SDFR: Synthetic Data for Face Recognition Competition

Hatef Otroshi Shahreza[1,2,†], Christophe Ecabert[1,†,*], Anjith George[1,†,*],
Alexander Unnervik[1,2,†,*], Sébastien Marcel[1,3,†], Nicolò Di Domenico[4,‡],
Guido Borghi[4,‡], Davide Maltoni[4,‡], Fadi Boutros[5,‡], Julia Vogel[5,‡],
Naser Damer[5,6,‡], Ángela Sánchez-Pérez[7,8,‡], Enrique Mas-Candela[7,8,‡],
Jorge Calvo-Zaragoza[7,‡], Bernardo Biesseck[9,10,‡], Pedro Vidal[9,‡], Roger Granada[11,‡],
David Menotti[9,‡], Ivan DeAndres-Tame[12,‡], Simone Maurizio La Cava[13,‡], Sara Concas[13,‡],
Pietro Melzi[12,‡], Ruben Tolosana[12,‡], Ruben Vera-Rodriguez[12,‡], Gianpaolo Perelli[13,‡],
Giulia Orrù[13,‡], Gian Luca Marcialis[13,‡], Julian Fierrez[12,‡]

[1]Idiap Research Institute, Switzerland [2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]Université de Lausanne (UNIL), Switzerland [4]University of Bologna, Italy
[5]Fraunhofer Institute for Computer Graphics Research (IGD), Germany [6]TU Darmstadt, Germany
[7]Universidad de Alicante, Spain, [8]Facephi Biometria SA, R&D Centre, Spain,
[9]Federal University of Paraná (UFPR), Curitiba, PR, Brazil
[10]Federal Institute of Mato Grosso (IFMT), Pontes e Lacerda, Brazil [11]unico - idTech, Brazil
[12]Universidad Autonoma de Madrid (UAM), Spain [13]Università degli Studi di Cagliari (UNICA), Italy

[†]Competition organizer   [‡]Competition participant
[*]Equal Contribution

Abstract— Large-scale face recognition datasets are collected by crawling the Internet and without individuals' consent, raising legal, ethical, and privacy concerns. With the recent advances in generative models, recently several works proposed generating synthetic face recognition datasets to mitigate concerns in web-crawled face recognition datasets. This paper presents the summary of the Synthetic Data for Face Recognition (SDFR) Competition held in conjunction with the 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024) and established to investigate the use of synthetic data for training face recognition models. The SDFR competition was split into two tasks, allowing participants to train face recognition systems using new synthetic datasets and/or existing ones. In the first task, the face recognition backbone was fixed and the dataset size was limited, while the second task provided almost complete freedom on the model backbone, the dataset, and the training pipeline. The submitted models were trained on existing and also new synthetic datasets and used clever methods to improve training with synthetic data. The submissions were evaluated and ranked on a diverse set of seven benchmarking datasets. The paper gives an overview of the submitted face recognition models and reports achieved performance compared to baseline models trained on real and synthetic datasets. Furthermore, the evaluation of submissions is extended to bias assessment across different demography groups. Lastly, an outlook on the current state of the research in training face recognition models using synthetic data is presented, and existing problems as well as potential future directions are also discussed.

I. INTRODUCTION

Recent advancements in state-of-the-art face recognition models are driven in part by the availability of large-scale datasets and deep learning models [14], [16], [23], [32].

Competition Website: www.idiap.ch/challenge/sdfr

Meanwhile, large-scale face recognition datasets, such as VGGFace2 [9], MS-Celeb-1M [17], WebFace260M [49], etc., were collected by crawling images from the Internet, raising legal, ethical, and privacy concerns. In light of these concerns and the fact that these datasets are collected without the individuals' consent, some dataset owners decided to take their datasets down; for example, Microsoft retracted MS-Celeb-1M [1]. To address such concerns, recently, several studies have proposed generating synthetic face recognition datasets and using synthetic face images for training face recognition models [2], [5], [11], [34]. The generated synthetic face recognition datasets should be composed of different synthetic subjects with several samples per subject. As a face recognition dataset, each generated synthetic face dataset should include sufficient *intra-class* variations to model challenging scenarios in the task of face recognition including variations caused by pose, aging, expressions, occlusions, environmental conditions, etc. Moreover, the generated dataset needs to have adequate *inter-class* variation so that the trained face recognition models can have generalization for unseen subjects when used in a face recognition system. Given these challenges, generating synthetic face datasets with sufficient inter-class and intra-class variations is an active area of research.

As the main motivation for generating synthetic face recognition datasets, the generation of synthetic face recognition datasets should not rely on large-scale web-crawled datasets which have privacy concerns. In addition, since the synthetic face recognition datasets should eventually replace large-scale real face recognition datasets, face recognition models trained using synthetic datasets should lead to a

high recognition accuracy. However, there is still a gap in the recognition accuracy between training using large-scale real face recognition datasets and existing synthetic face recognition datasets [8], [30], [38].

