An Unbiased Benchmark for Domain Generalization Face Anti-Spoofing

Raul Almeida*, Bruno Kamarowski*, Bernardo Biesseck*, Luiz Coelho†, Roger Granada†, David Menotti*
* Federal University of Paraná, Curitiba, PR, Brazil {rgpalmeida, bhkcarvalho, bernardo, menotti}@inf.ufpr.br
‡unico - idTech, Brazil {roger.granada, luiz.coelho}@unico.io

***Abstract*—Domain Generalization for Face Anti-Spoofing (DG-FAS) is an area of growing interest due to its importance in fraud prevention in Face Recognition Systems. Current benchmarks used for DG-FAS in evaluating state-of-the-art methods allow verification of test set performance during training, which causes bias towards test data. Consequently, practitioners cannot properly translate research conclusions to real-world applications, since there is no access to labels for production data in practice. We propose as an alternative an unbiased benchmark where a validation dataset is used so that the model's generalization capability is evaluated without compromising restricted data and the scientific rigor of research. Our experiments show that model performance benefits from current biased benchmarks and that introducing a new validation dataset makes for more challenging and scientifically rigorous benchmarks that also better represent real-world performance. We additionally experiment with training on current standard benchmarks and testing on WFAS, a recent in-the-wild large FAS dataset with more attack types than the standard datasets for DG-FAS, and similarly observe poor generalization capabilities for state-of-the-art models.**

I. Introduction

With the ubiquitous usage of face recognition systems in diverse applications, it has become an important matter to ensure these systems are reliant and not easily fooled by malicious users, who can produce spoofs to impersonate another person when authenticating themselves or to simply confuse the system and not have their identity associated with the session. An example of spoofing is placing someone else's photo in front of a phone camera to log in via facial recognition. See Figure 1 for a visual example. The recognition system might identify the person in the picture, but it is not a real person (live access). In Computer Vision, the task of identifying spoofs is called Face Anti-Spoofing (FAS), Presentation Attack Detection (PAD), or Face Liveness Detection [1].

Of particular importance to products that rely on FAS is that the used methods generalize well. It is not enough for a model to effectively detect spoofs in the training set, as in production environments it will be faced with samples very different from its training ones. The task of generalization to unseen domains is called Domain Generalization FAS (DG-FAS). Even though for most research datasets there are already models that perform very well in intra-dataset benchmarks, the task of DG-FAS is still a difficult one [3] and the state of the art is far from sufficient for real-world applications.

Currently commonly used evaluation benchmarks for DG-FAS models [1], [4], [5], [6], [3] are biased, as they rely on verifying test performance at each epoch of training (see Figure 2), not employing a early stopping strategy. In particular, the model performance in the test set is obtained after each epoch of training, and afterward, the best performance is reported as the final one. This is biased because in real-world applications there is no access to the "testing set": here, the final model performance reached on the test set is a direct product of verifying the very same performance during training, i.e., if there was no access to the test set during training, model performance would be worse.

