Privacy-enhancing Sclera Segmentation Benchmarking Competition: SSBC 2025

M. Vitek[1,*], D. Tomašević[1,*], A. Das[2], S. Nathan[4], G. Özbulak[5,6], G. A. T. Özbulak[7,8], J. P. Calbimonte[8,9], A. Anjos[5,6], H. H. Bhatt[10], D. D. Premani[10], J. Chaudhari[10], C. Wang[11], J. Jiang[12], C. Zhang[11], Q. Zhang[12], I. I. Ganapathi[13], S. S. Ali[13], D. Velayudan[13], M. Assefa[13], N. Werghi[13], Z. A. Daniels[14], L. John[15], R. Vyas[15], J. N. Khiarak[16], T. A. Saeed[17], M. Nasehi[18], A. Kianfar[19], M. P. Panahi[20], G. Sharma[21], P. R. Panth[21], R. Ramachandra[22], A. Nigam[21], U. Pal[3], P. Peer[1], V. Štruc[1]

[1]University of Ljubljana (UL, SI), [2]Birla Institute of Technology & Science Pilani (BITS Pilani, IN), [3]Indian Statistical Institute (ISI, IN), [4]Couger Inc. (JP), [5]École Polytechnique Fédérale de Lausanne (EPFL, CH), [6]Idiap Research Institute (CH), [7]Université de Lausanne (UNIL, CH), [8]University of Applied Sciences and Arts Western Switzerland (HES-SO, CH), [9]The Sense Innovation and Research Center (CH), [10]Ahmedabad University (AU, IN) [11]Beijing University of Civil Engineering and Architecture (BUCEA, CN), [12]People's Public Security University of China (PPSUC, CN), [13]Khalifa University of Science and Technology (KU, AE), [14]SRI International (US), [15]Pandit Deendayal Energy University (PDPU, IN), [16]Warsaw University of Technology (WUT, PL), [17]Pirogov Russian National Research Medical University (RNRMU, RU), [18]Amirkabir University of Technology (AUT, IR), [19]Institute for Advanced Studies in Basic Sciences (IASBS, IR), [20]Seraj Institute (SI, IR), [21]Indian Institute of Technology Mandi (IIT-M, IN), [22]Norwegian University of Science and Technology (NTNU, NO)

Abstract

This paper presents a summary of the 2025 Sclera Segmentation Benchmarking Competition (SSBC), which focused on the development of privacy-preserving sclera-segmentation models trained using synthetically generated ocular images. The goal of the competition was to evaluate how well models trained on synthetic data perform in comparison to those trained on real-world datasets. The competition featured two tracks: (i) one relying solely on synthetic data for model development, and (ii) one combining/mixing synthetic with (a limited amount of) real-world data. A total of nine research groups submitted diverse segmentation models, employing a variety of architectural designs, including transformer-based solutions, lightweight models, and segmentation networks guided by generative frameworks. Experiments were conducted across three evaluation datasets containing both synthetic and real-world images, collected under diverse conditions. Results show that models trained entirely on synthetic data can achieve competitive performance, particularly when dedicated training strategies are employed, as evidenced by the top performing models that achieved F_1 scores of over 0.8 in the synthetic data track. Moreover, performance gains in the mixed track were often driven more by methodological choices rather than by the inclusion of real data, highlighting the promise of synthetic data for privacy-aware biometric development.

1. Introduction

Ocular biometrics represent a popular branch of research that focuses on computer-aided techniques, capable of inferring the identity of individuals based on distinctive ocular traits. While iris recognition has historically dominated this field, recent research is increasingly looking into other ocular characteristics that can either complement or substitute the iris. Among these, the sclera, the white region of the eye, has emerged as a promising candidate due to its distinctiveness, long-term consistency, and resistance to presentation attacks [12,46,59]. Unlike the iris, the sclera is less susceptible to degradations from common visual conditions and can be imaged under normal lighting with standard cameras, making it especially practical for real-world applications.

Research into sclera biometrics has intensified in recent years, covering a wide range of topics including recognition algorithms [2,10,15,20,21,43,59,60,65], segmentation methods [1,41,46,56,57], detection of presentation attacks

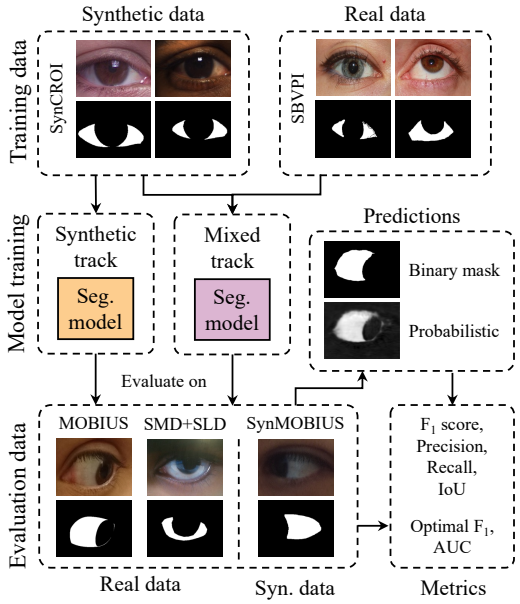


Figure 1. **Overview of SSBC 2025**. The competition entailed two tracks, where sclera segmentation models were trained on *(i)* synthetic data and *(ii)* a mix of synthetic and real data. Models were evaluated on sequestered real and synthetic datasets. Participants were required to submit binary and probabilistic predictions.

*M. Vitek and D. Tomašević are first authors with equal contributions.

[15, 46, 59], adaptability to user variability [8, 11], fusion techniques [17, 26], and lightweight methods [25, 56, 60]. However, with research in this area generally moving towards data-hungry deep learning models, a key issue that has arisen is privacy preservation, since large-scale datasets of ocular images compiled with privacy-protecting measures in mind are largely unavailable. A possible solution to this issue is the use of synthetically generated data [9, 51–54], which maintains the characteristics required to develop biometric models, but does not belong to (or contain identifying information about) real-world individuals and is, thus, not open to abuse or breach of privacy.

To investigate the effectiveness of synthetic data for sclera biometrics, the 2025 Sclera Segmentation Benchmarking Competition (SSBC), summarized in Fig. 1, was organized in the scope of the International Joint Conference on Biometrics (IJCB 2025). The competition focused on sclera segmentation, which is a crucial step in any sclera recognition pipeline, as the segmentation quality directly impacts all subsequent stages, including normalization, feature extraction, and identity matching. The competition aimed at answering key research questions, such as: How well does synthetic data mimic real-world datasets? What effect does the use of synthetic training data have on segmentation performance? What techniques can be used to adapt models to the use of synthetically generated images during training? Nine segmentation models submitted by visible research teams from around the world were evaluated to provide insights into these and related questions. The joint effort of the organizers and the participating teams resulted in the following contributions:

- A comprehensive evaluation of contemporary sclera segmentation models on real-world and synthetically generated evaluation data.
- A differential performance analysis, studying the impact of the use of synthetic and real-world training data in model development.
- A study of the impact of model size and computational complexity on the final segmentation performance and on the model's adaptability to synthetic training data.

2. Related Work

SSBC 2025 is the 9th iteration of the sclera segmentation benchmarking competition, originally started at the BTAS conference in 2015. The SSBC series of competitions has significantly pushed forward the development of sclera segmentation models, with each iteration addressing a different research problem. The 1st and 3rd SSBC (SSBC 2015 and SSBC 2016), studied the segmentation performance of various models and additionally introduced new datasets for sclera segmentation (i.e., MASD and SMD) [14]. The 2nd iteration (SSRBC 2016) studied recognition approaches in addition to sclera segmentation techniques [16]. The 4th iteration, SSERBC 2017, again included the recognition task, but additionally explored the impact of gaze direction on successful segmentation and recognition [18]. The 5th competition, SSBC 2018, studied the impact of cross-sensor image capture on the performance of sclera segmentation [19], while the 6th iteration, SSBC 2019, investigated how cross-resolution environments affect segmentation performance [13]. The 7th edition, SSBC 2020, introduced a novel dataset (MOBIUS), compiled specifically for mobile sclera biometrics, and, consequently, focused on sclera segmentation in the mobile domain [58]. A follow-up effort to SSBC 2020 [57] then explored various types of biases present in contemporary sclera segmentation models. Finally, the 8th iteration (SSRBC 2023) looked into segmentation and recognition performance individually, as well as the interplay between the two tasks [7].

