FRCSyn-onGoing: Benchmarking and Comprehensive Evaluation of Real and Synthetic Data to Improve Face Recognition Systems

Pietro Melzi[a], Ruben Tolosana[a,*], Ruben Vera-Rodriguez[a], Minchul Kim[b], Christian Rathgeb[c], Xiaoming Liu[b], Ivan DeAndres-Tame[a], Aythami Morales[a], Julian Fierrez[a], Javier Ortega-Garcia[a], Weisong Zhao[d,e], Xiangyu Zhu[f,g], Zheyu Yan[f], Xiao-Yu Zhang[d,e], Jinlin Wu[h], Zhen Lei[f,g,h], Suvidha Tripathi[i], Mahak Kothari[i], Md Haider Zama[i], Debayan Deb[i], Bernardo Biesseck[j,k], Pedro Vidal[j], Roger Granada[l], Guilherme Fickel[l], Gustavo Führ[l], David Menotti[j], Alexander Unnervik[m,n], Anjith George[m], Christophe Ecabert[m], Hatef Otroshi Shahreza[m,n], Parsa Rahimi[m,n], Sébastien Marcel[m,o], Ioannis Sarridis[p], Christos Koutlis[p], Georgia Baltsou[p], Symeon Papadopoulos[p], Christos Diou[q], Nicolò Di Domenico[r], Guido Borghi[r], Lorenzo Pellegrini[r], Enrique Mas-Candela[s], Ángela Sánchez-Pérez[s], Andrea Atzori[t], Fadi Boutros[u,v], Naser Damer[u,v], Gianni Fenu[t], Mirko Marras[t]

[a] *Universidad Autonoma de Madrid, Spain*
[b] *Michigan State University, US*
[c] *Hochschule Darmstadt, Germany*
[d] *IIE, CAS, China*
[e] *School of Cyber Security, UCAS, China*
[f] *MAIS, CASIA, China*
[g] *School of Artificial Intelligence, UCAS, China*
[h] *CAIR, HKISI, CAS, China*
[i] *LENS, Inc., US*
[j] *Federal University of Paraná, Curitiba, PR, Brazil*
[k] *Federal Institute of Mato Grosso, Pontes e Lacerda, Brazil*
[l] *unico - idTech, Brazil*
[m] *Idiap Research Institute, Switzerland*
[n] *École Polytechnique Fédérale de Lausanne, Switzerland*
[o] *Université de Lausanne, Switzerland*
[p] *Centre for Research and Technology Hellas, Greece*
[q] *Harokopio University of Athens, Greece*
[r] *University of Bologna, Cesena Campus, Italy*
[s] *Facephi, Spain*
[t] *University of Cagliari, Italy*

[*] Corresponding author. Email address: ruben.tolosana@uam.es; postal address: Universidad Autonoma de Madrid, C. Francisco Tomás y Valiente, 11, 28049 Madrid, Spain

[u]*Fraunhofer IGD, Germany*
[v]*TU Darmstadt, Germany*

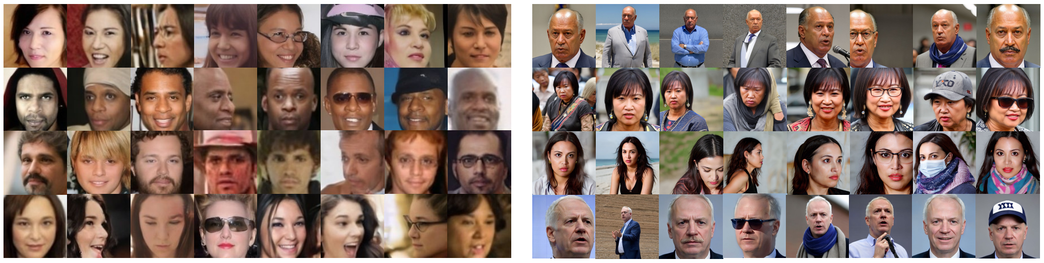
Abstract

This article presents FRCSyn-onGoing, an ongoing challenge for face recognition where researchers can easily benchmark their systems against the state of the art in an open common platform using large-scale public databases and standard experimental protocols. FRCSyn-onGoing is based on the Face Recognition Challenge in the Era of Synthetic Data (FRCSyn) organized at WACV 2024. This is the first face recognition international challenge aiming to explore the use of real and synthetic data independently, and also their fusion, in order to address existing limitations in the technology. Specifically, FRCSyn-onGoing targets concerns related to data privacy issues, demographic biases, generalization to unseen scenarios, and performance limitations in challenging scenarios, including significant age disparities between enrollment and testing, pose variations, and occlusions. To enhance face recognition performance, FRCSyn-onGoing strongly advocates for information fusion at various levels, starting from the input data, where a mix of real and synthetic domains is proposed for specific tasks of the challenge. Additionally, participating teams are allowed to fuse diverse networks within their proposed systems to improve the performance. In this article, we provide a comprehensive evaluation of the face recognition systems and results achieved so far in FRCSyn-onGoing. The results obtained in FRCSyn-onGoing, together with the proposed public ongoing benchmark, contribute significantly to the application of synthetic data to improve face recognition technology.

Keywords: FRCSyn-onGoing, Face Recognition, Generative AI, Demographic bias, Benchmark

1. Introduction

Facial images are the predominant data for biometric recognition nowadays, widely employed in various fields such as surveillance, government offices, and smartphone authentication [1], among others. Numerous studies in the literature have played a crucial role in advancing state-of-the-art (SOTA) Face Recognition (FR) technologies, demonstrating remarkable performance



(a) DCFace [6].

(b) GANDiffFace [7].

Figure 1: Examples of synthetic identities (one for each row) and their intra-class variations provided by two generative frameworks: (a) DCFace and (b) GANDiffFace. The synthetic identities represent different demographic groups considered in the FRCSyn Challenge.

on established benchmarks [2, 3]. The success of these technologies can be attributed to the emergence of Deep Learning (DL) and the development of highly effective loss functions based on margin loss, capable of producing exceptionally discriminative features [4]. Consequently, FR systems have made significant advances, achieving impressive results on well-recognized databases, such as LFW [5].

Nevertheless, FR continues to deal with numerous challenges, stemming from factors such as variations in facial images related to pose, aging, expressions, and occlusions. These challenges cause significant issues within the field [1, 8, 9]. The integration of DL brings forth additional concerns, including limited training data, noisy labeling, imbalanced data pertaining to diverse identities and demographic groups, and low resolution, among other issues [10]. Numerous studies indicate that DL models, even when trained on extensive databases, experience notable performance drops when confronted with previously unseen conditions [11, 12]. Deploying FR systems that can effectively overcome these challenges and generalize well to unforeseen conditions remains a difficult task. Notably, training data often exhibit significant imbalances across demographic groups [4], and they may fail to adequately represent the full range of possible occlusions in real-world scenarios [13]. Various limitations associated with established databases and benchmarks are extensively discussed in [14]. For instance, LFW [5] is considered to have a limited number of images per subject for SOTA challenges such as illumination, pose, and occlusion invariants.

In recent years, the literature has introduced various approaches for generating synthetic face content [15, 16, 17] intended for different applications, including FR [6, 18, 19] and digital face manipulations, commonly known as DeepFakes [20, 21, 22]. Additionally, synthetic content has been created for other biometric modalities [23, 24, 25]. The utilization of synthetic data presents several advantages compared to real-world databases. Firstly, synthetic databases offer a promising solution to address privacy concerns associated with real data, which are often collected from individuals without their knowledge or consent through various online sources [26]. Secondly, synthetic face generators have the capacity to generate large amounts of data, a particularly valuable property following the discontinuation of established databases due to privacy concerns [27] and the implementation of regulations such as the EU-GDPR, which mandates informed consent for collecting and using personal data [28]. Finally, when the synthesis process is controllable, it becomes relatively straightforward to create databases with specific characteristics (*e.g.,* demographic groups, age, pose, etc.) and their corresponding labels, without requiring additional human efforts [6, 7]. This is in contrast to real-world databases, which may not comprehensively represent diverse demographic groups [29], among various other aspects.

These advantages have motivated an initial exploration into the application of synthetic face data in current FR systems. Furthermore, synthetic data have proven successful when combined with domain adaptation techniques across various image applications, such as semantic segmentation [30], super-resolution [31], and image dehazing [32]. Innovative generative frameworks, including Generative Adversarial Networks (GANs) [33, 34] and 3D models [16], have been introduced to synthesize databases suitable for training FR systems. While these synthetic databases propel advancements in the field, some exhibit limitations that impact the performance of FR systems compared to those trained with real data. Specifically, databases synthesized with GANs offer limited representations of intra-class variations [33], and those synthesized with 3D models lack realism. Recently, Diffusion models have been employed to generate synthetic databases with enhanced intra-class variations, effectively addressing some limitations observed in prior synthetic databases [6, 7]. This is also supported by various recent works involving Diffusion models [17, 35, 36].

To evaluate the effectiveness of novel synthetic databases generated using Diffusion models for training FR systems, this article describes the experimental framework and results of FRCSyn-onGoing, which is based on the “Face Recognition Challenge in the Era of Synthetic Data (FRCSyn)” organized at WACV 2024[1]. This challenge is designed to comprehensively analyze the following research questions:

1. To what degree can synthetic data effectively replace real data for training FR systems, and what are the limits of FR technology exclusively trained with synthetic data?
2. Can the fusion of real and synthetic data be beneficial in addressing and mitigating the existing limitations within FR technology?

These research questions have gained significant relevance, especially in light of the discontinuation of FR databases due to privacy concerns [27] and the observed limitations in FR technology across demographic groups [18, 37] and challenging conditions [10]. In this study, we comprehensively evaluate the performance provided by SOTA FR systems for different demographic groups, utilizing diverse databases to also represent challenging conditions such as pose variations, aging, and presence of occlusions.

In FRCSyn-onGoing, we have designed specific tasks and sub-tasks to address the aforementioned questions. This enables the investigation of using synthetic data to train FR systems, incorporating domain generalization techniques and synthetic-to-real transfer learning as discussed in [11]. Some of the proposed sub-tasks specifically focus on analyzing the benefits provided by the fusion of databases belonging to the real and synthetic domain when training FR systems, a strategy similar to the domain mixup proposed in [33] and further explored in subsequent studies [16] to bridge the gap between synthetic and real face domains. In addition, we have released to the participants two novel synthetic databases created using two SOTA Diffusion methods: DCFace [6] and GANDiffFace [7]. These databases have been generated with a particular focus on tackling common challenges in FR, including imbalanced demographic distributions, pose variation, expression diversity, and the presence of occlusions (see Figure 1). FRCSyn-onGoing offers valuable insights into the future of FR and the use of synthetic data, with a particular focus on quantifying the performance disparity between training FR systems with real and synthetic data. Additionally, FRCSyn-onGoing introduces standardized benchmarks that are readily reproducible for the wider research community.

A preliminary version of this article was previously published in [38]. The present article significantly enhances [38] in the following ways: *i)* offering a

[1] https://frcsyn.github.io/

more detailed overview of the context of FR and synthetic data, including the new Section 2 Related Works to comprehensively discuss the current SOTA, *ii)* providing a more extensive description of the top FR systems presented so far in FRCSyn-onGoing, including key graphical representations of the proposed systems to improve the understanding, *iii)* incorporating additional metrics in the evaluation of the proposed FR systems in order to analyze different operational scenarios, and *iv)* presenting an in-depth analysis of the performance achieved for various demographic groups and databases used for evaluation, accompanied by novel figures and tables.

The reminder of the article is organized as follows. In Section 2, we provide an overview of the limitations of current FR technology in the literature and the current role of synthetic data. In Section 3, we delve into the databases considered in FRCSyn-onGoing. Following that, Section 4 provides an overview of the proposed tasks and sub-tasks, detailing the experimental protocol and metrics employed in the challenge. In Section 5, we provide a comprehensive description of the top-5 FR systems proposed so far in FRCSyn-onGoing for each sub-task. Section 6 presents the results achieved in the different tasks and sub-tasks of the challenge, accompanied by a thorough analysis of the FR system performance across demographic groups and challenging conditions. Finally, in Section 7, we draw the conclusions from FRCSyn-onGoing and highlight potential future research directions in the field.

2. Related Works

2.1. Limitations in Current Face Recognition Technology

The main limitations in current FR technology have been thoroughly explored in extensive surveys [4, 10, 14, 39]. Notably, pose variation emerges as a major challenge, with algorithms experiencing a performance degradation of over 10% when verifying faces from a frontal-profile perspective compared to frontal-frontal verification [40]. In fact, the variability between two images of the same individual in different poses can be greater than between two images of different individuals [14]. In unconstrained scenarios, such as surveillance, faces captured may exhibit large pose variations. Images of the same individual should ideally be captured in various poses at earlier times to facilitate recognition [41, 42]. However, training data typically contain far more frontal faces than profiles. Aging is also considered as another significant challenge for FR systems, given the changes in unique facial characteristics over time. DL methods have been studied to learn age-invariant features and distinguish them from age-related factors in the representation of facial images [43, 44]. According to [45], a significant loss in face recognition accuracy for SOTA FR systems occurs beyond a time lapse of 8.5 years. Facial occlusion presents a challenge, as there is often no prior knowledge available about the obstructed part of the face, whether intentionally or unintentionally obscured by items like hats, sunglasses, hands, scarves, masks, or makeup. A systematic categorization of methods for occluded FR is provided in [13]. The occluded facial part is frequently treated as noise and subtracted from the provided face image, enabling a comparison of the remaining information with the stored images [14]. An interesting approach in this line was presented in [46], where the authors designed a novel GAN for natural de-occlusion, ensuring that resulting faces can retain the attributes of the input faces. Since training data typically fail to represent challenging conditions, generative models have been proposed to synthesize identity-preserving faces with various poses [47, 48, 49], occlusions [50], and aging images [51], with identity preservation providing a significant challenge. Particularly for pose variations, generative framework composed of 3D model and GAN refiners to improve the realism of the generated images have been proposed in [52, 53], featuring identity perception loss to preserve identity information.

