SSBC 2020: Sclera Segmentation Benchmarking Competition in the Mobile Environment

M. Vitek[1,*], A. Das[2,*], Y. Pourcenoux[3], A. Missler[4], C. Paumier[4], S. Das[5], I. De Ghosh[6], D. R. Lucio[7], L. A. Zanlorensi Jr.[7], D. Menotti[7], F. Boutros[8,9], N. Damer[8,9], J. H. Grebe[8], A. Kuijper[8,9], J. Hu[10], Y. He[10], C. Wang[11], H. Liu[10], Y. Wang[10], Z. Sun[10], D. Osorio-Roig[12], C. Rathgeb[12], C. Busch[12], J. Tapia[13], A. Valenzuela[14], G. Zampoukis[15], L. Tsochatzidis[15], I. Pratikakis[15], S. Nathan[16], R. Suganya[17], V. Mehta[18], A. Dhall[18,19], K. Raja[20], G. Gupta[20], J. N. Khiarak[21], M. Akbari-Shahper[22], F. Jaryani[23], M. Asgari-Chenaghlu[22], R. Vyas[24], S. Dakshit[25], S. Dakshit[26], P. Peer[1], U. Pal[2], V. Štruc[1]

[1]University of Ljubljana (UL, SI), [2]Indian Statistical Institute (ISI, IN), [3]Grenoble Institute of Technology (INP, FR), [4]École Nationale Supérieure de l'Électronique et de ses Applications (ENSEA, FR), [5]University of Engineering and Management (UEM, IN), [6]Barrackpore Rastraguru Surendranath College (BRSNC, IN), [7]Federal University of Paraná (UFPR, BR), [8]Fraunhofer Institute for Computer Graphics Research (IGD, DE), [9]Technische Universität Darmstadt (TUD, DE), [10]Chinese Academy of Sciences (CAS, CN), [11]Beijing University of Civil Engineering and Architecture (BUCEA, CN), [12]Hochschule Darmstadt (HDA, DE), [13]Universidad de Santiago (USACH, CL), [14]TOC Biometrics (CL), [15]Democritus University of Thrace (DUTH, GR), [16]Couger Inc. (JP), [17]Thiagarajar College of Engineering (TCE, IN), [18]Indian Institute of Technology Ropar (IIT-RPR, IN), [19]Monash University (MU, AU), [20]Norwegian University of Science and Technology (NTNU, NO), [21]Warsaw University of Technology (WUT, PL), [22]Tabriz University (TU, IR), [23]Arak University (AU, IR), [24]Bennet University (BU, IN), [25]Calcutta Institute of Engineering and Management, Maulana Abul Kalam University of Technology (MAKAUT, IN), [26]University of Texas at Dallas (UTD, US)

Abstract

The paper presents a summary of the 2020 Sclera Segmentation Benchmarking Competition (SSBC), the 7th in the series of group benchmarking efforts centred around the problem of sclera segmentation. Different from previous editions, the goal of SSBC 2020 was to evaluate the performance of sclera-segmentation models on images captured with mobile devices. The competition was used as a platform to assess the sensitivity of existing models to i) differences in mobile devices used for image capture and ii) changes in the ambient acquisition conditions. 26 research groups registered for SSBC 2020, out of which 13 took part in the final round and submitted a total of 16 segmentation models for scoring. These included a wide variety of deep-learning solutions as well as one approach based on standard image processing techniques. Experiments were conducted with three recent datasets. Most of the segmentation models achieved relatively consistent performance across images captured with different mobile devices (with slight differences across devices), but struggled most with low-quality images captured in challenging ambient conditions, i.e., in an indoor environment and with poor lighting.

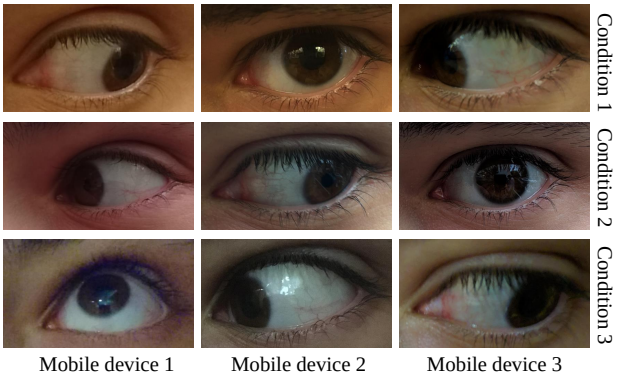


Figure 1: The quality of ocular images captured by mobile devices depends heavily on the imaging sensor used and the ambient acquisition conditions present during the capturing process - as illustrated by the sample images above. The goal of SSBC 2020 is to evaluate the performance of different sclera segmentation models across different capturing devices and acquisition conditions in the largest group benchmarking effort in this problem domain so far.

1. Introduction

Ocular biometrics represent a branch of biometric recognition technology that exploits various eye characteristics for identity inference. Recognition techniques based on ocular traits are particularly important for authentication schemes on mobile devices, where reliable (and convenient) mechanisms for ascertaining the identity of users are of paramount importance. While the field has long been dominated by iris recognition techniques, recent research is increasingly looking at alternative traits that can either be used as stand-alone modalities or, alternatively, in combination with the iris. Among the different ocular traits available, the sclera region is particularly appealing due to desir-

*M. Vitek and A. Das are first authors with equal contributions.

978-1-7281-9186-7/20/$31.00

able characteristics [1], such as high discriminability, permanence and robustness to presentation attacks [2, 3].

An increasing amount of research directed towards exploring sclera as a biometric trait has been presented in the literature in the last few years. This includes techniques for recognition [4, 5, 6, 7, 8, 3], segmentation [2, 9, 10], presentation attacks detection [2, 3, 6], adaptability of the trait [11, 12], information fusion with the iris [13, 14] and synthetic sclera generation [15]. However, despite this research effort, comprehensive studies investigating the characteristics of sclera biometrics in mobile scenarios are still limited in the literature. As illustrated in Figure 1, the quality of images captured in such scenarios very much depends on the specific device (and consequently imaging sensors) used and the external conditions present during the acquisition step. It is important to understand how these sources of variability affect various components of sclera-based recognition techniques and what kind of performance degradation can be expected due to changes in the mobile device used and ambient conditions present during image capture.