Table 1. Summary of the real-world and synthetic datasets used for SSBC 2025. Reported is the number of images and subjects, the main sources of variability and the purpose in the competition.

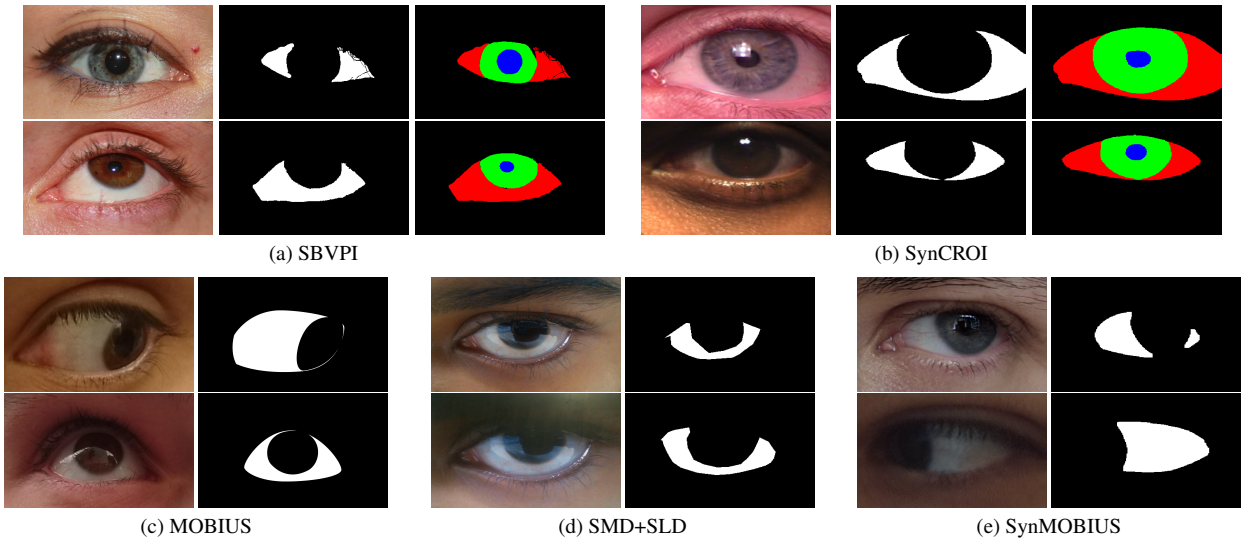
Origin	Dataset	#Images	#IDs	Variability	Purpose
Real-world	SBVPI	1840	55	GZ, CLR	Training
	SMD+SLD	489	52	CN	Testing
	MOBIUS	3540	35	GZ, CLR, CN	Testing
Synthetic	SynCROI (CE)	5500	N/A	GZ, CLR	Training
	SynCROI (PU)	5500	N/A	CN	Training
	SynMOBIUS	4772	N/A	GZ, CLR, CN	Testing

(GZ) - gaze, (CLR) - eye color, (CN) - acquisition condition

Differently from past SSBC editions, SSBC 2025 focuses on privacy-preserving sclera segmentation models, developed with the use of synthetically generated (identityless) training data. The main (distinct) goal of the competition is to study how well such models perform in relation to models trained on real-world data from actual individuals.

3. SSBC 2025 Competition Data

The aim of SSBC is to study the use of synthetic data for training and testing of sclera segmentation models and investigate what impact synthetic data has on model performance when compared to real-world sclera imagery. To this end, several real-world (R) and synthetic datasets (S) were used for the competition, i.e., *(i)* the Sclera Blood Vessels, Periocular and Iris (SBVPI) dataset (R) [59], *(ii)* the Mobile Ocular Biometrics in Unconstrained Settings (MOBIUS) dataset (R) [58], *(iii)* the combined Sclera Mobile Dataset [6] and Sclera Liveness Dataset (R) [7] (i.e., SMD+SLD), as well as *(iv)* the Synthetic Cross-Racial Ocular Image (SynCROI) dataset (S), and *(v)* the Synthetic MOBIUS (SynMOBIUS) dataset (S). SBVPI and SynCROI represent the primary training datasets used throughout SSBC 2025, while MOBIUS, SMD+SLD, and SynMOBIUS were used for evaluation purposes only. A summary of the datasets is provided in Table 1, while sample images of each dataset and their corresponding segmentation masks



(a) SBVPI (b) SynCROI

(c) MOBIUS (d) SMD+SLD (e) SynMOBIUS

Figure 2. **Samples from the training and testing datasets used in SSBC 2025**. The top row contains ocular images and available ground truth segmentation masks of the training datasets, while the bottom row depicts testings samples and their reference masks. SSBC 2025 participant were free to train their models either on the complete 4-class mark-up, or only on the region corresponding to the sclera.

are presented in Figure 2. As part of SSBC 2025, images of all datasets were resized to a resolution of 400×300.

3.1. Real–World Datasets

SBVPI. The first real-world dataset of the competition, SBVPI, consists of 1840 high-resolution RGB images from 55 Caucasian subjects (29 females and 26 males) captured using a Canon EOS 60D DSLR camera with a macro lens under controlled conditions. The images contain four gaze directions (straight, left, right, and up) of each eye. 4 samples per eye and gaze direction were acquired at varying distances and camera positions. Subjects span an age range of 15–80 years and exhibit diverse eye colors. All images have manually annotated segmentation masks of the sclera and periocular regions, while more detailed masks, which also include the iris and pupil are available for 100 images.

SMD+SLD. The second real-world dataset, SMD+SLD, combines the Sclera Mobile Dataset [6] and the Sclera Liveness Dataset (SLD) [57], comprising 381 RGB images of 25 subjects from SMD and 108 images of 27 subjects from SLD. The images were captured by different mobile phones with an 8-mega pixel rear camera and under various acquisition conditions to increase data variety, resulting in blurry images, images with blinking eyes, and images taken at different times of the day under different lighting conditions. Consequently, the dataset enables evaluation of sclera segmentation models under non-ideal or challenging conditions. The dataset also contains manually generated ground truth segmentation masks for the sclera region.

MOBIUS. The third real-world dataset of the competition, MOBIUS, was designed specifically with mobile ocular biometrics in mind. In total, the dataset features 16,717 high-resolution RGB images of both eyes from 100 male and female subjects of Caucasian origin. The images were captured using three commercial mobile devices (Sony Xperia Z5 Compact, Apple iPhone 6s, and Xiaomi Pocophone F1) under varying gaze directions (straight, left, right, up) and lighting conditions (natural daylight, indoor light, low light), resulting in high image variety. For the purposes of SSBC 2025 only the part of the dataset with manually annotated segmentation masks was utilized, consisting of 3542 images from 35 subjects with annotated sclera, iris, and pupil regions. However, only the sclera region was considered during the evaluation process of SSBC 2025.

3.2. Synthetic Datasets

As SSBC 2025 focuses on the development of segmentation models with privacy-preserving synthetic data, we construct two large-scale synthetic datasets, SynCROI and SynMOBIUS, that are used for training and evaluation purposes, following the approach outlined in Figure 3.