In addition to these limitations, FR systems often exhibit biases linked to the demographic attributes of individuals [29, 37]. These biases primarily come from training databases that inadequately represent diverse demographic groups. In popular large-scale databases [54, 55, 56], male, white, and middle-aged individuals are disproportionately over-represented compared to other demographic groups. FR systems trained on such data unintentionally replicate these biases, resulting in significant performance disparities among demographic groups [4, 18]. The magnitude of this issue becomes even more pronounced when examining the intersectionality of certain demographic attributes [57]. Efforts to correct these biases have primarily concentrated on balancing training databases [58]. However, additional disparities may exist among demographic groups [59]. Certain groups may need more extensive data representation than others, and identifying the optimal representation for each demographic group to prevent biases is a challenging task [18].

2.2. Synthetic Data in Face Recognition

Several approaches have been introduced to create synthetic databases for training FR systems. Their applicability has been investigated in [60], to compensate for the lack of publicly available large-scale test databases, and in [19], to provide a taxonomy and further discussion. Several synthetic databases for training have been synthesized using generative frameworks relying on GANs. The advantageous property of linear separability offered by StyleGAN networks [61] has been widely employed to generate databases with desired demographic distributions [7, 62] and obtain multiple images of the same individuals while modifying attributes such as pose, illumination, and expression [63]. Other databases created using alternative generative frameworks based on GANs include *SYNFace* [33], which generates face images by sampling random noise from multiple normal distributions to control different facial attributes, and *SFace* [34], a privacy-friendly database based on StyleGAN2-ADA [64] and identity labels. However, these databases present limitations in terms of intra-class variations in the former and unrealistic mated score distributions in the latter. A large-scale synthetic database, named *DigiFace-1M*, has been recently presented by rendering digital faces through a computer graphics pipeline [16]. Identities in *DigiFace-1M* are defined as unique combinations of facial geometry, texture, eye color, and hair style, while other parameters (*i.e.* pose, expression, environment, and camera distance) are adjusted to render multiple images. Although DigiFace-1M notably diminishes the synthetic-to-real domain gap in training FR systems with synthetic data, it produces images with unrealistic textures compared to real images and lacks an analysis of demographic distributions.

More recently, Diffusion models have emerged for synthesizing more realistic databases for FR, with the first generative frameworks being *DCFace* [6] and *GANDiffFace* [7]. Examples of face images synthesized with both DCFace and GANDiffFace are included in Figure 1. DCFace offers improved intra-class variations compared to previous databases and achieves SOTA performance in training FR systems, surpassing DigiFace-1M. On the other hand, GANDiffFace is specifically designed to target demographic distributions and approximate the similarity score distributions provided by real databases. The synthetic database created using GANDiffFace have proven to be successful in mitigating demographic bias in FR by fine-tuning existing systems [18]. Both DCFace and GANDiffFace databases are used in the proposed FRCSyn-onGoing, and additional details about them are provided in Section 3. For completeness, we would like to highlight also other syn-

Table 1: Details of the databases considered in FRCSyn-onGoing. Id = Identities, Img = Images.

Database	**Framework**	**Use**	**# Id**	**# Img/Id**
DCFace [6]	DCFace	Train	10K	50
GANDiffFace [7]	GANDiffFace	Train	10K	50
CASIA-WebFace [54]	Real-world	Train	10.5K	47
FFHQ [65]	Real-world	Train	70K	1
BUPT-BalancedFace [58]	Real-world	Eval	24K	45
AgeDB [66]	Real-world	Eval	570	29
CFP-FP [40]	Real-world	Eval	500	14
ROF [67]	Real-world	Eval	180	31

thesis approaches recently presented in the literature [17, 35, 36]. One of them, named *IDiff-Face* [35], relies on conditional latent Diffusion models for the synthetic generation of identities with realistic variations. FR systems trained with IDiff-Face achieve a benchmark accuracy of 88.20%, not far from the accuracy of 89.56% provided by DCFace [6]. Inclusive text-to-image models generate images based on human-written prompts and ensure the resulting images are uniformly distributed across attributes of interest have been proposed in [17]. Finally, the stochastic nature of the denoising diffusion process is leveraged in [36] to produce high-quality, identity-preserving face images with various backgrounds, lighting, poses, and expressions.

3. FRCSyn-onGoing: Databases

Table 1 provides the details of the public databases considered in FRCSyn-onGoing. Participants are instructed to download all necessary databases for FRCSyn-onGoing upon registration. Permission for redistributing these databases was obtained from the owners.

3.1. Synthetic Databases

For the training of the proposed FR systems, we provide access to two synthetic databases generated using recent frameworks based on Diffusion models:

- **DCFace** [6]. This framework comprises: *i)* a sampling stage for generating synthetic identities, and *ii)* a mixing stage for generating images with the same identities from the sampling stage and styles selected from a "style bank" of images.

- **GANDiffFace** [7]. This framework combines GANs and Diffusion models to generate fully-synthetic FR databases with desired properties such as human face realism, controllable demographic distributions, and realistic intra-class variations.

Figure 1 provides examples of the synthetic face images created using DCFace and GANDiffFace approaches. These synthetic databases represent a diverse range of demographic groups, including variations in ethnicity, gender, and age. The synthesis process considers typical variations in FR, including pose, facial expression, illumination, and occlusions. In FRCSyn-onGoing, synthetic data are exclusively utilized in the training stage, replicating realistic operational scenarios.

3.2. Real Databases

For the training of FR systems (depending on the sub-task, please see Section 4), participants are allowed to use two real databases: *i)* **CASIA-WebFace** [54], a database containing face images of real identities collected from the web, and *ii)* **FFHQ** [65], a database designed for face applications, containing high-quality face images with considerable variation in terms of age, ethnicity and image background. These real databases are chosen as they are used to train the generative frameworks of DCFace and GANDiffFace, respectively. This strategy enables a direct comparison between the traditional approach of training FR systems using only real data and the novel approach explored in this challenge, using only synthetic data or fusion of both real and synthetic data. Despite not being specifically designed for FR, the FFHQ database can be considered in the proposed challenge for various purposes, such as training a model for feature extraction and applying domain adaptation, among other possibilities.

For the final evaluation of the proposed FR systems, we consider four real databases: *i)* **BUPT-BalancedFace** [58], *ii)* **AgeDB** [66], *iii)* **CFP-FP** [40], and *iv)* **ROF** [67]. BUPT-BalancedFace [58] is designed to address performance disparities across different ethnic groups. We relabel it according to the FairFace classifier [68], which provides labels for ethnicity and gender. We then consider the eight demographic groups obtained from all possible combinations of four ethnic groups (Asian, Black, Indian, and White) and two genders (Female and Male). We recognize that these groups do not comprehensively represent the entire spectrum of real world ethnic diversity. The selection of these categories, while imperfect, is primarily driven by the need to align with the demographic categorizations used in BUPT-BalancedFace [58] for facilitating easier and more consistent evaluation. A list of $8,000$ random comparison pairs is generated from identities in the BUPT-BalancedFace database to evaluate the proposed FR systems, with $1,000$ comparisons equally divided into matching and non-matching pairs representing each of the eight demographic groups considered.

The other three databases, *i.e.,* AgeDB [66], CFP-FP [40], and ROF [67], are real-world databases widely employed to benchmark FR systems in terms of age variations, pose variations, and presence of occlusions. It is important to highlight that, as different real databases are considered for training and evaluation, we also intend to analyse the generalization ability of the proposed FR systems. For AgeDB, we consider all the comparisons outlined in the original evaluation protocol, comprising $6,000$ comparisons for each of the four age intervals considered, *i.e.,* 5, 10, 20, and 30 years. This results in a total of $24,000$ comparison pairs. For CFP-FP, we exclusively consider the frontal-profile comparisons specified in the original evaluation protocol, excluding all frontal-frontal comparisons. This results in a total of $7,000$ comparison pairs. Finally, for ROF, we randomly generate $1,600$ comparisons between individuals without occlusions and wearing a mask, and $2,000$ comparisons between individual without occlusions and wearing sunglasses. This results in a total of $3,600$ comparison pairs. Comparisons for each database are equally divided into matching and non-matching pairs.

4. FRCSyn-onGoing: Setup

FRCSyn-onGoing is hosted on Codalab[2], a robust open-source framework for running scientific competitions and benchmarks. The proposed tasks and sub-tasks, experimental protocol, and metrics are described in the following.

4.1. Tasks

FRCSyn-onGoing aims to explore the application of synthetic data into the training of FR systems, with a specific focus on addressing two critical

[2] https://codalab.lisn.upsaclay.fr/competitions/15485

Table 2: Tasks and sub-tasks proposed in FRCSyn-onGoing with their respective metrics and databases. AVG = Average, SD = Standard Deviation, FNMR = False Non-Match rate, FMR = False Match Rate, AUC = Area Under Curve, GAP = Gap to Real.

Task 1: synthetic data for **demographic bias mitigation**
Baseline: training only with CASIA-WebFace [54] and FFHQ [65];
Metrics: accuracy, FNMR@FMR=1%, AUC, GAP;
Ranking: AVG (across demographic groups) vs SD of accuracy, see Section 4.3 for more details.

Sub-Task 1.1: training exclusively with **synthetic** databases Train: DCFace [6] and GANDiffFace [7]; Eval: BUPT-BalancedFace [58].
Sub-Task 1.2: training with **real and synthetic** databases Train: CASIA-WebFace, FFHQ, DCFace, and GANDiffFace; Eval: BUPT-BalancedFace.

Task 2: synthetic data for **overall performance improvement**
Baseline: training only with CASIA-WebFace and FFHQ;
Metrics: accuracy, FNMR@FMR=1%, AUC, GAP;
Ranking: AVG accuracy (across databases).

Sub-Task 2.1: training exclusively with **synthetic** databases Train: DCFace and GANDiffFace; Eval: BUPT-BalancedFace, AgeDB [66], CFP-FP [40], and ROF [67].
Sub-Task 2.2: training with **real and synthetic** databases Train: CASIA-WebFace, FFHQ, DCFace, and GANDiffFace; Eval: BUPT-BalancedFace, AgeDB, CFP-FP, and ROF.

aspects in current FR technology: *i)* mitigating demographic bias, and *ii)* enhancing overall performance under challenging conditions that include variations in age and pose, the presence of occlusions, and diverse demographic groups. To investigate these two areas, in FRCSyn-onGoing we consider two distinct tasks, each comprising two sub-tasks. Sub-tasks have been designed to consider different approaches for training FR systems: *i)* utilizing solely synthetic data, and *ii)* involving a fusion of real and synthetic data. Consequently, FRCSyn-onGoing comprises a total of four sub-tasks. A summary is provided in Table 2. For each sub-task, we specify the databases allowed for training FR systems. Nevertheless, participants have the flexibility to decide whether and how to utilize each database in the training process.

4.1.1. Task 1

The first proposed task explores the use of synthetic data to address demographic biases in FR systems. To evaluate the proposed systems, we create lists of mated and non-mated comparisons derived from individuals in the BUPT-BalancedFace database [58]. We consider the eight demographic groups described in Section 3, obtained from the combination of four ethnic groups with two genders. For non-mated comparisons, we exclusively focus on pairs of individuals belonging to the same demographic group, as these are more relevant than non-mated comparisons between individuals of different demographic groups.

4.1.2. Task 2

The second proposed task explores the application of synthetic data to enhance overall performance in FR under challenging conditions. To assess the proposed systems, we use lists of mated and non-mated comparisons derived from individuals included in the four databases indicated in Section 3, namely BUPT-BalancedFace [58], AgeDB [66], CFP-FP [40], and ROF [67]. Each database allows the evaluation of specific challenging conditions for FR, including diverse demographic groups, aging, pose variations, and presence of occlusions.

4.2. Experimental protocol

4.2.1. Training

The four sub-tasks proposed in FRCSyn-onGoing are mutually independent. This means that participants have the freedom to participate in any number of sub-tasks of their choice. For each selected sub-task, participants are expected to propose a FR system and train it twice: *i)* using authorized real databases only, *i.e.,* CASIA-WebFace [54] and FFHQ [65], and *ii)* in accordance with the specific requirements of the chosen sub-task, as summarized in Table 2. According to this protocol, participants provide both the *baseline system* and the *proposed system* for the specific sub-task. The baseline system plays a critical role in evaluating the impact of synthetic data on training and serves as a reference point for comparing against the conventional practice of training solely with real databases. To maintain consistency, the baseline FR system, trained exclusively with real data, and the proposed FR system, trained according to the specifications of the selected sub-task, must have the same architecture.