To explore these issues, the 2020 Sclera Segmentation Benchmarking Competition (SSBC) was organised as part of the International Joint Conference on Biometrics (IJCB 2020). The competition focused on the task of sclera segmentation, which is a key component and typically the first step in sclera biometric pipeline. Improper segmentation affects the performance of all downstream tasks, including image normalisation, feature extraction and recognition. Consequently, the competition was organised around a novel dataset the *Mobile Ocular Biometrics in Unconstrained Settings (MOBIUS)*, that features challenging ocular/sclera images captured in unconstrained settings with three mobile devices and in three acquisition conditions. The dataset allowed to explore a number of research questions within SSBC 2020, such as: How do contemporary segmentation models perform with challenging images captured with mobile devices? What impact do changes in imaging sensors have on segmentation performance? How does the external imaging conditions affect segmentation accuracy? To answer these questions 16 segmentation models were contributed to the competition from 13 different research groups and analysed for their performance.

The joint effort of all participating groups resulted in the following contributions that are presented in this paper:

- A rigorous evaluation of several contemporary (sclera) segmentation models using images captured in challenging mobile scenarios.
- A comprehensive sensitivity analysis of the segmentation models to variations in acquisition device and external capturing conditions.
- A novel dataset of ocular images designed for research into segmentation models for ocular biometrics in mobile scenarios.

2. Related Work

SSBC 2020 represents the 7th edition of the annual sclera segmentation competition started initially in 2015 as part of the BTAS 2015. The series of competitions had a considerable impact on the state of technology in sclera segmentation and over the years provided popular benchmarks for the community.

The 1st and 3rd iteration of SSBC introduced a novel dataset for sclera segmentation [16]. The 2nd iteration (SSRBC 2016) also studied sclera recognition techniques in addition to segmentation models [17]. The 4th iteration, SSERBC 2017, focused on sclera segmentation and eye recognition with different gaze directions [18]. The 5th, SSBC 2018, explored the impact of cross senor image capture on sclera segmentation, and the 6th iteration, SSBC 2019, investigated the performance of segmentation models in cross-resolution settings [19, 20].

The current edition of SSBC aims to continue the series of competitions with a new segmentation problem relevant in the context of mobile biometrics. By introducing a novel dataset for this task, it also makes an important contribution to the community.

3. Benchmarking methodology

This section presents the benchmarking methodology used for SSBC 2020. It first describes the datasets used in the challenge and then elaborates on the logistics of the competition and performance metrics utilised for the final comparative evaluation.

3.1. SSBC 2020 datasets

Three datasets were used for SSBC 2020: *i)* the *Multi-Angle Sclera Dataset* (MASD) [21], *ii)* the *Sclera Mobile Dataset* (SMD) [22] and *iii)* the *Mobile Ocular Biometrics in Unconstrained Settings* (MOBIUS) dataset. MASD and SMD were made available to the participants together with the ground truth segmentation masks at the start of challenge and served as the main training data for the challange. However, using additional (external) dataset for the training process was also allowed. A small sample of 36 images from the MOBIUS dataset was also released. The MOBIUS dataset acted as the testing data for the competition. This dataset was sequestered and made available (without the ground truth masks) three weeks before the result-submission deadline. A few sample images from the datasets are shown in Figure 2 and a summary of the main characteristics of the datasets is provided in Table 1. A detailed description of the competition data is provided below.

MASD. The first dataset used in SSBC 2020 contains high-quality ocular images acquired with a DSLR camera, i.e., a NIKON D 800 with 28-300mm lenses. The dataset features 2624 RGB images taken from 82 subjects. The

(a) MASD (b) SMD (c) MOBIUS

Figure 2: Sample images and ground truth masks for the datasets used in the competition. The figure displays images and sclera ground truth masks from (a) MASD and (b) SMD, as well as image examples from the segmentation part of (c) MOBIUS along with the ground truth segmentation masks for the sclera and other eye parts available with the dataset. Note the variability across image qualities and gaze directions present in the datasets.

MASD images were collected from both eyes of each subject, so there are 164 different eyes in the dataset in total. For each subject, four gaze directions were captured (looking straight, left, right and up) and for each direction 4 images were taken. The subjects in the dataset comprise both male and female subjects, some subjects are wearing contact lenses. Images were acquired at different times of the day, which affected the ambient illumination conditions and consequently the quality of the captured images. Some of the images were acquired while blinking and with closed eyes. Images in the MASD dataset are stored in JPEG format and are 7500×5000 pixels in size. Ground truth segmentation masks of the sclera region were generated manually for all images in the dataset.

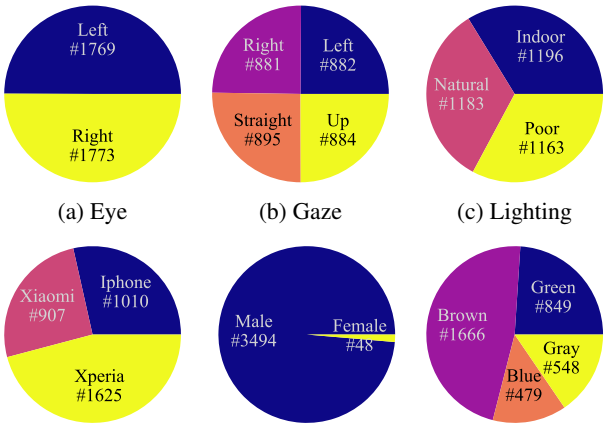
SMD. The second dataset of the competition, SMD, was captured by a mobile phone (with a 8-mega pixel rear camera) and consists of 500 RGB images of both eyes of 25 individuals (in other words, 50 different eyes). For each eye, 10 sample images were captured. The dataset contains blurred images and images with blinking eyes. The subjects in the dataset comprise both male and female subjects (12 males and 13 females) of different ages and different skin colours. Two subjects have contact lenses. Similarly to MASD, the images from SMD were also taken at different times of the day. Different acquisition conditions were (intentionally) considered when capturing the dataset to generate variations in image quality (blur, lighting condition, etc.) and facilitate investigations into the performance of (sclera) segmentation models in non-ideal conditions. The images in SMD are stored in JPEG format with a resolution of 3264×2448 pixels. The dataset also contains manually generated ground truth segmentation masks.