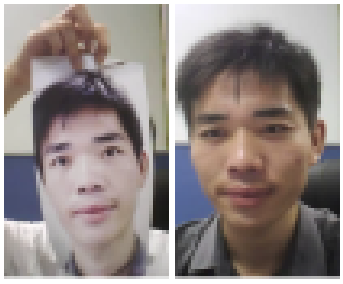
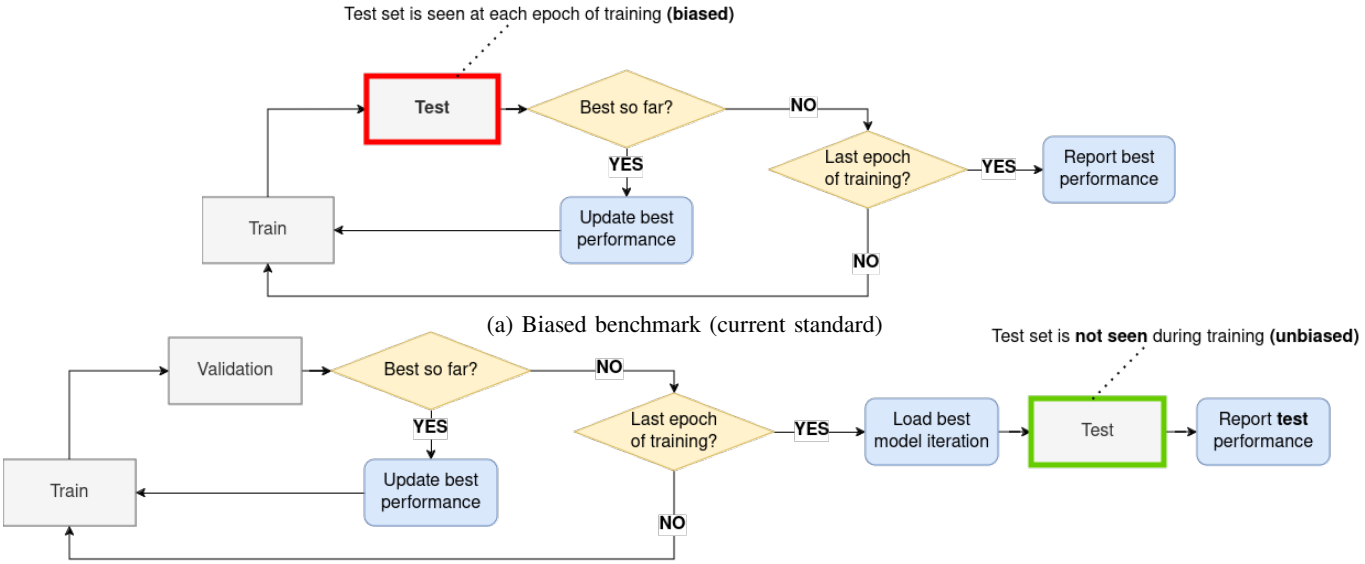


Fig. 1. Examples of attack (left-a print attack) and bona fide (right) images from the CASIA-FASD dataset [2]. Besides more direct traces of spoofs (hands holding a picture and paper distortion), the spoof-presented face has very contrasting texture and color to those found in a bona fide access.

In this paper, we propose a new benchmark to solve the aforementioned bias issue, with a single difference from its biased counterpart: a validation dataset. Instead of evaluating the model in the test set after each epoch, we do it in a validation dataset, and then evaluate the best model iteration on the test set after training is complete, as a machine learning approach aiming for generalization should be built. This way the best model is selected without requiring access to the test data.

This work contributes to the DG-FAS research field in four key aspects:

- Bringing attention to the **bias issue** in DG-FAS benchmarks;
- Providing an **unbiased alternative** that is coherent with the community's directions in terms of dataset preference and evaluation format;
- Benchmarking DG-FAS state-of-the-art models against WFAS [7], a huge in-the-wild FAS dataset.
- Providing public code access after publication, with the aim of allow complete reproducibility of reported results



(a) Biased benchmark (current standard)

(b) Proposed unbiased benchmark

Fig. 2. Biased benchmark (top) vs unbiased (bottom). In the biased benchmark, there is access to the test set in every epoch of training, which leads to biased decision-making (both in the research process and model choice for final evaluation). In our proposed unbiased benchmark, the test set is only used after training is finished, so there is no bias in decision-making.

[1];

- **Improving the performance of the state-of-the-art method GAC-FAS** in DG-FAS by removing (and not adding) unfair advantages during training.

Previous works have touched on the first point but never directly tackled it [4]. To the best of our knowledge, this is the first work to focus on the bias issue in DG-FAS benchmarks and propose an unbiased solution to it. Also, this is the first work to use the WFAS dataset [7] in the context of DG-FAS.

Many FAS works fail to release source code, hindering reproducibility and progress [8]–[10]. We contribute by publicly releasing our complete codebase. Surprisingly, while more challenging, our unbiased benchmark sometimes improves test set performance compared to biased counterparts.

