang Jaryani is with the Faculty of Engineering, Arak University (AU), Arak 802111, Iran.

Juan E. Tapia is with the Biometrics and Internet-Security Research Group (da/sec), Hochschule Darmstadt (HDA), 64295 Darmstadt, Germany.

Andres Valenzuela is with R&D Center, TOC Biometrics (TOC), Santiago 7510099, Chile.

Caiyong Wang is with the School of Electrical and Information Engineering, University of Civil Engineering and Architecture (BUCEA), Beijing 100044, China.

Yunlong Wang and Zhenan Sun are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China.

Zhaofeng He is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China.

Fadi Boutros, Naser Damer, and Arjan Kuijper are with the Smart Living and Biometric Technologies Department, Fraunhofer Institute for Computer Graphics Research (IGD), 64283 Darmstadt, Germany, and also with the Department of Computer Science, Technical University of Darmstadt (TUD), 64289 Darmstadt, Germany.

Jonas Henry Grebe is with the Smart Living and Biometric Technologies Department, Fraunhofer Institute for Computer Graphics Research (IGD), 64283 Darmstadt, Germany.

Kiran Raja and Gourav Gupta are with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway.

Georgios Zampoukis, Lazaros Tsochatzidis, and Ioannis Pratikakis are with the Department of Electrical and Computer Engineering, Democritus University of Thrace (DUTH), 67100 Xanthi, Greece.

S. V. Aruna Kumar is with the Department of Computer Science and Engineering, Malnad College of Engineering, Hassan, Karnataka 573202, India.

B. S. Harish is with the Department of Information Science and Engineering, JSS Science and Technology University (JSS STU), Mysuru 570006, India.

Umapada Pal is with the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute (ISI) Kolkata, Kolkata 700108, India.

Vitomir Štruc is with the Laboratory for Machine Intelligence, Faculty of Electrical Engineering, University of Ljubljana (UL), 1000 Ljubljana, Slovenia.

References

[1] M. Vitek et al., "SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, 2020, pp. 1–10.

[2] I. Nigam, M. Vatsa, and R. Singh, "Ocular biometrics: A survey of modalities and fusion approaches," *Inf. Fusion*, vol. 26, pp. 1–35, Nov. 2015.

[3] P. Rot, Z. Emersic, V. Struc, and P. Peer, "Deep multi-class eye segmentation for ocular biometrics," in *Proc. IEEE Int. Work Conf. Bioinspired Intell. (IWOBI)*, Jul. 2018, pp. 1–8.

[4] S. Das, I. D. Ghosh, and A. Chattopadhyay, "An efficient deep sclera recognition framework with novel sclera segmentation, vessel extraction and gaze detection," *Signal Process., Image Commun.*, vol. 97, Sep. 2021, Art. no. 116349.

[5] M. Vitek, P. Rot, V. Štruc, and P. Peer, "A comprehensive investigation into sclera biometrics: A novel dataset and performance study," *Neural Comput. Appl.*, vol. 32, pp. 1–15, Feb. 2020.

[6] D. Yadav, N. Kohli, J. Doyle, R. Singh, M. Vatsa, and K. Bowyer, "Unraveling the effect of textured contact lenses on iris recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 5, pp. 851–862, Mar. 2014.

[7] P. Radu, J. Ferryman, and P. Wild, "A robust sclera segmentation algorithm," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Mar. 2015, pp. 1–6.

[8] S. Alkassar, W. Woo, S. Dlay, and J. Chambers, "Sclera recognition: On the quality measure and segmentation of degraded images captured under relaxed imaging conditions," *IET Biometrics*, vol. 6, no. 4, pp. 266–275, Jul. 2017.

[9] P. Rot, M. Vitek, K. Grm, V. Z. Emeršič, P. Peer, and V. Štruc, "Deep sclera segmentation and recognition," in *Handbook of Vascular Biometrics*. Cham, Switzerland: Springer, 2020, pp. 395–432.

[10] D. Riccio, N. Brancati, M. Frucci, and D. Gragnaniello, "An unsupervised approach for eye sclera segmentation," in *Proc. Iberoamerican Congr. Pattern Recognit.* New York, NY, USA: Springer, 2017, pp. 550–557.

[11] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein, "A decision-level fusion strategy for multimodal ocular biometric in visible spectrum based on posterior probability," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 794–798.