MOBIUS. The third dataset used for SSBC 2020 represents a recent dataset designed for research in mobile ocular biometrics. The complete MOBIUS dataset contains over 16,000 RGB ocular images of 100 subjects. However, only the segmentation part of the dataset is utilised for SSBC 2020. This part consists of 3542 manually annotated RGB images belonging to 35 subjects (70 eyes). Images from the MOBIUS dataset were captured at four different gaze directions (straight, left, right and up) using three different mobile phones, i.e,. Sony Xperia Z5 Compact, Apple iPhone 6s, Xiaomi Pocophone F1 (shown in this order in the columns in Figure 1), and in three different acquisition conditions, i.e., under good lighting inside (*Indoor*), under good lighting outside (*Neutral*), and under bad lighting inside (*Poor*). The effect of the acquisition conditions on the image quality (in the order listed above) is shown in the rows of Figure 1. The MOBIUS images correspond to male and female subjects of Caucasian origin and vary considerably in terms of quality (blur, noise, etc.) both due to the changes in mobile devices used as well as due to differences in the acquisition conditions. All images in the dataset were manually annotated with segmentation masks for the sclera, iris and pupil, as illustrated in Figure 2. Imperfections in the annotated masks were then corrected using a semi-automatic post-processing procedure. For SSBC 2020, only masks for the sclera region were considered.

Table 1: Overview of the datasets used for SSBC 2020. Information on the number of images, number of subjects, the image resolution (in pixels), main sources of variability and purpose in the competition is provided.

Dataset	#Images	#IDs	Resolution	Variability	Purpose
MASD	2624	82	7500×5000	GZ, BL	Training
SMD	500	25	3264×2448	BL, CN	Training
MOBIUS	3542	35	3000×1700	MD, CN, GZ, BL	Testing

†GZ - gaze, BL - blur, CN - acquisition condition, MD - mobile device.



(a) Eye (b) Gaze (c) Lighting

(d) Phone (e) Gender (f) Eye Colour

Figure 3: Distribution of images in the segmentation part of MOBIUS across different sources of appearance variability. The pie graphs show the distribution over: (a) eyes, (b) gaze directions, (c) lighting conditions, (d) capturing devices, (e) genders, and (f) eye colours. Best viewed in colour.

To allow for an appropriate interpretation of the results of SSBC 2020, the distribution of images across various sources of data variability (i.e, left or right eye, gaze direction, acquisition conditions, mobile device used, gender and eye colour) is presented in Figure 3. Note that the number of image samples across the different categories is not always balanced perfectly, as some samples had to be discarded during the quality check when finalising the dataset. The segmentation part of the MOBIUS dataset is currently publicly available on request[1].

3.2. Evaluation protocol

SSBC 2020 was held in two separate stages. During the first stage participants were given the training datasets, MASD and SMD, including the ground truth segmentation masks and were asked to design and train their segmentation models. In the second stage, the complete MOBIUS data (without the ground truth segmentation masks) was provided to the participants, who then had another three weeks to generate the final segmentation results. The three-week time constraint was put in place to limit the time available for experiments with

To facilitate a detailed analysis, the SSBC participants were asked to submit two types of results for scoring: *i)* binary segmentation masks (with non-zero valued pixels representing the sclera and zero valued pixels representing everything else), and *ii)* grayscale segmentation maps (with pixel intensities representing probabilities of the pixels belonging to the sclera region). Results had to be submitted for all 3542 test images from the MOBIUS dataset. Examples of the requested results for a sample image are shown Figure 4. The submitted binary masks were used to determine the performance scores for the final ranking (as described in the next section), while the probability maps allowed for more detailed analysis of the approaches and generation of performance curves.

[1]For details, visit: sclera.fri.uni-lj.si.

Figure 4: Illustration of results to be submitted (from left to right): original image, generated binary segmentation mask, probabilistic (grey-scale) segmentation prediction.

3.3. Performance metrics

The segmentation performance of the participating models was evaluated by the organisers of SSBC 2020 based on the submitted segmentation predictions. Specifically, the following performance metrics were computed based on the binary segmentation masks for each model:

- **Precision**, which is defined as the fraction of correctly retrieved sclera pixels *w.r.t.* the overall number of retrieved sclera pixels ($\frac{TP}{TP+FP}$). In other words, precision measures how many of the pixels identified are relevant for the segmentation task [23, 24, 25, 26].
- **Recall**, which is defined as the fraction of correctly retrieved sclera pixels by the segmentation model *w.r.t.* the overall number of actual sclera pixels ($\frac{TP}{TP+FN}$). Recall, hence, measures how many of the relevant pixels are retrieved [23, 24, 25, 26].
- $\boldsymbol{F_1}$ **score**, which is defined as the harmonic mean between precision and recall ($2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}}$). The score is computed to have a single performance metric for the competition and balance the trade-off between precision and recall. The F_1 score is used as the primary metric for ranking participants in SSBC 2020.
- **Intersection over union (IoU)** – or Jaccard index, which is defined as the ratio between *i)* the size of the intersection of the retrieved and ground truth sclera regions, and (ii) the size of the union of the retrieved and ground truth sclera regions ($\frac{TP}{TP+FP+FN}$).

In the above equations TP denotes *true positives*, i.e., the number of correctly retrieved sclera pixels, FP denotes *false positives*, i.e., the number of background pixels incorrectly retrieved as sclera pixels, and FN denotes *false negatives*, i.e., the number of sclera pixels incorrectly retrieved as background pixels.

Because the results computed based on the binary masks only show a partial picture of the performance of the segmentation models, complete *precision-recall curves* were

Table 2: Summary of participants and list of submitted approaches to SSBC 2020. The table lists the abbreviations of the models, as used in the experimental section.