Synthetic-Data Generation. To generate synthetic datasets, we rely on BiOcularGAN [53], a recent deep generative framework for creating bimodal ocular images with corresponding (synthetic, ground truth) segmentation masks. BiOcularGAN [53] extends the StyleGAN2 [33] approach and achieves better image quality, by utilizing a dual-branch synthesis network, which creates aligned visible and near-infrared (NIR) ocular images based on the noise-based style information provided by the mapping network, along with two discriminator networks, one for each light spectrum, which form a joint training objective. In addition, BiOcularGAN relies on a separate ensemble pixel classifier to interpret latent features of the synthesis network

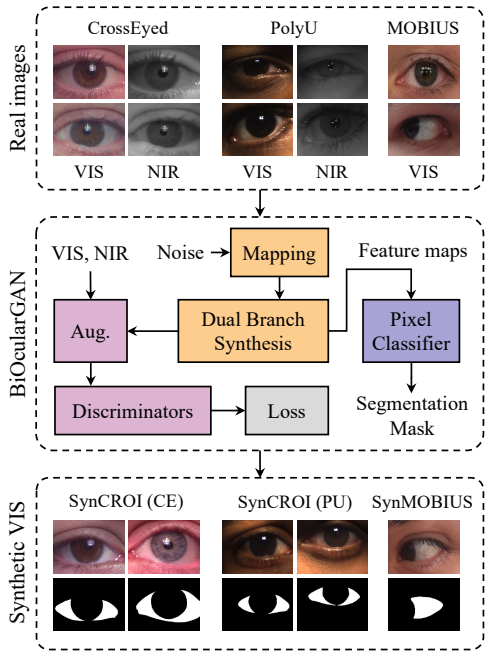


Figure 3. **Overview of data synthesis process**. To generate synthetic datasets for SSBC 2025 we train separate instances of BiOcularGAN [53] on three real-world datasets (Cross-Eyed [49], PolyU Cross-Spectral Iris, and MOBIUS [58]).

for accurate mask generation of synthetic images [64]. Crucially, the model also employs the adaptive discriminator augmentation (ADA) to enable stable training even in low-data regimes [32]. We use only the generated RGB images for the competition and discard the generated NIR images.

The specific BiOcularGAN [53] implementation we employ consists of an 8-layer mapping network, a synthesis network with 7 synthesis blocks that outputs VIS-NIR data at resolutions from 4×4 to 256×256, and two discriminators, each with 7 downsampling blocks. We train the model in batches of 16 with a learning rate of 0.0025 and the Adam optimizer [34] with $\beta_1 = 0$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. The ensemble pixel classifier, used to generate ground truth masks for the synthetic data, consists of 10 multi-layer perceptrons trained after the initial generative model on randomly sampled pixels of manually annotated synthetic images. Training uses the cross-entropy loss and the Adam optimizer [34] with a learning rate of 10^{-3} in batches of 64, which stops after 50 steps with no improvement beyond the third epoch [64]. With BiOcularGAN [53], we generate the following synthetic datasets for SSBC 2025:

- **SynCROI.** The first synthetic dataset, SynCROI, comprises 11,000 synthetic ocular images, each accompanied by a segmentation mask of the sclera, iris, and pupil region. The dataset is divided into two large subsets with subjects of different origin, each generated with a separate instance of BiOcularGAN [53], trained on real-world visible (VIS) and near-infrared (NIR) image pairs. The first subset, **SynCROI (CE)**, is produced by training BiOcularGAN [53] on the real-world CrossEyed (CE) dataset, which includes 3840 VIS-NIR image pairs of 120 Caucasian subjects with diverse eye colours. Differently, the second subset, **SynCROI (PU)**, is based on the Hong Kong Polytechnic University (PolyU) Cross-Spectral Iris Image Database, which features 12,540 VIS-NIR image pairs of 209 Asian subjects. Both models are trained to convergence, for 1120 and 1600 $kimgs$, respectively for CE and PU. Once trained, 8 synthetic samples are generated with each model and manually annotated for training the ensemble pixel classifier. Afterwards, random sampling of the noise input is used with each model to produce 5500 aligned VIS and NIR images of a resolution 256×256 with corresponding segmentation masks. For the competition, only VIS images and their masks are considered and are resized to 400×300 to retain the aspect ratio of real-world training data.
- **SynMOBIUS.** The second synthetic dataset, SynMOBIUS, is derived from the MOBIUS [58] dataset and is used specifically for evaluation in SSBC 2025. Differently from the previous setup, the BiOcularGAN [53] model is first adapted to accommodate training on single (VIS) spectrum images by removing the second synthesis branch and its associated discriminator. The modified model is trained on MOBIUS [58] to convergence over 2240 $kimgs$. To then train the pixel classifiers, we utilize 45 manual annotations from MOBIUS [58], by projecting their corresponding images to the latent space of the generative model [33], along with 5 manually annotated synthetic samples that contain eye occlusions. After training, the model is used to generate 5000 images-mask pairs, which are then resized from 256×256 to 400×300.

Characteristics of Synthetic Data. To obtain better insight into the generated synthetic datasets, we utilized a ResNet101-based [28] classifier to determine the visual characteristics of each synthetic image, including the eye side (left or right), the gaze direction (straight, left, right, or up) and the eye colour (brown, blue, gray, or green). For use on the SynCROI dataset, we trained the classifier on the combined data of SBVPI and MOBIUS, which include the required annotations. Training was performed on a $9:1$ data split, over 20 epochs in batches of 32 and the Adam optimizer [34], with an initial learning rate of 10^{-4} that was reduced by a factor of 10 if no improvements were observed in 4 epochs. Conversely, the classifier for SynMOBIUS was trained solely on the MOBIUS dataset to minimize the domain gap, with the same parameters. The distributions of the predicted characteristics for each dataset are presented in Figure 4. Note that both subsets of SynCROI mainly contain images with a straight gaze direction and that SynCROI (PU) only contains brown eyes, due to prevelance of brown eyes in the training data of the generative model.

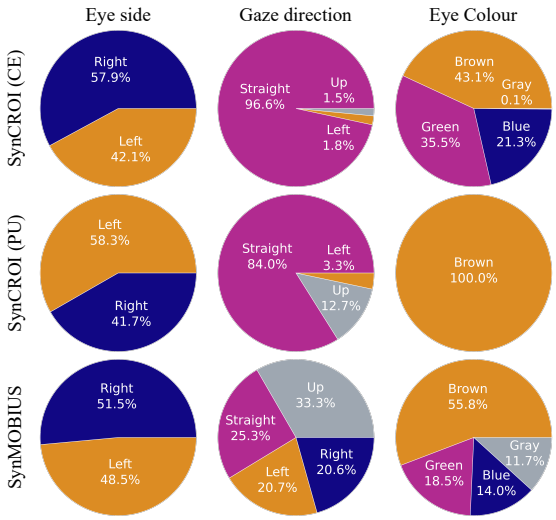


Figure 4. **Distributions of synthetic samples** over different characteristics: eye side (left, right), gaze direction (straight, left, right, up), and eye color (brown, blue, gray, green). Values are estimated with a ResNet101-based [28] classifier trained on annotations of MOBIUS and SBVPI.

4. Benchmarking Methodology

4.1. Competition Protocol

SSBC 2025 entailed two separate phases. As part of the first phase, participants were provided with real-world and synthetic training data in the form of ocular images and ground truth segmentation masks from the SBVPI and the SynCROI dataset. The datasets were split into training and validation subsets, however, the participants were not required to use the preassigned splits. To familiarize participants with the final submission format, a small sample of testing data was also released, comprising images of a single identity from MOBIUS and SMD+SLD, along with 10 random samples from SynMOBIUS. These sequestered test datasets were provided to the participants 9 weeks after the start of the competition, without ground truth segmentation masks. The participants then only had a single week to submit their segmentation predictions, in order to limit potential fine-tuning on test data. Participants were required to deliver two types of outputs for each of the 14,772 images from MOBIUS, SMD+SLD, and SynMOBIUS, including: *(i)* a **binary segmentation mask**, where non-zero pixels denote the sclera region and zero-valued pixels denote the background, and *(ii)* **grayscale probability maps**, where pixel intensities represent the confidence of that pixel belonging to the sclera region. An example of expected output formats is provided in Figure 5. The submitted binary masks served as the basis for performance ranking of participating models, while the probability maps were used for obtaining deeper insight into model behavior and enabled the derivation of detailed performance curves.