The Synthetic Data for Face Recognition (SDFR) competition held in conjunction with the 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2024) was organized to accelerate research in synthetic data generation for privacy-friendly face recognition models and to bridge the gap between real and synthetic face datasets. Teams were invited to propose clever ways to use synthetic face recognition datasets (either existing or new synthetic face datasets) to train face recognition models. The competition was split into two tasks, where the first task involved a predefined face recognition backbone and a limit on the dataset size to focus on the quality of synthesized face datasets, while the second task provided almost complete freedom on the model backbone, the dataset, and the training pipeline. The submitted trained face recognition models using synthetic data were evaluated based on a diverse set of seven benchmarking datasets. The SDFR competition is the first competition which is specifically focused on generating synthetic face recognition datasets and allows training face recognition models using new methods or existing datasets from the literature. The competition rules were also designed to allow exploring ideas for generating privacy-friendly datasets, while preventing the application of large-scale web-crawled datasets. This paper presents a summary of the SDFR competition and discusses submissions and findings in this competition.

The remainder of the paper is organized as follows. Section II provides details about the SDFR competition, including the definition of tasks, and evaluation metrics used in the competition. In Section III, we provide a description of the final submitted face recognition systems proposed in the SDFR competition for each task. Section IV presents the results achieved in the different tasks and compare them with several baselines. Section V extends the analyses for final submission and evaluates their bias on different demography groups. Section V also discusses submissions with not qualified datasets, which are excluded from the leaderboard, and discusses their limitations and findings. Furthermore, Section V presents a further discussion on training face recognition using synthetic data and highlights potential future research directions in the field. Finally, in Section VI, we draw the conclusions from the SDFR competition.

II. SDFR Competition

In the SDFR competition, participants were expected to train face recognition models using synthetic data and submit their trained models (in ONNX format) through a submission platform. The organizers regularly evaluated submitted models on benchmarking datasets, and the results appeared in the leaderboard on the competition website throughout the competition. In the following, a brief overview of the definition of competition tasks, rules, and evaluation is presented. Finally, the novelties of the SDFR competition

compared to previous competitions on face recognition are discussed.

A. Tasks

The SDFR competition was divided into two tasks, and each team could use existing synthetic datasets or a newly generated dataset to participate in either or both tasks:

- **Task 1 [Constrained]:** In this task, the generated synthetic dataset could have up to one million synthesized images (for example, 10,000 identities and 100 images per identity). The backbone is also fixed to iResNet-50 [14].
- **Task 2 [Unconstrained]:** In this task, participants could use synthetic data with no limit on the number of synthesized images. Participants were also allowed to use any network architecture and train their best model with state-of-the-art techniques, but only using synthetic data.

Submitted models in each task were evaluated and ranked in the leaderboard of the corresponding task.

B. Rules

The competition was organized with several rules which allowed participants to generate new synthetic face recognition datasets and/or use (a subset, the complete set, or an extension of) existing ones to train face recognition models. Participants were not allowed to use a real dataset with identity labels for training the face generator model and generate synthetic images. As a result, some of the existing synthetic datasets for the literature, such as DCFace [24], SFace [6], and GANDiffFace [29], were not qualified for the competition. Further details about such synthetic datasets are discussed in Section V. However, the rules allowed to use real datasets without identity labels (such as FFHQ[1] [22]) to train the face generator model. This could be extended to real datasets with identity labels if the identity labels were not used in training the face generator model[2]. Participants were also allowed to use a pre-trained face recognition model for controlling and generating synthetic datasets. However, they were not allowed to directly learn embeddings of a pre-trained face recognition model. The complete definition of the rules is available on the competition website[3].

---

[1] The FFHQ dataset [22] contains 70,000 images without identity labels with variation in terms of age, ethnicity, etc. Especially, each individual image in this dataset was published on Flickr by their respective authors under licenses that allow free use, redistribution, and adaptation for non-commercial purposes.

[2] Given the limitations in available datasets without identity labels, this rule relaxed the restriction in available datasets and allowed participants to use more available datasets. Indeed, this relaxation in the rules was considered to provide the opportunity for participants to investigate limitations in existing datasets without identity labels (such as FFHQ [22]). However, outside of the scope of the competition, for a responsible and privacy-friendly synthetic face recognition dataset, large-scale web-crawled datasets (even without labels) should not be used to train generative models.

[3] www.idiap.ch/challenge/sdfr

C. Evaluation

The submitted models were evaluated on a diverse set of seven benchmarking face recognition datasets. The benchmarking datasets include:

- **one high-quality unconstrained dataset:** Labeled Faces in the Wild (LFW) [18],
- **two cross-pose datasets:** Celebrities in Frontal-Profile in the Wild (CFP-FP) and Cross-Pose LFW (CPLFW) [45],
- **two cross-age datasets**: AgeDB-30 [33] (30 years age gap) and Cross-age LFW (CALFW) [46],
- **two challenging mixed-quality datasets:** IARPA Janus Benchmark–B (IJB-B) [41] and IARPA Janus Benchmark–C (IJB-C) [28].

The submitted models were ranked on each dataset separately, and the final ranking was based on the Borda count.