No.	Group	Model Acronym	DL/Other
1.	Hochschule Darmstadt (HDA)	Multi-Deeplab	DL
2.	Hochschule Darmstadt (HDA)	Multi-FCN	DL
3.	Chinese Academy of Sciences (CAS)	SaSSNet	DL
4.	Bennett University (BU)	SSIP	Other
5.	Warsaw University of Technology (WUT)	MU-Net	DL
6.	Fraunhofer Institute for Computer Graphics Research (IGD)	RGB-SS-Eye-MS	DL
7.	Fraunhofer Institute for Computer Graphics Research (IGD)	Y-SS-Eye-MS	DL
8.	Universidad de Santiago, Chile (USACH)	Mask2020CL	DL
9.	Universidad de Santiago, Chile (USACH)	CGANs2020CL	DL
10.	Couger Inc.	AB Sclera Net	DL
11.	Norwegian University of Science and Technology (NTNU)	ScleraMaskRCNN	DL
12.	University of Engineering and Management (UEM)	UNet-P	DL
13.	Federal University of Paraná (UFPR)	FCN8	DL
14.	Democritus University of Thrace (DUTH)	ScleraU-Net	DL
15.	Indian Institute of Technology Ropar (IIT-RPR)	Color RITNet	DL
16.	Maulana Abul Kalam University of Technology (MAKAUT)	S-Net	DL

[†]For details on the participants from the institutions see the author list.

generated from the submitted probabilistic (grey-scale) segmentation predictions [27, 28]. An optimal F_1 score (F_1^{opt}) was determined on the computed performance curves and the *Area Under the precision-recall Curve* (AUC) was also calculated as another performance indicator [29].

4. Summary of submitted approaches

Thirteen groups entered SSBC 2020 and submitted 16 segmentation models for scoring. The majority of submitted models (15 in total) relied on deep learning and only one used standard image processing techniques. Table 2 presents a summary of the participating groups, while a brief description of the submitted models is provided below.

Multi-Deeplab (HDA) represents a variant of the Deeplab v3+ model from [30]. It is trained for multi-class segmentation (different eye parts) on the MCIS dataset [31] and samples from the MASD and SMD dataset that were manually annotated with additional eye components.

Multi-FCN (HDA) exploits a Fully Convolutional Neural Network (FCN) and is designed for multi-class eye segmentation. The model is initialised with weights provided by fcn8s-at-once for the Pascal 2012 VOC dataset [32] and then fine-tuned on the MCIS, MASD and SMD datasets.

SaSSNet (CAS) or Shape-aware Sclera SegNet is a deep learning solution that uses Deeplab v3+ [30] for its backbone model. During training, SaSSNet utilises heavy data augmentation [33] and a pixel-wise cross-entropy loss combined with a shape-aware loss to better preserve the shape of the segmented sclera.

SSIP (BU) is the only model from SSBC 2020 that employs standard image processing techniques for segmentation. It first processes input images with a single-scale retinex technique [34, 35] and then analyses image intensities for segmentation. A sequence of morphological operations is used to generate the final segmentation results.

MU-Net (WUT) represents a U-Net inspired model [36] conditioned on MobileNetV2 [37] class features to segment the sclera region from the background in an eye. A two-stage fine-tuning procedure is utilised with MobileNetV2 using the SSBC training data. To avoid overfitting different techniques are used for data augmentation.

RGB-SS-Eye-MS (IGD) is based on the Multi-scale segmentation solutions (Eye-MS) from [38]. The model is a simple CNN that refines the segmentation results across different image resolutions and is trained with an IoU loss. The model processes three channel RGB image.

Y-SS-Eye-MS (IGD) is identical conceptually to the RGB-SS-Eye-MS model described above, but processes only a single channel of the input images, i.e., the luma component (Y) of the YUV representation. Thus, differently from RGB-SS-Eye-MS, Y-SS-Eye-MS does not rely on colour information during segmentation.

Mask2020CL (USACH) is a modified Mask R-CNN [39, 40], extended from Faster-R-CNN by adding a branch for predicting segmentation masks on each Region of Interest (RoI), in parallel with the existing branch for classification and bounding box regression. The mask branch is a small fully connected network applied to each RoI, predicting a segmentation mask in a per pixel manner.

CGANs2020CL (USACH) formulates the segmentation problem as a patch-to-patch translation task and uses a Conditional Generative Adversarial Network [41] for the segmentation process. A Resnet-101 model is utilised as the backbone of the solution and trained from scratch. Aggressive data augmentation is used during training.

AB Sclera Net (Couger) is inspired by the recent (encoder-decoder) EyeNet model from [42]. Efficient Net B4 [43] is used as the encoder in the model. The decoder consists of two residual blocks and an upsampling layer, the output of which is passed to a CBAM attention layer [44]. The model is trained using a combination of categorical cross entropy and dice losses.

ScleraMaskRCNN (NTNU) uses a MaskRCNN model for instance segmentation. A ResNet-101 model is used as the backbone. The segmentation procedure involves two stages, where the first generates region proposals, while the second predicts the class of the objects and refines the bounding box needed for generating segmentation masks.

UNet-P (UEM) is a modified version of U-Net [36]. At the core of this submission is a novel pre-processing procedure that normalises the input RGB images before feeding them to the segmentation model. The pre-processing procedure helps with the model convergence during training and enhances performance by reducing data variability.

FCN8 (UFPR) uses a Fully Convolutional Network (FCN) for segmentation and is, hence, applicable to arbitrarily-sized input images [45]. It relies on ideas from [46] and uses a VGG-16 model (without the FC layers) in

Table 3: Comparative assessment on the MOBIUS dataset. The results are ordered according to the achieved F_1 scores. The F_1, Precision, Recall and IoU scores were computed from the submitted binary masks. The optimal F_1 score on the precision-recall curve (F_1^{opt}) and AUC values were calculated from the probabilistic segmentation predictions.