Figure 5. **Illustration of results to be submitted** (from left to right): original image, generated binary segmentation mask, probabilistic (grey-scale) segmentation prediction.

The evaluation was performed in two distinct tracks, reflecting the focus of SSBC 2025 on the privacy-preserving use of synthetic biometric data in model development and evaluation. The *(i)* **Synthetic track** focused on models trained on synthetic data only, studying how such models perform on (different, sequestered) synthetic, as well as real-world evaluation data. The *(ii)* **Mixed track**, on the other hand, allowed the models to be trained on synthetic data, but also on the additional real-world SBVPI dataset, while again evaluating the trained models on both synthetic and real-world evaluation data.

4.2. Performance Measures

The overall performance of the segmentation models was evaluated both through the submitted binary segmentation masks and the probabilistic predictions. For the binary masks, standard segmentation metrics were derived from true positives TP (i.e. correctly detected sclera pixels), false positives FP (i.e. background pixels incorrectly detected as sclera) and false negatives FN (i.e. sclera pixels incorrectly detected as background), as follows:

- **Precision**, which measures the proportion of correctly predicted sclera pixels relative to the total number of pixels predicted as sclera, i.e., $\frac{TP}{TP+FP}$ [38, 39, 45].
- **Recall**, which quantifies the proportion of true sclera pixels that were successfully identified by the model, computed as $\frac{TP}{TP+FN}$ [23, 38, 39, 45].
- $\boldsymbol{F_1}$ **score**, which represents the harmonic mean of precision and recall, defined as $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$. This aggregated metric balances the trade-off between precision and recall and serves as the main criterion for ranking participating models.
- **Intersection over union (IoU)**, or the Jaccard index, captures the overlap between the predicted and the ground truth sclera regions, normalized by their union, calculated as $\frac{TP}{TP+FP+FN}$.

To provide a more comprehensive assessment beyond binary classifications, the grayscale (probabilistic) segmentation maps were also used to produce **precision-recall curves** [40, 48]. From the curves, the optimal F_1 score (F_1^{opt}) and the **Area Under the Curve** (AUC) [4] can be computed, offering a more nuanced comparison of models across varying confidence thresholds.

Table 2. **List of submitted entries to SSBC 2025** and their respective institutions. The abbreviations used for the models in this table correspond to the ones used in the experimental section.

Team	Model Acronym	Base†
Indian Institute of Technology Mandi (IIT-M)	SEG-U-Sclera	SAM
Beijing University of Civil Engineering and Architecture (BUCEA)	SAM2-UNet	SAM/UN
Warsaw University of Technology (WUT)	AEOS	SAM/SF
Khalifa University (KU)	KU-CVML	DL
Ahmedabad University (AU)	ShapeGAN-DLV3+	DL
Couger Inc.	SwinDANet	ST/DN
SRI International	UL-VMUNet	M/UN
Idiap – University of Applied Sciences and Arts Western Switzerland (HES-SO)	SAM-Iris	SAM
Pandit Deendayal Energy University (PDPU)	UNet++_Binary	UN

For details on the participants from the institutions, see the author list.
†Base architecture: DL – DeepLab [5], DN – DenseNet [30], M – Mamba [27], SAM – Segment Anything Model [35] SF – SegFormer [62], ST – Swin Transformer [37], UN – U-Net [44].

5. Summary of Submitted Approaches

A total of 9 teams submitted their entries to SSBC 2025. Table 2 presents a summary of the submissions, while a brief description of each of the entries is provided below.

SEG-U-Sclera (IIT-M) is a variant of SAM2 [42] trained with uncertainty-weighted binary cross-entropy (BCE). This loss addresses the difference between real-world and synthetic data by focusing on the uncertain regions where synthetic and real-world images vary.

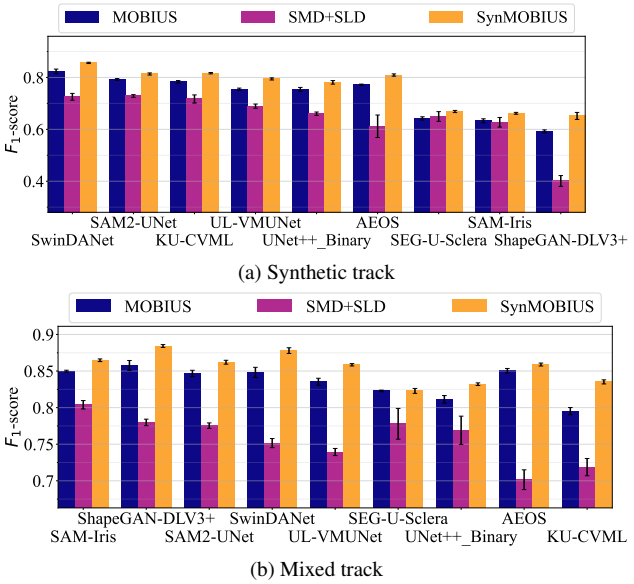
SAM2-UNet (BUCEA) is another SAM2-based [42] model, utilizing a U-Net-like architecture [63] with a Hiera encoder [47] and a U-shaped 3-block decoder. Lightweight adapters are inserted into the parameter-frozen Hiera backbone to ensure parameter-efficient fine-tuning for the sclera segmentation task. The model is trained with weighted IoU and BCE losses and deep supervision strategies.

AEOS (WUT): the WUT team used a hybrid approach, with their architecture depending on the amount of training data. Their SAM2-based [42] architecture tended to underperform on small amounts of training data, so for the Synthetic track, they relied on a SegFormer-like [62] model. This hybrid approach resulted in better generalization, as it explicitly addressed small/large training datasets.

KU-CVML (KU) relies on the DeepLabV3+ [5] architecture and the EfficientNet-B4 [50] encoder and is trained with a hybrid loss, which is partially inspired by semi-supervised training approaches and SAM [35], and also addresses class imbalance and contour accuracy.

ShapeGAN-DLV3+ (AU) is an extension of DeepLabV3+ [5] that consists of several sclera-segmentation-specific modules, i.e.: (i) a combination of CoordConv [36] and deformable stem convolution [67] to provide spatial awareness and adaptive receptive fields; (ii) a FourierLoss [24] to enforce shape fidelity; (iii) an InverseFormLoss [3] to promote fine boundary alignment; and (iv) a PatchGAN discriminator that is attached to the end of the encoder and trained in an adversarial manner, to provide a bridge crossing the synthetic-to-real domain gap.

SwinDANet (Couger) is a hybrid encoder-decoder model. The encoder is based on the Swin Transformer V2 [37] with shifted-window self-attention, which enables the modeling of long-range dependencies and extracts multi-scale features in a hierarchical contextual representation. The decoder is DenseNet-based [30], augmented with CSSE attention [31], emphasizing salient spatial and channel-wise information, and adaptive skip-connections that preserve spatial details from the encoder.



(a) Synthetic track

(b) Mixed track

Figure 6. **Performance comparison** of the submissions trained on (a) synthetic and (b) mixed training data, in terms of the F_1 score achieved over all the images in the evaluation datasets.

UL-VMUNet (SRI) is a lightweight Mamba [27] variant, combining the state-space Mamba model with a U-Net-like architecture [61] to attain compute- and parameter-efficient high-quality segmentations. To generalize better from synthetic data, hand-crafted data augmentation is performed to mimic different lighting, poses, and camera artifacts.

SAM-Iris (Idiap – HES-SO) is a SAM2-based [42] model, fine-tuned with a sequential synthetic-to-real training process inspired by [55]. The model additionally contains a prompt point in the center of the input eye image, which ensures that the model focuses specifically on the eye region.

UNet++_Binary (PDPU) adopts the UNet++ [66] architecture, using a ResNet-style [29] encoder. It is fine-tuned for the task of sclera segmentation on images normalized using ImageNet [22] statistics.

6. Benchmarking Results

In this section, we present the results of SSBC 2025, both for the synthetic and the mixed track. Additionally, we perform a detailed analysis of the performance of the submitted models across different segmentation thresholds and study the performance differentials between the corresponding models from the two competition tracks.

