Label-Efficient Sclera Segmentation with Foundation Models: SSBC 2026

M. Vitek[1,*], D. D. Premani[2,*], A. Das[3], U. N. Modi[2], M. S. Raval[2], H. H. Bhatt[2], P. Baumstimler[5], Z. A. Daniels[6], A. H. Sheth[2], M. Gondalia[2], K. J. Shah[2], C. Zhang[7], C. Wang[7], A. Shin[8], K. Shigematsu[8], H. Shirono[9], K. Inoue[8], G. Gupta[10], A. Gupta[10], G. Sharma[11], T. Gupta[11], A. Nigam[11], R. Ramachandra[12], J. Nourmohammadi Khiarak[13], T. Akbari Saeed[13], F. Nourmohammadi Khiarak[13], A. Kumar[11], U. Pal[4], P. Peer[1], V. Štruc[1]

[1]University of Ljubljana (UL, SI), [2]Ahmedabad University (AU, IN), [3]Birla Institute of Technology & Science Pilani (BITS Pilani, IN), [4]Indian Statistical Institute (ISI, IN), [5]Polytechnique Montréal (PM, CA), [6]SRI International (SRI, USA), [7]Beijing University of Civil Engineering and Architecture (BUCEA, CN), [8]National Institute of Technology, Oita College (NIT Oita, JP), [9]Kyhusu Institute of Technology (KIT, JP), [10]Arogya Pandit (AP, IN), [11]Indian Institute of Technology Mandi (IIT Mandi, IN), [12]Norwegian University of Science and Technology (NTNU, NO), [13]Kartal Ol NGO (KO, PL)

Abstract

This paper presents a summary of the 2026 Sclera Segmentation Benchmarking Competition (SSBC), which focused on the adaptation of foundation models for the task of sclera segmentation and the implementation of label-efficient training paradigms. The goal of the competition was to evaluate whether massive, generalized architectures can surpass domain-specific segmentation networks, and to assess the capacity of label-efficient strategies to overcome severe data scarcity. The competition featured two tracks, i.e.: (1) the Foundation Model Adaptation track, which provided fully annotated training data to adapt pre-trained (foundation) backbones, and (2) the Label-Efficient Learning track, which restricted access to ground truth annotations but supplied a massive collection of unlabelled imagery for training arbitrary sclera segmentation models, not just foundation model. The submitted approaches employed a variety of architectural designs, including parameter-efficient adaptations, vision-language guidance, and semi-supervised multi-architecture ensembles. To score the competition entries, experiments were conducted across three sequestered evaluation datasets containing real-world and synthetic images collected under diverse conditions. Results show that label-efficient methodologies can achieve remarkable accuracy, with the top Track 2 model performing comparably to the fully supervised Track 1 winner. Moreover, comparisons with SSBC 2025 results suggest that properly (domain) adapted foundation backbones significantly outperform previous approaches under the same training/evaluation protocol and thus represent a new state-of-the-art in sclera segmentation.

1. Introduction

Inferring human identity through the computer-aided analysis of unique eye features forms the core of ocular biometrics. Historically, the field has been dominated by iris recognition; however, the research focus is now expanding to include alternative traits that can operate independently or alongside traditional methods. The sclera, i.e., the visible white of the eye, stands out in this regard. Its high resistance to presentation attacks, temporal stability, and distinct vascular structures make it a highly promising biometric modality [12,54,68]. Furthermore, practical deployment in real-world scenarios is significantly easier than with the iris, as scleral imaging requires only standard cameras operating under visible light and remains robust against typical visual degradations [54,68].

Scientific interest in sclera biometrics has escalated significantly, encompassing diverse areas such as recognition algorithms [2, 10, 15, 20, 21, 51, 68, 69, 82], segmentation techniques [1, 47, 54, 65, 66], presentation attack detection [15, 54, 68], user adaptability [8, 11], fusion strategies [17, 27], and lightweight architectures [26, 65, 69]. At the same time, recent advances in Vision-Language Models (VLMs) [24, 58] and Foundation Models (FMs) [25, 43, 58, 74] are reshaping modern computer vision and increasingly influencing biometric research. These large-scale architectures demonstrate remarkable zero-shot and transfer-learning capabilities, while also offering substantial potential for adaptation to specialised biometric tasks. Consequently, an important open question is how to effectively adapt such generalized models for fine-grained ocular analysis tasks. Simultaneously, the growing reliance on data-intensive deep neural networks has amplified the long-standing challenge of obtaining pixel-perfect ground truth annotations. Although previous efforts to mitigate data scarcity primarily relied on synthetically generated imagery [9,61–64], recent progress in semi-supervised [43,73] and un-/self-supervised learning [41, 73] has demonstrated that large collections of unlabelled data can be leveraged effectively through label-efficient training paradigms.

To investigate these emerging paradigms within ocular biometrics, the 2026 Sclera Segmentation Benchmarking Competition (SSBC 2026) was organised around two

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

complementary research tracks. The first track, *Foundation Model Adaptation*, focuses on evaluating how effectively modern foundation architectures can be adapted for fine-grained biometric segmentation tasks. Rather than relying solely on their zero-shot capabilities, the track aims to assess whether domain adaptation and task-specific refinement strategies can transform large general-purpose models into robust sclera segmentation systems. The second track, *Label-Efficient Learning*, addresses the persistent challenge of annotation scarcity in ocular biometrics. Its primary goal is to evaluate whether semi-supervised, unsupervised, and self-supervised learning paradigms can leverage large quantities of unlabelled ocular imagery to reduce dependence on costly manual annotations. This problem is particularly important for highly specialised domains such as sclera (vessel) segmentation, where obtaining large-scale pixel-level annotations remains exceptionally difficult.

Through the collaborative efforts of the organisers and participating teams, this work makes the following main contributions:

- A comprehensive and independent evaluation of foundation model adaptation strategies for fine-grained sclera segmentation under a unified benchmarking protocol and across identical evaluation datasets;
- A rigorous analysis of the effectiveness of label-efficient learning techniques under conditions of severe (annotated) data scarcity using the same experimental and evaluation framework; and
- An investigation into the inherent trade-offs between computational complexity, parameter count, and task-specific generalization across varied architectural frameworks, while also establishing a new state of the art in sclera segmentation.

2. Related Work

SSBC 2026 represents the 10th edition of the sclera-segmentation competition series, initiated at the BTAS conference in 2015. Since its inception, the competition series has provided a standardized experimental framework for the systematic evaluation of sclera segmentation methods under diverse acquisition conditions and application scenarios. Each edition has focused on a specific methodological challenge, including cross-sensor generalisation, image resolution variability, mobile-domain segmentation, privacy-preserving training, and synthetic data generation, thus enabling the controlled assessment of emerging segmentation paradigms and their applicability to ocular biometrics.