Segmentation Model	From binary masks				From probabilistic predictions	
	F_1	Precision	Recall	IoU	F_1^{opt}	AUC
UNet-P	0.868 ± 0.003	0.909 ± 0.004	0.831 ± 0.003	0.868 ± 0.003	0.870 ± 0.003	0.930 ± 0.003
FCN8	0.854 ± 0.004	0.820 ± 0.004	0.890 ± 0.004	0.853 ± 0.003	0.865 ± 0.004	0.936 ± 0.003
RGB-SS-Eye-MS	0.836 ± 0.004	0.917 ± 0.002	0.769 ± 0.005	0.841 ± 0.003	0.842 ± 0.004	0.872 ± 0.003
Y-SS-Eye-MS	0.823 ± 0.005	0.930 ± 0.004	0.738 ± 0.006	0.830 ± 0.004	0.836 ± 0.005	0.868 ± 0.005
SaSSNet	0.821 ± 0.004	0.885 ± 0.004	0.765 ± 0.006	0.827 ± 0.003	0.818 ± 0.004	0.893 ± 0.004
Multi-Deeplab	0.806 ± 0.005	0.915 ± 0.001	0.719 ± 0.008	0.816 ± 0.004	0.821 ± 0.004	0.896 ± 0.004
CGANs2020CL	0.803 ± 0.006	0.771 ± 0.009	0.838 ± 0.003	0.810 ± 0.005	0.803 ± 0.006	0.828 ± 0.006
ScleraU-Net	0.795 ± 0.005	0.941 ± 0.002	0.689 ± 0.007	0.809 ± 0.003	0.805 ± 0.004	0.848 ± 0.003
AB Sclera Net	0.786 ± 0.008	0.785 ± 0.016	0.787 ± 0.007	0.797 ± 0.006	0.793 ± 0.008	0.878 ± 0.005
Color RITNet	0.774 ± 0.007	0.898 ± 0.006	0.680 ± 0.011	0.791 ± 0.005	0.783 ± 0.006	0.793 ± 0.012
ScleraMaskRCNN†	0.763 ± 0.011	0.828 ± 0.008	0.707 ± 0.015	0.782 ± 0.008	n/a	n/a
Multi-FCN	0.760 ± 0.007	0.941 ± 0.003	0.638 ± 0.009	0.782 ± 0.005	0.716 ± 0.006	0.786 ± 0.007
Mask2020CL†	0.717 ± 0.010	0.833 ± 0.013	0.629 ± 0.007	0.749 ± 0.006	n/a	n/a
MU-Net	0.651 ± 0.013	0.638 ± 0.012	0.665 ± 0.014	0.698 ± 0.008	0.659 ± 0.013	0.554 ± 0.014
SSIP	0.595 ± 0.007	0.762 ± 0.009	0.489 ± 0.008	0.672 ± 0.005	0.596 ± 0.007	0.524 ± 0.007
S-Net	0.462 ± 0.008	0.348 ± 0.008	0.687 ± 0.019	0.552 ± 0.005	0.598 ± 0.010	0.662 ± 0.011

† F_1^{opt} and AUC scores are not reported for Mask2020CL and ScleraMaskRCNN because of issues with the submitted probabilistic results.

the encoder and a three-layer decoder for upsampling.

ScleraU-Net (DUTH) is a modified U-Net [36] enhanced with regularisation and normalisation layers. Compared to the original U-Net, the model has a light-weight architecture. Along the encoding path of the ScleraU-Net, 2D-dropout and batch-normalisation is applied in selected locations to facilitate training and improve generalisation.

Color RITNet (IIT-RPR) represents a variant of the recent RITNet model [47] applied to colour images. The model is trained using a boundary-aware loss and heavy data augmentation over the MASD and SMD datasets.

S-Net (MAKAUT) is a light-weight variant of U-Net [36] trained with a cross entropy loss. Prior to training images from MASD and SMD are processed through a noise removal filter and augmented via rotation and flipping.

5. Benchmarking results with analysis and discussion

In this section the results of SSBC 2020 are presented. A comprehensive analysis in conducted to analyse the performance of the submitted models and study their sensitivity *w.r.t.* the acquisition device used and ambient conditions present during the acquisition process.

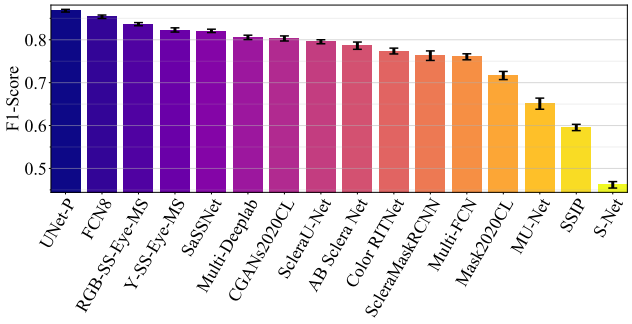


Figure 5: Performance comparison of the submitted models in terms of the F_1 score achieved over all test images from the MOBIUS dataset. Best viewed in colour.

5.1. Comparative assessment

To evaluate the submitted models, average performance scores were computed over the submitted segmentation masks. Confidence estimates for the reported averages were also calculated for all experiments by partitioning the test data into 5 data splits (in a subject disjoint manner) and computing standard deviations over the five data splits.