D. Novelties

Considering the wide application of face recognition systems and the increasing demand for high-performing face recognition models, several competitions have been organized in conjunction with different conferences [4], [13], [15], [25], [31]. However, most of these competitions were based on large-scale web-crawled face recognition datasets, which raise privacy concerns. The FRCSyn challenge held in WACVW 2024 [30], [31] was the first competition on training face recognition models based on synthetic data; however, participants were limited to using two existing synthetic datasets from the literature (i.e., DCFace [24] and GANDiffFace [29]). The SDFR competition is the first competition focused on generating synthetic face recognition datasets, and allows training face recognition models using new methods or existing datasets from the literature (as long as they follow competition rules). The rules of the SDFR competition were also designed to allow exploring ideas for generating privacy-friendly datasets, while preventing the application of large-scale web-crawled datasets. Furthermore, the SDFR competition is based on a thorough evaluation using seven benchmarking datasets.

III. DESCRIPTION OF SUBMISSIONS

The SDFR competition received significant interest, with 12 international teams registered. In total, we received 17 submissions from 5 teams. Table I provides a summary overview of the participating teams, including the team name, their affiliation, and the tasks they participated. In the following section, we briefly describe the proposed method used by each team.

BioLab (Task 1 and Task 2): The BioLab team used iResNet-50 [14] for task 1 and an iResNet-100 [14] for task 2. Both models were trained using the margin-based AdaFace loss [14], chosen for its state-of-the-art results and robustness during training, especially when employing low-quality images. For task 1, they trained their model on the IDiff-Face (Uniform) [5] dataset, which contains about 500K images across 10K identities; for task 2, they extended their training dataset in task 1 with the DigiFace-1M [2] dataset,

TABLE I: A summary of the participant teams, their affiliations, and their participating tasks.

Team Name	Affiliations[†]	Country	Tasks
BioLab	4	Italy	1,2
IGD-IDiff-Face	5,6	Germany	1
APhi	7,8	Spain	1
BOVIFOCR-UFPR	9-11	Brazil	1,2
BiDA-PRA	12,13	Spain, Italy	1,2

[†]The numbers reported in the "Affiliations" column are provided in the title page.

bringing the total up to 1.7M images spanning 120k classes. To improve the performance of the models, they applied data augmentation techniques to the images during training. In particular, they sequentially applied random horizontal flips, color jittering, random crops by blacking out the rest of the image, random JPEG compression, random grayscale, and random down-and-upsampling with different interpolation algorithms (i.e. nearest neighbor, bilinear, bicubic, area, Lanczos). When employing images from DigiFace-1M [2], they applied a more aggressive data augmentation pipeline to bridge the quality gap between synthetic and real images; as suggested in [2], they extended the above-mentioned pipeline by adding more photometric augmentation steps i.e., Gaussian blur (20%), random noise (20%), and motion blur (10%). They built the validation set by generating random matching and non-matching pairs by selecting images from the first classes of IDiff-Face and DigiFace-1M, generating 12K and 84K pairs, respectively, for task 1 and task 2. The classes used to generate such pairs are excluded from training. With this validation set, they monitored the verification accuracy during training and applied an early stopping policy with patience of 10 epochs to prevent overfitting. Both models are optimized with SGD with a batch size of 512, and the initial learning rate of 0.1 divided by a factor of 10 at prefixed epochs to ensure better training stability. Code is available: https://github.com/ndido98/sdfr

IGD-IDiff-Face (Task 1): The submitted solution of the IGD-IDiff-Face team was based on IDiff-Face [5]. The training dataset was obtained by combining two subsets, IDiff-Face CPD25 (Uniform) and IDiff-Face CPD50 (Two-Stage), of the IDiff-Face dataset. These subsets contain 10,049 and 10,050 identities, respectively, and 50 images per identity. They selected the first exact 10k from each subset. Therefore, the total number of identities is 20K, each contains 50 images. They trained the face recognition model using CosFace [39] with a margin of 0.35 and a scale factor of 64. During the training, they augmented the training data using RandAugment [12] with parameters (operation is 16 and magnitude is 4) described in [7]. The model is trained using the SGD optimizer with a batch size of 512. The model is trained using a plateau-based learning scheduler. The initial learning rate is set to 0.1 and divided by 10 when the average validation accuracy does not improve for 10 consecutive epochs where the patience parameter is set to 10 and the factor to 0.4. The total number of epochs is set to 200 with an early stopping parameter of 10

epochs, i.e., the training will stop if the average validation accuracy does not improve for 10 epochs. Code is available: https://github.com/fdbtrs/IDiff-Face

APhi (Task 1): The APhi team used iResNet-50 as the backbone of their model and trained it with MagFace [32] loss function. As their training dataset, they used the IDiff-Face (Uniform) [5], which consists of about 550K images and 10K identities. They used data augmentation techniques, including random scale and then resize, Gaussian blur, Hue saturation, and horizontal flip. They trained their model for 25 epochs using the SGD optimizer with a batch size of 512 and a learning rate initialized from 0.1 and divided by 10 at epochs 10, 18, and 22. The weight decay was set to 5×10^{-4} and the momentum is 0.9. In addition, 200 users have been taken from the training set and used for validation (a total of 10K images). The validation was performed by evaluating the equal error rate (EER) over all the possible pairs in the validation set. They validated the trained model every 1,000 iterations, and stored the model with the lowest EER.