The 1st and 3rd events (SSBC 2015 and SSBC 2016) assessed the baseline segmentation efficacy of various models while providing novel datasets, i.e. MASD and SMD [14]. Extending this foundation, the 2nd iteration (SSRBC 2016) also evaluated recognition frameworks alongside segmentation techniques [16]. The 4th edition, SSERBC 2017, retained the recognition task but further analysed how gaze direction impacts both segmentation and matching accuracy [18]. The influence of cross-sensor image acquisition was the primary focus of the 5th competition, SSBC 2018 [19], whereas the 6th iteration, SSBC 2019, examined algorithmic performance across varying image resolutions [13]. Addressing the shift towards mobile biometrics, the 7th edition, SSBC 2020, released the MOBIUS dataset and concentrated on mobile-domain sclera segmentation [67]. Building upon that work, a subsequent study [66] quantified the distinct biases present in contemporary segmentation architectures. The 8th iteration (SSRBC 2023) subsequently evaluated segmentation and recognition indicators both independently and in conjunction [7]. Most recently, the 9th edition (SSBC 2025) investigated the viability of privacy-preserving segmentation model development utilising synthetically generated training datasets [70].

Table 1. **Summary of SSBC 2026 datasets.** Reported are the number of images and subjects, label availability across tracks, main sources of variability, and the overall purpose in the competition. T1 denotes track 1 and T2 denotes track 2.

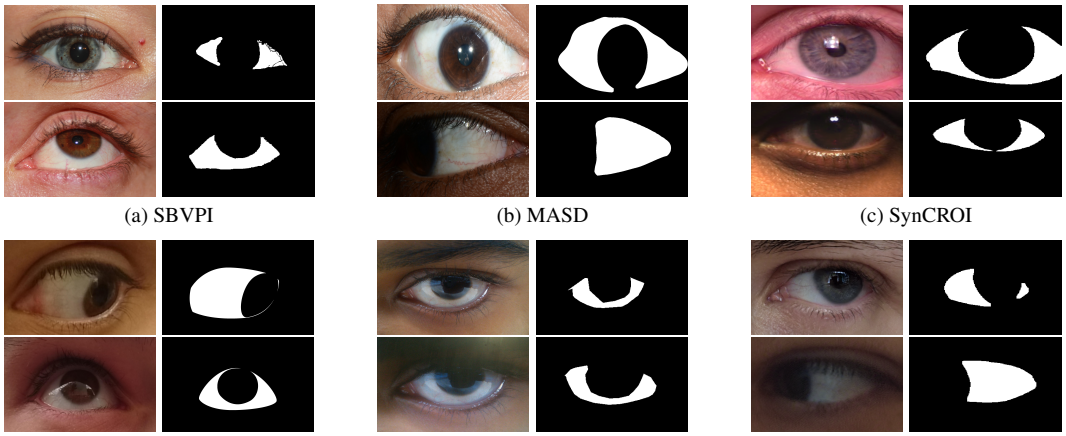
Dataset	#Images	#IDs	Annotations	Variability	Purpose
MASD	2606	82	Full (T1) / 82 (T2)	GZ, CN, BL	Training
SBVPI	1856	55	Full (T1) / 55 (T2)	GZ, CLR	Training
SynCROI	11000	N/A	Full (T1) / 220 (T2)	GZ, CLR, CN	Training
MOBIUS (Unlabelled)	13158	100	0	GZ, CLR, CN	Training (T2)
SMD+SLD	489	52	Full	CN, BL	Testing
MOBIUS (Annotated)	3539	35	Full	GZ, CLR, CN	Testing
SynMOBIUS	4771	N/A	Full	GZ, CLR, CN	Testing

(GZ) - gaze, (CLR) - eye colour, (CN) - acquisition condition, (BL) - blur

Building on the success of the previous editions, SSBC 2026 focuses on two emerging methodological directions in modern computer vision: foundation model adaptation and label-efficient learning. The primary objective of this edition is to evaluate how effectively large-scale pre-trained (foundational) architectures can be adapted for fine-grained sclera segmentation beyond their standard zero-shot capabilities, and whether such models can outperform conventional task-specific segmentation networks under a unified evaluation framework. In parallel, the competition investigates the effectiveness of semi-supervised, unsupervised, and self-supervised learning strategies in reducing dependence on large-scale annotated datasets. This problem is particularly relevant to related domains such as sclera vessel segmentation, where obtaining pixel-level annotations is substantially more difficult.

3. Competition Data

SSBC 2026 uses a combination of fully annotated and unlabelled datasets that are partitioned differently across the two competition tracks. Track 1 (Foundation Model Adaptation) relies on fully annotated training data to support the supervised adaptation of large pre-trained architectures. In contrast, Track 2 (Label-Efficient Learning) restricts access to ground truth annotations and instead emphasises



(a) SBVPI (b) MASD (c) SynCROI

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

Figure 1. **Samples from the training and testing datasets used in SSBC 2026**. The top row contains ocular images and available ground truth segmentation masks of the training datasets, while the bottom row depicts testing samples and their reference masks.

the use of large-scale unlabelled ocular imagery for semi-supervised, unsupervised, and self-supervised training.

Across both tracks, model evaluation is conducted using the same three sequestered test datasets: SMD+SLD, MOBIUS (annotated subset), and SynMOBIUS. All images are resized to 400×300 for the final evaluation. This unified evaluation protocol enables a direct comparison between foundation model adaptation and label-efficient learning approaches under identical testing conditions. A summary of all datasets, their role within the competition, and the availability of annotations across tracks is provided in Table 1, while sample image are shown in Fig. 1. Detailed descriptions of the individual datasets are provided below:

- **MASD.** The Multi-Angle Sclera Dataset (MASD) contains 2606 high-resolution RGB images acquired from 82 identities using a DSLR camera (Nikon D800) [14]. Captured across multiple gaze directions under varying environmental conditions, it serves as a primary training source. Track 1 utilises all available masks, whereas Track 2 restricts annotations via a one-per-subject selection strategy, resulting in 82 labels.
- **SBVPI.** The Sclera Blood Vessels, Periocular, and Iris (SBVPI) dataset provides 1856 high-resolution RGB images from 55 subjects, captured under highly controlled conditions using a Canon EOS 60D camera [68]. For SSBC 2026, it is utilised as a core training dataset. Track 1 employs the fully annotated dataset, while Track 2 is limited to one mask per person, yielding 55 labels.
- **SynCROI.** SynCROI consists of 11,000 synthetically generated images. It comprises two demographic subsets, SynCROI (CE) and SynCROI (PU), generated by training the generative BiOcularGAN framework [63] on Caucasian (CrossEyed) and Asian (PolyU) VIS-NIR image pairs. For Track 1, all labels are provided, while Track 2 makes a 2% annotated subset (220 labels) available. Combined with MASD and SBVPI, this yields a highly restricted total of just 357 ground truth labels for Track 2 model development.
- **MOBIUS.** The MOBIUS dataset features high-resolution imagery captured via mobile devices (e.g. Sony Xperia, Apple iPhone, Xiaomi) in real-world, unconstrained settings [67]. It serves a dual purpose in this competition. A subset of 13,158 unlabelled images is provided exclusively for Track 2 training, facilitating label-efficient domain adaptation. A sequestered subset of 3539 fully annotated images is reserved for the evaluation.
- **SMD+SLD.** Combining the Sclera Mobile Dataset [6] and the Sclera Liveness Dataset [66], this evaluation dataset comprises 489 RGB images. Captured via mobile devices, the dataset intentionally includes non-ideal scenarios, such as motion blur, blinking, and extreme lighting variations, to rigorously test model robustness.
- **SynMOBIUS.** Also used strictly for evaluation, this dataset contains 4771 image-mask pairs. It is synthesised by a modified, single-spectrum BiOcularGAN model trained directly on the MOBIUS dataset to assess segmentation performance on mobile-domain imagery. Because synthetic data generation was extensively covered in previous iterations, we refer the reader to the SSBC 2025 summary for comprehensive generation details [70].