4.2.2. Evaluation

In each sub-task, participants are provided with comparison files containing both mated and non-mated comparisons, which are used to evaluate the performance of their proposed FR system. In Task 1 there is a single comparison file containing balanced comparisons of different demographic groups, while in Task 2 there are four comparison files, one for each real database considered. The evaluation process occurs twice for each sub-task to assess: *i)* the baseline system trained exclusively with real databases, and *ii)* the proposed system trained in accordance with the sub-task specifications. For the evaluation of each sub-task, participants must submit through Codalab platform two files per database (one for the baseline and one for the proposed system), including the score and the binary decision (mated/non-mated) for each comparison listed in the comparison files. The organizers retain the right to disqualify participants to uphold the integrity of the evaluation process if anomalous results are detected or if participants fail to adhere to the challenge's rules.

4.2.3. Restrictions

Participants have the freedom to choose the FR system for each task, provided that the system's number of Floating Point Operations Per Second (FLOPs) does not exceed 25 GFLOPs. This threshold has been established to facilitate the exploration of innovative architectures and encourage the use of diverse models while preventing the dominance of excessively large models. Participants are also free to utilize their preferred training modality, with the requirement that only the specified databases are used for training. This means that no additional databases can be employed during the training phase, such as to establish verification thresholds. Generative models cannot be utilized to generate supplementary data. Participants are allowed to use non-face databases for pre-training purposes and employ traditional data augmentation techniques using the authorized training databases.

4.3. Metrics

We evaluate FR systems using a protocol based on lists of mated and non-mated comparisons for each sub-task and database. From the binary decisions provided by participants, we calculate verification accuracy. This approach is straightforward and allows participants to choose the preferred threshold for their systems. From the scores provided by participants, we can compute other interesting metrics. Specifically, we calculate False Non-Match

Rate (FNMR) at a fixed False Match Rate (FMR) of 1% (FNMR@FMR=1%) and Area Under Curve (AUC). These metrics are widely used for the analysis of FR systems in real-world applications. In the Face Recognition Technology Evaluation (FRTE) 1:1 Verification by NIST [69], FR algorithms are ranked based on FNMR@FMR=10^{-4}%. This metric involves a substantial number of comparisons to provide a statistically significant value of FMR=10^{-4}%, with FR algorithms tested against multiple face images of more than 8 million people. In the context of FRCSyn-onGoing, we consider a number of comparisons for each demographic group and database in the order of 10^3. This approach enables participating teams with less resources to carry out a streamlined evaluation process by facilitating the download of selected public databases for evaluation and executing a significantly reduced number of comparisons. Consequently, we consider a fixed operational point at FMR=1% to calculate statistically significant metrics. We calculate accuracy, FNMR@FMR=1%, and AUC for each of the eight demographic groups defined in Section 3 in Sub-Tasks 1.1 and 1.2, as well as for each of the four evaluation databases described in Section 3 in Sub-Tasks 2.1 and 2.2. Furthermore, all these metrics are averaged across demographic groups or databases, respectively, to provide summarized metrics for each participating team.

Additionally, we calculate the gap to real (GAP) metric as follows: $\text{GAP} = (\text{REAL} - \text{SYN})/\text{SYN}$, where REAL represents a metric computed on the baseline system, and SYN represents the same metric computed on the proposed system trained with synthetic (or real + synthetic) data. The GAP metric, introduced in [6], quantifies the difference in verification accuracy between a FR system trained with synthetic and real data. In this study, we extend the calculation of the GAP to metrics beyond accuracy while maintaining the same underlying concept. In the following, we explain how participants are ranked in the different tasks.

4.3.1. Task 1

To rank participants and determine the winners of Sub-Tasks 1.1 and 1.2, we closely examine the trade-off between the average (AVG) and standard deviation (SD) of the verification accuracy across the eight demographic groups defined in Section 3. We define the trade-off metric (TO) as follows: $\text{TO} = \text{AVG} - \text{SD}$. This metric corresponds to plotting the average accuracy on the x-axis and the standard deviation on the y-axis in 2D space. We draw multiple 45-degree parallel lines to find the winning team whose performance

Table 3: Description of the top-5 best teams ordered by the affiliation letters. The letters reported in the column 'affiliations' refer to the ones provided in the title page. For each team, we report the ranking metric across all the sub-tasks. The top-3 results of each sub-task are remarked in bold. TO = Trade-off, AVG = average accuracy.

Team	**Affiliations**	**Country**	**Task 1.1** **TO**	**Task 1.2** **TO**	**Task 2.1** **AVG**	**Task 2.2** **AVG**
CBSR	d, e, f, g, h	China	-	**95.25 (1)**	-	**94.95 (1)**
LENS	i	USA	**92.25 (1)**	**95.24 (2)**	**88.18 (2)**	**92.40 (2)**
BOVIFOCR-UFPR	j, k, l	Brazil	**90.51 (3)**	93.15 (4)	**90.50 (1)**	91.34 (4)
Idiap	m, n, o	Switzerland	**91.88 (2)**	87.22 (6)	**86.39 (3)**	**91.74 (3)**
MeVer	p, q	Greece	87.51 (4)	**93.97 (3)**	83.45 (5)	87.50 (5)
BioLab	r	Italy	-	-	83.93 (4)	-
Aphi	s	Spain	82.24 (5)	-	80.53 (6)	-
UNICA-FRAUNHOFER IGD	t, u, v	Italy, Germany	-	91.03 (5)	-	84.86 (6)

falls to the far right side of these lines. With this proposed metric, we reward FR systems that achieve good levels of performance and fairness simultaneously, unlike common benchmarks based only on recognition performance. The standard deviation of verification accuracy across demographic groups is a common metric for assessing bias and should be reported by any work addressing demographic bias mitigation.

4.3.2. Task 2

To rank participants and determine the winners of Sub-Tasks 2.1 and 2.2, we consider the average verification accuracy across the four databases used for evaluation, described in Section 3. This approach allows us to evaluate simultaneously four challenging aspects of FR systems: *i)* pose variations, *ii)* aging, *iii)* presence of occlusions, and *iv)* diverse demographic groups, providing a comprehensive evaluation of FR systems in real operational scenarios.

5. FRCSyn-onGoing: Description of Systems

FRCSyn-onGoing has received so far significant interest, with 67 international teams correctly registered, comprising research groups from both industry and academia. These teams work in various domains, including FR, generative AI, and other aspects of computer vision, such as demographic fairness and domain adaptation. Until now, we have received submissions from 15 teams, receiving all sub-tasks high attention. The submitting teams are geographically distributed, with six teams from Europe, five

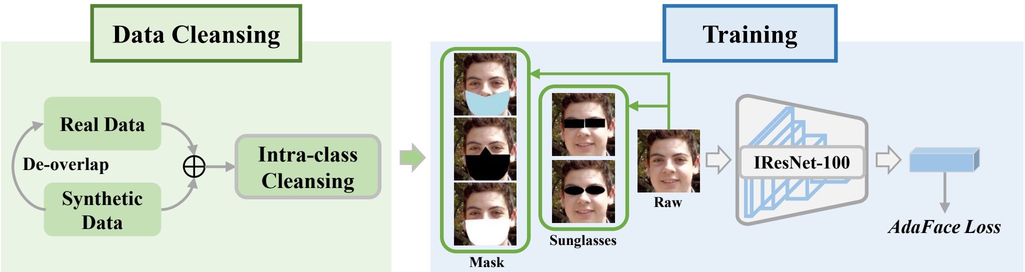


Figure 2: Architecture proposed by the CBSR team.

teams from Asia, and four teams from America. Table 3 provides a comprehensive overview of the top-5 best teams for each sub-task, showcasing their performance across all the sub-tasks in which they participated. Next, we provide a description of the FR systems proposed by each of these teams.

5.1. CBSR

This team comprises members of the IIE, CAS; School of Cyber Security, UCAS; MAIS, CASIA; School of Artificial Intelligence, UCAS; and CAIR, HKISI, CAS. They participated in Sub-Tasks 1.2 and 2.2. The proposed architecture is described in Figure 2. They first trained a FR system using CASIA-WebFace [54]. They extracted features for images in FFHQ [65] and clustered them using the DBSCAN [70] for pseudo labels since the FFHQ is unlabeled. Then, they removed the samples in FFHQ that are similar to CASIA-WebFace with a cosine similarity of 0.6 and merged the two as the training database to train a new recognition model F. Subsequently, they utilized F to extract the features for DCFace [6] and GANDiffFace [7], and de-overlapped the images that are similar to CASIA-WebFace and FFHQ using a similarity threshold of 0.6. They conducted the intra-class clustering for the training database using DBSCAN with a similarity threshold of 0.3 and removed the samples that were separate from the class center. Next, they merged all cleansed data and trained IResNet-100 with AdaFace loss [3]. For data augmentation, they adopted mask occlusion augmentation via the methods introduced in [71], consisting of surgical-style and N95-style masks, with colors blue, black and white. In addition, they also added sunglasses via detected face landmarks. Note that the face landmarks were detected via FaceX-Zoo [72]. Also, they used random flipping with a probability of 50% on the images. They trained two recognition models by adding occlusion augmentation with 10% and 30% probability, respectively. They finally used the average similarity prediction of the two models as the final prediction and verified the pairs in the test set with the 10-fold optimal threshold determined in the validation set.

They constructed different validation sets for different evaluation tasks. For AgeDB [66], they randomly sampled image pairs from the training data since the training databases consist of facial images with plenty of age variations. For CFP-FP [40], they added randomly positioned vertical bar occlusions to the images to simulate the self-occlusion problem due to pose. For ROF [67], they detected face keypoints by FaceX-Zoo [72] and added mask occlusions to images as in [71]. Also, they filled the eye regions with rectangular and elliptical occlusions to simulate an image of a face with sunglasses. For BUPT-BalancedFace [58], they randomly sampled image pairs from DCFace with GANDiffFace because they have balanced demographic groups. All validation sets consisted of $12,000$ image pairs containing $6,000$ positive pairs and $6,000$ negative pairs. Code available[3].

5.2. LENS

This team comprises members of LENS, Inc. They participated in all the proposed sub-tasks. The proposed architecture is described in Figure 3. Keeping in mind the challenges of all the sub-tasks and the databases that can be used for training, they adopted the architecture of ResNet-50 [73] (R50) backbone for all the sub-tasks, due to less number of parameters and suitability when the size of the databases is limited. For sub-tasks using only synthetic data, they observed that since the test data are real databases, they needed an architecture that increased the robustness to domain shifts between synthetic training data and real test data. To this end, they incorporated various augmentation techniques and the AdaFace loss [3]. Augmentation techniques included cropping, rescaling, and photometric jittering (each selected with a probability of 0.2). Database augmentation aided in bringing synthetic images closer to the real image distribution. This *i)* reduced the effect of synthetic noises, *ii)* reduced the domain gap between synthetic training data and the real test data, and *iii)* significantly improved recognition rates. They further improved the performance by using a fusion of two models with the same R50 architecture. The second model was trained with a different style of augmenting databases, inspired by [35]. For

[3] https://github.com/zws98/wacv_frcsyn

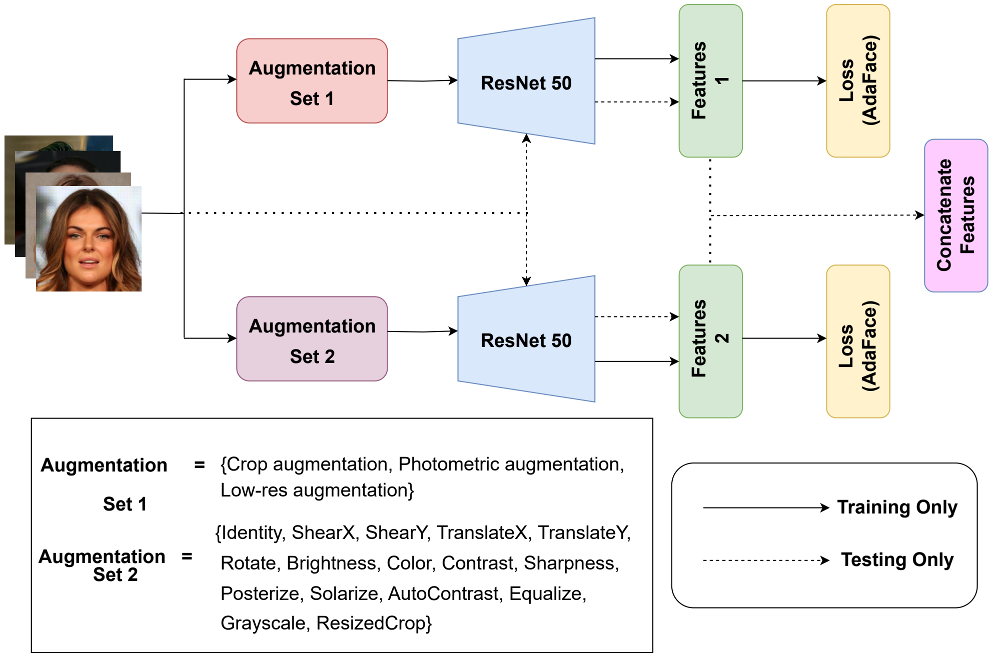

Augmentation Set 1
ResNet 50
Features 1
Loss (AdaFace)
Concatenate Features
Augmentation Set 2
ResNet 50
Features 2
Loss (AdaFace)
Augmentation Set 1 = {Crop augmentation, Photometric augmentation, Low-res augmentation}
Augmentation Set 2 = {Identity, ShearX, ShearY, TranslateX, TranslateY, Rotate, Brightness, Color, Contrast, Sharpness, Posterize, Solarize, AutoContrast, Equalize, Grayscale, ResizedCrop}
Training Only
Testing Only

Figure 3: Architecture proposed by the LENS team.

each image, they chose four random augmentations from the following set: Identity, ShearX, ShearY, TranslateX, TranslateY, Rotate, Brightness, Color, Contrast, Sharpness, Posterize, Solarize, AutoContrast, Equalize, Grayscale, ResizedCrop. The experiments conducted in [35, 33, 16] evaluated the impact of data augmentation on the performance of their FR model. The features of the two models were then combined to create a feature set length of $1,024$. In addition, incorporating AdaFace loss helped create robust embeddings. The same method was repeated for Sub-Tasks 1.2 and 2.2.