The paper is organized as follows: Section II reviews FAS methods and benchmarks; Section III presents our benchmark; Section IV details validation experiments; Section V analyzes results; and Section VI concludes.

II. Related Works

Figure 3 illustrates how FAS methods can be divided into three main categories. First are those that make use of both deep learning (DL) models and specialist-handcrafted features for classification. In the second category are the more direct works that rely on DL methods only; these models are typically trained on binary supervision but some works use auxiliary tasks such as face depth estimation with pixel-wise supervision to reinforce characteristics such as locality. The final category is domain generalization (DG), where models can be trained

[1]The source code of this work is available at https://github.com/BOVIFOCR/unbiased-benchmarks-aggregator

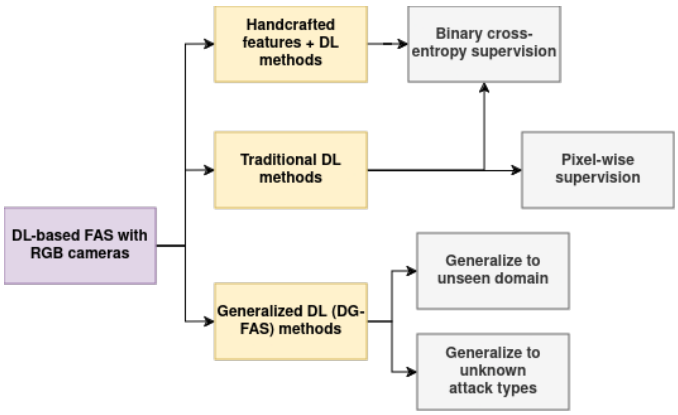


Fig. 3. Topology of FAS methods. This work focuses on the *Generalize to unseen domain* leaf node. Figure inspired by [1] (Figure 2)

specifically to generalize well to unseen domains, attack types, or both. This work focuses on unseen domain generalization. Due to the relevance of this field, there are many works in FAS and specifically in DG-FAS [1]. In this work, we choose to focus on recent state-of-the-art methods with code made available by the authors to ensure reproducibility of our experiments. We list these methods in Section IV.

DG-FAS research tries to bridge the gap between the training domain and production domains, usually through auxiliary modules and objective engineering. IADG [5] whitens domain characteristics from samples to reach a generalized embedding capability, while SA-FAS [4] embraces domain-variant information and focuses instead on learning a generalized transition from the live to the spoof class. GAC-FAS [3] is a more recent work somewhat similar to SA-FAS that focuses on reaching

a flat minimum during optimization, which leads to better generalization.

DG-FAS is usually evaluated with benchmarks that consist of sums of datasets. In particular, the most common benchmarks involve training on three datasets and testing on a fourth, where datasets are picked from CASIA-FASD [2], Replay-Attack [11], MSU-MFSD [12] and Oulu-NPU [13]. These datasets might not be particularly challenging for an intra-dataset scenario, but together they present variations in illumination, subject expression, background, camera, and attack presentation, besides other nuances, that make for challenging cross-domain benchmarks. Model performance is reported with the Equal Error Rate (EER), Half-Total Error Rate (HTER) and Area Under the Curve (AUC) metrics [1], but recent intra-dataset benchmarks such as WFAS [7] also use Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER) and Average Classification Error Rate (ACER) [1].

However, current model evaluation for cross-domain benchmarks is biased towards the test set, which makes for inaccurate representations of how well a model performs in real-world usage. Previous works have tried to come up with alternative evaluations by using the TPR@FPR metric [6] or by reporting average error rates for the last 10 epochs of training instead of choosing the best model iteration [4]. While both of these solutions are better representations of real-world performance, they do not tackle the main issue of bias, which allows the representation problem to persist. To the best of our knowledge, this is the first work to propose an unbiased benchmark for DG-FAS.

III. Proposed Benchmark

Current DG-FAS evaluation benchmarks suffer from a critical flaw: models are assessed on the test set after every training epoch, with only the best performance being reported. This practice introduces significant bias, as real-world applications cannot leverage test data during training. Consequently, existing benchmarks fail to properly guide the development of practical FAS systems.