BOVIFOCR-UFPR (Task 1 and Task 2): The BOVIFOCR-UFPR team used iResNet-50 [14] for both task 1 and task 2 and trained their models using ArcFace loss [14], chosen for its performance for deep face recognition [13]. For task 1, they used the uniform version of the IDiff-Face [5] as the training dataset containing $10,049$ identities, with 55 images per identity ($502,450$ images in total). The training images were augmented using Random Flip with a probability of 50%. They also applied random erasing [47] and randaugment [12] as additional augmentation. They used the SGD optimizer and set the momentum to 0.9 and weight decay to 5×10^{-4}. The learning rate was set to 0.02 and reduced at each iteration following the equation:

$$\left(\frac{1.0 - \frac{l}{t}}{1.0 - \frac{l-1}{t}} \right), \tag{1}$$

where l and t represent the current iteration and the total number of iterations, respectively.

The model was trained for 20 epochs within a batch size of 128, running for approximately 78K iterations using the Insightface library[4]. Code for task 1 is available: https://github.com/PedroBVidal/insightface

For task 2, they also used IDiff-Face (Uniform) [5], and applied random flip and also random face pose augmentation [10] based on 3D face reconstruction [48] to manipulate face pose in yaw axis. About 40K faces were randomly profiled in left ($-60°$) or right ($+60°$) direction, corresponding to $\sim 8\%$ from the whole dataset. Some sample augmented face images are shown in Fig. 1. The used SGD optimizer with learning rate 0.1, momentum 0.9, batch size 256, and weight decay of 5×10^{-4}. The model was trained during 20 epochs, corresponding to 84k iterations. Code for task 2 is also available: https://github.com/BOVIFOCR/insightface_SDFR_at_FG2024

BiDA-PRA (Task 1 and Task 2): The BiDA-PRA team submitted the same model for both tasks. They used iResNet-

Fig. 1: Sample original and augmented face images from the IDiff-Face (Uniform) dataset obtained by synthetically rotating face pose in the yaw axis used by the BOVIFOCR-UFPR team in task 2.

50 [14] and trained using the AdaFace [23] loss function. For preparing their training dataset, they used the information about demographic distribution across the competition evaluation datasets, focusing on ethnicity, gender, and age, available from the datasets meta-data or from the analyses reported in previous studies, such as [27]. For the training dataset, they used pre-generated IDiff-Face synthetic dataset (both Uniform and Two-stage subsets, containing about 1M images in total), and created a "cleaned" version of the dataset, discarding each identity image whose labels were different from the average ethnicity and gender predictions. These attributes were automatically extracted for each face image in the IDiff-Face dataset using the FairFace [19] model. Finally, after the cleaning step, they considered 600K total face images of the IDiff-Face dataset. In addition, to generate more training samples, they trained[5] StyleGAN3 on CASIA-WebFace [43] (without using its identity labels) and used it to generate 20K images of unique identities. They applied the method in [29] to generate pose variations in the latent space of those images, resulting in 20 images per identity, and in total 400K face images generated by StyleGAN3. However, they did not perform an analysis of the demographic distribution for the images generated by Style-GAN3. During training, they augmented data using random histogram equalization (10%), random horizontal flip (50%), random grayscale conversion (20%), and Gaussian blur (80% with kernel 1, and 20% with kernel 5). In addition, they applied a custom augmentation technique involved cutting off half of the face, either the left or the right side, from some of the frontal or near-frontal images[6]. In particular, for each selected image, they checked if the face could be considered frontal or not through a landmark-based estimation. If the image was considered frontal, they randomly selected the left or the right side to be removed. They detected the landmarks of the nose and the eyes using MTCNN [44], and then they computed the difference between the angles on the eyes from the triangle composed of these three landmarks. Specifically, they selected a maximum angle difference of 15 degrees to

TABLE II: Leaderboard of Task 1 and Task 2 compared to several baselines.