All the databases were first cropped and aligned using the landmarks detected by Retinaface [74], resulting in a size of 112×112. For training, they divided their total data (respective of sub-tasks) in the ratio $80 : 20$ where 80% of the data was a training set and the rest was validation. For training the baseline model and Sub-Tasks 1.2 and 2.2, they utilized CASIA-WebFace [54] for the real database and skipped FFHQ [65]. They adopted the training hyperparameters of [3] with $lr = 0.1$ and trained for 30 epochs from scratch. The AdaFace loss function [3] approximates image quality using feature norms and assigns different importance to easy or hard samples based on their image quality. This adaptive margin function enhanced

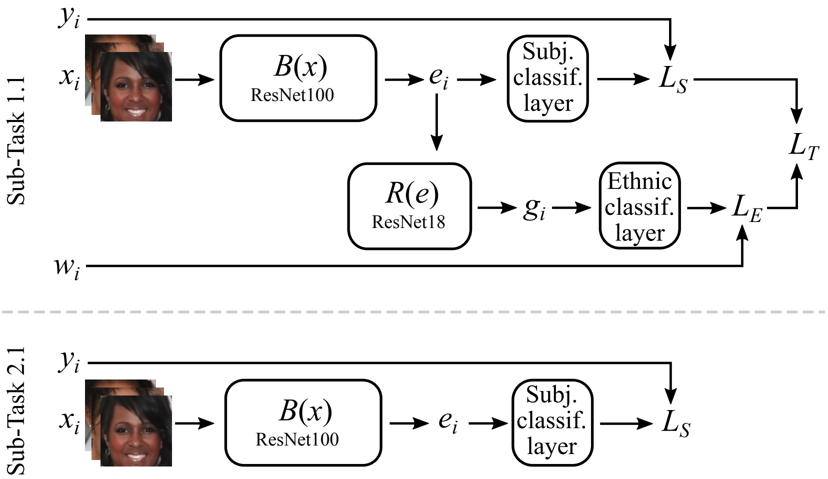


Figure 4: Architecture proposed by the BOVIFOCR-UFPR team.

the discriminability of learned features and achieved SOTA performance on multiple FR databases.

5.3. BOVIFOCR-UFPR

This team comprises members of the Federal University of Paraná, Federal Institute of Mato Grosso, and unico - idTech. They participated in all the proposed sub-tasks and provided a description of the systems proposed for Sub-Tasks 1.1 and 2.1, in which they ranked in top-3. The proposed architecture is described in Figure 4. To reduce demographic bias, in Sub-Task 1.1, they proposed to enforce a FR model to increase similarities between people from the same ethnic group while learning to discriminate between different subjects. Inspired by Zhang *et al.* [75], they created a multi-task collaborative model composed of two backbones $B(x)$ and $R(e)$, which produced the embeddings $e \in R^{512}$ and $g \in R^{256}$, respectively, containing the subject and its ethnic group features. This schema is shown on the top of Figure 4 and forces the main backbone $B(x)$ to learn less biased features. ResNet100 and ResNet18 [73] architectures were used as $B(x)$ and $R(e)$, respectively. Training databases were organized as $X = x_i, y_i, w_i$, where x_i is the input face image, y_i is the subject label used to compute the subject classification loss L_S [2] and w_i is the ethnic group label used to compute the ethnic group classification loss L_E [2]. The total loss L_T was computed as $L_T = \lambda_S L_S + \lambda_E L_E$. Experiments using their strategy on the synthetic

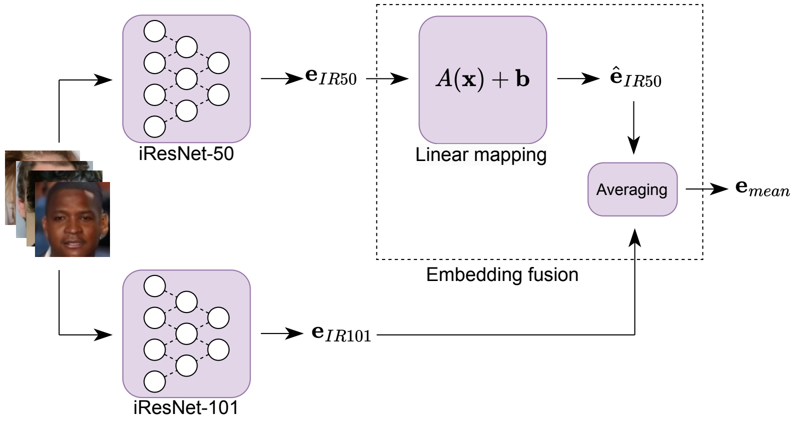


Figure 5: Architecture proposed by the Idiap team.

databases DCFace [6] and GANDiffFace [7] increased the average verification accuracy in the database BUPT-BalancedFace [58] while reducing the standard deviation between demographic groups.

For Sub-Task 2.1, they normalized and preprocessed the images by cropping and aligning the database images using Retina Face [74]. Then, they employed ArcFace [2] as their loss function and ResNet-100, which is one of the top-performing methods for deep FR [76]. They trained the network using the InsightFace library for 26 epochs. All images from the training set were augmented using Random Flip with a probability of 0.5. For this task, they used DCface as the training set, which has $10,000$ identities and $550,000$ images, and was the database that provided the most accurate feature vectors on the validation set. The validation consisted of a training database subsample, with genuine and impostor pairs. Using the validation set, they selected the best threshold to classify the output scores for the validation set.

5.4. Idiap

This team comprises members of the Idiap Research Institute, École Polytechnique Fédérale de Lausanne, and Université de Lausanne. They participated in all the proposed sub-tasks. The proposed architecture is described in Figure 5. For all tasks and sub-tasks, the main architecture chosen is the fusion of features of two models, as the ensemble of models can lead to improved accuracy and bias mitigation. In this case, the ensemble was composed of two models, based on the iResNet-50 and iResNet-101 architectures [2], which were used jointly with a linear mapping [77]. The linear mapping was performed on the embedding $\mathbf{e}_{IR50}$ from the iResNet-50 model, arbitrarily selected, followed by a feature fusion approach [78] by embedding averaging, to compute a mean feature vector. Both models underwent slightly different training processes to allow for differences to emerge and improve the feature fusion. Preprocessing was performed for training, validation, and testing as follows [79]: the face landmarks were detected using RetinaFace [74] for all the evaluation sets. Then, five facial points (both eyes, nose tip, and both mouth corners) were used to compute an alignment with a predefined template. The images were cropped and resized to 112×112 afterward. Each pixel was normalized in the range $[-1, 1]$. Additionally, specific to the training step, additional data augmentation was performed. It involved randomized cropping, resolution augmentation, and adjustments to brightness, contrast, and saturation.

The models were trained on all permitted task-specific databases: DCFace [6] and GANDiffFace [7] for the synthetic tracks and CASIA-WebFace [54], DCFace, and GANDiffFace for the mixed tracks. For the iResNet-101, the models were trained using the CosFace loss function [79], whereas for the iResNet-50, the AdaFace [3] loss function was used. The training was performed for around $60,000$ batches, of size 256, with MultiStepLR and learning-rate 0.1, with a reduction factor of 10 at $24,000$, $40,000$, and $48,000$ steps. The checkpoint selected was the last checkpoint after the training of their model reached the maximum number of steps. No other database than those detailed above could be used, so the entirety of the databases (corresponding to each sub-task) was dedicated to training in order to maximize the training set. The threshold was determined on a split of DCFace for the synthetic track, or CASIA-WebFace for the mixed track. The split involved 150 identities chosen at random and with approximately $10,000$ genuine and $10,000$ zero-effort impostors comparisons thereof. The threshold was set such as to maximize the verification accuracy in a 10-fold cross-validation setup from those selected comparisons.

Regarding the linear mapping, it was composed of a linear layer, with input and output dimensions set to the dimension of $\mathbf{e}_{IR50}$ and $\mathbf{e}_{IR101}$ respectively. The layer was independently trained using FFHQ with both models trained, with $\mathbf{e}_{IR101}$ as labels and $\mathbf{e}_{IR50}$ as input. Notably, no identity labels are required for training the linear layer. The loss function was set to be the mean cosine distance between $\hat{\mathbf{e}}_{50}$, the output of the linear layer, and $\mathbf{e}_{101}$. In effect, this linear layer allows for an estimated projection of the embedding from the iResNet-50 embedding space into the iResNet-101 embedding

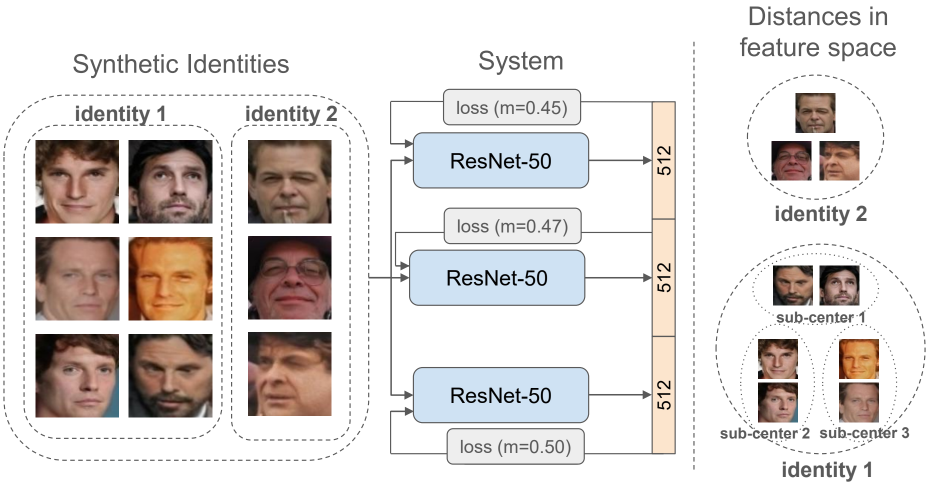


Figure 6: Architecture proposed by the MeVer team.

space, allowing both embeddings to be evaluated in a common embedding space. The average of these embeddings $\mathbf{e}_{mean}$ provides for a better common estimate of an ideal embedding.

5.5. MeVer

This team comprises members of the Centre for Research and Technology Hellas and the Harokopio University of Athens. They participated in all the proposed sub-tasks. The proposed architecture is described in Figure 6. The MeVer team utilized the sub-center ArcFace [80] loss as a pivotal methodology to mitigate the impact of label noise that often arises in large-scale databases [81]. Specifically, the methodology considers K sub-centers for each identity, allowing the training samples to closely align with any K positive sub-center, rather than exclusively with a single positive center. This approach encourages the dominance of one primary sub-class housing the majority of clean faces, alongside non-dominant sub-classes that contain noisier or more challenging facial data. In scenarios involving synthetic data, errors in generative models can cause some of the generated images to be different from each other, even though they should be similar. Using a less strict form of margin-based losses, like the sub-center ArcFace, can help address the problem by allowing the model to create clusters of similar identities for each synthetic identity without being penalized. Furthermore, the proposed system includes three CNNs, using different margins in the ArcFace loss, aligning with the relevant literature [82, 58] highlighting that distinct demographic groups exhibit varying margin requisites for effective and fair FR systems. In particular, it consists of three ResNet-50 [73] models (19.05 GFLOPs in total), each trained separately with 4, 5, and 5 sub-centers K per identity and margins m equal to 0.45, 0.47, and 0.50, respectively. These hyperparameters were tuned through a grid search on $K = \{3, 4, 5, 6\}$ and $m = \{0.40, 0.43, 0.45, 0.47, 0.50, 0.52\}$. Notably, relevant research [82] also suggests margin values less than 0.5 for specific demographic groups.