Table 3. **Comparative assessment of the models trained on synthetic data**. The submissions are ranked according to the harmonic means of the achieved F_1 scores over the three evaluation datasets. 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. The harmonic means are also reported for the rest of the performance measures in the column next to the individual dataset results. Note that the ranking is quite consistent across performance indicators.

Rank	Segmentation Model	Evaluation Dataset	From binary masks								From probabilistic predictions			
			F_1		Precision		Recall		IoU		F_1^{opt}		AUC	
		MOBIUS	0.824		0.825		0.842		0.722		0.850		0.887	
1	SwinDANet	SMD+SLD	0.725	0.798	0.640	0.754	0.866	0.868	0.587	0.680	0.758	0.824	0.809	0.870
		SynMOBIUS	0.856		0.830		0.897		0.757		0.875		0.921	
		MOBIUS	0.792		0.792		0.822		0.677		0.821		0.843	
2	SAM2-UNet	SMD+SLD	0.729	0.776	0.636	0.728	0.872	0.856	0.584	0.649	0.756	0.803	0.779	0.826
		SynMOBIUS	0.813		0.777		0.876		0.698		0.837		0.861	
		MOBIUS	0.783		0.711		0.914		0.667		-		-	
3	KU-CVML	SMD+SLD	0.717	0.770	0.596	0.676	0.926	0.925	0.571	0.641	-	-	-	-
		SynMOBIUS	0.816		0.738		0.937		0.700		-		-	
		MOBIUS	0.755		0.822		0.736		0.633		0.817		0.859	
4	UL-VMUNet	SMD+SLD	0.689	0.744	0.600	0.728	0.839	0.793	0.537	0.609	0.729	0.791	0.772	0.836
		SynMOBIUS	0.794		0.808		0.811		0.674		0.836		0.885	
		MOBIUS	0.754		0.851		0.720		0.643		0.832		0.878	
5	UNet++_Binary	SMD+SLD	0.660	0.728	0.690	0.784	0.668	0.716	0.513	0.599	0.746	0.806	0.762	0.840
		SynMOBIUS	0.781		0.834		0.769		0.666		0.848		0.893	
		MOBIUS	0.772		0.914		0.714		0.683		0.871		0.921	
6	AEOS	SMD+SLD	0.612	0.720	0.731	0.847	0.589	0.677	0.511	0.622	0.731	0.822	0.763	0.867
		SynMOBIUS	0.809		0.927		0.749		0.712		0.883		0.941	
		MOBIUS	0.642		0.655		0.669		0.512		0.716		0.642	
7	SEG-U-Sclera	SMD+SLD	0.650	0.653	0.569	0.621	0.782	0.725	0.521	0.524	0.685	0.709	0.593	0.632
		SynMOBIUS	0.669		0.644		0.733		0.541		0.728		0.666	
		MOBIUS	0.633		0.639		0.657		0.487		0.675		0.683	
8	SAM-Iris	SMD+SLD	0.627	0.640	0.574	0.619	0.716	0.693	0.489	0.497	0.658	0.677	0.631	0.672
		SynMOBIUS	0.662		0.648		0.710		0.516		0.700		0.707	
		MOBIUS	0.592		0.743		0.539		0.482		0.729		0.779	
9	ShapeGAN-DLV3+	SMD+SLD	0.401	0.525	0.430	0.612	0.517	0.545	0.275	0.396	0.483	0.629	0.492	0.661
		SynMOBIUS	0.652		0.813		0.583		0.534		0.755		0.817	

The probabilistic results for KU-CVML are not reported due to issues in the submission.

6.1. Overall Results and Performance Ranking

To evaluate and rank the SSBC 2025 submissions, we computed **average performance scores** over the submitted segmentation masks. The standard errors reported in this section were obtained by partitioning the test data into 5 subject-disjoint folds and computing the corresponding standard deviation. The harmonic mean F_1 score computed across the three evaluation datasets was used as the criterion for model ranking. This ensured that the scoring system rewards the submissions that maintain a steady high performance across all evaluation datasets, rather than performing well on some datasets and failing on others.

Synthetic Track. Since SSBC 2025 focuses on the use of synthetically generated data in model development, the first track of the competition benchmarked models trained solely on synthetic data. The results of this track are presented in Fig. 6a and Table 3. The winner of the Synthetic track is **SwinDANet**, which outperformed all other approaches in all the computed performance metrics. Most models, however, achieved reasonably competitive results, with all but 3 of the submissions landing in the 0.72–0.8 F_1 score range.

The first row of Fig. 7 shows how the performance of the models varies with different choices of the binarization threshold in the grayscale predictions. Observe from Table 3 that several models exhibit significant discrepancies between their binary F_1 scores and the optimal F_1 scores obtained from the grayscale predictions. This implies that the models may benefit greatly from employing a dedicated binarization threshold selection algorithm.

It is interesting to see that the models' performance, in general, matched well between the MOBIUS data and MOBIUS-like synthetic data (SynMOBIUS), demonstrating a considerable correspondence between the generated synthetic images and the original real-world data. F_1 scores of over 0.8 for the best performing model indicate that it is possible to learn a competitive segmentation model even from synthetic data with quite different characteristics than the test data. When looking at the performance on SMD+SLD, we observe consistently weaker results for all models compared to the respective performance on MOBIUS or SynMOBIUS, suggesting that the domain shift between the synthetic training data and the SMD+SLD images significantly impacts segmentation results - even for the best performing models.

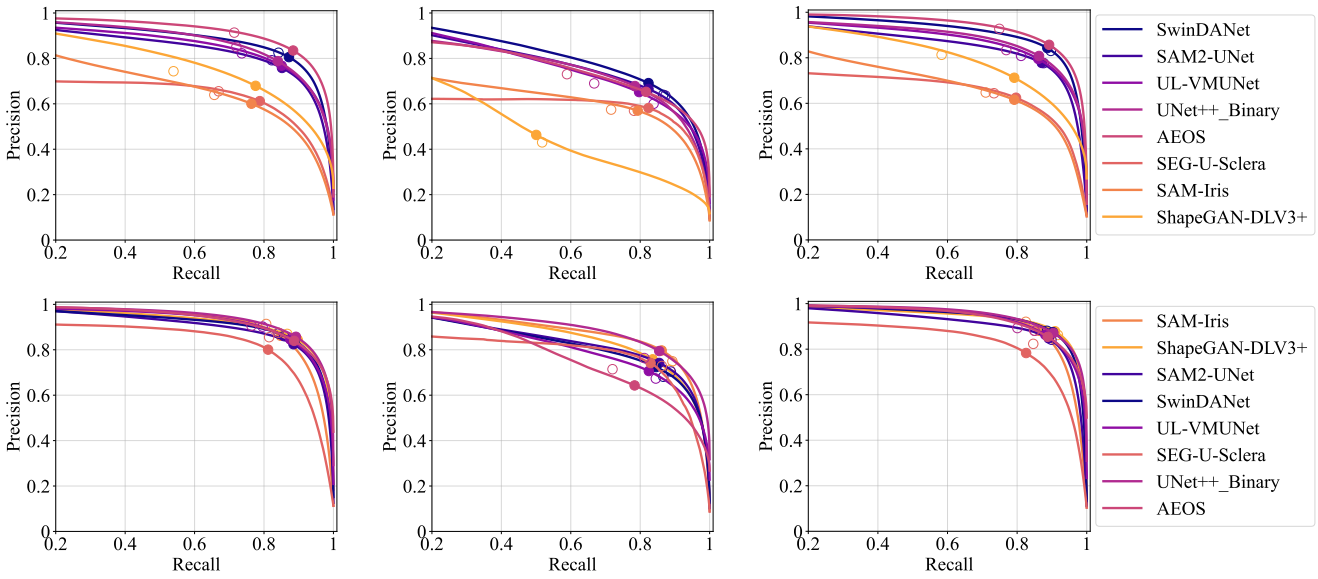


Figure 7. **Precision-recall curves for the submitted SSBC 2025 models**. The operating points denoted with a full circle represent the best possible F_1 score (F_1^{opt}), whereas the empty circle denotes the precision-recall point produced by the binary masks. The first row displays the results of the submissions trained on synthetic data, whereas the second row contains submissions trained in the mixed-data scenario. The columns show the results on the evaluation datasets in following order: MOBIUS, SMD+SLD and SynMOBIUS. The legends are sorted by the corresponding track's ranking. The figure is best viewed in color and zoomed in.