Task / Baseline	Team / Method	No. Images	LFW	CALFW	CPLFW	AgeDB30	CFP-FP	IJB-B (TAR@10^{-4})	IJB-C (TAR@10^{-4})	Rank
Baselines (real)	MS-Celeb [17]	5.8M	99.82	95.92	92.52	97.62	96.01	94.88	96.23	-
	WebFace-4M [49]	4M	99.78	96.02	93.90	97.52	97.19	95.52	97.02	-
	Casia-WebFace [43]	490K	99.08	92.88	89.23	93.72	94.67	45.91	52.43	-
Baselines (synthetic)	SynFace [34]	1M	85.55	69.85	59.42	58.38	62.11	14.62	12.85	-
	DigiFace [2]	1.2M	90.63	74.02	71.38	65.03	78.11	38.89	45.09	-
	IDNet [26]	1.05M	84.90	70.28	68.32	63.88	70.34	27.57	32.28	-
Task 1 (leaderboard)	IGD-IDiff-Face	1M	98.07	90.60	81.28	87.60	84.76	64.36	68.04	1
	APhi	500K	97.45	89.95	78.03	84.75	80.04	58.18	60.85	2
	BOVIFOCR-UFPR	500K	97.53	89.38	80.07	83.90	84.37	12.70	13.71	3
	BioLab	500K	96.97	89.12	76.80	83.77	77.34	60.21	63.56	4
	BiDA-PRA	1M	96.88	87.95	78.13	83.85	78.90	58.08	58.33	5
Task 2 (leaderboard)	BioLab	1.7M	98.33	90.87	84.45	87.85	88.11	76.94	81.25	1
	BiDA-PRA	1M	96.88	87.95	78.13	83.85	78.90	58.08	58.33	2
	BOVIFOCR-UFPR	500K	96.38	88.38	75.98	81.55	76.97	40.97	45.93	3

The values reported in the table for each benchmarking dataset are in percentage. The values reported on the LFW, CALFW, CPLFW, AgeDB30, and CFP-FP datasets are accuracy. The values reported on the IJB-B and IJB-C datasets are True Accept Rate (TAR) at the False Accept Rate (FAR) of 10^{-4}.

All baseline models are trained with iResNet-50 and AdaFace [23] loss function. Baselines with real data are pretrained models available on the AdaFace GitHub repository, and baseline models with synthetic data are trained by organizers using the same hyperparamters.

consider the face as frontal or near-frontal, while setting the image selection probability to 5%. Code is available: https://github.com/BiDAlab/SDFR-FRModels

IV. LEADERBOARDS

As mentioned in Section II, submitted models are evaluated on a diverse set of benchmarking datasets. Table II reports the leaderboard of each task and also compares the results of submissions with some baselines from the literature. The values in the table for each benchmarking dataset are in percentage. The values reported on the LFW, CALFW, CPLFW, AgeDB30, and CFP-FP datasets are accuracy and the values reported on the IJB-B and IJB-C datasets are True Accept Rate (TAR) at the False Accept Rate (FAR) of 10^{-4}. The baselines were trained by the corresponding dataset mentioned in the table using iResNet-50 [14] network with AdaFace [23] loss function. As can be seen in Table II, there is a gap in the recognition performance of baselines with real datasets and baselines with synthetic datasets. This difference is more significant for challenging datasets such as IJB-B, and IJB-C. However, the submitted models could improve the performance compared to baselines with synthetic datasets. In particular, the submitted model by the BioLab team for task 2 achieved considerable performance on all datasets, even on challenging datasets such as IJB-B and IJB-C. This submission used two different datasets, DigiFace [2] (a synthetic dataset based on computer graphic pipeline) and IDiff-Face [5] (a synthetic dataset based on diffusion models), showing potentials in improving performance using synthetic datasets with different types. While task 1 received significant interest from the participating teams, task 2 received fewer submissions, indicating the difficulty of task 2 which allowed unconstrained training with synthetic data.

In task 1, the IGD-IDiff-Face team used two subsets of IDiff-Face (Uniform and Two-stage), resulting in larger training data, compared to the APhi, BOVIFOCR-UFPR, and BioLab teams, who used only the Uniform subset of IDiff-Face with about 500K images. Meanwhile, the submission of BiDA-PRA also benefits from both subsets of IDiff-Face and in addition used StyleGAN generated images, using the complete limit of 1M images in task 1. While using 1M training samples helped the performance of their submission by the IGD-IDiff-Face team, it could not help the performance of the BiDA-PRA team. This might be caused by not enough variation in their StyleGAN-generated images, which only had pose variation.

All the teams also used different data augmentation techniques, to further improve their trained models. The data augmentation transformations used by participants contain various techniques and include traditional methods (such as random flip, Gaussian blur, etc), state-of-the-art techniques (such as random erasing [47], randaugment [12], and pose augmentation [10]), as well as custom methods (such as cutting off half of the face). This highlights the importance of data augmentation in training with synthetic data, which lack intra-class variations.

Last but not least, all final submissions have used the IDiff-Face dataset as part of their training, with different sampling strategies and training approaches. In fact, the IDiff-Face dataset is the best-performing synthetic dataset in the literature which is aligned with the competition rules. We should note that participating teams also submitted other models, which had competitive performance with the final leaderboard but had a conflict with the competition rules. Such submissions with not qualified datasets are discussed in detail in Section V. To facilitate reproducibility of results in the SDFR competition, all final submission which appeared on the final leaderboard are publicly available on the competition website: www.idiap.ch/challenge/sdfr

TABLE III: Evaluation results of the final submissions on the RFW dataset.