4. Benchmarking Methodology

4.1. Competition Protocol

SSBC 2026 was structured into two primary phases. Initially, participating teams were granted access to the designated training datasets corresponding to their chosen track. To acquaint participants with the required submission format, a small set of sample testing data was also made avail-

Figure 2. **Illustration of expected submission formats**. The figure depicts an original input image alongside its corresponding generated binary segmentation mask and greyscale probabilistic segmentation prediction, from left to right.

able early in the timeline. Subsequently, the sequestered evaluation datasets were distributed without their corresponding ground truth segmentation masks. To mitigate the risk of extensive fine-tuning on the test imagery, participants were given a strictly limited time frame to return their final segmentation predictions. For each image from the MOBIUS, SMD+SLD, and SynMOBIUS evaluation datasets, teams were instructed to submit two distinct formats: *(i)* a **binary segmentation mask**, where non-zero pixels designate the sclera region and zero-valued pixels indicate the background; and *(ii)* a **greyscale probability map**, where the pixel intensities reflect the predicted confidence of belonging to the sclera. An illustration of the requested outputs is presented in Fig. 2. The binary masks were utilised to establish the ranking of the submitted models in SSBC 2026, whereas the probabilistic maps enabled a deeper examination of their behaviour through the generation of performance curves.

The benchmarking procedure was divided into two tracks aligned with the primary research objectives of the competition. **Track 1 (Foundation Model Adaptation)** focused on adapting large pre-trained vision-language and foundation architectures for robust sclera segmentation using the fully annotated MASD, SBVPI, and SynCROI datasets. In contrast, **Track 2 (Label-Efficient Learning)** evaluated semi-supervised, unsupervised, and self-supervised learning strategies under severely limited annotation availability, while leveraging a large collection of unlabelled MOBIUS imagery for domain adaptation. Regardless of the track, all submitted models were evaluated under the same protocol using identical sequestered test datasets.

4.2. Performance Indicators

The quantitative evaluation of the participating segmentation algorithms relied on both the binary masks and the continuous probabilistic predictions provided by the teams. Using the binary outputs, we computed several performance indicators derived from the counts of true positives (TP, representing accurately identified sclera pixels), false positives (FP, representing background regions erroneously labelled as sclera), and false negatives (FN, representing true sclera pixels missed by the model). The indicators include:

- **Precision**, which defines the proportion of accurately predicted sclera pixels relative to the total number of pixels classified as sclera by the algorithm, computed as $\frac{TP}{TP+FP}$ [39, 40, 53];
- **Recall**, which quantifies the proportion of true sclera pixels that were successfully identified by the model, calculated as $\frac{TP}{TP+FN}$ [23, 39, 40, 53];
- **F_1 score**, which represents the harmonic mean of precision and recall via the formula $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$ [67, 70] and serves as the primary indicator for ranking the submissions by managing the inherent trade-off between the aforementioned components;
- **Intersection over union (IoU)**, also known as the Jaccard index, which evaluates the spatial overlap between the predicted segmentation and the ground truth mask, normalised by their combined area and defined as $\frac{TP}{TP+FP+FN}$ [67, 70].

To expand the evaluation beyond discrete binary outputs, the submitted greyscale probability maps were utilised to generate continuous **precision–recall curves** [44, 57]. By analysing these curves, we extracted both the **area under the curve (AUC)** [3] and the optimal F_1 score (F_1^{opt}) as additional performance indicators, yielding a highly detailed comparative perspective of the algorithms across a full spectrum of confidence thresholds.

Table 2. **List of submitted SSBC 2026 entries** and their respective institutions, organised by track 1 (T1 and T2).

Team	T1	T2
ArogyaPandit (AP)	SAM-LoRA	DINOv2 MT Ensemble
Ahmedabad University Team 1 (AU-1)	SAM2-UNet+FDA	R3-Ensemble
Ahmedabad University Team 2 (AU-2)	ScleraCLIP	-
Ahmedabad University Team 3 (AU-3)	VLS-SegNet	U2Net-SSL
Beijing University of Civil Engineering and Architecture (BUCEA)	SegDINO-Sclera	UniMedCLIPSeg
Indian Institute of Technology Mandi (IIT-M)	CGDSeg_sclera	-
Indian Institute of Technology Mandi (IIT-M)	SAMSeg_sclera	-
National Institute of Technology, Oita College (NIT-OC)	DINOv2Refine	UCalShapeEns
Polytechnique Montréal (PM)	SAU-DINOv3	FreeSAU-DINOv3
SRI International (SRI)	-	FNOReg-ULVM-UNet
Kartal Ol NGO (KO)	ScleraSAM3-CLIP	-

5. Summary of Submitted Approaches

A total of 10 teams submitted 10 approaches to Track 1 and 7 approaches to Track 2. Table 2 provides an overview of the participating approaches, while a brief description of each is given below.

5.1. Track 1: Foundation Model Adaptation

DINOv2Refine (NIT-OC) is a fully fine-tuned DINOv2-large model [42] featuring a task-specific gated refinement decoder and an auxiliary edge prediction head. This approach introduces a domain-specific bias towards sharp sclera boundaries while preserving the robust visual representations of the foundation backbone.

CGDSeg_sclera (IIT-M) adapts a DINOv2 ViT-S/14 backbone [42] using LoRA [29] and a frozen CLIP ViT-B/32 text encoder [46]. Sclera-specific adaptation is achieved

through cross-attention grounding of visual tokens with text prompts, followed by parameter-efficient decoding via Dense-CSSE blocks [30, 55] within an FPN [35].

SAMSeg_sclera (IIT-M) adapts the SAM2.1 foundation model [49] utilising a Hiera-Large backbone [56]. To specialise the architecture for sclera segmentation while preventing overfitting, the image encoder is kept completely frozen, and only the mask and prompt encoders are fine-tuned. The network relies on a simple image-centre point prompt during inference to localise the sclera region.

SAM-LoRA (AP) adapts the Segment Anything Model (SAM) ViT-B backbone [32] for sclera segmentation using Low-Rank Adaptation (LoRA) [29]. To achieve parameter-efficient domain adaptation, the core encoder is frozen while LoRA layers are injected into its multi-head attention blocks. Additionally, the original prompt-based SAM decoder is replaced with a custom, lightweight, prompt-free progressive convolutional upsampling decoder.