The final embeddings were derived by concatenating the three backbones' outputs and the predictions were made by comparing the Euclidean distance between the feature vectors with thresholds 1.5 and 1.35 for the tasks considering synthetic-only and mixed synthetic-real training data, respectively. During training, a batch size of 256 was employed. The initial learning rate was 0.1 and decayed by a factor of 10 at training steps 75k, 127.5k, and 165k, while the total training steps were 180k. Furthermore, the stochastic gradient descent (SGD) optimizer, with 0.9 momentum and 0.0005 weight decay was employed. Concerning data preprocessing, face crops (112×112) were derived from MTCNN [83] predictions, and color jittering and random horizontal flip augmentations were applied. Both synthetic databases were used for all tasks, while additionally the CASIA-WebFace [54] database was considered for Sub-Tasks 1.2 and 2.2. 800 identities from the synthetic databases and 1000 from the CASIA-WebFace were used for validation for the sub-tasks involving synthetic-only and mixed synthetic-real databases, respectively. The experiments were conducted on two RTX 3090-ti GPUs using the MXNet framework. Code available[4].

5.6. BioLab

This team comprises members of the University of Bologna. They participated in Sub-Task 2.1. The proposed architecture is described in Figure 7. The model selected for the Sub-Task 2.1 is a ResNet-101 [73] customized as indicated in [2], which has been trained using the margin-based AdaFace loss [3]. One notable advantage of this loss is its resilience when training data contains low-quality images with unrecognizable faces. According to their assumption, this feature ensured that the model's performance remained unaffected when exposed to GAN-related visual glitches and artifacts that usually

[4] https://github.com/gsarridis/fair-face-verification-with-synthetic-data

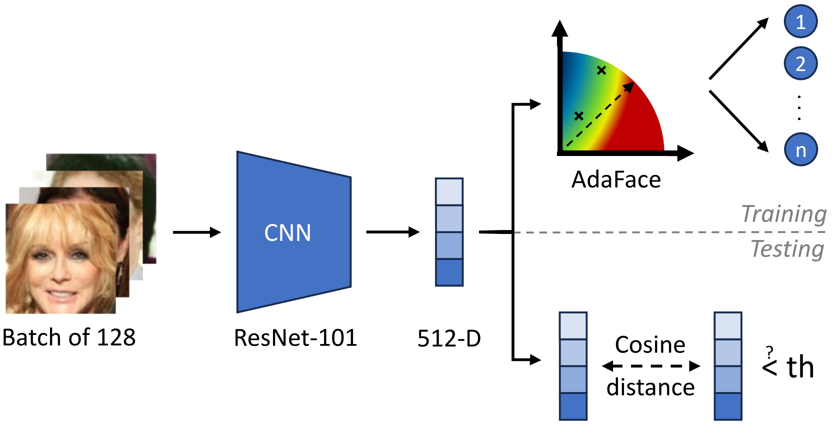


Figure 7: Architecture proposed by the BioLab team.

affect classifier performance [84]. Their baseline model was trained employing the CASIA-WebFace database [54]; the proposed model employed both DCFace [6] and GANDiffFace [7]. They built the validation set by generating matching and non-matching couples from the first 100 classes of CASIA-WebFace, or the first 4 of each ethnicity/gender combination in DCFace and GANDiffFace. The selected classes were excluded from training.

They applied data augmentation on the training set. Following the findings in [3], the resulting pipeline consisted of random horizontal flips, random crop-and-resize, and random color jittering on the saturation and value channels. Each transformation had a probability of 20% of being applied. The model was optimized with SGD using cross entropy loss with batch size of 128. The initial learning rate of 0.05 was divided by a factor of 10 at prefixed epochs to ensure better training stability. For face verification, the dissimilarity between the embeddings was measured employing the cosine distance. Its threshold was computed to maximize the mean accuracy on 10 separate folds of the validation set (*i.e.,* using a non-overlapping partition of the training databases), following the same idea described in the LFW protocol [5]. Code available[5].

5.7. Aphi

This team comprises members of Facephi. They participated in Sub-Tasks 1.1 and 2.1. The proposed architecture is described in Figure 8. In

[5] https://github.com/ndido98/frcsyn

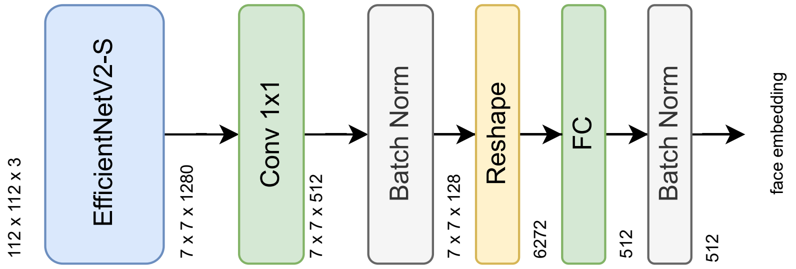


Figure 8: Architecture proposed by the Aphi team.

their approach, they used an EfficientNetV2-S [85] architecture to produce a 512-D deep embedding trained with ArcFace [2] loss function. They modified the backbone network by reducing the first layer's stride from 2 to 1 to enhance the preservation of spatial features. The output of the backbone network was projected with a 1 × 1 convolutional layer and normalized with batch normalization. These features were flattened and fed into a fully connected layer which produces the deep embedding. The weights of the model were optimized through the SGD algorithm with a momentum of 0.9 and a weight decay of $1e^{-4}$ during 20 epochs and a learning rate starting at 0.1 and decayed through a polynomial scheduler. The model was trained with the images aligned using a proprietary algorithm, resized to 112 × 112, and normalized in the range of -1 to 1. To prevent overfitting, they applied data augmentation techniques during training, including Gaussian Blur, Random Scale, Hue-Saturation adjustments, and Horizontal Flip transformations as well as dropout with a rate of 0.2 before the deep embedding projection. To train the baseline model, they made use of CASIA-WebFace [54] and for their proposed model, they employed the synthetic database DCFace [6].

5.8. UNICA-FRAUNHOFER IGD

This team comprises members of the University of Cagliari, Fraunhofer IGD, and TU Darmstadt. They participated in Sub-Task 1.2 and 2.2. The proposed architecture is described in Figure 9. The presented solution utilized ResNet100 [73] as network architecture as it is one of the most widely used architectures in SOTA FR approaches [86]. The training and validation images were aligned and cropped to 112 × 112 using five landmark points extracted with MTCNN. The outputs of the network were 512-D feature representations.

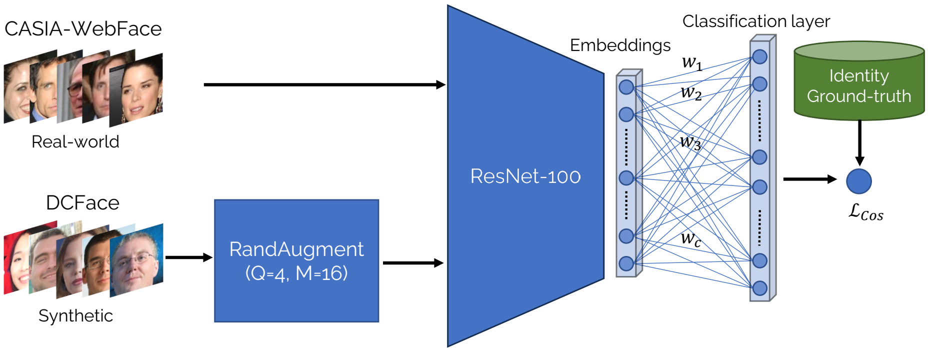


Figure 9: Architecture proposed by the UNICA-FRAUNHOFER IGD team.

The presented solution was based on training the ResNet100 network [73] with a margin-penalty softmax loss. Specifically, the presented solution used CosFace as a loss function with a margin penalty value of 0.35, and a scale parameter of 64 [79]. The model was trained for 40 epochs with a batch size of 512 and an initial learning rate of 0.1. The learning rate was divided by 10 after 10, 22, 30 and 40 training iterations. During the training phase the training databases, CASIA-Webface [54] and DCFace [6], provided by the competition organizers, were merged into one database with a total number of identities equal to 20.572. During the training phase, an extensive set of data augmentation operations based on RandAugment [87, 88] was applied only to the synthetic samples. The real samples were only augmented with horizontal flipping. Code available[6].

6. FRCSyn-onGoing: Results

In Table 4, we present the current rankings for the four different sub-tasks considered in FRCSyn-onGoing, determined according to the criteria outlined in Section 4.3. The metrics reported for accuracy (named AVG), FNMR@FMR=1%, and AUC represent the average metrics calculated across the eight demographic groups (for Sub-Tasks 1.1 and 1.2) and the four databases (for Sub-Tasks 2.1 and 2.2). For completeness, in Table 4, we also provide alternative rankings based on FNMR@FMR=1% and AUC, enclosed in brackets in the respective columns. SD is the standard deviation of

[6] https://github.com/atzoriandrea/FRCSyn

Table 4: Ranking for the four sub-tasks proposed in FRCSyn-onGoing. GAP quantifies the difference between AVG of the baseline and proposed systems. For each sub-task, we highlight in bold the best team according to the ranking metric (*i.e.,* TO for Sub-Tasks 1.1 and 1.2, AVG for Sub-Tasks 2.1 and 2.2). For completeness, we also highlight in bold the best results achieved according to the other metrics. TO = Trade-Off, AVG = Average accuracy, SD = Standard Deviation of accuracy, FNMR = False Non-Match Rate, FMR = False Match Rate, AUC = Area Under Curve, GAP = Gap to Real.

Sub-Task 1.1 (Bias Mitigation): Synthetic Data

Pos.	**Team**	**TO [%]**	**AVG [%]**	**SD [%]**	**FNMR@ FMR=1%**	**AUC [%]**	**GAP [%]**
1	**LENS**	**92.25**	**93.54**	**1.28**	**15.25** (2)	**98.01** (2)	**-0.74**
2	Idiap	91.88	93.41	1.53	**13.97 (1)**	**98.30 (1)**	-3.80
3	BOVIFOCR	90.51	92.35	1.84	16.35 (3)	97.98 (3)	4.23
4	MeVer	87.51	89.62	2.11	32.57 (5)	96.06 (5)	5.68
5	Aphi	82.24	86.01	3.77	23.80 (4)	97.06 (4)	0.84

Sub-Task 1.2 (Bias Mitigation): Synthetic + Real Data

Pos.	**Team**	**TO [%]**	**AVG [%]**	**SD [%]**	**FNMR@ FMR=1%**	**AUC [%]**	**GAP [%]**
1	**CBSR**	**95.25**	**96.45**	**1.20**	**8.68** (4)	**99.33** (3)	**-2.10**
2	LENS	95.24	96.35	1.11	6.35 (2)	**99.38 (1)**	-5.67
3	MeVer	93.87	95.44	1.56	9.50 (5)	99.00 (5)	-0.78
4	BOVIFOCR	93.15	95.04	1.89	10.00 (6)	99.14 (4)	1.28
5	UNICA	91.03	94.06	3.03	6.85 (3)	99.36 (2)	-10.62
6	Idiap	87.22	91.54	4.32	**5.50 (1)**	99.33 (3)	-0.65

Sub-Task 2.1 (Overall Improvement): Synthetic Data

Pos.	**Team**	**AVG [%]**	**FNMR@ FMR=1%**	**AUC [%]**	**GAP [%]**
1	**BOVIFOCR**	**90.50**	**20.83 (1)**	**96.04 (1)**	**2.66**
2	LENS	88.18	33.25 (3)	93.55 (3)	3.75
3	Idiap	86.39	30.73 (2)	93.96 (2)	6.39
4	BioLab	83.93	49.51 (5)	91.78 (4)	6.88
5	MeVer	83.45	50.05 (6)	91.47 (5)	3.20
6	Aphi	80.53	46.09 (4)	88.14 (6)	9.12

Sub-Task 2.2 (Overall Improvement): Synthetic + Real Data

Pos.	**Team**	**AVG [%]**	**FNMR@ FMR=1%**	**AUC [%]**	**GAP [%]**
1	**CBSR**	**94.95**	**10.82 (1)**	**97.92 (1)**	**-3.69**
2	LENS	92.40	17.67 (4)	96.58 (5)	-1.63
3	Idiap	91.74	23.27 (5)	96.87 (4)	0.00
4	BOVIFOCR	91.34	16.51 (2)	97.03 (3)	1.77
5	MeVer	87.60	17.10 (3)	97.40 (2)	-1.57
6	UNICA	84.86	39.35 (6)	91.46 (6)	-27.43

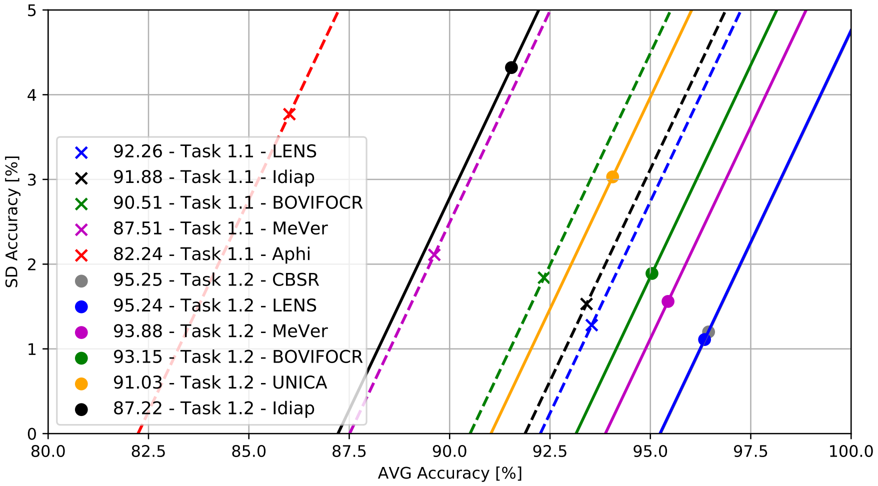


Figure 10: Graphical representation of the trade-off metric (TO) between average accuracy (AVG) and standard deviation (SD) across the eight demographic groups, calculated for the top-5 teams in both Sub-Tasks 1.1 and 1.2.

accuracy calculated across the eight demographic groups, and GAP quantifies the difference between AVG of the baseline and proposed systems.