Task/Baseline	Team / Method	Caucasian	Asian	Indian	African	Avg.	Std.
Baselines (real)	MS-Celeb [17]	99.33	97.73	98.23	98.32	98.40	0.68
	WebFace-4M [49]	98.98	97.10	97.72	97.63	97.86	0.80
	Casia-WebFace [43]	93.05	85.07	88.20	85.57	87.97	3.66
Baselines (synthetic)	SynFace [34]	65.60	64.48	61.48	57.27	62.21	3.73
	DigiFace [2]	71.93	68.30	69.02	64.80	68.52	2.94
	IDNet [26]	70.03	64.22	65.77	59.30	64.83	4.44
Task 1	IGD-IDiff-Face	85.98	80.98	82.45	76.85	81.57	3.78
	APhi	84.55	79.45	80.70	73.85	79.64	4.43
	BOVIFOCR-UFPR	83.65	78.92	80.18	74.72	79.37	3.69
	BioLab	81.70	77.22	78.97	71.20	77.27	4.45
	BiDA-PRA	82.12	75.87	77.88	70.00	76.47	5.04
Task 2	BioLab	86.32	81.47	82.22	77.30	81.82	3.70
	BiDA-PRA	82.12	75.87	77.88	70.00	76.47	3.48
	BOVIFOCR-UFPR	80.37	76.47	77.65	71.08	76.39	3.91

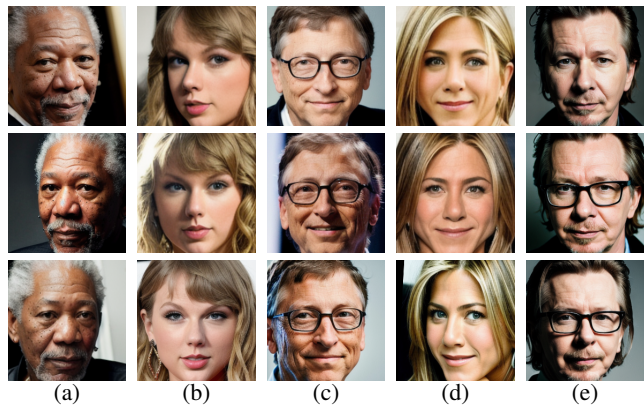The values reported in the table are accuracy and in percentage.

Fig. 2: Sample face images from the new custom generated dataset by the BioLab team generated for different famous people: (a) Morgan Freeman, (b) Taylor Swift, (c) Bill Gates, (d) Jennifer Aniston, (e) Gary Oldman.

V. DISCUSSION

In addition to our evaluation of submitted models in Table II, we evaluated each of the models in the leaderboard on the Racial Faces in-the-Wild (RFW) [40] dataset, to investigate their performance across different demography groups. The RFW dataset contains four testing subsets corresponding to Caucasian, Asian, Indian, and African groups, and each is composed of about 3,000 subjects with 6,000 image pairs for the face verification task. Table III reports the performance of the final submissions in the leaderboard as well as baseline models trained on real and synthetic datasets on the different subsets of the RFW dataset. As the results in this table show, all final submissions from the leaderboard achieved their best performance in the Caucasian group and worst performance in the African group. In each task, the best submission in the leaderboard in Table II also achieves the highest average and lowest standard deviation across different demography groups on the RFW dataset. The results in Table III also show that there is similarly bias in the baselines trained on real and synthetic datasets. Hence, the results in this table motivates future research on generating responsible synthetic face recognition datasets to alleviate bias in recognition accuracy across different demography groups.

Some of the synthetic datasets in the literature, such as DCFace [24], SFace [6], and GANDiffFace [29], have a conflict with the competition rules, and therefore participants were not allowed to use them for their final submission. DCFace [24] used CASIA-WebFace [43] to train the dual condition generator, where training the generator requires two samples of the same subject, and thus it requires an identity-labeled dataset. SFace [6] also used CASIA-WebFace with identity labels to train a conditional version of StyleGAN2-ADA [20] (conditioned on identity). Therefore, DCFace and SFace rely on CASIA-WebFace, as a large-scale web-crawled dataset. GANDiffFace [29] used Style-GAN to generate different synthetic images and then used DreamBooth [36] to generate different samples of images that were generated by StyleGAN in the first place. However,

DreamBooth [36] is based on Stable Diffusion [35], which is trained on the LAION [37] dataset. The LAION dataset is a large-scale web-crawled dataset, which includes the identity label of famous people in the caption of images. As a result, Stable Diffusion (a text-to-image model) can generate images of famous people given the name of the person in the input text [42], raising privacy and ethical issues. Therefore, since the generator in GANDiffFace is trained and benefits from identity labels in the LAION dataset, the GANDiffFace dataset was not allowed in the competition.