ScleraSAM3-CLIP (KO) adapts the SAM 3 foundation model [4] utilizing LoRA [29] alongside a frozen CLIP vision-language semantic reranker [46]. To eliminate manual prompting, a lightweight ocular prompt generator automatically predicts bounding boxes and points, which are combined with text prompts. Multiple candidate masks generated by SAM 3 are scored by CLIP against positive and negative ocular concepts, fused, and finally processed by a boundary-refinement decoder.

SegDINO-Sclera (BUCEA) pairs a frozen DINOv3 ViT-S/16 backbone [59] with a DPT-style decoder [48]. It addresses the transition from synthetic to real training data via two-stage training, optimising the decoder using BCE, Dice, and boundary-aware losses.

SAU-DINOv3 (PM) employs a frozen DINOv3 ViT-Small backbone [59] with a custom step-attention decoder [79]. It refines multi-scale features using MLP blocks and convolutional upsampling [48, 52].

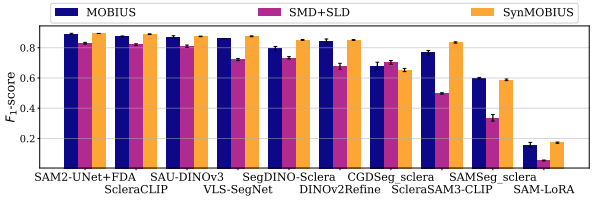
SAM2-UNet+FDA (AU-1) employs a two-stage DoRA-finetuned SAM2 foundation model [36, 49] integrated into a U-Net architecture. It addresses the training/evaluation distribution gap by applying Fourier Domain Adaptation (FDA) [80] for amplitude swapping, followed by a pseudo-labelling strategy to further improve robustness.

ScleraCLIP (AU-2) adapts the SAM2.1 Hiera-B+ backbone [49, 56] using shared and domain-adaptive LoRA [29]. It incorporates a text-semantic feature gating mechanism driven by frozen CLIP text embeddings [46] to modulate visual features. The model follows a two-stage synthetic-to-real curriculum, with final masks generated via a boundary-aware FPN [35] decoder.

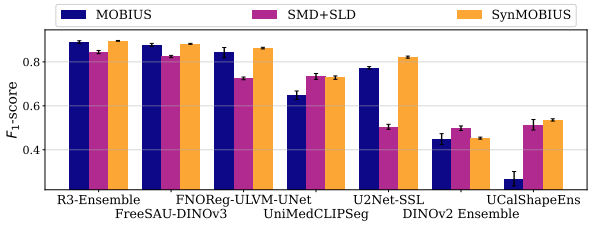
VLS-SegNet (AU-3) uses a Swin-Tiny backbone [38] with GSC/FUE skip blocks [77] and frozen BiomedCLIP semantic embeddings [81]. It fuses YOLOv8-generated [50] boxes and heuristic point prompts into the decoder via multi-head cross-attention.

Table 3. **Comparative assessment of Track 1 (Foundation Models) submissions, ranked by harmonic-mean F_1** across the three evaluation datasets (MOBIUS, SMD+SLD, SynMOBIUS). F_1, Precision, Recall, and IoU were computed from the submitted binary masks; F_1^{opt} and AUC from the probabilistic predictions. Per-dataset scores are shown in Fig. 3, and the full per-dataset numerical breakdown is provided in Table S2 of the supplementary material.

Rank	Model	F_1	Precision	Recall	IoU	F_1^{opt}	AUC
1	SAM2-UNet+FDA	0.870	0.831	0.927	0.776	0.896	0.941
2	ScleraCLIP	0.860	0.853	0.881	0.763	0.883	0.939
3	SAU-DINOv3	0.851	0.857	0.859	0.749	0.874	0.909
4	VLS-SegNet	0.813	0.782	0.880	0.707	0.844	0.887
5	SegDINO-Sclera	0.790	0.891	0.743	0.675	0.865	0.908
6	DINOv2Refine	0.781	0.849	0.761	0.665	0.816	0.851
7	CGDSeg_sclera	0.678	0.814	0.653	0.568	0.757	0.814
8	ScleraSAM3-CLIP	0.665	0.614	0.819	0.537	0.763	0.808
9	SAMSeg_sclera	0.473	0.454	0.518	0.341	0.530	0.516
10	SAM-LoRA	0.098	0.058	0.398	0.054	0.210	0.113



(a) Track 1 (Foundation Models)



(b) Track 2 (Label-Efficient Learning)

Figure 3. **Per-dataset F_1 scores** for submissions in (a) Track 1 (Foundation Models) and (b) Track 2 (Label-Efficient Learning), evaluated over MOBIUS, SMD+SLD, and SynMOBIUS.

5.2. Track 2: Label-Efficient Learning

UCalShapeEns (NIT-OC) ensembles Unet++ [83] and shape-aware refinement CNNs. It addresses label scarcity via semi-supervised self-training, employing a teacher-student framework with confidence-filtered pseudo-labels [34] and uncertainty-aware ensemble aggregation.

DINOv2 MT Ensemble (AP) utilizes an ensemble of DINOv2 ViT-B/14 and ViT-L/14 backbones [42] with FPN decoders [35]. To mitigate label scarcity, it employs a semi-supervised Mean Teacher framework [60], minimizing MSE consistency loss between student predictions and EMA-teacher pseudo-labels on unlabelled data.

UniMedCLIPSeg (BUCEA) adapts a UniMedCLIP ViT-B/16 backbone [31] and BiomedBERT encoder [28] via the MedCLIPSeg framework [33]. It addresses label scarcity through supervised transfer learning from pretrained medical vision-language foundation models, and fine-tuning on the limited-label dataset.

FNOReg-ULVM-UNet (SRI) couples an ultra-light Mamba-based U-Net [75] with an FNOReg neural operator [22]. It utilises a Mean Teacher semi-supervised framework [60] using confidence-filtered pseudo-labels and registration-based learned augmentation, where FNOReg warps labelled images/masks to mimic unlabelled anatomy, enhanced by Fourier-domain style transfer [80].

FreeSAU-DINOv3 (PM) is a semi-supervised segmentation framework based on the SAU-DINOv3 architecture. It addresses label scarcity using FreeMatch [72], which employs self-adaptive thresholding to generate reliable pseudo-labels from unlabelled data, ensuring the model effectively leverages the unlabelled dataset package.

R3-Ensemble (AU-1) employs a multi-architecture ensemble consisting of SegFormer-B2 [76], SAM2-UNet [78], and DeepLabV3+ [5] with a ConvNeXt-Large [37] encoder. It addresses label scarcity through three rounds of semi-supervised self-training, utilizing an ensemble teacher to generate confidence-filtered pseudo-labels on unlabeled MOBIUS data. Robustness to out-of-distribution captures is further improved by Fourier amplitude swapping (FDA) [80] used for both data augmentation and proxy validation.