Results on binary segmentation masks. Binary segmentation masks are commonly produced by segmentation models through a thresholding procedure (may be integrated in the model) that typically defines the trade-off be-

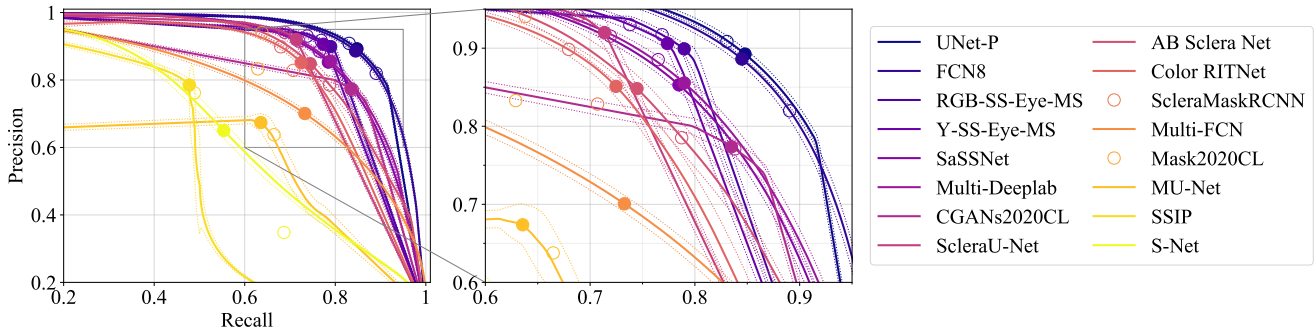


Figure 6: Precision-recall curves of the submitted models. The operating points denoted with a full circle represent the best possible F_1 score (F_1^{opt}), whereas the empty circle denotes the precision-recall point produced by the binary masks. The dotted lines denote the standard deviation. The figure is best viewed in colour and zoomed in.

tween precision and recall and determines a fixed F_1 score. Such binary masks are typically the default output of contemporary (deep) segmentation models. The first analysis in this section, therefore, compares the submitted models based on scores from the binary masks. The results of the comparison are presented in Table 3 and Figure 5.

As can be seen, UNet-P is the best performer of the competition but with an overall result very close to FCN8 considering the F_1 and IoU scores. Five more models result in F_1 scores above 0.8, i.e., RGB-SS-Eye-MS, Y-SS-Eye-MS, SaSSNet, Multi-DeepLab and CGANs2020CL. The next five solutions, i.e., ScleraU-Net, AB Sclera Net, Color RITNet, ScleraMaskRCNN and Multi-FCN, still produce competitive results with F_1 scores above 0.76, but are somewhat behind the top performers of SSBC 2020. Among the deep learning models, Mask2020CL still achieves an F_1 score of 0.717, whereas MU-Net and S-Net result in less competitive performance indicators with F_1 values of 0.651 and 0.462, respectively. SSIP generates weaker results than most other models with an F_1 score of 0.595. However, SSIP is the only SSBC 2020 approach not based on deep learning.

It is interesting to note that the performance of the models is not necessarily related to their size/complexity. UNet-P, for example, is among the smaller models with close to 2 million parameters, whereas FCN8 and Multi-FCN are the largest models with around 135 million parameters. The former two of these models are the top performers of the competition, while the latter is less competitive. In summary, the results show that most models produced solid results on the MOBIUS images, but also that there is considerable variability in the results among the best and worst performing models, even if similar model topologies were considered for for the segmentation solutions.

Results on probabilistic segmentation predictions. To get better insight into the performance of the submitted models, the next analysis centres around the submitted probabilistic segmentation predictions. From the right part

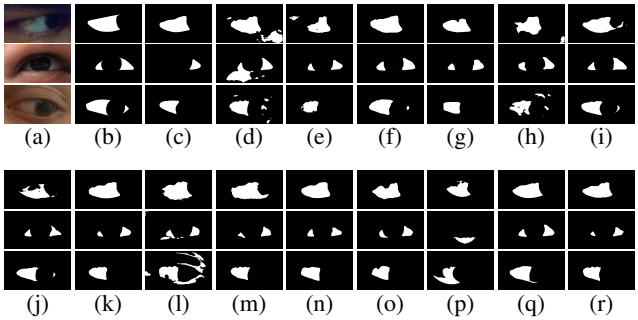


Figure 7: Qualitative comparison of the submitted models (in terms of binary masks) on selected MOBIUS images. Observe the difference in the segmentation quality across the evaluated models. The figure shows (a) the original image; (b) the ground truth mask; and the submitted binary masks from: (c) AB Sclera Net, (d) CGANs2020CL, (e) Color RITNet, (f) FCN8, (g) Mask2020CL, (h) MU-Net, (i) Multi-Deeplab, (j) Multi-FCN, (k) RGB-SS-Eye-MS, (l) S-Net, (m) SaSSNet, (n) ScleraMaskRCNN, (o) ScleraU-Net, (p) SSIP, (q) UNet-P, (r) Y-SS-Eye-MS.

of Table 3 and the precision-recall curves in Figure 6 it can be observed that the two top performing models, UNet-P and FCN8, are now even closer together in terms of performance than in the case when the binary segmentation masks were considered. The precision-recall curves for the two models in fact overlap considerably for the most part of the curves. Most of the remaining models are also clustered together tighter, with the optimal F_1 scores ranging from 0.841 for the third-place model, RGB-SS-Eye-MS, and 0.783 for the Color RITNet model that ranked 10th. The rest of the models, including the SSIP approach, perform similarly as with the binary masks and overall somewhat weaker than the top performers.

Qualitative comparison. Figure 7 presents a qualitative comparison of the submitted models in terms of binary segmentation masks produced. Here, three challeng-

ing samples corresponding to the three ambient acquisition conditions present in the MOBIUS dataset were selected to demonstrate the varying segmentation performance of the submitted models. Note that despite the solid performance of most models in terms of scores (Table 3) there are noticeable differences between the best and worst performing models when looking at the quality of the segmentation.

5.2. Sensitivity analysis

The last set of investigations focuses on the impact of different acquisition devices and ambient acquisition conditions on the submitted segmentation models. The F_1 scores generated by the submitted submitted models across different devices and conditions are presented in Figure 8.

Impact of acquisition device. The top graph in Figure 8 shows that there are observable differences in the segmentation performance across the different acquisition devices (and consequently imaging sensors). The majority of submitted models consistently performs best with images captured with the Xiaomi phone. The performance with images from the remaining two phones is comparable for most models, except for Color RITNet and Multi-FCN, which both favour images from one of the two other phones. Notably, the SSIP and S-Net approaches are the only ones, where the Xiaomi phone did not result in the best performance. Overall, these results show that there are notable differences in the sensitivity of the segmentation models *w.r.t.* different acquisition devices, but better performing models typically also perform better with all devices.