To address this limitation, we propose an unbiased benchmark design with two fundamental requirements: (i) Complete isolation of the test set during model training; (ii) Rigorous evaluation of both classification accuracy and generalization capability

Since the biased benchmark we want to improve consists of four datasets (three for training and one for testing), we propose introducing a fifth dataset as either validation or test. This way, during training the model's generalization and classification capabilities are still evaluated, but the test set does not influence the choice of best-performing model iteration from the training process. When the final model is evaluated on the test set, after training, we have a better representation of its real-world performance. By making the benchmark unbiased, we introduce significant changes to the effects training and testing have on current DG-FAS model research.

TABLE I

Comparison of FAS datasets CASIA-FASD, Replay-Attack, MSU-MFSD, Oulu-NPU, SiW, and Rose-Youtu. Columns correspond, in order, to dataset name, year of publication, number of subjects, number of live samples, number of spoof samples, and attack types (the last being a subset of (P)rint, (R)eplay, and (M)ask attacks). All datasets are video datasets.

Dataset	Year	Sub	Live	Spoof	Types
CASIA	2012	50	150	450	P,R
Replay	2012	50	200	1000	P,R
MSU	2014	35	70	210	P,R
Oulu	2017	55	720	2880	P,R
SiW	2018	165	1320	3300	P,R
Rose	2018	20	897	1601	P,R,M

This approach maintains continuous performance assessment during training while eliminating test set contamination. The resulting evaluation provides a more accurate measure of real-world performance with minimal implementation overhead. Notably, our method requires no modifications to existing training procedures, as demonstrated in our experiments with state-of-the-art models. Figure 2 visually contrasts the conventional and proposed benchmarking approaches.

IV. Methodology

We evaluate the proposed benchmark in comparison with four already well-established DG-FAS benchmarks that consist of training on three datasets and testing on a fourth, where the datasets are CASIA-FASD [2], Replay-Attack [11], MSU-MFSD [12] and Oulu-NPU [13]. Aditionally we use the WFAS [7] for auxiliary experiments on another unbiased benchmark. WFAS differs from other FAS datasets in both number of attacks and volume of data.

In considering options for the fifth dataset, we must consider in which aspects they are similar or different from the four currently used datasets. Particular aspects of interest are the data volume, types of attacks, number of subjects, and variability of samples.

We consider two datasets as options for the fifth dataset due to their similarities and differences to the four currently used datasets: SiW [14] and Rose-Youtu [15]. Both involve print and replay attacks, which is the case in the other four datasets used. Rose-Youtu also includes paper mask attack samples, which we discard to keep the task intra-type (no new attacks on the test set). Besides the attack type consideration, we choose these two datasets specifically because they are both reasonably sized (number of samples comparable to that of Oulu-NPU), recent (both more recent than the other four), and varied (both datasets feature variations in illumination, environments, attacks, and subjects; SiW features variations in pose and expression as well). From these characteristics, we hypothesize that they can be at least as useful for DG evaluation as currently used datasets are. Table I compares the characteristics of these datasets.

For evaluation, we use four models. First are ResNet18 and ResNet50 [16] to serve as baselines. The other two are SA-FAS [4] and GAC-FAS [3], both recent methods with public code made available by the authors and with which we were able to

obtain results close to those reported in their respective papers. For ResNet models we use the Torch Image Models library which contains model implementations as well as ImageNet pre-trained weights [17].

We start with baseline experiments on the biased benchmarks that have already been used in previous works. From those experiments, we can establish a reference point for model performance and also compare our own local results with those reported by the authors (in the case of SA-FAS and GAC-FAS).

Afterward, we execute three categories of experiments, which we call *biased*, *unbiased with added validation*, *and unbiased with added test*. They correspond, respectively, to: (1) biased experiments on the additional datasets (Figure I), (2) unbiased experiments where we add a new dataset as validation in a standard benchmark (Figure I), and (3) unbiased experiments where we replace the test set with a fifth dataset in a standard benchmark (keeping its previous test set as validation - Figure I). From these results, we expect to validate our hypothesis that model performance on biased benchmarks (i.e., performance on the test set when there is access to the same test set during training) is not representative of the model's generalization capability, and from the different options of designed benchmarks we draw suggestions for how evaluation should instead be carried out in future work.