U2Net-SSL (AU-3) utilizes an enhanced U2-Net backbone [45] within a two-stage semi-supervised pipeline [71]. It addresses label scarcity by leveraging 28k unlabeled images via photometric invariance (Stage 1) and spatial consistency using affine transformations (Stage 2).

6. Benchmarking Results

In this section, we present the results of SSBC 2026 for Track 1 (Foundation Model Adapatation) and Track 2 (Label-Efficient Learning). The submissions are ranked by the harmonic-mean F_1 score computed across the three evaluation datasets (MOBIUS, SMD+SLD, and SynMOBIUS), consistent with SSBC 2025 [70], to maintain year-over-year comparability.

6.1. Overall Results and Performance Ranking

Submissions to SSBC 2026 were evaluated and ranked based on the **average performance scores** derived from the provided segmentation masks. Standard errors are reported as the standard deviation across 5 subject-disjoint folds created from the test data. For the primary ranking criterion, the harmonic mean of the F_1 score over the three evaluation datasets was utilised. The rationale for applying the harmonic mean is to prioritize solutions that maintain stable, elevated performance across the entire evaluation spectrum, thereby discouraging models that overfit to specific datasets at the expense of generalisability.

Table 4. **Comparative assessment of Track 2 (Label-Efficient Learning) submissions, ranked by harmonic-mean F_1** across the three evaluation datasets (MOBIUS, SMD+SLD, SynMOBIUS). F_1, Precision, Recall, and IoU were computed from the submitted binary masks; F_1^{opt} and AUC from the probabilistic predictions. Per-dataset scores are shown in Fig. 3, and the full per-dataset numerical breakdown is provided in Table S3 of the supplementary material.

Rank	Model	F_1	Precision	Recall	IoU	F_1^{opt}	AUC
1	R3-Ensemble	0.876	0.873	0.889	0.785	0.894	0.928
2	FreeSAU-DINOv3	0.861	0.861	0.872	0.762	0.881	0.921
3	FNOReg-ULVM-UNet	0.805	0.857	0.782	0.690	0.830	0.871
4	UniMedCLIPSeg	0.701	0.774	0.683	0.595	0.750	0.811
5	U2Net-SSL	0.668	0.701	0.722	0.546	0.729	0.757
6	DINOv2 Ensemble	0.465	0.346	0.854	0.314	0.604	0.646
7	UCalShapeEns	0.398	0.632	0.324	0.294	0.617	0.599

Track 1: Foundation Model Adaptation. The first track benchmarked the capacity of massive pre-trained architectures to act as backbone models for sclera segmentation under a fully-supervised regime. Results are presented in Fig. 3a and Table 3. The winner of Track 1 is **SAM2-UNet+FDA** (AU-1), with a harmonic mean F_1 of 0.870. The runner-up, **ScleraCLIP** (AU-2), achieved $F_1 = 0.860$, and **SAU-DINOv3** (PM) placed third at 0.851. The top six submissions cluster tightly in the 0.78–0.87 range, demonstrating that well-adapted foundation backbones achieve strong and consistent sclera-segmentation results.

However, the bottom four submissions fell sharply below this cluster: **CGDSeg_sclera** (0.678), **ScleraSAM3-CLIP** (0.665), **SAMSeg_sclera** (0.473), and **SAM-LoRA** (0.098). The near-zero F_1 of SAM-LoRA represents an obvious adaptation failure: its near-zero precision reveals that the model predicted sclera pixels across virtually the entire image. This points to an important insight, in that foundation models only succeed with principled adaptation, while vanilla LoRA fine-tuning alone is often insufficient.

The precision–recall curves (first row of Fig. 4) reveal a binary-vs.-optimal F_1 gap in several models, indicating that even Foundation Models could benefit substantially from a dedicated binarisation threshold selection strategy.

Track 2: Label-Efficient Learning. The second track directly addressed data scarcity. Participants were restricted to a severely limited annotation budget, but given access to a large pool of unlabelled MOBIUS images. Results are shown in Fig. 3b and Table 4. The winner, **R3-Ensemble** (AU-1), achieved a harmonic mean F_1 of 0.876, marginally above the Track 1 winner, demonstrating that label-efficient strategies can match or even surpass fully supervised foundation model adaptation when given sufficient additional unlabelled data. **FreeSAU-DINOv3** (PM) placed second at 0.861, and **FNOReg-ULVM-UNet** (SRI) third at 0.805.

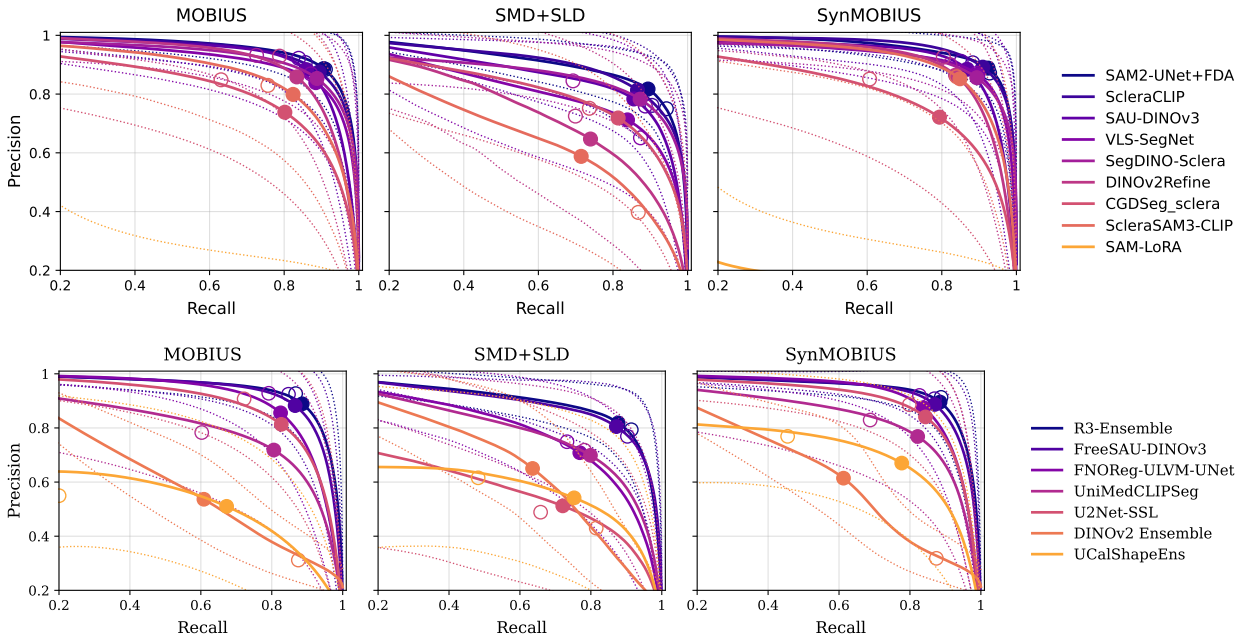
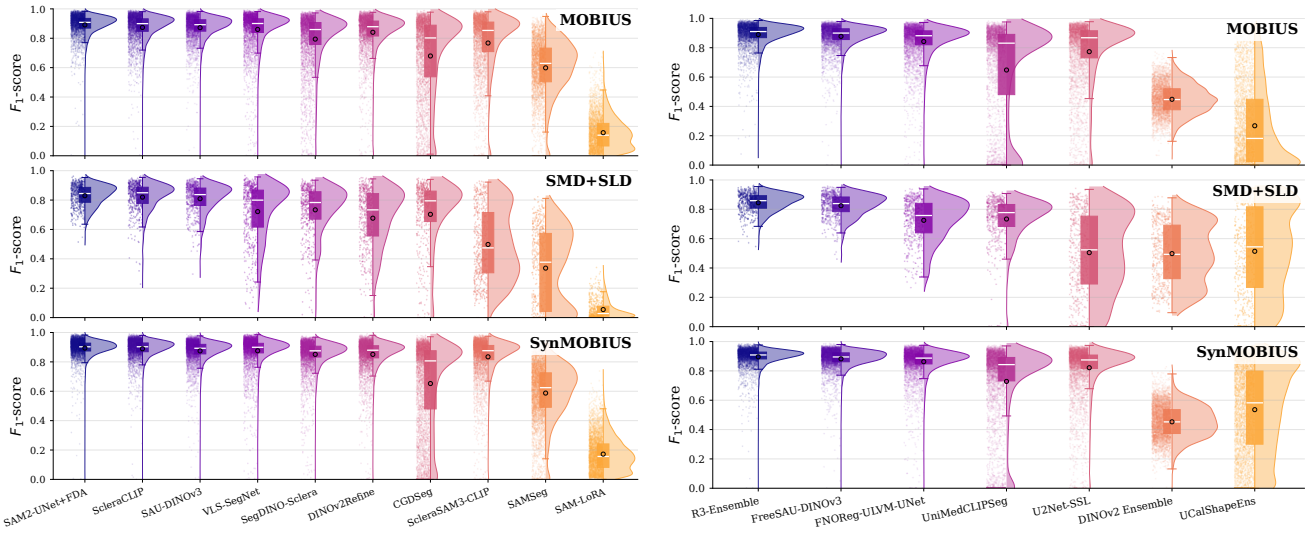