Impact of acquisition conditions. As shown in the bottom graph of Figure 8, the ambient conditions present during acquisition have a considerable impact on the performance of the evaluated models, much more so that the acquisition devices used. It is interesting to note that most of the performance differences between the models (observed in Section 5.1) can be attributed to the performance in the *Poor* setting, i.e., inside with poor lighting. All models (except MU-Net, SSIP and S-Net) are relatively close in the *Indoor* setting and exhibit slight differences with the *Natural* acquisition conditions - with F_1 scores of all deep learning models above 0.8. However, the differences are considerably larger with the most challenging scenario, where the best performing model, UNet-P, is the only one with an F_1 score above 0.8. Interestingly, in the *Poor* setting, the SSIP model is even able to outperform some of the deep models. In summary, the external acquisition conditions seem to present a considerable challenge for many of the submitted segmentation models, suggesting that more work is needed to improve performance in such conditions.

6. Conclusion

The 2020 edition of the Sclera Segmentation Benchmarking Competition (SSBC 2020) was organised in an effort to evaluate and benchmark the performance of contemporary sclera segmentation models with ocular image captured with mobile devices and explore the robustness of existing models *w.r.t.* to changes in the mobile devices used for image acquisition as well as changes in the external acquisition conditions (inside vs. outside, good vs. bad lighting). A total of 13 groups from 22 institutions participated in the competition and contributed 16 segmentation models for the group evaluation. The submitted models ensured solid segmentation results in most experimental scenarios. The biggest performance differences across the tested models were observed with images captured in the most challenging conditions (inside and with bad lighting), suggesting that changes in image quality due to ambient conditions represent one of the biggest challenges for existing model that will need to be addressed going forward.

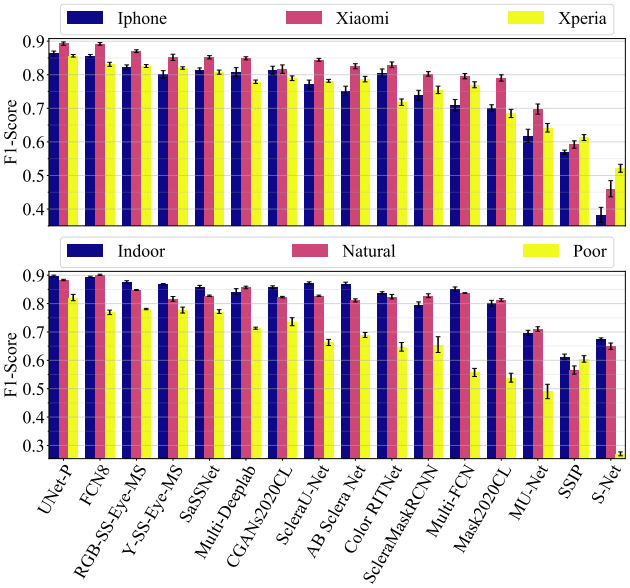


Figure 8: Sensitivity of the submitted models to differences in the mobile device used during image capture (top) and the ambient acquisition conditions (bottom). The graphs show F_1 scores generated based on the submitted binary segmentation masks. Best viewed online and in colour.

Acknowledgments

Supported in parts by the ARRS Research Programmes P2-0250 (B) Metrology and Biometric Systems, P2-0214 (A) Computer Vision and the ARRS young researcher program, by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860813 - TReSPAsS-ETN and the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Abhijit Das, Umapada Pal, Michael Blumenstein, and Miguel Angel Ferrer Ballester. Sclera recognition-a survey. In *2013 2nd IAPR Asian Conference on Pattern Recognition*, pages 917–921. IEEE, 2013. 2

[2] Peter Rot, Matej Vitek, Klemen Grm, Žiga Emeršič, Peter Peer, and Vitomir Štruc. Deep sclera segmentation and recognition. In *Handbook of vascular biometrics*, pages 395–432. Springer, Cham, 2020. 2

[3] Matej Vitek, Peter Rot, Vitomir Štruc, and Peter Peer. A comprehensive investigation into sclera biometrics: a novel dataset and performance study. *Neural Computing and Applications*, pages 1–15, 2020. 2

[4] Zhi Zhou, Eliza Yingzi Du, N. Luke Thomas, and Edward J. Delp. A new human identification method: Sclera recognition. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(3):571–583, 2011. 2

[5] Daniel Riccio, Nadia Brancati, Maria Frucci, and Diego Gragnaniello. An unsupervised approach for eye sclera segmentation. In *Iberoamerican Congress on Pattern Recognition*, pages 550–557. Springer, 2017. 2

[6] Abhijit Das, Umapada Pal, Miguel Angel Ferrer, and Michael Blumenstein. A framework for liveness detection for direct attacks in the visible spectrum for multimodal ocular biometrics. *Pattern Recognition Letters*, 82:232–241, 2016. 2

[7] Abhijit Das, Umapada Pal, Miguel A. Ferrer Ballester, and Michael Blumenstein. Sclera recognition using dense-sift. In *2013 13th International Conference on Intellient Systems Design and Applications*, pages 74–79. IEEE, 2013. 2

[8] Sinan Alkassar, Wai Lok Woo, Satnam Singh Dlay, and Jonathon A Chambers. Robust sclera recognition system with novel sclera segmentation and validation techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(3):474–486, 2015. 2

[9] Petru Radu, James Ferryman, and Peter Wild. A robust sclera segmentation algorithm. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6. IEEE, 2015. 2

[10] Sinan Alkassar, Wai-Lok Woo, Satnam Dlay, and Jonathon Chambers. Sclera recognition: on the quality measure and segmentation of degraded images captured under relaxed imaging conditions. *IET Biometrics*, 6(4):266–275, 2016. 2

[11] Abhijit Das, Rituraj Kunwar, Umapada Pal, Miguel A. Ferrer, and Michael Blumenstein. An online learning-based adaptive biometric system. In *Adaptive Biometric Systems*, pages 73–96. Springer, 2015. 2