[12] V. Gottemukkula, S. Saripalle, S. P. Tankasala, and R. Derakhshani, "Method for using visible ocular vasculature for mobile biometrics," *IET Biometrics*, vol. 5, no. 1, pp. 3–12, 2016.

[13] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Trans. Technol. Soc.*, vol. 1, no. 2, pp. 89–103, Jun. 2020.

[14] EuropeanCommission. (Apr. 2021). *The Artificial Intelligence Act*. [Online]. Available: https://artificialintelligenceact.eu/the-act/

[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022.

[16] O. A. Osoba and W. Welser, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, CA, USA: Rand Corporation, 2017.

[17] P. Grother, M. Ngan, and K. Hanaoka, *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. Gaithersburg, MD, USA: NIST, 2019.

[18] J. P. Robinson, C. Qin, Y. Henon, S. Timoner, and Y. Fu, "Balancing biases and preserving privacy on balanced faces in the wild," 2021, *arXiv:2103.09118*.

[19] V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer, "Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 127–137, 2021.

[20] A. Puc, V. Struc, and K. Grm, "Analysis of race and gender bias in deep age estimation models," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 830–834.

[21] Z. Babnik and V. Struc, "Assessing bias in face image quality assessment," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2022, pp. 1–5.

[22] B. Meden et al., "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4147–4183, 2021.

[23] R. Ramachandra, K. Raja, and C. Busch, "Algorithmic fairness in face morphing attack detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 410–418.

[24] S. K. Modi, S. J. Elliott, J. Whetsone, and H. Kim, "Impact of age groups on fingerprint recognition performance," in *Proc. IEEE Workshop Autom. Identificat. Adv. Technol.*, Apr. 2007, pp. 19–23.

[25] S. Yoon and A. K. Jain, "Longitudinal study of fingerprint recognition," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 28, pp. 8555–8560, Jul. 2015.

[26] P. Drozdowski et al., "Demographic bias: A challenge for fingervein recognition systems?" in *Proc. EUSIPCO*, 2021, pp. 825–829.

[27] A. Uhl and P. Wild, "Comparing verification performance of kids and adults for fingerprint, palmprint, hand-geometry and digitprint biometrics," in *Biometrics: Theory, Applications, and Systems (BTAS)*. Piscataway, NJ, USA: IEEE Press, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5339069

[28] H. Wu, V. Albiero, K. S. Krishnapriya, M. C. King, and K. W. Bowyer, "Face recognition accuracy across demographics: Shining a light into the problem," 2022, *arXiv:2206.01881*.

[29] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 330–347.

[30] P. Terhörst, "Mitigating soft-biometric driven bias and privacy concerns in face recognition systems," Ph.D. dissertation, Dept. Comput. Sci., Univ. Darmstadt, Darmstadt, Germany, 2021.

[31] A. Das, A. Dantcheva, and F. Bremond, "Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach," in *Proc. ECCV Workshops*, Aug. 2018, pp. 1–5.

[32] A. Krishnan, A. Almadan, and A. Rattani, "Investigating fairness of ocular biometrics among young, middle-aged, and older adults," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2021, pp. 1–7.

[33] A. Krishna, "Probing fairness of mobile ocular biometrics methods across gender on VISOB 2.0 dataset," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 229–243.

[34] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper, "Demographic bias in presentation attack detection of iris recognition systems," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 835–839.

[35] D. O. Gorodnichy and M. P. Chumakov, "Analysis of the effect of ageing, age, and other factors on iris recognition performance using Nexus scores dataset," *IET Biometrics*, vol. 8, no. 1, pp. 29–39, Jan. 2019.

[36] A. Das, U. Pal, M. A. F. Ballester, and M. Blumenstein, "Multiangle based lively sclera biometrics at a distance," in *Proc. IEEE Symp. Comput. Intell. Biometrics Identity Manag. (CIBIM)*, Dec. 2014, pp. 22–29.

[37] A. Das, "Towards multi-modal sclera and iris biometric recognition with adaptive liveness detection," Ph.D. dissertation, School Inf. Commun. Technol., Griffith Univ., Brisbane, QLD, Australia, 2017. [Online]. Available: https://www.griffith.edu.au/griffith-sciences/school-information-communication-technology/contact-us

[38] R. Derakhshani, A. Ross, and S. Crihalmeanu, "A new biometric modality based on conjunctival vasculature," in *Proc. Artif. Neural Netw. Eng.*, 2006, pp. 1–8.