Figure 4. **Precision–recall curves for the submitted SSBC 2026 models**. The operating points denoted with a full circle represent the best possible F_1 score (F_1^{opt}), whereas the empty circle denotes the precision–recall point produced by the binary masks. The top row displays Track 1 (Foundation Models) submissions, and the bottom row shows Track 2 (Label-Efficient Learning) submissions. The columns correspond to MOBIUS, SMD+SLD, and SynMOBIUS. Legends are sorted by track ranking. Best viewed in colour and zoomed in.



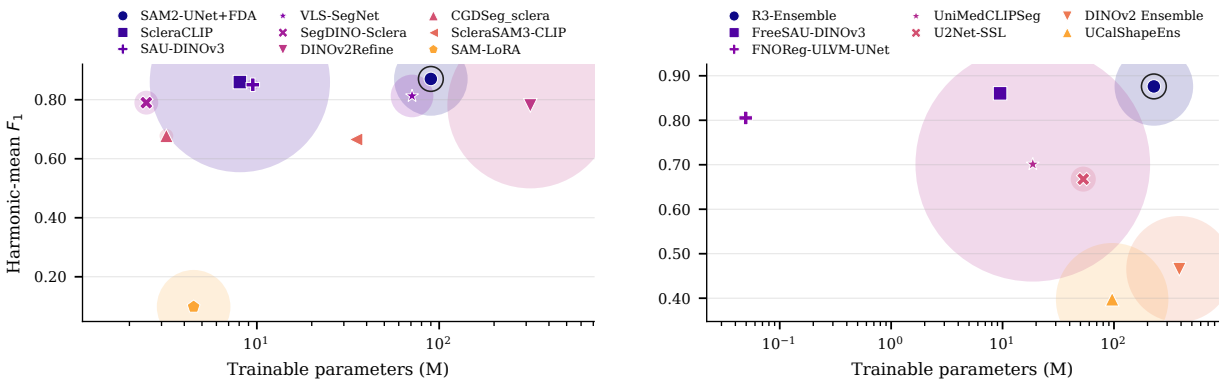
(a) Track 1 (Foundation Models)

(b) Track 2 (Label-Efficient Learning)

Figure 5. **Per-image F_1 distribution (raincloud plots).** Each column corresponds to one submission in ranking order. The cloud shows the kernel density estimate of F_1 scores across all test images; the box shows quartiles; the strip plots individual image scores. Best viewed in colour and zoomed in.

Performance dropped sharply after rank 3: **UniMedCLIPSeg** (0.701), **U2Net-SSL** (0.668), **DINOv2 Ensemble** (0.465), and **UCalShapeEns** (0.398). More than half the Track 2 submissions failed to achieve competitive performance under the label-efficient regime, emphasising that effective semi-supervised and self-supervised approaches require careful design beyond naïve adaptation. The precision–recall curves for Track 2 (second row of Fig. 4) reflect a wider spread in model behaviour.

Notably, the performance delta between the top track entries (SAM2-UNet+FDA in Table 3; R3-Ensemble in Table 4) peaked on the SMD+SLD dataset, i.e., the evaluation set most distinct from the MOBIUS training data. This divergence, absent in other cross-track pairs with similar F_1



(a) Track 1

(b) Track 2

Figure 6. **Performance vs. trainable parameters** on a log-scale x-axis. The translucent halo area is proportional to model FLOPs; the bold ring marks the top-ranked submission in each track. ScleraSAM3-CLIP is plotted without a FLOPs halo because the submission report did not provide a FLOPs estimate. Best viewed in colour and zoomed in.

scores, suggests that robust label-efficient strategies leverage unlabelled data for true generalisation, whereas weaker methods primarily overfit to the supplementary distribution.

6.2. Robustness Analysis

Harmonic-mean F_1 scores determine the competition ranking, but do not reveal model behaviour on individual images. This is critical, as in certain applications, a model that fails silently on difficult images may be unsuitable for deployment even if it achieves a high mean score. Here, we analyse the *distribution* of per-image F_1 scores using raincloud plots (Fig. 5), characterising each model by its median, interquartile range (IQR), and 5th-percentile F_1.

While most models excel on straightforward images, top-performing methods consistently exhibit significantly tighter distributions, indicating that performance gaps primarily emerge on challenging samples. Nonetheless, even leading models ($F_1 \approx 0.87$) produce a notable long tail of low-F_1 outliers. Although this leaves room for further refinement on these difficult images, such edge cases may also partially stem from inherent label noise or inter-annotator inconsistencies rather than purely algorithmic deficiencies.