We report for each experiment both the Half-Total Error Rate and the Area Under the Curve percentages.

Additionally, we execute similar experiments when considering WFAS [7] as the test set. WFAS is a more recent dataset created from web scraping techniques with the proposal of being larger and more varied than previous FAS datasets and focuses on *in-the-wild* classification. Unbiased experiments with WFAS differ from those previously described because they consist of a *cross-type cross-domain* task (a union of the Generalized DL leaf nodes in Figure 3), i.e., the WFAS dataset has attack types other than those present in the training datasets. WFAS also has a fundamental difference in size, being huge in comparison to all other datasets (see Table I). The rationale for considering this scenario is to evaluate how well-adapted these models are to a slightly different task. In particular, WFAS consists of $1.383.246$ image samples of 469.920 subjects in unconstrained settings, with 17 different types of presentation attacks. Since WFAS includes a range of other attack types and consists of an unconstrained scenario, we consider experiments involving it to be fundamentally different from other experiments in this work. With our unbiased benchmark formulation (refer to Figure I), we evaluate SA-FAS and GAC-FAS on the WFAS test set after training on standard benchmarks (using the test set as validation) and, in the case of GAC-FAS, after training on the WFAS training set. The latter is reserved to GAC-FAS for performance reasons - for comparison, WFAS is almost 300 times larger than SiW, the largest dataset referenced in Table I.

With the WFAS experiments, we aim to verify how applicable these DG-FAS models are to an in-the-wild scenario that compares in many aspects to real-world applications of (even constrained) face recognition. If these models do not indeed perform well in this case, it would reinforce our general hypothesis that current evaluation benchmarks are limited in representing real-world model performance.

For experiments with standard DG-FAS benchmarks and our proposed unbiased variations involving SiW and Rose-Youtu, we report the HTER% and AUC% values. In experiments involving the WFAS dataset as test set, since it consists of a closed-set evaluation, the APCER%, BPCER% and ACER% results provided by the WFAS ongoing contest [7] are shared.

V. Results

We now share and discuss the results of the experiments carried out. We start with biased benchmarks and then move to their unbiased counterparts (i.e. the proposed benchmark).

A. Biased benchmarks

Table II compares our SA-FAS and GAC-FAS results with prior work, including author-provided GAC-FAS weights. Despite using original code and datasets, we note some discrepancies in error rates.

Table III shows performance on biased benchmarks (where Val=Test). Testing on Rose-Youtu proves consistently more challenging than SiW across all models, including ResNet baselines (though ResNet shows less consistent trends). Both datasets offer reasonable difficulty - neither trivial nor impractical - compared to CASIA-FASD, Replay-Attack, MSU-MFSD, and Oulu-NPU.

B. Unbiased benchmarks

Table III shows results for unbiased benchmarks using SiW/Rose-Youtu as validation or test sets (rows where Val $\neq$ Test). Unbiased benchmarks generally yield higher HTER, confirming bias removal exposes defective learning. Rose-Youtu again shows steeper performance drops than SiW.

While adding a fifth dataset as test or validation produces similar effects, the validation approach may be preferred as it preserves reporting on standard test sets (Oulu-NPU, etc.) [1]. Notably, test-set additions effectively evaluate biased-trained models on new data.

Crucially, in 3/4 benchmarks, GAC-FAS (our top model) with SiW validation outperformed biased benchmarks—achieving 7.91% HTER (vs. 8.60% SOTA) in ICMO [3]. Similar gains occur for ResNet50 in ICMO/OCIM, though we emphasize GAC-FAS as the current leader.

1) Validation EER analysis: The reported EER matches the HTER in biased benchmarks, as state-of-the-art benchmarks derive thresholds from validation sets. While our focus remains on test performance, EER offers valuable training insights.