In general, the rankings for Sub-Tasks 1.1 and 1.2 (bias mitigation), corresponding to the descending order of TO, closely align with the ascending order of SD (*i.e.,* from less to more biased FR systems). In Figure 10, we visually represent the trade-off between average (AVG) and standard deviation (SD) of the accuracy obtained for the eight demographic groups in both Sub-Tasks 1.1 and 1.2. The trend observed in Figure 10 suggests that a higher accuracy usually comes with lower standard deviation. The top-ranked FR systems are predominantly located in the lower right corner of the graph. Unlike accuracy, which depends on the threshold selected by each team, FNMR@FMR=1% measures the performance of FR systems at a fixed operational point that remains unchanged across teams. Additionally, AUC measures the performance of FR systems across all possible thresholds, offering a comprehensive evaluation of system performance. For completeness, we also analyze next the alternative rankings considering these popular metrics. Notably, Idiap is the team that achieves the best FNMR@FMR=1% in Sub-Tasks 1.1 and 1.2 (13.97% and 5.50%, respectively), along with the highest AUC in Sub-Task 1.1 (98.30%). This suggests that their proposed systems achieve superior performance at thresholds different from the ones selected. In Sub-Task 1.2, LENS achieves the best AUC (99.38%), but all the top six teams demonstrate similar AUCs, ranging from 99% to 99.38%.

In Sub-Task 1.1, the top two classified teams, LENS (92.25% TO) and Idiap (91.88% TO), exhibit negative GAP values (-0.74% and -3.80%, respectively), indicating higher accuracy when training the FR system with synthetic data compared to real data. These results highlight the potential of DCFace [6] and GANDiffFace [7] synthetic data to reduce bias in current FR technology. As shown in Figure 10, adding real data to the training process (*i.e.,* Sub-Task 1.2) generally causes the AVG and SD to increase and decrease respectively simultaneously. The CBSR team is the winner with a 95.25% TO (*i.e.,* 3% TO general improvement between Sub-Tasks 1.1 and 1.2). In addition, and as it happens in Sub-Task 1.1, we can observe in Sub-Task 1.2 negative GAP values for the top teams (*e.g.,* -2.10% and -5.67% for the CBSR and LENS teams, respectively), evidencing that the combination of synthetic and real data (proposed system) outperforms FR systems trained only with real data (baseline system).

For Task 2, it is evident that the average accuracy across databases in Sub-Tasks 2.1 and 2.2 is lower than the accuracy achieved for BUPT-BalancedFace [58] in Sub-Tasks 1.1 and 1.2, emphasizing the additional challenges introduced by the other real databases considered for evaluation. Also, although good results are achieved in Sub-Task 2.1 when training only with synthetic data (90.50% AVG for BOVIFOCR-UFPR), the positive GAP values provided by the top teams indicate that synthetic data alone currently struggles to completely replace real data for training FR systems in challenging conditions. Nevertheless, the negative GAP values provided by the top-2 teams in Sub-Task 2.2 (-3.69% and -1.63%, respectively) also suggest that synthetic data combining with real data can mitigate existing limitations within FR technology. Unlike Task 1, for both Sub-Tasks 2.1 and 2.2, the winning teams (*i.e.,* BOVIFOCR-UFPR and CBSR, respectively) are also the ones that provide the best FNMR@FMR=1% (20.83% and 10.82%, respectively) and AUC (96.04% and 97.92%, respectively). This suggests that their proposed systems comprehensively obtain the best performance in the overall improvement of FR under challenging conditions.

Finally, analyzing the description of the FR approaches proposed by the eight top teams, a notable trend emerges, showing the prevalence of well-established methodologies. ResNet backbones [73] were chosen by seven teams, except for Aphi, which opted for EfficientNet [85]. The AdaFace [3]

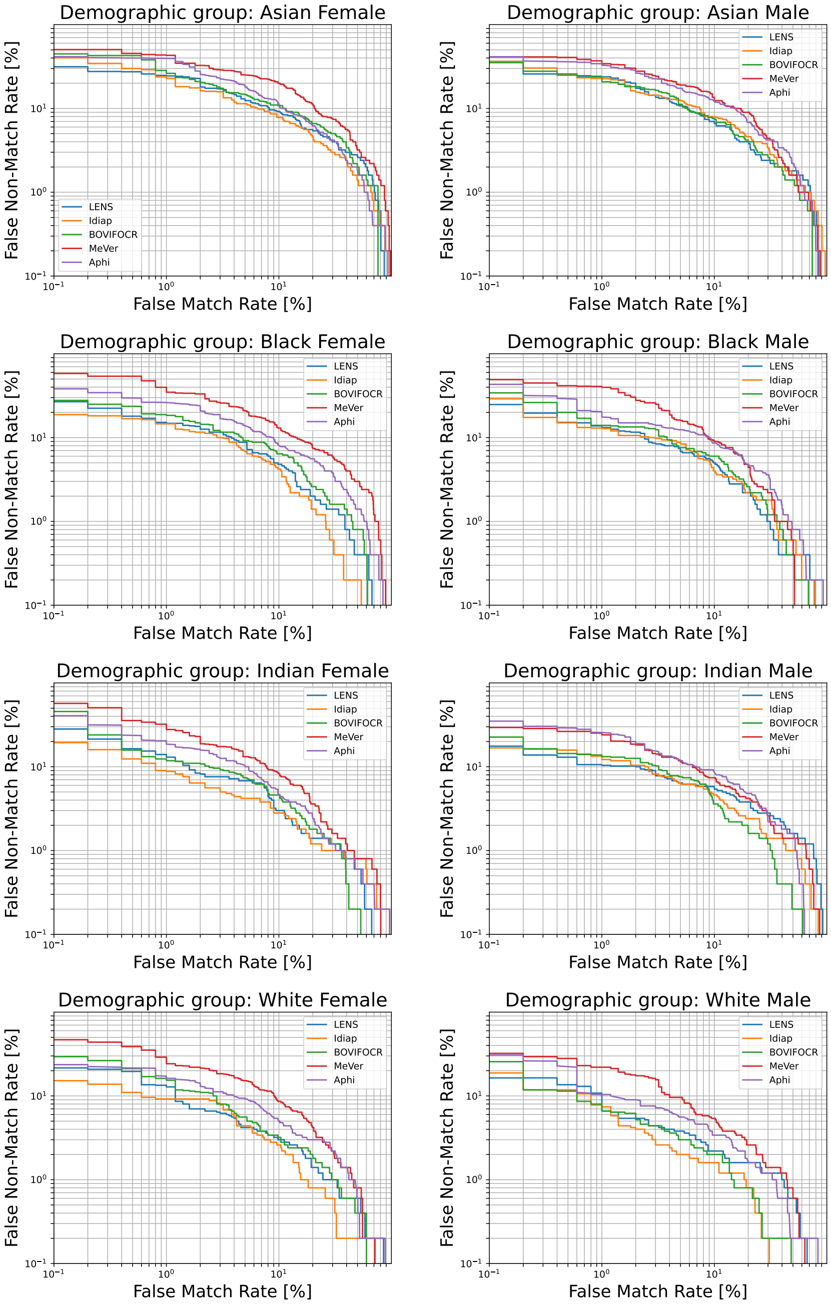


Figure 11: Comparison of the DET curves provided for each demographic group of interest by top-5 teams in **Sub-Task 1.1**. DET = Detection Error Trade-off.

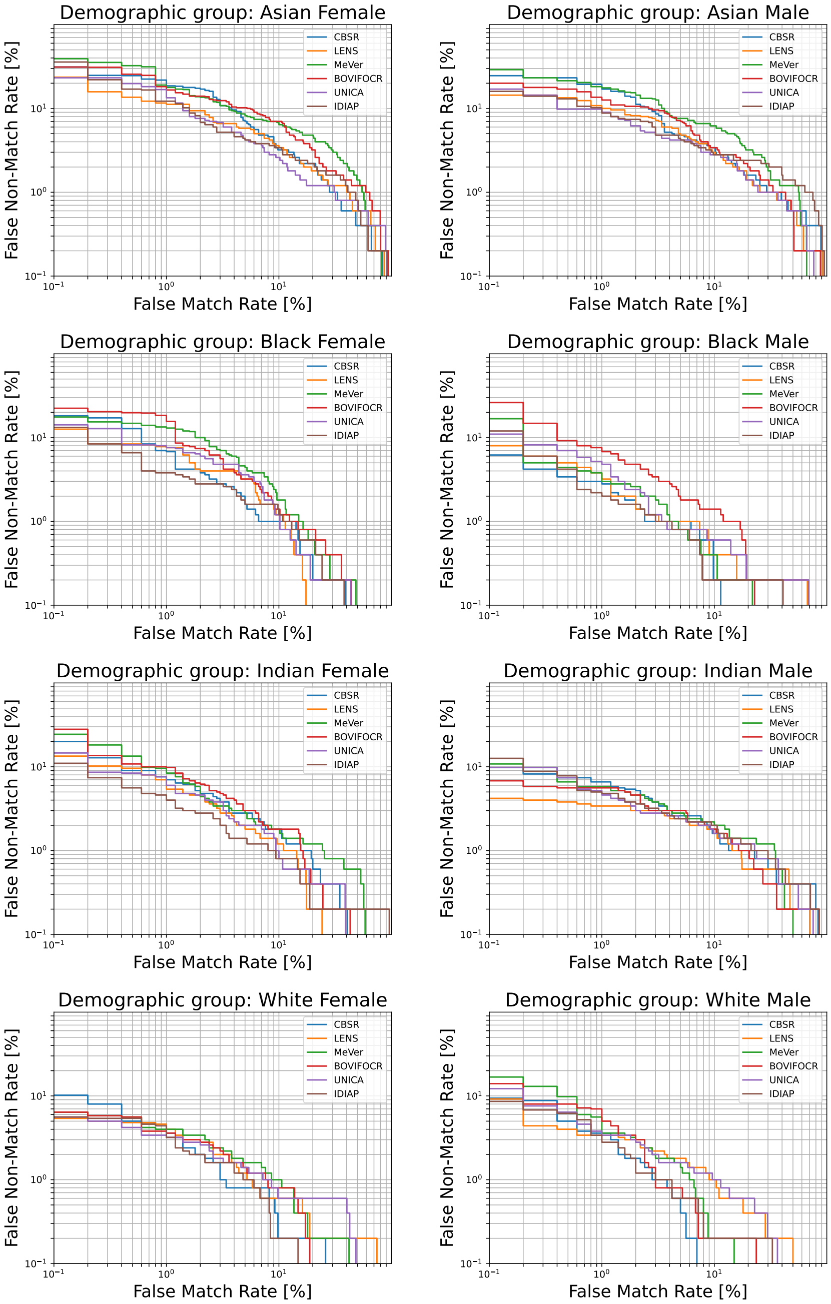


Figure 12: Comparison of the DET curves provided for each demographic group of interest by top-6 teams in **Sub-Task 1.2**. DET = Detection Error Trade-off.