6.3. Performance vs. Trainable Parameters

The trade-off between segmentation accuracy and model complexity is a key factor for real-world deployment, particularly on limited-hardware systems. The model complexity can be measured both in terms of the number of trainable parameters the model is required to learn, as well as the number of floating point operations (FLOPs) required to process a single image. Fig. 6 shows a comparison of the trade-off between segmentation performance (y-axis), model size (x-axis) and FLOPs (circle area).

In Track 1 (Fig. 6a), the winning **SAM2-UNet+FDA** occupies the higher end of the parameter range ($\sim$90M trainable parameters). However, **ScleraCLIP** and **SAU-DINOv3** achieve nearly identical accuracy at $\sim$10M parameters, showing that parameter-efficient adaptation (LoRA, partial fine-tuning) is competitive when correctly implemented. The cautionary example is **SAM-LoRA** ($\sim$4M parameters, $F_1 \approx 0.10$), confirming that parameter efficiency alone does not guarantee functional adaptation.

In Track 2 (Fig. 6b), the Pareto frontier is dominated by three submissions: **R3-Ensemble** ($\sim$200M parameters, top accuracy), **FreeSAU-DINOv3** ($\sim$10M), and the efficiency leader **FNOReg-ULVM-UNet** ($<$0.1M parameters, $F_1 \approx 0.805$). All other Track 2 submissions lie strictly inside the frontier. Under label-efficient supervision, small lightweight models sacrifice surprisingly little accuracy, an important result for resource-constrained deployment.

6.4. Qualitative Evaluation

In Fig. 7, we show best and worst cases of the generated segmentation masks for both tracks. The worst-case performance is mainly caused by: eyelid occlusion (partially closed eye), external occluders (e.g. a finger across the eye), lighting-induced skin discolouration (skin appearing brighter than the surrounding sclera), poorly-lit sclera, and specular reflection in the iris bleeding into the mask.

Track 1 foundation-backbone models did not eliminate these failure modes; they raised the performance floor without resolving the underlying difficulty. The failure taxonomy has not changed, only the baseline has improved. Track 2 reveals a complementary failure pattern: semi-supervised models trained with limited labels tend to exhibit over-confident under-segmentation on unfamiliar acquisition conditions, producing tight but spatially incorrect masks. This contrasts with the over-segmentation failures seen in some Track 1 models, pointing to the distinct challenges of each track's training paradigm.

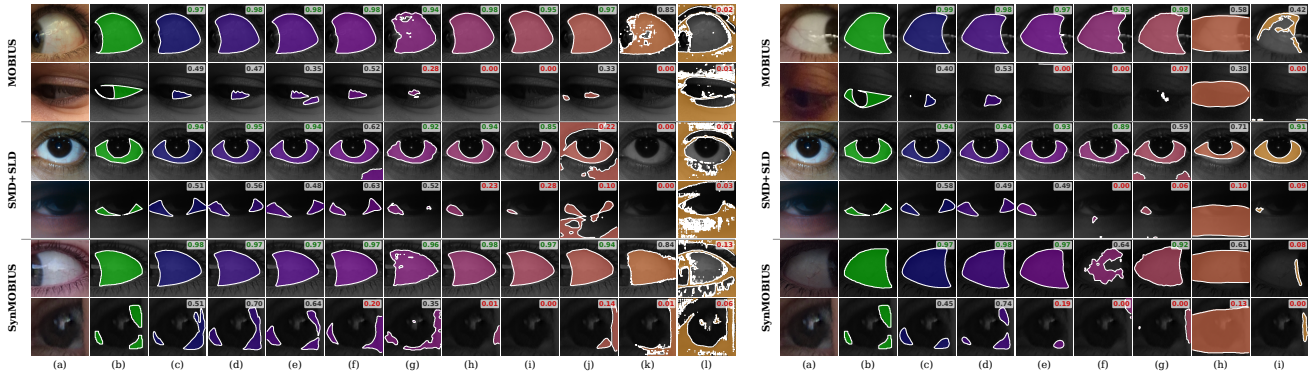


Figure 7. **Qualitative comparison of submitted models** for Track 1 (left) and Track 2 (right). In each panel, every dataset (MOBIUS, SMD+SLD, SynMOBIUS) is shown as two rows: the upper row is a well-segmented sample and the lower row is a failure case. Columns are labelled (a)–(l) for Track 1 and (a)–(i) for Track 2, corresponding to (a) the original image, (b) the ground truth mask, and the binary masks of each submission in ranking order. **Track 1 (left):** (c) SAM2-UNet+FDA, (d) ScleraCLIP, (e) SAU-DINOv3, (f) VLS-SegNet, (g) SegDINO-Sclera, (h) DINOv2Refine, (i) CGDSeg_sclera, (j) ScleraSAM3-CLIP, (k) SAMSeg_sclera, and (l) SAM-LoRA. **Track 2 (right):** (c) R3-Ensemble, (d) FreeSAU-DINOv3, (e) FNOReg-ULVM-UNet, (f) UniMedCLIPSeg, (g) U2Net-SSL, (h) DINOv2 Ensemble, and (i) UCalShapeEns. Best viewed in colour and zoomed in.

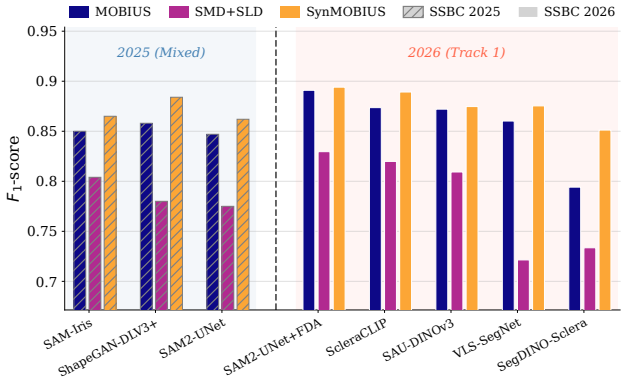


Figure 8. **Year-over-year progress: SSBC 2025 vs. SSBC 2026.** Per-dataset F_1 of the 2025 and 2026 track winners. Best viewed in colour.

6.5. Progress Over SSBC 2025

The aligned protocols of Track 1 and the SSBC 2025 Mixed Track [70] enable a direct longitudinal comparison, revealing substantial year-over-year progress, as seen in Fig. 8. This year's leading model, **SAM2-UNet+FDA** ($F_1 = 0.87$), outperforms the 2025 winner, **SAM-Iris** ($F_1 = 0.84$), by $+0.03$, a significant margin that would place the previous winner only in 4$^{\text{th}}$ position under the current benchmark. Furthermore, the top two Track 2 methods also surpass all 2025 entries, despite operating under substantially more restrictive annotation settings.

Nevertheless, while performance gains on MOBIUS and SynMOBIUS are pronounced, SMD+SLD continues to present a significant challenge. This suggests that current methods still struggle to generalise robustly under severe image degradations, motion blur, occlusions, and unconstrained acquisition conditions, despite the substantial representational capacity of modern foundation architectures.

Overall, the results demonstrate the strong potential of both foundation model adaptation and label-efficient learning for sclera segmentation. At the same time, the persistent failure cases observed across the benchmark indicate that improving robustness under challenging real-world conditions remains an important open research problem.