Typically, validation performance exceeds test performance [1], consistent with ML literature since validation sets resemble accessible test sets. However, experiments using Rose-Youtu show the opposite trend - likely due to domain gaps between validation and test data.

This supports our hypothesis that SiW's domain aligns closer to test datasets than Rose-Youtu's does.

TABLE II

REPRODUCTION RESULTS: TEST HTER% AND AUC% RESULTS ON STANDARD BIASED DG-FAS BENCHMARKS REPORTED IN RELATED WORKS (TOP, ROWS MARKED WITH †) AND OBTAINED IN OUR OWN EXPERIMENTS (BOTTOM). WE INCLUDE THE GAC-FAS EVALUATION WITH PROVIDED WEIGHTS (*GAC-FAS^w*). BENCHMARK NAMES CORRESPOND TO THE TRAINING DATASETS (FIRST THREE LETTERS) AND THE TEST DATASET (LAST LETTER), WHERE I = REPLAY-ATTACK, C = CASIA-FASD, M = MSU-MFSD, AND O = OULU-NPU. THE VALIDATION EER FOR OUR LOCAL EXPERIMENTS THAT INVOLVED TRAINING MODELS (TWO LAST ROWS) IS AVAILABLE IN TABLE III.

GAC-FAS	ICM→O		OCM→I		OCI→M		OMI→C	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
SSAN†	13.72	93.63	8.88	96.79	6.67	98.75	10.00	96.67
IADG†	8.86	97.14	10.62	94.50	5.41	98.19	8.70	96.40
SA-FAS†	10.00	96.23	6.58	97.54	5.95	96.55	8.78	95.37
GAC-FAS†	8.60	97.16	4.29	98.87	5.00	97.56	8.20	95.16
GAC-FAS^w	9.72	96.87	5.13	99.02	5.00	97.54	10.00	94.76
GAC-FAS	10.52	96.27	5.00	98.36	7.92	96.31	12.22	93.09
SA-FAS	12.03	94.78	6.55	97.66	8.57	96.96	13.33	92.29

TABLE III

TEST HTER AND VALIDATION EER (T-HTER AND V-EER, REPORTED IN PERCENTAGES) FOR RESNET18, RESNET50, SA-FAS AND GAC-FAS ON DIFFERENT STANDARD AND PROPOSED BENCHMARKS (DESCRIBED BY THE DATASETS USED FOR TRAINING, VALIDATING AND TESTING, WHERE I=REPLAY-ATTACK, C=CASIA-FASD, M=MSU-MFSD, O=OULU-NPU, S=SIW AND R=ROSE-YOUTU). ROWS WHERE THE VALIDATION AND TEST VALUES ARE THE SAME CORRESPOND TO BIASED BENCHMARKS. THE FIRST ROW OF EACH GROUP CORRESPONDS TO STANDARD BENCHMARKS ALREADY USED IN PREVIOUS WORKS, AND ALL OTHER ROWS CORRESPOND TO SOME BENCHMARK PROPOSED IN THIS WORK. ALSO, WE HIGHLIGHT UNBIASED BENCHMARKS WITH INTRODUCED TEST SETS SINCE THEY DIFFER IN THE FINAL EVALUATION DATASET FROM STANDARD BIASED BENCHMARKS.