In addition to the submissions reported in the leaderboard, some participating teams trained models with datasets which had a conflict with the competition rules. The BOVIFOCR-UFPR team also used DCFace [24] with 1.2M images, and dropped 5,454 identities with 55 images per id, reducing the total number of images to 999,975 to fulfill the maximum allowed number of images in task 1. Then, they trained iResNet-50 with the same data augmentation and loss function as their final submission for task 1 described in Section III. The APhi team submitted two models for task 1 trained on DCFace [24]. In the first model, they selected arbitrary 1M images with 59,968 identities and trained the iResNet-50 network with AdaFace [23] loss function. In their second model, they selected 1M based on the face quality. To this end, they extracted the MagFace [32] score of each image, and then they selected the best quality images, resulting in 1M images and 59,894 identities. Then, they trained the iResNet-50 network with ArcFace [14] loss function. The BioLab team submitted a model trained with a newly generated dataset based on Stable Diffusion as well as with DCFace. Their newly generated dataset based on Stable Diffusion, was built by querying Wikidata[7] for all people born between 1900 and 2004 that have an associated picture. To remove subjects that may be difficult to generate with Stable Diffusion, they filtered the list by keeping only the ones with at least 30 translations of their Wikipedia page.

[7]www.wikidata.org

TABLE IV: Submissions with not qualified datasets.

Team (method)	LFW	CALFW	CPLFW	AgeDB30	CFP-FP	IJB-B (TAR@10^{-4})	IJB-C (TAR@10^{-4})
BOVIFOCR-UFPR	98.98	92.42	86.38	91.38	90.89	79.62	84.11
APhi (model 1)	98.83	92.25	82.93	90.95	87.36	78.41	82.38
APhi (model 2)	98.38	92.22	84.27	90.88	87.57	79.48	83.61
BioLab (custom dataset)	93.90	82.30	71.23	74.42	72.04	49.87	52.76
BioLab (DCFace)	98.52	92.48	83.35	91.12	87.63	30.17	24.38

After gathering the list, they generated 64 images per subject by employing the Realistic Vision 5.1 model, fine-tuned from Stable Diffusion 1.5. Then, they filtered the generated images by computing ArcFace [14] features for each image, computing the distance between each pair of images of each subject, and finally keeping the largest subset so that the cosine distance between all feature vectors is less than a given threshold. Moreover, if after this filtering step, there were not enough images for a given subject, that subject was removed entirely from the dataset. This process led to a dataset with around 825K images across 18K identities. Fig. 2 illustrates sample images from their generated dataset. For task 1, they used their newly generated dataset[8] and trained iResNet-50 with the same data augmentation and loss function as their final submission described in Section III. For task 2, they also used the same training approach to train the iResNet-100 network, but using the DCFace dataset with 1.2M images. Table IV reports the performance of these submissions on benchmarking datasets. As the results in this table show, these submissions achieve competitive performance with the final submissions on the leaderboards in Table II.

In drawing our discussion to a close, the submissions in this competition used creative approaches for training face recognition using synthetic data and could achieve improvements compared to baselines based on synthetic datasets. However, as the results in Table II show, we still observe a significant gap between models trained with synthetic data and models trained with large-scale web-crawled datasets, demanding more research on generation and training face recognition with synthetic data. In fact, one aspect of this challenge is to increase inter-class and intra-class variation in generating synthetic face datasets. Compared to large-scale datasets, such as MS-Celeb which has 5.8M images, the largest synthetic face recognition dataset in the literature has 1.2M images (DigiFace). Therefore, another potential aspect of the topic for future work can be to explore if we can scale synthetic face datasets to generate more images and investigate if it can contribute to the recognition performance of face recognition models. In fact, scaling synthetic face recognition datasets may not be straightforward, as we may achieve the generation capacity of face generator networks [3], which also requires further studies. We should

---

[8]As this approach relies on Stable Diffusion (trained on the LAION [37] dataset) for image generation, this submission had a conflict with competition rules, which prevent the use of large-scale datasets with identity labels to train face generator network.

also note that in this competition and in most methods for generating synthetic face recognition datasets, still a pre-trained face recognition model is used in the dataset generation process. Therefore, if we agree on using a pre-trained face recognition model in data generation, it has been shown that directly learning embeddings of a pre-trained face recognition model using synthetic data, without any identity class, achieves superior performance than training using existing synthetic face recognition datasets [38]. However, in a *fully* privacy-friendly and responsible synthetic face recognition dataset, no pre-trained face recognition model based on large-scale web-crawled face datasets should be used, otherwise, the generated dataset still relies on large-scale web-crawled datasets. This is in fact an important future research direction that requires more attention from the research community.

VI. CONCLUSION

Existing large-scale face recognition datasets are collected without individuals' consent and by crawling the Internet, raising legal, ethical, and privacy concerns. While generating a synthetic face recognition dataset can be a potential solution, it is challenging to generate a synthetic dataset with sufficient inter-class and intra-class variations. As a result, there is still a gap between training face recognition models with real and synthetic datasets. The submitted models in the SDFR competition deployed different techniques using newly generated and existing datasets to improve the performance of face recognition models trained with synthetic datasets. In this paper, different submissions from participating teams in the SDFR competition were presented and their trained models were evaluated on a diverse set of benchmarking datasets. We also outlined the current state of research on this topic and discussed open problems, such as scaling synthetic datasets as well as increasing the variations in generated images, which require more research in the future.

REFERENCES

[1] The rise and fall (and rise) of datasets. *Nature Machine Intelligence*, 4(1):1–2, 2022.