7. Conclusion

The 10$^{\text{th}}$ edition of the Sclera Segmentation Benchmarking Competition (SSBC 2026) investigated two emerging research directions in modern ocular biometrics: the adaptation of large-scale foundation models for fine-grained sclera segmentation and the use of label-efficient learning strategies under severe annotation constraints. To this end, the competition was organised into two complementary tracks differing in annotation availability and access to unlabelled training imagery. In total, seventeen submissions from ten research groups were evaluated under a unified benchmarking protocol using identical sequestered test datasets.

The winning submission of Track 1 (Foundation Model Adaptation) was **SAM2-UNet+FDA**, a DoRA-adapted SAM2 framework with Fourier-domain adaptation and pseudo-labelling components developed by the AU-1 team. Track 2 (Label-Efficient Learning) was won by **R3-Ensemble**, a multi-architecture semi-supervised ensemble model also developed by the AU-1 team. Notably, the Track 2 winner marginally outperformed the fully supervised Track 1 winner ($F_1 = 0.876$ vs. $F_1 = 0.870$), demonstrating that carefully designed label-efficient methodologies can achieve performance comparable to, or even exceeding, fully supervised foundation-model adaptation.

More broadly, the results show that modern foundation architectures, when combined with principled adaptation and semi-supervised training strategies, establish a

new state of the art in sclera segmentation and substantially outperform approaches evaluated in previous SSBC editions. At the same time, the persistent performance degradation observed on the SMD+SLD dataset indicates that robust generalisation under unconstrained acquisition conditions, image degradations, and severe domain shifts remains unresolved. Addressing these challenges will likely require future advances in domain generalisation, uncertainty-aware learning, and robustness-oriented adaptation strategies specifically tailored to ocular biometrics.

Acknowledgments

Supported in parts by the ARIS Research Programmes P2-0250 Metrology and Biometric Systems and P2-0214 Computer Vision.

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] 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. 4

[4] N. Carion, L. Gustafson, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 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. 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. 1

[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. 1

[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. 1

[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, 3

[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

[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. 1

[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] N. A. Drozdov and D. V. Sorokin. Fnoreg: resolution-robust medical image registration method based on fourier neural operator. In *International Conference on Pattern Recognition*, pages 163–177. Springer, 2024. 6

[23] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer. Pixel-wise ear detection with convolutional encoder-decoder networks. *IET Biometrics*, 2017. 4

[24] P. Farmanifard and A. Ross. Chatgpt meets iris biometrics. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024. 1

[25] P. Farmanifard and A. Ross. Iris-sam: Iris segmentation using a foundation model. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 394–409. Springer, 2024. 1

[26] 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. 1

[27] 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. 1

[28] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. 6

[29] E. J. Hu, Y. Shen, P. Wallis, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 4, 5

[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. 5

[31] M. U. Khattak, S. Kunhimon, M. Naseer, S. Khan, and F. S. Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities, 2024. 6

[32] 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. 5

[33] T. Koleilat, H. Asgariandehkordi, O. N. Manzari, B. Barile, Y. Xiao, and H. Rivaz. Medclipseg: Probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation, 2026. 6

[34] D.-H. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013. 5

[35] T.-Y. Lin, P. Dollár, R. Girshick, et al. Feature pyramid networks for object detection. In *CVPR*, 2017. 5

[36] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen. Dora: Weight-decomposed low-rank adaptation. In *ICML*, 2024. 5

[37] Z. Liu et al. A convnet for the 2020s. In *CVPR*, 2022. 6

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[39] 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. 4

[40] 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. 4

[41] Z. Mahmud, P. Hungler, and A. Etemad. Gaze estimation with eye region segmentation and self-supervised multistream learning. *arXiv preprint arXiv:2112.07878*, 2021. 1

[42] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 5

[43] H. Peng, S. Chen, N. Niu, J. Wang, Q. Yao, L. Wang, Y. Wang, J. Tang, G. Wang, M. Huang, et al. Diffusion model in semi-supervised scleral segmentation. In *2023 IEEE 6th International conference on pattern recognition and artificial intelligence (PRAI)*, pages 376–382. IEEE, 2023. 1

[44] 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. 4

[45] X. Qin et al. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 6

[46] A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 5

[47] 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

[48] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 5

[49] 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. 5

[50] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 5

[51] 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

[52] 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. 5

[53] 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. 4

[54] 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

[55] A. G. Roy, N. Navab, and C. Wachinger. Recalibrating fully convolutional networks with spatial and channel squeeze and excitation blocks. *IEEE Transactions on Medical Imaging*, 38(2):540–549, 2019. 5

[56] 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. 5

[57] 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. 4

[58] H. O. Shahreza and S. Marcel. Foundation models and biometrics: A survey and outlook. *IEEE Transactions on Information Forensics and Security*, 2025. 1

[59] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski. DINOv3, 2025. 5

[60] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, volume 30, 2017. 5, 6

[61] 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. 1

[62] 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. 1

[63] 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. 1, 3

[64] 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. 1

[65] 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

[66] 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

[67] 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, 3, 4

[68] 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, 3

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

[70] M. Vitek, D. Tomašević, A. Das, S. Nathan, G. Özbulak, G. A. T. Özbulak, J.-P. Calbimonte, A. Anjos, H. H. Bhatt, D. D. Premani, J. Chaudhari, C. Wang, J. Jiang, C. Zhang, Q. Zhang, I. I. Ganapathi, S. S. Ali, D. Velayudan, M. Assefa, N. Werghi, Z. A. Daniels, L. John, R. Vyas, J. N. Khiarak, T. A. Saeed, M. Nasehi, A. Kianfar, M. P. Panahi, G. Sharma, P. R. Panth, R. Ramachandra, A. Nigam, U. Pal, P. Peer, and V. Štruc. Privacy-enhancing sclera segmentation benchmarking competition: SSBC 2025. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–13, 2025. 2, 3, 4, 6, 9

[71] G. Wang et al. Boosting sclera segmentation through semi-supervised learning with fewer labels, 2025. 6

[72] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Pan, L. Wang, Y. Wu, H. Zhou, and M.-L. Lukniw. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2023. 6

[73] J. Wei, R. He, and Z. Sun. Contrastive uncertainty learning for iris recognition with insufficient labeled samples. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 1

[74] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, and Y. Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025. 1

[75] R. Wu, Y. Liu, G. Ning, P. Liang, and Q. Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *Patterns*, 6(11), 2025. 6

[76] 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

[77] Z. Xing, T. Ye, Y. Yang, D. Cai, B. Gai, X.-J. Wu, F. Gao, and L. Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025. 5

[78] 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

[79] S. Yang, H. Wang, Z. Xing, S. Chen, and L. Zhu. Segdino: An efficient design for medical and natural image segmentation with dino-v3, 2025. 5

[80] Y. Yang and S. Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4085–4095, 2020. 5, 6

[81] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. 5

[82] 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

[83] 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. 5