Train	Val	Test	RN18		RN50		SA		GAC	
			V-EER	T-HTER	V-EER	T-HTER	V-EER	T-HTER	V-EER	T-HTER
ICM	**O**		30.56		33.33		12.13		10.52	
	S		21.55		20.57		7.17		7.62	
	R		22.50		23.09		20.97		19.29	
	S	**O**	21.55	32.50	20.57	32.78	7.17	13.81	7.62	7.91
	R	**O**	22.50	33.33	23.09	31.74	20.97	13.64	19.29	19.94
	O	**S**	30.56	25.06	33.33	22.93	12.13	8.76	10.52	12.04
	O	**R**	30.56	25.66	33.33	24.21	12.13	20.82	10.52	19.38
OCM	**I**		31.50		33.75		9.00		5.00	
	S		18.67		18.89		5.18		8.15	
	R		24.42		16.05		24.10		21.38	
	S	**I**	18.67	37.50	18.89	33.75	5.18	21.50	8.15	8.55
	R	**I**	24.42	41.25	16.05	45.00	24.10	20.10	21.38	21.41
	I	**S**	31.50	21.03	33.75	26.12	9.00	6.48	5.00	11.28
	I	**R**	31.50	27.10	33.75	22.06	9.00	26.86	5.00	29.86
OMI	**C**		17.96		22.22		14.00		12.22	
	S		19.36		19.20		7.93		9.52	
	R		19.18		20.29		21.50		19.97	
	S	**C**	19.36	18.89	19.20	25.56	7.93	26.67	9.52	9.14
	R	**C**	19.18	22.22	20.29	23.33	21.50	40.67	19.97	20.62
	C	**S**	17.96	24.07	22.22	17.78	14.00	14.09	12.22	12.95
	C	**R**	17.96	20.62	22.22	21.41	14.00	27.42	12.22	32.10
OCI	**M**		24.58		32.08		11.43		7.92	
	S		22.40		19.65		6.32		7.08	
	R		20.18		19.18		19.18		22.98	
	S	**M**	22.40	32.50	19.65	25.42	6.32	11.43	7.08	7.31
	R	**M**	20.18	32.50	19.18	30.00	19.18	15.48	22.98	22.42
	M	**S**	24.58	25.68	32.08	22.24	11.43	10.05	7.92	9.58
	M	**R**	24.58	22.74	32.08	21.82	11.43	25.18	7.92	32.74

C. WFAS-based experiments

Table IV presents the ACER% performance of SA-FAS and GAC-FAS on WFAS's [7][2] closed test set (only APCER, BPCER, and ACER metrics are available). Unlike other benchmarks using HTER%, WFAS reports ACER% for its worst-case representability, calculated per attack type subset [1].

Both models show comparable cross-dataset performance, with each outperforming in two benchmarks. However, their results fall far short of the WFAS challenge winner's 2.82% ACER [7], highlighting both WFAS's distinct task nature and our benchmark's greater difficulty. Table V shows validation EER% for cross-dataset benchmarks. For WFAS intra-dataset evaluation (Table IV), we use the final model iteration [4] to leverage the full training set. These results expose DG-FAS methods' weaknesses when faced with WFAS's diverse, less constrained data. While not designed for WFAS's specific attacks, these models should ultimately work in real-world

[2]Since we are dealing with a closed test set, we only have access to test set performance in three metrics: APCER, BPCER, and ACER.

TABLE IV

SA-FAS and GAC-FAS performance (ACER%) in the WFAS test set when trained (biased) on given train and validation datasets, where I = Replay-Attack, C = CASIA-FASD, M = MSU-MFSD and O = Oulu-NPU. For time constraint reasons, we only experiment with training on WFAS (last row) with GAC-FAS. The last row corresponds to an **intra-dataset** evaluation, while the others are **cross-dataset**.

Train	Val	SA	GAC
ICM	**O**	30.14	31.99
OCM	**I**	32.54	28.62
OMI	**C**	28.67	28.76
OCI	**M**	31.22	30.76
WFAS	**WFAS**	-	21.51

TABLE V

SA-FAS and GAC-FAS validation performance (EER%) for cross-dataset experiments in Table IV, where I = Replay-Attack, C = CASIA-FASD, M = MSU-MFSD and O = Oulu-NPU.

Train	Val	SA	GAC
ICM	**O**	12.13	10.52
OCM	**I**	9.00	5.00
OMI	**C**	14.00	12.22
OCI	**M**	11.43	7.92

systems—emphasizing the need for evaluation methods that better predict production performance.