Mixed Track. In the mixed track, participants could use both the synthetic as well as real-world data for model training. As can be seen from Fig. 6b and Table 4, the results of this track were closer (considering the strongest and weakest model) than those of the synthetic track. This suggests that it is still easier to train a well performing model with a mix of real and synthetic data than with synthetic data alone, where dedicated mechanisms need to be incorporated into the learning procedure to account for the domain shift w.r.t. real-world data. The winner of the mixed track, **SAM-Iris**, and the runner-up, **ShapeGAN-DLV3+**, were within 0.001 of each other in terms of their F_1 scores, which is well within the margin of error. As such, they are considered the joint winners of the SSBC 2025 Mixed track. Notably, all the approaches resulted in competitive performance, with F_1 scores in the range of 0.78–0.84.

It is interesting to note that the two worst performers from the Synthetic track performed the best in the Mixed track, driven chiefly by their significant performance improvement on the SMD+SLD evaluation data. The two best performers from the Synthetic track maintained their high performance, taking the 3rd and 4th place in the Mixed track.

The second row of Fig. 7 shows the Mixed track results over the grayscale predictions. We observe the significantly tighter precision-recall curves relative to the Synthetic track in the top row, again implying a much closer performance of the submitted models. The discrepancies between the binary F_1 and optimal grayscale F_1 scores are still present, but significantly lower. This implies that a good thresholding strategy is of less importance with mixed-trained models. The exception is UNet++_Binary, which achieves the top performance in its probabilistic metrics, but only the 7th place in the binary ranking.

6.2. Differential Performance

The goal of SSBC 2025 was to establish how viable the use of synthetic data is in developing sclera segmentation models. As such, we are interested in the performance differentials between the models trained on the synthetic data and their mixed-trained counterparts.

From Fig. 8, we can see that every single model's performance improved from the Synthetic to the Mixed track. However, for most of the models, the performance boost was on the lower end, mostly under 0.05 in terms of F_1 score. However, for the three worst-performing models of the Synthetic track, the performance boost from adding the real-world training data was substantial. This implies that the choice of architecture is important when training on synthetic data alone, as certain model architectures perform significantly worse without real-world samples to complement the synthetic data. However, with many of the performance boosts being small, the synthetic training data has shown to be a feasible approach for training most models in cases where privacy protection is essential.

The largest difference in performance was seen with the ShapeGAN-DLV3+ model, which features a generator-discriminator architecture, albeit different from classical GANs, since the discriminator is attached to the generator

Table 4. **Comparative assessment of the models trained on mixed data.** The submissions are ranked according to the harmonic means of the achieved F_1 scores over the three evaluation datasets. 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. The harmonic means are also reported for the rest of the performance measures in the column next to the individual dataset results. Note that the ranking is quite consistent across performance indicators.

Rank	Segmentation Model	Evaluation Dataset	From binary masks								From probabilistic predictions			
			F_1		Precision		Recall		IoU		F_1^{opt}		AUC	
		MOBIUS	0.850		0.914		0.806		0.751		0.876		0.918	
1	SAM-Iris	SMD+SLD	0.804	0.839	0.749	0.853	0.893	0.840	0.684	0.733	0.833	0.865	0.881	0.910
		SynMOBIUS	0.865		0.920		0.825		0.769		0.889		0.933	
		MOBIUS	0.858		0.870		0.867		0.769		0.882		0.926	
1	ShapeGAN-DLV3+	SMD+SLD	0.780	0.838	0.712	0.808	0.885	0.889	0.649	0.733	0.811	0.863	0.862	0.909
		SynMOBIUS	0.884		0.862		0.917		0.800		0.901		0.944	
		MOBIUS	0.847		0.860		0.858		0.756		0.871		0.906	
3	SAM2-UNet	SMD+SLD	0.775	0.826	0.704	0.795	0.888	0.881	0.649	0.720	0.806	0.852	0.842	0.887
		SynMOBIUS	0.862		0.841		0.899		0.767		0.882		0.918	
		MOBIUS	0.848		0.873		0.842		0.756		0.875		0.916	
4	SwinDANet	SMD+SLD	0.752	0.822	0.681	0.800	0.866	0.864	0.623	0.715	0.786	0.850	0.833	0.896
		SynMOBIUS	0.878		0.881		0.886		0.790		0.896		0.945	
		MOBIUS	0.835		0.902		0.796		0.732		0.873		0.927	
5	UL-VMUNet	SMD+SLD	0.739	0.808	0.673	0.804	0.845	0.830	0.600	0.690	0.770	0.838	0.820	0.893
		SynMOBIUS	0.859		0.881		0.850		0.760		0.881		0.942	
		MOBIUS	0.823		0.855		0.815		0.722		0.843		0.834	
5	SEG-U-Sclera	SMD+SLD	0.778	0.807	0.729	0.799	0.862	0.841	0.659	0.699	0.811	0.833	0.786	0.820
		SynMOBIUS	0.823		0.823		0.846		0.721		0.846		0.840	
		MOBIUS	0.811		0.906		0.764		0.710		0.889		0.941	
7	UNet++_Binary	SMD+SLD	0.769	0.803	0.764	0.849	0.812	0.792	0.641	0.692	0.835	0.872	0.892	0.928
		SynMOBIUS	0.832		0.893		0.801		0.730		0.895		0.954	
		MOBIUS	0.850		0.875		0.842		0.754		0.874		0.934	
8	AEOS	SMD+SLD	0.701	0.797	0.714	0.816	0.720	0.799	0.569	0.682	0.733	0.823	0.782	0.880
		SynMOBIUS	0.859		0.881		0.850		0.762		0.880		0.945	
		MOBIUS	0.795		0.725		0.928		0.689		-		-	
9	KU-CVML	SMD+SLD	0.719	0.780	0.601	0.688	0.928	0.937	0.576	0.658	-	-	-	-
		SynMOBIUS	0.835		0.758		0.954		0.730		-		-	

The probabilistic results for KU-CVML are not reported due to issues in the submission.

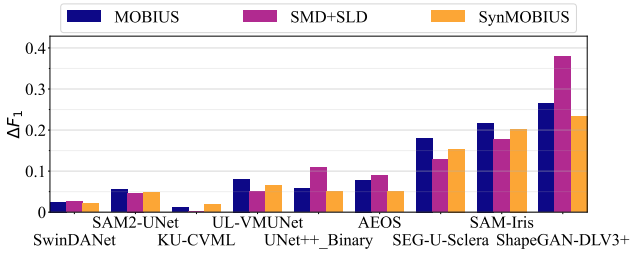
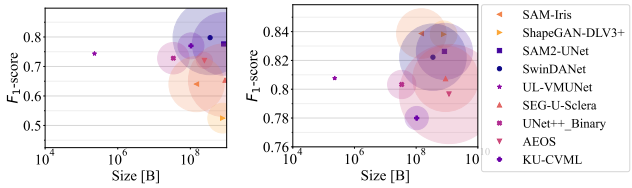


Figure 8. **Performance differentials** of the submitted models from the Synthetic track to the Mixed track on three test datasets, i.e., MOBIUS, SMD+SLD and SynMOBIUS. The graph shows differences in the achieved F_1 scores, i.e., $F_1^{mixed} - F_1^{synthetic}$. Positive values denote a better performance in the Mixed track.