[12] Abhijit Das, Umapada Pal, Miguel A. Ferrer Ballester, and Michael Blumenstein. A new efficient and adaptive sclera recognition system. In *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 1–8. IEEE, 2014. 2

[13] Abhijit Das, Umapada Pal, Miguel A. Ferrer, and Michael Blumenstein. A decision-level fusion strategy for multimodal ocular biometric in visible spectrum based on posterior probability. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 794–798. IEEE, 2017. 2

[14] Vikas Gottemukkula, Sashi Saripalle, Sriram P. Tankasala, and Reza Derakhshani. Method for using visible ocular vasculature for mobile biometrics. *IET Biometrics*, 5(1):3–12, 2016. 2

[15] Abhijit Das, Prabir Mondal, Umapada Pal, Michael Blumenstein, and Miguel A. Ferrer. Sclera vessel pattern synthesis based on a non-parametric texture synthesis technique. In *Proceedings of international conference on computer vision and image processing*, pages 241–250. Springer, 2017. 2

[16] Abhijit Das, Umapada Pal, Miguel A. Ferrer, and Michael Blumenstein. SSBC 2015: Sclera Segmentation Benchmarking Competition. In *Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, pages 742–747, 2015. 2

[17] Abhijit Das, Umapada Pal, Miguel A. Ferrer, and Michael Blumenstein. SSRBC 2016: sclera segmentation and recognition benchmarking competition. In *International Conference on Biometrics (ICB)*, pages 1–6, 2016. 2

[18] Abhijit Das, Umapada Pal, Miguel A. Ferrer, Michael Blumenstein, Dejan Štepec, Peter Rot, Ziga Emeršič, Peter Peer, Vitomir Štruc, and S.A. Kumar. SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition. In *International Joint Conference on Biometrics (IJCB)*, pages 742–747, 2017. 2

[19] Abhijit Das, Umapada Pal, Miguel A. Ferrer, Michael Blumenstein, Dejan Štepec, Peter Rot, Peter Peer, and Vitomir Štruc. SSBC 2018: Sclera Segmentation Benchmarking Competition. In *International Conference on Biometrics (ICB)*, pages 303–308, 2018. 2

[20] Abhijit Das, Umapada Pal, Michael Blumenstein, Caiyong Wang, Yong He, Yuhao Zhu, and Zhenan Sun. Sclera segmentation benchmarking competition in cross-resolution environment. In *IAPR International Conference on Biometrics. IEEE*, 2019. 2

[21] Abhijit Das, Umapada Pal, Miguel A. Ferrer Ballester, and Michael Blumenstein. Multi-angle based lively sclera biometrics at a distance. In *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 22–29. IEEE, 2014. 2

[22] Abhijit Das. *Towards Multi-modal Sclera and Iris Biometric Recognition with Adaptive Liveness Detection*. PhD thesis, Griffith University, 2017. 2

[23] Žiga Emeršič, Luka Lan Gabriel, Vitomir Štruc, and Peter Peer. Pixel-wise ear detection with convolutional encoder-decoder networks. *IET Biometrics*, 2017. 4

[24] Peter Rot, Žiga Emeršič, Vitomir Štruc, and Peter Peer. Deep multi-class eye segmentation for ocular biometrics. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–8. IEEE, 2018. 4

[25] Juš Lozej, Blaž Meden, Vitomir Štruc, and Peter Peer. End-to-end iris segmentation using U-Net. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6. IEEE, 2018. 4

[26] Juš Lozej, Dejan Štepec, Vitomir Štruc, and Peter Peer. Influence of segmentation on deep iris recognition performance. In *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6, 2019. 4

[27] David Martin Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2:37–63, 2011. 5

[28] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015. 5

[29] Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 451–466. Springer, 2013. 5

[30] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5

[31] Dailé Osorio-Roig, Christian Rathgeb, Marta Gomez-Barrero, Annette Morales-González, Eduardo Garea-Llano, and Christoph Busch. Visible wavelength iris segmentation: a multi-class approach using fully convolutional neuronal networks. In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2018. 5

[32] Mark Everingham, Andrew Zisserman, Christopher KI Williams, Luc Van Gool, Moray Allan, Christopher M Bishop, Olivier Chapelle, Navneet Dalal, Thomas Deselaers, Gyuri Dorkó, et al. The 2005 pascal visual object classes challenge. In *Machine Learning Challenges Workshop*, pages 117–176. Springer, 2005. 5

[33] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 5

[34] Ritesh Vyas, Tirupathiraju Kanumuri, Gyanendra Sheoran, and Pawan Dubey. Smartphone based iris recognition through optimized textural representation. *Multimedia Tools and Applications*, 79(19-20):14127–14146, 2020. 5

[35] Zijing Zhao and Ajay Kumar. An accurate iris segmentation framework under relaxed imaging constraints using total variation model. In *IEEE International Conference on Computer Vision*, pages 3828–3836, 2015. 5

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI): 18th International Conference, Proceedings, Part III*, pages 234–241. Springer International Publishing, Cham, 2015. 5, 6

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. pages 4510–4520, 2018. 5

[38] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Eye-mms: Miniature multi-scale segmentation network of key eye-regions in embedded applications. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 5

[39] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5

[40] Sebastian Gonzalez, Claudia Arellano, and Juan E. Tapia. Deepblueberry: Quantification of blueberries in the wild using instance segmentation. *IEEE Access*, 7:105776–105788, 2019. 5

[41] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 5

[42] Priya Kansal and Sabarinathan Devanathan. EyeNet: Attention Based Convolutional Encoder-Decoder Network for Eye Region Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3688–3693, 2019. 5

[43] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, 2019. 5

[44] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*. Springer, 2018. 5

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015. 5

[46] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1013–1020. IEEE, 2018. 5

[47] Aayush K Chaudhary, Rakshit Kothari, Manoj Acharya, Shusil Dangi, Nitinraj Nair, Reynold Bailey, Christopher Kanan, Gabriel Diaz, and Jeff B. Pelz. RITnet: Real-time Semantic Segmentation of the Eye for Gaze Tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3698–3702. IEEE, 2019. 6