D. Discussion

Since the community is likely to keep using CASIA-FASD, Replay-Attack, MSU-MFSD, and Oulu-NPU as test sets in DG-FAS benchmarks, we expect our proposed benchmarks to be better accepted with the added validation set variant. We tested two validation datasets, SiW and Rose-Youtu, observing similar results. Given SiW's domain similarity to the four test sets, it may be preferable for validation, though Rose-Youtu is also suitable. We encourage using both datasets in separate evaluations when possible. In summary, we recommend experimenting with all unbiased benchmarks proposed here, or for brevity, using the unbiased benchmarks with SiW validation as listed in Table VI.

VI. Conclusions

We identified bias in current DG-FAS evaluation benchmarks and its negative impact. As a solution, we proposed a new benchmark with validation datasets that proves more challenging than biased alternatives, sometimes even improving test performance. The benchmark is easy to integrate into existing codebases.

We also tested our approach on the WFAS dataset, a distinct task from DG-FAS (featuring new attack types), highlighting the need for greater real-world variability in DG-FAS methods.

We encourage exploring alternative validation datasets and evaluation formats to improve rigor and advance the field.

Future work will refine benchmarks, focusing on two key questions: (1) Can DG-FAS models perform better on WFAS? (2) Can top WFAS models generalize to DG-FAS? These questions are compelling due to WFAS's unique characteristics.

TABLE VI

Benchmark name: train, validation and test datasets of recommended concise (reduced) battery of experiments. SiW is used as the validation set to make them unbiased. Here, I=Replay-Attack, C=CASIA-FASD, M=MSU-MFSD, O=Oulu-NPU and S=SiW.

Abbreviation	Train	Val	Test
ICM→S→O	ICM	S	O
OCM→S→I	OCM	S	I
OCI→S→M	OCI	S	M
OMI→S→C	OMI	S	C

Acknowledgment

This work was supported by a tripartite-contract, i.e., unico - idTech, UFPR (Federal University of Paraná), and FUNPAR (Fundação da Universidade Federal do Paraná). We also thank the *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)* (# 315409/2023-1) and *Fundação Araucária (Paraná)* for supporting Prof. David Menotti.

References

[1] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[2] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *2012 5th IAPR International Conference on Biometrics (ICB)*, 2012, pp. 26–31.

[3] B. M. Le and S. S. Woo, "Gradient alignment for cross-domain face anti-spoofing," in *CVPR*, 2024, pp. 188–199.

[4] Y. Sun, Y. Liu, X. Liu, Y. Li, and W.-S. Chu, "Rethinking domain generalization for face anti-spoofing: Separability and alignment," in *CVPR*, 2023.

[5] Q. Zhou, K.-Y. Zhang, T. Yao, X. Lu, R. Yi, S. Ding, and L. Ma, "Instance-aware domain generalization for face anti-spoofing," in *CVPR*, 2023.

[6] P. Zhang, X. Guo, Y. Zhang, J. Yang, W. Zhang, and Q. Li, "Domain generalization via shuffled style assembly for face anti-spoofing," in *CVPR*, 2022, pp. 16 066–16 075.

[7] D. Wang, J. Guo, Q. Shao, H. He, Z. Chen, C. Xiao, A. Liu, S. Escalera, H. J. Escalante, Z. Lei, J. Wan, and J. Deng, "Wild face anti-spoofing challenge 2023: Benchmark and results," in *CVPRW*, 2023.

[8] Q. Zhou, K.-Y. Zhang, T. Yao, X. Lu, S. Ding, and L. Ma, "Test-time domain generalization for face anti-spoofing," in *CVPR*, 2024, pp. 175–187.

[9] R. Cai, Z. Yu, C. Kong, H. Li, C. Chen, Y. Hu, and A. C. Kot, "S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 8385–8397, 2024.

[10] Z. Li, H. Li, K.-Y. Lam, and A. C. Kot, "Unseen face presentation attack detection with hypersphere loss," in *ICASSP*, 2020, pp. 2852–2856.

[11] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, 2012, pp. 1–7.

[12] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[13] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017.

[14] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *CVPR*, June 2018.

[15] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[17] R. Wightman, "Github: Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019.