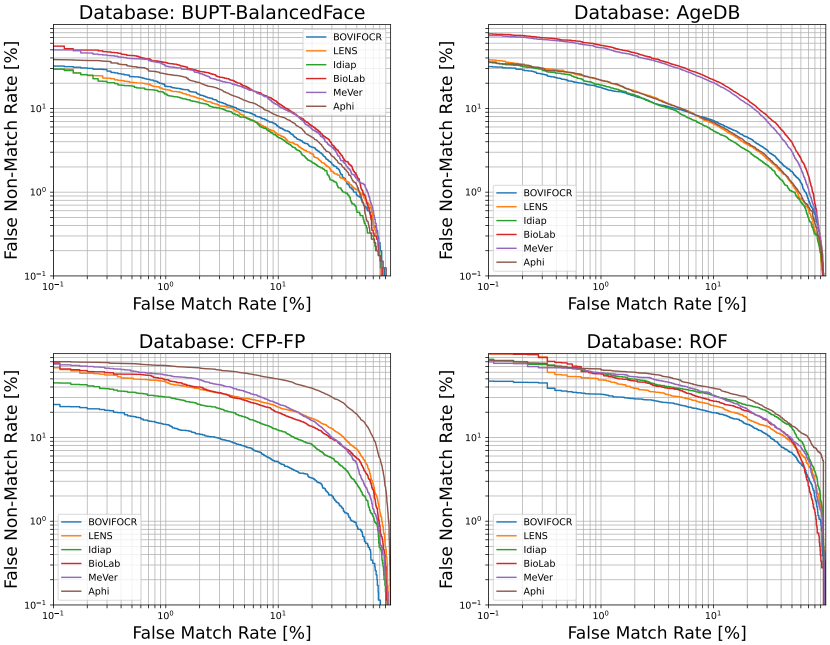


Figure 13: Comparison of the DET curves provided for each evaluation database of interest by top-6 teams in **Sub-Task 2.1**. DET = Detection Error Trade-off.

and ArcFace [2] loss functions were widely used, featuring in the approaches of CBSR, LENS, Idiap, and BioLab for the former, and BOVIFOCR-UFPR, MeVer, and Aphi for the latter. Idiap and UNICA-FRAUNHOFER IGD also considered the CosFace loss function [79]. Most of the teams integrated multiple networks into their proposed architectures for different objectives, *e.g.*, CBSR and LENS trained different networks with distinct augmentation techniques, while BOVIFOCR-UFPR and Idiap combined different loss functions. In these proposed architectures, the features extracted by different networks are fused before making a decision in the verification process, indicating the validity of information fusion at both the feature and score levels [89]. Some teams also addressed the challenges of domain shift between synthetic and real data, *e.g.*, LENS proposed solutions robust to domain shifts with consistent data augmentation, while CBSR implemented a range of strategies, including advanced data augmentation, identity clustering, and distinct thresholds for different databases. Notably, CBSR utilized all available databases for training, including FFHQ [65], unlike other teams. Excluding BOVIFOCR-UFPR, Aphi, and UNICA-FRAUNHOFER IGD, which

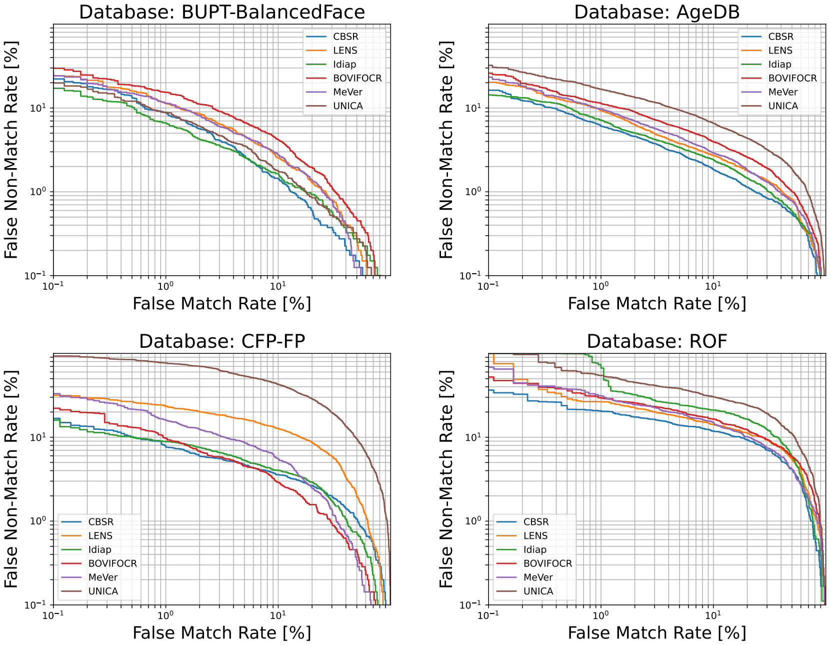


Figure 14: Comparison of the DET curves provided for each evaluation database of interest by top-6 teams in **Sub-Task 2.2**. DET = Detection Error Trade-off.

exclusively used DCFace [6], the majority of teams employed both DCFace [6] and GANDiffFace [7], demonstrating the suitability of both generative frameworks.

6.1. Analysis of Specific Demographic Groups and Databases

Detection Error Trade-off Curves. We plot the Detection Error Trade-off (DET) curves of the best classified teams for each demographic group (Figures 11 and 12, associated with Sub-Tasks 1.1 and 1.2, respectively) or database (Figures 13 and 14, associated with Sub-Tasks 2.1 and 2.2, respectively). This analysis offers a visual comparison of the proposed FR systems across the different demographic groups and databases for different operational points. For instance, analyzing Figure 11 we can observe that the FR system proposed by Idiap provides the best FNMR at FMR ranging from 0.1% to 10% for the demographic groups of Black Females and Indian Females in Sub-Task 1.1. Similarly, Idiap also achieves the best FNMR at FMR ranging from 0.1% to 1% for the same demographic groups in Sub-Task 1.2 (Figure 12), while LENS achieves the best FNMR at FMR ranging from

0.1% to 1% for the demographic groups of Indian Males and White Males. DET curves consistently overlap in the graphs provided for Sub-Tasks 1.1 and 1.2, indicating that ranking the proposed FR systems without fixing an operational point is challenging. On the other hand, it is easier to identify the best FR systems in terms of FNMR for large intervals of FMR in Sub-Tasks 2.1 and 2.2 (Figures 13 and 14, respectively). BUPT-BalancedFace [58] and AgeDB [66] emerge as the databases that yield the highest performance in evaluation. In Sub-Task 2.1 (Figure 13), the winning team BOVIFOCR-UFPR clearly outperforms the other teams when evaluated with the CFP-FP [40] and ROF [67] databases, showing a better generalization of the FR system against pose variations and occlusions. In Sub-Task 2.2 (Figure 14), the winning team CBSR outperforms the other teams in the evaluation of AgeDB (although the Idiap team achieves better FNMR results around the operational point of FMR=0.1%) and ROF databases, showing in general a more robust FR system against age variability and occlusions.

Top-5 Teams Average Metrics. To comprehensively quantify the trend of FR performance for different demographic groups and databases, we conduct an in-depth analysis focusing on the average metrics obtained from the top-5 teams in each sub-task. In Table 5, we present the averages (and standard deviations) of accuracy, FNMR@FMR=1%, and AUC computed using the values provided by the top-5 teams for each sub-task. We analyze both the baseline and proposed systems, presenting the average metrics for each demographic group in Sub-Tasks 1.1 and 1.2, and for each evaluation database in Sub-Tasks 2.1 and 2.2. Finally, for each of the considered metrics (*i.e.,* accuracy, FNMR@FMR=1%, and AUC), we compute the GAP between the average values obtained for the baseline and proposed systems. It's worth noting that we calculate the GAP for FNMR@FMR=1% with the opposite sign compared to the GAP calculated for the other metrics, following the formula described in Section 4.3. This is because improvements in FR systems are represented by increasing values for accuracy and AUC, and decreasing values for FNMR.

In both Sub-Tasks 1.1 and 1.2, we observe that the average performance for the two demographic groups representing the Asian population is consistently lower across all metrics (*i.e.,* accuracy, FNMR@FMR=1%, and AUC), both in the baseline and proposed systems, compared to the other demographic groups. The lower performance within the Asian population is a known issue, and previous efforts to mitigate this bias involved databases

Table 5: Analysis of specific demographic groups and databases for both baseline and proposed systems, averaged across the top-5 teams of each sub-task. GAP values are calculated for each metric (*i.e.,* accuracy, FNMR@FMR=1%, and AUC) according to the average values of the baseline and proposed systems reported in the table. The GAP of FNMR@FMR=1% has the opposite sign compared to the GAP of the other metrics, because improvements in FR systems are represented by increasing values for accuracy and AUC, and decreasing values for FNMR. All values are expressed in percentage. Acc. = Accuracy, FNMR = False Non-Match Rate, FMR = False Match Rate, AUC = Area Under Curve, GAP = Gap to Real.

Sub-Task 1.1 (Bias Mitigation): Synthetic Data

	Average of baseline systems			**Average of proposed systems**			**GAP**		
Demographic group	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**
Black Female	93.24±2.43	9.76±2.96	99.26±0.30	90.14±3.71	22.48±8.70	97.19±1.54	3.44	56.58	2.12
Black Male	95.10±2.66	5.96±2.05	99.61±0.06	90.38±3.87	19.80±10.32	97.75±0.73	5.22	69.90	1.90
Asian Female	87.30±5.46	16.44±3.02	97.93±0.52	88.06±2.99	31.56±8.28	95.46±1.34	-0.86	47.91	2.59
Asian Male	89.32±5.12	14.80±4.14	98.12±0.49	89.54±3.06	27.76±5.98	96.33±0.81	-0.25	46.69	1.86
Indian Female	87.84±6.94	8.20±1.91	99.32±0.34	90.64±3.92	16.68±7.41	98.20±0.65	-3.09	50.84	1.15
Indian Male	91.36±4.97	6.40±2.16	99.15±0.23	91.82±3.14	17.48±6.38	97.74±0.61	-0.50	63.39	1.44
White Female	96.00±1.33	4.80±1.82	99.64±0.16	92.92±2.06	16.28±5.79	98.29±0.75	3.31	70.52	1.37
White Male	96.58±0.91	4.64±1.64	99.70±0.12	94.40±1.62	11.08±5.57	98.88±0.55	2.31	58.12	0.82

Sub-Task 1.2 (Bias Mitigation): Synthetic + Real Data

	Average of baseline systems			**Average of proposed systems**			**GAP**		
Demographic group	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**
Black Female	92.44±4.90	13.92±2.78	98.61±0.83	95.90±0.74	10.60±4.03	99.42±0.17	-3.61	-31.32	-0.81
Black Male	92.70±5.55	10.56±4.12	98.29±2.04	97.52±0.58	4.44±1.68	99.72±0.11	-4.94	-137.84	-1.43
Asian Female	90.36±3.32	22.68±4.74	96.60±1.57	92.28±1.73	16.32±3.14	98.10±0.57	-2.08	-38.97	-1.53
Asian Male	90.66±4.34	19.84±4.53	97.12±1.73	94.10±1.19	13.68±3.62	98.50±0.37	-3.66	-45.03	-1.41
Indian Female	93.04±2.75	12.00±4.27	98.04±2.10	94.52±2.14	7.88±1.37	99.42±0.16	-1.57	-52.28	-1.39
Indian Male	92.56±4.19	11.72±6.67	96.75±3.99	95.78±1.81	5.24±1.08	99.31±0.08	-3.36	-123.66	-2.59
White Female	92.20±5.96	7.56±3.01	98.72±1.40	96.84±0.53	3.92±0.41	99.71±0.07	-4.79	-92.86	-0.99
White Male	92.26±6.22	7.80±2.86	98.68±1.45	96.80±0.61	4.12±0.92	99.75±0.09	-4.69	-89.32	-1.07

Sub-Task 2.1 (Overall Improvement): Synthetic Data

	Average of baseline systems			**Average of proposed systems**			**GAP**		
Database	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**
BUPT	91.98±2.01	12.40±3.14	98.82±0.30	91.55±1.84	23.54±8.53	97.13±1.02	0.47	47.31	1.74
AgeDB	94.58±1.39	9.75±3.00	98.57±0.41	89.44±3.97	33.63±17.38	95.34±2.51	5.75	71.00	3.39
CFP-FP	90.37±5.28	15.79±8.04	96.85±2.51	85.12±4.32	39.00±14.88	93.08±3.01	6.17	59.50	4.05
ROF	84.31±4.70	30.40±3.45	92.10±1.49	79.84±3.47	51.33±10.03	87.88±2.23	5.59	40.77	4.80

Sub-Task 2.2 (Overall Improvement): Synthetic + Real Data

	Average of baseline systems			**Average of proposed systems**			**GAP**		
Database	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**	**Acc.**	**FNMR@ FMR=1%**	**AUC**
BUPT	92.81±1.63	13.14±3.90	97.97±1.88	94.48±1.74	10.72±2.93	99.02±0.32	-1.77	-22.59	-1.06
AgeDB	94.96±1.31	9.30±2.25	98.70±0.20	95.33±0.97	8.74±1.88	98.75±0.30	-0.39	-6.43	-0.05
CFP-FP	90.77±5.28	15.04±7.47	96.96±2.48	91.54±4.65	13.19±5.91	97.89±1.24	-0.85	-13.99	-0.94
ROF	84.11±4.67	31.65±4.64	91.86±1.61	85.08±4.54	35.65±17.75	92.98±1.35	-1.14	11.22	-1.20

generated with GANDiffFace [7, 18]. Remarkably, analyzing the results of Sub-Task 1.1, the four demographic groups with the lowest average accuracy across the baseline systems (*i.e.,* Asian and Indian populations of both genders, with average accuracy between 87.30% and 91.36%), benefit from the use of synthetic data alone for training. This results in an improvement in average accuracy of the proposed systems, quantified with GAP values between -0.25% and -3.09%. Conversely, for the other demographic groups representing the Black and White populations, the average accuracy across the top-5 teams decreases from the baseline to the proposed systems, quantified with GAP values between 2.31% and 5.22%. To consistently achieve a negative GAP value for each demographic group and each metric, indicating therefore a comprehensive performance improvement, a combination of synthetic and real data is necessary for training, as can be seen in the results of Sub-Task 1.2. GAP values ranging from -1.57% to -4.94% are observed for accuracy across the various demographic groups. These results prove the potential of combining real and synthetic data to reduce the bias in FR technology.