[39] R. Derakhshani and A. Ross, "A texture-based neural network classifier for biometric identification using ocular surface vasculature," in *Proc. Int. Joint Conf. Neural Netw.*, Aug. 2007, pp. 2982–2987.

[40] S. Crihalmeanu, A. Ross, and R. Derakhshani, "Enhancement and registration schemes for matching conjunctival vasculature," in *Proc. Int. Conf. Biometrics*. Cham, Switzerland: Springer, 2009, pp. 1240–1249.

[41] S. P. Tankasala, P. Doynov, R. R. Derakhshani, A. Ross, and S. Crihalmeanu, "Biometric recognition of conjunctival vasculature using GLCM features," in *Proc. Int. Conf. Image Inf. Process.*, Nov. 2011, pp. 1–6.

[42] A. Das, U. Pal, M. A. F. Ballester, and M. Blumenstein, "A new method for sclera vessel recognition using OLBP," in *Proc. Chin. Conf. Biometric Recognit.*, 2013, pp. 370–377.

[43] K. Oh and K.-A. Toh, "Extracting sclera features for cancelable identity verification," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 245–250.

[44] R. A. Naqvi and W.-K. Loh, "Sclera-net: Accurate sclera segmentation in various sensor images based on residual encoder and decoder network," *IEEE Access*, vol. 7, pp. 98208–98227, 2019.

[45] D. R. Lucio, R. Laroca, E. Severo, A. S. Britto, and D. Menotti, "Fully convolutional networks and generative adversarial networks applied to sclera segmentation," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–7.

[46] J. E. Tapia, E. L. Droguett, A. Valenzuela, D. P. Benalcazar, L. Causa, and C. Busch, "Semantic segmentation of periocular near-infra-red eye images under alcohol effects," *IEEE Access*, vol. 9, pp. 109732–109744, 2021.

[47] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi, "OpenEDS: Open eye dataset," 2019, *arXiv:1905.03702*.

[48] A. Das et al., "Sclera segmentation benchmarking competition in cross-resolution environment," in *Proc. Int. Conf. Biometrics*, 2019, pp. 1–7.

[49] A. Dasa, U. Palb, M. A. Ferrerc, and M. Blumensteina, "SSBC 2015: Sclera segmentation benchmarking competition," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2015, pp. 1–10.

[50] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein, "SSRBC 2016: Sclera segmentation and recognition benchmarking competition," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–6.

[51] A. Das et al., "SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, 2017, pp. 742–747.

[52] A. Das et al., "SSBC 2018: Sclera segmentation benchmarking competition," in *Proc. Int. Conf. Biometrics (ICB)*, 2018, pp. 1–4.

[53] O. Golob, P. Peer, and M. Vitek, "Semi-automated correction of MOBIUS eye region annotations," in *Proc. IEEE Int. Electrotech. Comput. Sci. Conf. (ERK)*, Dec. 2020, pp. 344–347.

[54] V. Z. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer, "Pixel-wise ear detection with convolutional encoder–decoder networks," *IET Biometrics*, vol. 7, no. 3, pp. 1–13, Feb. 2017.

[55] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, Feb. 2011.

[56] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0118432.

[57] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Proc. JECMLKDD*, 2013, pp. 134–145.

[58] P. Saleiro et al., "Aequitas: A bias and fairness audit toolkit," 2018, *arXiv:1811.05577*.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[60] C. Wang, Y. Wang, Y. Liu, Z. He, R. He, and Z. Sun, "ScleraSegNet: An attention assisted U-Net model for accurate sclera segmentation," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 1, pp. 40–54, Jan. 2019.

[61] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[62] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[63] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[64] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[65] M. Teichmann, M. Weber, J. M. Zöllner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," 2016, *arXiv:1612.07695*.

[66] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[68] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Eye-MMS: Miniature multi-scale segmentation network of key eye-regions in embedded applications," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–6.

[69] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

[70] J. Lozej, B. Meden, V. Struc, and P. Peer, "End-to-end iris segmentation using U-Net," in *Proc. IEEE Int. Work Conf. Bioinspired Intell. (IWOBI)*, Jul. 2018, pp. 1–6.

[71] R. Baeza-Yates, "Bias on the web," *Commun. ACM*, vol. 61, no. 6, pp. 54–61, 2018.

[72] V. Albiero, K. Zhang, and K. W. Bowyer, "How does gender balance in training data affect face recognition accuracy?" in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[73] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 1877–1990.

[74] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12104–12113.

[75] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, "Scalable vision transformers with hierarchical pooling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 377–386.