[2] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen. DigiFace-1M: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3526–3535, 2023.

[3] V. N. Boddeti, G. Sreekumar, and A. Ross. On the biometric capacity of generative face models. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[4] F. Boutros, N. Damer, J. N. Kolf, K. Raja, F. Kirchbuchner, R. Ramachandra, A. Kuijper, P. Fang, C. Zhang, F. Wang, D. Montero, N. Aginako, B. Sierra, M. Nieto, M. E. Erakin, U. Demir, H. K. Ekenel, A. Kataoka, K. Ichikawa, S. Kubo, J. Zhang, M. He, D. Han, S. Shan, K. Grm, V. Štruc, S. Seneviratne, N. Kasthuriarachchi, S. Rasnayaka, P. C. Neto, A. F. Sequeira, J. R. Pinto, M. Saffari, and J. S. Cardoso. MFR 2021: Masked face recognition competition. In *Proceedings of the IEEE International joint conference on biometrics (IJCB)*, pages 1–10. IEEE, 2021.

[5] F. Boutros, J. H. Grebe, A. Kuijper, and N. Damer. IDiff-Face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19650–19661, 2023.

[6] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer. SFace: Privacy-friendly and accurate face recognition using synthetic data. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022.

[7] F. Boutros, M. Klemt, M. Fang, A. Kuijper, and N. Damer. Unsupervised face recognition using unlabeled synthetic data. In *Proceedings of the IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

[8] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, page 104688, 2023.

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 67–74. IEEE, 2018.

[10] S. Cheng, P. Ma, G. Tzimiropoulos, S. Petridis, A. Bulat, J. Shen, and M. Pantic. Towards pose-invariant lip-reading. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4357–4361. IEEE, 2020.

[11] L. Colbois, T. de Freitas Pereira, and S. Marcel. On the use of automatically generated synthetic image datasets for benchmarking face recognition. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

[12] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 702–703, 2020.

[13] J. Deng, J. Guo, X. An, Z. Zhu, and S. Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1437–1444, 2021.

[14] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[15] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[16] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. EdgeFace: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(2):158–168, 2024.

[17] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102. Springer, 2016.

[18] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[19] K. Karkkainen and J. Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558, 2021.

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

[21] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[22] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[23] M. Kim, A. K. Jain, and X. Liu. AdaFace: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022.

[24] M. Kim, F. Liu, A. Jain, and X. Liu. DCFace: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12715–12725, 2023.

[25] J. N. Kolf, F. Boutros, J. Elliesen, M. Theuerkauf, N. Damer, M. Alansari, O. A. Hay, S. Alansari, S. Javed, N. Werghi, K. Grm, V. Štruc, F. Alonso-Fernandez, K. H. Diaz, J. Bigun, A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, S. Marcel, I. Medvedev, B. Jin, D. Nunes, A. Hassanpour, P. Khatiwada, A. A. Toor, and B. Yang. EFaR 2023: Efficient face recognition competition. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–12. IEEE, 2023.

[26] J. N. Kolf, T. Rieber, J. Elliesen, F. Boutros, A. Kuijper, and N. Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 806–816, 2023.

[27] C. Kong, Q. Luo, and G. Chen. RSFAD: A large-scale real scenario face age dataset in the wild. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021.

[28] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. IARPA janus benchmark-c: Face dataset and protocol. In *Proceedings of the International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[29] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert. GANDiffFace: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3086–3095, 2023.

[30] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X.-Y. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M. H. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, H. O. Shahreza, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltsou, S. Papadopoulos, C. Diou, N. D. Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, Ángela Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, and M. Marras. FRCSyn-onGoing: Benchmarking

and comprehensive evaluation of real and synthetic data to improve face recognition systems. *Information Fusion*, 107:102322, 2024.

[31] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X.-Y. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M. H. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, H. O. Shahreza, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltsou, S. Papadopoulos, C. Diou, N. D. Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, Ángela Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, and M. Marras. FRCSyn challenge at WACV 2024: Face recognition challenge in the era of synthetic data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 892–901, 2024.

[32] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vsion and Pattern Recognition (CVPR)*, pages 14225–14234, 2021.

[33] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. AgeDB: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 51–59, 2017.

[34] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. SynFace: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10880–10890, 2021.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[36] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[37] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[38] H. O. Shahreza, A. George, and S. Marcel. SynthDistill: Face recognition with knowledge distillation from synthetic data. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[39] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

[40] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702, 2019.

[41] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, et al. IARPA janus benchmark-b face dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 90–98, 2017.

[42] Y. Yang, Y. lin, H. Liu, W. Shao, R. Chen, H. Shang, Y. Wang, Y. Qiao, K. Zhang, and P. Luo. Towards implicit prompt for text-to-image models, 2024.

[43] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[45] T. Zheng and W. Deng. Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.

[46] T. Zheng, W. Deng, and J. Hu. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[47] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

[48] X. Zhu, X. Liu, Z. Lei, and S. Z. Li. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):78–92, 2017.

[49] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502, 2021.