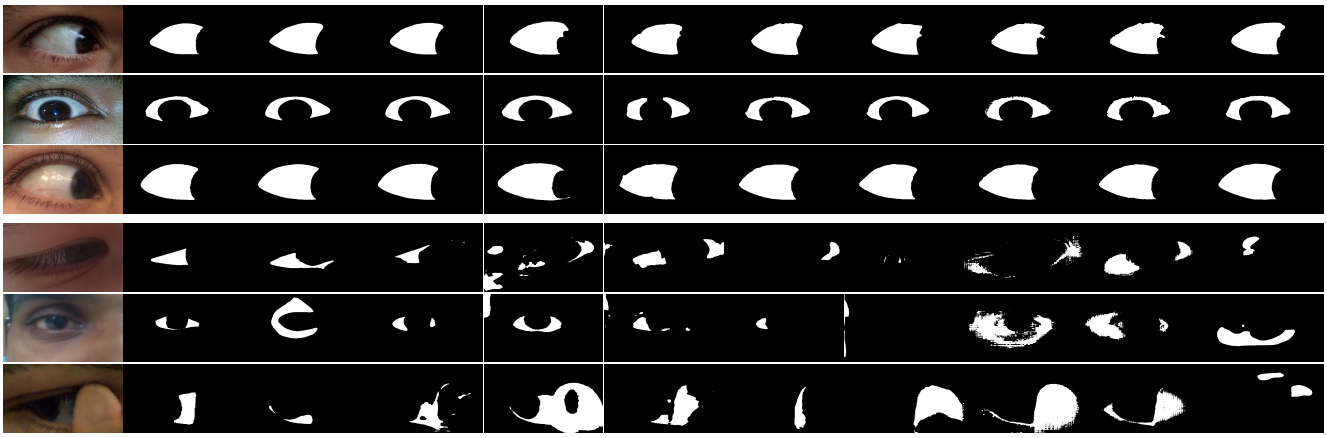
(a) Synthetic (b) Mixed

Figure 9. **Performance/complexity trade-off**. The areas of the circles correspond to the computational complexity of the models, expressed in FLOPs. Best viewed in color and zoomed in.

bottleneck. Since GANs are well known to require large amounts of training data, we can partly attribute the difference in performance simply to the increased size of the training dataset with additional real-world samples. However, we note that the real-world SBVPI dataset was significantly smaller than the synthetic training dataset. The second-largest difference came with the SAM-Iris model, which relies on a training pipeline with an explicit transition from synthetic to real training data, which explains the discrepancy in performance. Similarly, AEOS (which resulted in the fourth-largest difference) employed completely different architectures for the two tracks. As such, we note that three of the four biggest performance differentials between the Synthetic and the Mixed track are a result of explicit methodological choices rather than domain shifts.

6.3. Performance Across Different Complexities

Finally, we study the trade-off between model performance and complexity, which is a key factor for real-world



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Figure 10. Qualitative comparison of the submitted models (in terms of binary masks) on selected images. The top three rows contain the well-performing samples from the evaluation datasets (in order, MOBIUS, SMD+SLD, SynMOBIUS), while the bottom three rows show poor-performing samples. 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) SwinDANet, (d) SAM2-UNet, (e) KU-CVML, (f) UL-VMUNet, (g) UNet++_Binary, (h) AEOS, (i) SEG-U-Sclera, (j) SAM-Iris, and (k) ShapeGAN-DLV3+.

deployment. From Fig. 9, we see that most of the submitted models' complexities are roughly on par. However, a notable exception here is UL-VMUnet, which achieved competitive performance despite a comparatively tiny size (229 KB) and computational complexity (60 MFLOPs). This result is in line with recent works in sclera biometrics [56,60], where it was shown that typically employed network architectures could be significantly reduced in size and complexity without a noticeable degradation in performance.

It is also worth noting that UL-VMUNet ranked higher in the Synthetic than in the Mixed track, which may imply that smaller, less complex models generalize better when trained with synthetic data and are not disturbed as easily by artifacts produced by the data generation procedure.

6.4. Qualitative Evaluation

In Fig. 10, we show some of the best and worst cases of the generated segmentation masks. Note that even on the best-performing samples, models often struggled with minor artifacts, such as those resulting from specular reflection in the iris. Among the poor performing samples, we often find images where the sclera is occluded, either by the eyelid (in the case of a partially closed eye, such as in row 4 of Fig. 10) or by external sources (such as the finger in row 6 of Fig. 10). Another source of errors is lighting-induced skin discolouration (row 5 of Fig. 10), particularly when the skin around the eye appears significantly lighter than elsewhere. Additionally, a poorly-lit sclera (row 6 of Fig. 10) can also cause issues in segmentation with many of the models.

7. Conclusion

The 2025 edition of the Sclera Segmentation Benchmarking Competition (SSBC 2025) was organized to benchmark the performance of contemporary segmentation models in the task of sclera segmentation, and to establish the viability of using synthetically generated data to develop and train such models. The use of synthetic data ensures no identifying information is required to train segmentation models and facilitates the ethical development of biometric systems. To this end, SSBC 2025 was conducted in two tracks, which differed in the training data the contestants used – purely synthetic data for the first track and a mix of synthetic and real-world data for the second. Nine research groups participated in the competition.

The winner of the Synthetic track was **SwinDANet** (submitted by the Couger Inc. team), while the Mixed track had two joint winners, whose results were within 0.001 of each other: **SAM-Iris** (Idiap – HES-SO team) and **ShapeGAN-DLV3+** (AU team). The results of the competition point to the high fidelity of synthetic data and its viability in model training, as most approaches performed roughly as well in the Synthetic track as in the Mixed one, and the models that did substantially improve their performance between the tracks, did so mostly due to methodological decisions, rather than any infidelities or improprieties in the synthetic training data. This is an important result for future research, as synthetically generated datasets enable biometric research without the risks of privacy breaches or abuse, as well as being significantly easier to obtain or compile.

Acknowledgments

Supported in parts by the ARIS Research Programmes P2-0250 Metrology and Biometric Systems, P2-0214 Computer Vision, the ARIS young researcher program, and the SNSF Projects 214653 FairMI and 213369 StreamKG.

References

[1] S. Alkassar, W.-L. Woo, S. Dlay, and J. Chambers. Sclera recognition: on the quality measure and segmentation of degraded images captured under relaxed imaging conditions. *IET Biometrics*, 6(4):266–275, 2016. 1

[2] S. Alkassar, W. L. Woo, S. S. Dlay, and J. 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. 1

[3] S. Borse, Y. Wang, Y. Zhang, and F. Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5911, 2021. 6

[4] K. Boyd, K. H. Eng, and C. D. Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Springer Joint European conference on machine learning and knowledge discovery in databases (ECML PKDD)*, pages 451–466, 2013. 5

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2018. 6

[6] A. Das. *Towards Multi-modal Sclera and Iris Biometric Recognition with Adaptive Liveness Detection*. PhD thesis, Griffith University, 2017. 2, 3

[7] A. Das, S. Atreya, A. Mukherjee, M. Vitek, H. Li, C. Wang, G. Zhao, F. Boutros, P. Siebke, J. N. Kolf, et al. Sclera segmentation and joint recognition benchmarking competition: Ssrbc 2023. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2023. 2

[8] A. Das, R. Kunwar, U. Pal, M. A. Ferrer, and M. Blumenstein. An online learning-based adaptive biometric system. In *Adaptive Biometric Systems*, pages 73–96. Springer, 2015. 2

[9] A. Das, P. Mondal, U. Pal, M. Blumenstein, and M. A. Ferrer. Sclera vessel pattern synthesis based on a non-parametric texture synthesis technique. In *Springer International Conference on Computer Vision and Image Processing (CVIP)*, pages 241–250, 2017. 2

[10] A. Das, U. Pal, M. A. F. Ballester, and M. Blumenstein. Sclera recognition using dense-sift. In *IEEE International Conference on Intellient Systems Design and Applications (ISDA)*, pages 74–79, 2013. 1

[11] A. Das, U. Pal, M. A. F. Ballester, and M. Blumenstein. A new efficient and adaptive sclera recognition system. In *IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 1–8, 2014. 2

[12] A. Das, U. Pal, M. Blumenstein, and M. A. F. Ballester. Sclera recognition-a survey. In *IEEE IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 917–921, 2013. 1

[13] A. Das, U. Pal, M. Blumenstein, C. Wang, Y. He, Y. Zhu, and Z. Sun. Sclera segmentation benchmarking competition in cross-resolution environment. In *IEEE/IAPR International Conference on Biometrics*, 2019. 2

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

[15] A. Das, U. Pal, M. A. Ferrer, and M. Blumenstein. A framework for liveness detection for direct attacks in the visible spectrum for multimodal ocular biometrics. *Elsevier Pattern Recognition Letters*, 82:232–241, 2016. 1, 2

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

[17] 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 *IEEE International Joint Conference on Biometrics (IJCB)*, pages 794–798, 2017. 2

[18] A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, Z. Emeršič, P. Peer, V. Štruc, and S. Kumar. SSERBC 2017: Sclera segmentation and eye recognition benchmarking competition. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 742–747, 2017. 2

[19] A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, P. Peer, and V. Štruc. SSBC 2018: Sclera Segmentation Benchmarking Competition. In *International Conference on Biometrics (ICB)*, pages 303–308, 2018. 2

[20] S. Das, I. De Ghosh, and A. Chattopadhyay. An efficient deep sclera recognition framework with novel sclera segmentation, vessel extraction and gaze detection. *Signal Processing: Image Communication*, 97:116349, 2021. 1

[21] S. Das, I. De Ghosh, and A. Chattopadhyay. Sclera biometrics in restricted and unrestricted environment with cross dataset evaluation. *Displays*, 74:102257, 2022. 1

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6

[23] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer. Pixel-wise ear detection with convolutional encoder-decoder networks. *IET Biometrics*, 2017. 5

[24] M. B. Erden, S. Cansiz, O. Caki, H. Khattak, D. Etiz, M. C. Yakar, K. Duruer, B. Barut, and C. Gunduz-Demir. Fourierloss: Shape-aware loss function with fourier descriptors. *Neurocomputing*, 638:130155, 2025. 6

[25] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019. 2

[26] V. Gottemukkula, S. Saripalle, S. P. Tankasala, and R. Derakhshani. Method for using visible ocular vasculature for mobile biometrics. *IET Biometrics*, 5(1):3–12, 2016. 2

[27] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 6

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4, 5

[29] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017. 6

[31] P. Kansal and S. Devanathan. Eyenet: Attention based convolutional encoder-decoder network for eye region segmentation. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3688–3693, 2019. 6

[32] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12104–12114, 2020. 4

[33] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 3, 4

[34] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[35] A. Kirillov, E. Mintun, N. Ravi, H. Mao, L. Rolland, R. Gustafson, C. Xiao, S. Whitehead, A. Cevallos, Z. Howald, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 6

[36] R. Liu, J. Lehman, et al. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 6

[37] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. 6

[38] J. Lozej, B. Meden, V. Štruc, and P. Peer. End-to-end iris segmentation using U-Net. In *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6, 2018. 5

[39] J. Lozej, D. Štepec, V. Štruc, and P. Peer. Influence of segmentation on deep iris recognition performance. In *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6, 2019. 5

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

[41] P. Radu, J. Ferryman, and P. Wild. A robust sclera segmentation algorithm. In *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, 2015. 1

[42] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6

[43] D. Riccio, N. Brancati, M. Frucci, and D. Gragnaniello. An unsupervised approach for eye sclera segmentation. In *Springer Iberoamerican Congress on Pattern Recognition*, pages 550–557, 2017. 1

[44] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Springer International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 6

[45] P. Rot, Ž. Emeršič, V. Štruc, and P. Peer. Deep multi-class eye segmentation for ocular biometrics. In *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–8, 2018. 5

[46] P. Rot, M. Vitek, K. Grm, Ž. Emeršič, P. Peer, and V. Štruc. Deep sclera segmentation and recognition. In *Handbook of vascular biometrics*, pages 395–432. Springer, 2020. 1, 2

[47] C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *PMLR International Conference on Machine Learning (ICML)*, pages 29441–29454, 2023. 6

[48] 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*, 10(3):e0118432, 2015. 5

[49] A. Sequeira, L. Chen, J. Ferryman, P. Wild, F. Alonso-Fernandez, J. Bigun, K. Raja, R. Raghavendra, C. Busch, T. F. Pereira, et al. Cross-eyed 2017: Cross-spectral iris/periocular recognition database and competition. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2017. 4

[50] M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning (ICML)*, 2019. 6

[51] D. Tomašević, F. Boutros, N. Damer, P. Peer, and V. Štruc. Generating bimodal privacy-preserving data for face recognition. *Engineering Applications of Artificial Intelligence (EAAI)*, 133:108495, 2024. 2

[52] D. Tomašević, F. Boutros, C. Lin, N. Damer, V. Štruc, and P. Peer. ID-Booth: Identity-consistent face generation with diffusion models. *arXiv preprint arXiv:2504.07392*, 2025. 2

[53] D. Tomašević, P. Peer, and V. Štruc. BiOcularGAN: Bimodal synthesis and annotation of ocular images. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2022. 2, 3, 4

[54] D. Tomašević, P. Peer, and V. Štruc. BiFaceGAN: Bimodal Face Image Synthesis. In *Face Recognition Across the Imaging Spectrum (FRAIS)*, pages 273–311. Springer, 2024. 2

[55] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 969–977, 2018. 6

[56] M. Vitek, M. Bizjak, P. Peer, and V. Štruc. Ipad: Iterative pruning with activation deviation for sclera biometrics. *Journal of King Saud University-Computer and Information Sciences*, 35(8):101630, 2023. 1, 2, 10

[57] M. Vitek, A. Das, D. R. Lucio, L. A. Zanlorensi, D. Menotti, J. N. Khiarak, M. A. Shahpar, M. Asgari-Chenaghlu,

F. Jaryani, J. E. Tapia, et al. Exploring bias in sclera segmentation models: A group evaluation approach. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18:190–205, 2022. 1, 2, 3

[58] M. Vitek, A. Das, Y. Pourcenoux, A. Missler, C. Paumier, S. Das, I. De Ghosh, D. R. Lucio, L. A. Zanlorensi Jr., D. Menotti, F. Boutros, N. Damer, J. H. Grebe, A. Kuijper, J. Hu, Y. He, C. Wang, H. Liu, Y. Wang, Z. Sun, D. Osorio-Roig, C. Rathgeb, C. Busch, J. Tapia Farias, A. Valenzuela, G. Zampoukis, L. Tsochatzidis, I. Pratikakis, S. Nathan, R. Suganya, V. Mehta, A. Dhall, K. Raja, G. Gupta, J. N. Khiarak, M. Akbari-Shahper, F. Jaryani, M. Asgari-Chenaghlu, R. Vyas, S. Dakshit, S. Dakshit, P. Peer, U. Pal, and V. Štruc. SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2020. 2, 4

[59] M. Vitek, P. Rot, V. Štruc, and P. Peer. A comprehensive investigation into sclera biometrics: a novel dataset and performance study. *Springer Neural Computing and Applications*, pages 1–15, 2020. 1, 2

[60] M. Vitek, V. Štruc, and P. Peer. Gazenet: A lightweight multitask sclera feature extractor. *Alexandria Engineering Journal*, 112:661–671, 2025. 1, 2, 10

[61] R. Wu, Y. Liu, P. Liang, and Q. Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*, 2024. 6

[62] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 6

[63] X. Xiong, Z. Wu, S. Tan, W. Li, F. Tang, Y. Chen, S. Li, J. Ma, and G. Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *ArXiv*, abs/2408.08870, 2024. 6

[64] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10145–10155, 2021. 4

[65] Z. Zhou, E. Y. Du, N. L. Thomas, and E. 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. 1

[66] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, 2018. 6

[67] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convolutional networks v2: More deformable, better results. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6