Analyzing Sub-Task 2.1, we observe that synthetic data alone are insufficient to improve the average performance of baseline systems for any of the four considered databases. The combination of synthetic and real databases (Sub-Task 2.2) is necessary to achieve improvements between the averages of the metrics provided by baseline and proposed systems. Consistent with our previous discussion, the average performance of the top-5 teams in both Sub-Tasks 2.1 and 2.2 emphasizes that BUPT-BalancedFace [58] and AgeDB [66] are the databases yielding the highest performance during evaluation, in both the baseline and proposed systems, and across all metrics (*i.e.,* accuracy, FNMR@FMR=1%, and AUC). BUPT-BalancedFace [58] also stands out as the database with the lowest GAP values for accuracy in both Sub-Tasks 2.1 and 2.2 (0.47% and -1.77%, respectively), for AUC in Sub-Task 2.1 (1.74%), and for FNMR@FMR=1% in Sub-Task 2.2 (-22.59%). This confirms that using DCFace [6] and GANDiffFace [7] for FR system training, particularly when fused with real data, enhances performance across diverse demographic groups. Similar results are observed for the GAP values calculated for the three other databases (*i.e.,* AgeDB [66], CFP-FP [40], and ROF [67]) and across all metrics (*i.e.,* accuracy, FNMR@FMR=1%, and AUC). The results provide positive GAP values in Sub-Task 2.1 and negative GAP values in Sub-Task 2.2, except for the GAP in FNMR@FMR=1% calculated for the ROF database, indicating that the fusion of real and synthetic data also enhances performance in presence of pose variations, aging, and occlusions.

7. Conclusion

The proposed FRCSyn-onGoing represents a significant step forward in evaluating the application of synthetic data to FR, addressing current limitations in the field. Information fusion played a crucial role in this study at various levels. Notably, the fusion of synthetic and real data emerged as the optimal configuration for training, resulting in the proposed FR systems outperforming baseline systems exclusively trained with real-world databases. Additionally, numerous participating teams adopted an approach that involved fusing information from different networks to enhance FR performance. These networks were trained with diverse loss functions or differently augmented data, allowing the extraction of distinct features from input images that could be fused before conducting face verification.

Within FRCSyn-onGoing, various approaches from different research groups were proposed and compared across different sub-tasks. A detailed analysis of the performance across demographic groups and databases representing different challenges revealed notable findings. Specifically, the proposed FR systems exhibited lower performance when evaluated on demographic groups representing the Asian population, compared to other groups, in both the baseline and proposed systems of Sub-Tasks 1.1 and 1.2. Nevertheless, the BUPT-BalancedFace database [58] substantially benefits from the training of FR systems with the proposed synthetic databases, *i.e.,* DCFace [6] and GANDiffFace [7]. It is important to observe that BUPT-BalancedFace evaluates FR performance in presence of demographic diversity within the test population, utilizing comparisons between individuals of the same demographic group, considered more challenging compared to comparisons between individuals of different demographic groups.

FRCSyn-onGoing provides a reproducible ongoing benchmark accessible to all researchers in the field for evaluating their deployed FR systems. The material provided by many participating teams hold promise for advancing the application of synthetic data to enhance FR technology. Future work will focus on maintaining the ongoing competition and introducing new tasks to evaluate additional aspects of interest. Potential new tasks may involve exploring the feasibility of training FR systems with additional synthetic databases, to evaluate their applicability in the field.

Acknowledgements

Special thanks to Mei Wang and Stylianos Moschoglou for authorizing the distribution of their databases. This study has received funding from the European Union's Horizon 2020 TReSPAsS-ETN (No 860813) and is supported by INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER), R&D Agreement DGGC/UAM/FUAM for Biometrics and Cybersecurity, Autonomous Community of Madrid (ELLIS Unit Madrid), and Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-0002). It is also supported by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. BioLab acknowledge Andrea Pilzer from the NVIDIA AI Technology Center, EMEA, for his support. MeVer was supported by the EU Horizon Europe project MAMMOth (Grant Agreement 101070285).

References

[1] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang, Biometrics recognition using deep learning: A survey, Artificial Intelligence Review (2023) 1–49. 2, 3

[2] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 3, 20, 21, 24, 26, 33

[3] M. Kim, A. K. Jain, X. Liu, AdaFace: Quality Adaptive Margin for Face Recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 3, 17, 18, 19, 22, 24, 25, 30

[4] M. Wang, W. Deng, Deep face recognition: A survey, Neurocomputing 429 (2021) 215–244. 3, 6, 7

[5] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, in: Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, 2008. 3, 25

[6] M. Kim, F. Liu, A. Jain, X. Liu, DCFace: Synthetic Face Generation with Dual Condition Diffusion Model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 3, 4, 5, 8, 9, 12, 15, 17, 21, 22, 25, 26, 27, 30, 34, 37, 38

[7] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, 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, 2023. 3, 4, 5, 8, 9, 10, 12, 17, 21, 22, 25, 30, 34, 37, 38

[8] I. Adjabi, A. Ouahabi, A. Benzaoui, A. Taleb-Ahmed, Past, present, and future of face recognition: A review, Electronics 9 (8) (2020) 1188. 3

[9] D. Wanyonyi, T. Celik, Open-source face recognition frameworks: A review of the landscape, IEEE Access 10 (2022) 50601–50623. 3

[10] H. Du, H. Shi, D. Zeng, X.-P. Zhang, T. Mei, The elements of end-to-end deep face recognition: A survey of recent advances, ACM Computing Surveys 54 (10s) (2022) 1–42. 3, 5, 6

[11] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, C. C. Loy, Domain generalization: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022). 3, 5

[12] Y. Shi, X. Yu, K. Sohn, M. Chandraker, A. K. Jain, Towards universal representation learning for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6817–6826. 3

[13] D. Zeng, R. Veldhuis, L. Spreeuwers, A survey of face recognition techniques under occlusion, IET Biometrics 10 (6) (2021) 581–606. 3, 7

[14] W. Ali, W. Tian, S. U. Din, D. Iradukunda, A. A. Khan, Classical and modern face recognition approaches: a complete review, Multimedia tools and applications 80 (2021) 4825–4880. 3, 6, 7

[15] Y. Deng, J. Yang, D. Chen, F. Wen, X. Tong, Disentangled and controllable face image generation via 3d imitative-contrastive learning, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2020. 3

[16] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, 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, 2023. 3, 4, 5, 8, 19

[17] C. Zhang, X. Chen, S. Chai, C. H. Wu, D. Lagun, T. Beeler, F. De la Torre, ITI-GEN: Inclusive Text-to-Image Generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 3, 4, 9

[18] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, A. Morales, D. Lawatsch, F. Domin, M. Schaubert, Synthetic Data for the Mitigation of Demographic Biases in Face Recognition, in: Proceedings of the IEEE International Joint Conference on Biometrics, 2023. 4, 5, 7, 8, 37

[19] F. Boutros, V. Struc, J. Fierrez, N. Damer, Synthetic data for face recognition: Current state and future prospects, Image and Vision Computing (2023) 104688. 4, 8

[20] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, Information Fusion 64 (2020) 131–148. 4

[21] C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, C. Busch, Handbook of digital face manipulation and detection: from DeepFakes to morphing attacks, Springer Nature, 2022. 4

[22] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, J. Fierrez, GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection, IEEE Journal of Selected Topics in Signal Processing 14 (5) (2020) 1038–1048. 4

[23] E. H. Salazar-Jurado, R. Hernández-García, K. Vilches-Ponce, R. J. Barrientos, M. Mora, G. Jaswal, Towards the generation of synthetic images of palm vein patterns: A review, Information Fusion 89 (2023) 66–90. 4

[24] F. Boutros, N. Damer, K. Raja, R. Ramachandra, F. Kirchbuchner, A. Kuijper, Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation, Image and Vision Computing 104 (2020) 104007. 4

[25] P. Kang, S. Jiang, P. B. Shull, Synthetic EMG Based on Adversarial Style Transfer Can Effectively Attack Biometric-Based Personal Identification Models, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2023). 4

[26] M. Murgia, M. Harlow, Who's using your face? The ugly truth about facial recognition, Financial Times 19 (2019). 4

[27] J. Harvey, Adam. LaPlace, Exposing.ai, https://exposing.ai (2021). 4, 5

[28] P. Voigt, A. Von dem Bussche, The EU General Data Protection Regulation (GDPR), A Practical Guide, 1st Ed., Cham: Springer International Publishing 10 (3152676) (2017) 10–5555. 4

[29] A. Morales, J. Fierrez, R. Vera-Rodriguez, R. Tolosana, SensitiveNets: Learning agnostic representations with application to face images, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (6) (2020) 2158–2164. 4, 7

[30] M. Kim, H. Byun, Learning texture invariant representation for domain adaptation of semantic segmentation, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2020, pp. 12975–12984. 4

[31] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, C. Zhu, Real-world single image super-resolution: A brief review, Information Fusion 79 (2022) 124–145. 4

[32] Y. Shao, L. Li, W. Ren, C. Gao, N. Sang, Domain adaptation for image dehazing, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2020, pp. 2808–2817. 4

[33] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, D. Tao, SynFace: Face Recognition With Synthetic Data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 4, 5, 8, 19

[34] F. Boutros, M. Huber, P. Siebke, T. Rieber, N. Damer, Sface: Privacy-friendly and accurate face recognition using synthetic data, in: Proceedings of the IEEE International Joint Conference on Biometrics, 2022. 4, 8

[35] F. Boutros, J. H. Grebe, A. Kuijper, 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, 2023. 4, 9, 18, 19

[36] M. Kansy, A. Raël, G. Mignone, J. Naruniec, C. Schroers, M. Gross, R. M. Weber, Controllable Inversion of Black-Box Face Recognition Models via Diffusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2023. 4, 9

[37] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, A. Kuijper, A comprehensive study on face recognition biases beyond demographics, IEEE Transactions on Technology and Society 3 (1) (2021) 16–30. 5, 7

[38] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, et al., 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, 2024. 5

[39] M. K. Rusia, D. K. Singh, A comprehensive survey on techniques to handle face identity threats: challenges and opportunities, Multimedia Tools and Applications 82 (2) (2023) 1669–1748. 6

[40] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, D. W. Jacobs, Frontal to profile face verification in the wild, in: Proceedings of the IEEE Winter conference on Applications of Computer Vision, 2016. 6, 9, 10, 11, 12, 13, 18, 35, 37

[41] M. O. Oloyede, G. P. Hancke, H. C. Myburgh, A review on face recognition systems: recent approaches and challenges, Multimedia Tools and Applications 79 (2020) 27891–27922. 6

[42] M. De-la Torre, E. Granger, P. V. Radtke, R. Sabourin, D. O. Gorodnichy, Partially-supervised learning from facial trajectories for face recognition in video surveillance, Information Fusion 24 (2015) 31–53. 6

[43] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, B. Zheng, Age-invariant face recognition by multi-feature fusionand

decomposition with self-attention, ACM Transactions on Multimedia Computing, Communications, and Applications 18 (1s) (2022) 1–18. 7

[44] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, T. Zhang, Orthogonal deep features decomposition for age-invariant face recognition, in: Proceedings of the European Conference on Computer Vision, 2018. 7

[45] D. Deb, L. Best-Rowden, A. K. Jain, Face recognition performance under aging, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 46–54. 7

[46] J. Cai, H. Han, J. Cui, J. Chen, L. Liu, S. K. Zhou, Semi-Supervised Natural Face De-Occlusion, IEEE Transactions on Information Forensics and Security 16 (2020) 1044–1057. 7

[47] J. Deng, S. Cheng, N. Xue, Y. Zhou, S. Zafeiriou, UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018. 7

[48] Y. Hu, X. Wu, B. Yu, R. He, Z. Sun, Pose-guided photorealistic face rotation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018. 7

[49] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, et al., Towards pose invariant face recognition in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2207–2216. 7

[50] K. T. Voo, L. Jiang, C. C. Loy, Delving into high-quality synthetic face occlusion segmentation datasets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4711–4720. 7

[51] G. Antipov, M. Baccouche, J.-L. Dugelay, Face aging with conditional generative adversarial networks, in: Proceedings of the IEEE International Conference on Image Processing, 2017. 7

[52] J. Zhao, L. Xiong, P. Karlekar Jayashree, J. Li, F. Zhao, Z. Wang, P. Sugiri Pranata, P. Shengmei Shen, S. Yan, J. Feng, Dual-Agent GANs

for Photorealistic and Identity Preserving Profile Face Synthesis, in: Proceedings of the Advances in Neural Information Processing Systems, Vol. 30, 2017. 7

[53] J. Zhao, L. Xiong, Y. Cheng, Y. Cheng, J. Li, L. Zhou, Y. Xu, J. Karlekar, S. Pranata, S. Shen, et al., 3D-Aided Deep Pose-Invariant Face Recognition, in: Proceedings of the International Joint Conferences on Artificial Intelligence, Vol. 2, 2018, p. 11. 7

[54] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923 (2014). 7, 9, 10, 12, 13, 17, 19, 22, 24, 25, 26, 27

[55] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, VGGFace2: A dataset for recognising faces across pose and age, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2018. 7

[56] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: A dataset and benchmark for large-scale face recognition, in: Proceedings of the European Conference on Computer Vision, 2016. 7

[57] I. Sarridis, C. Koutlis, S. Papadopoulos, C. Diou, Towards Fair Face Verification: An In-depth Analysis of Demographic Biases, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases Workshops, 2023. 7

[58] M. Wang, W. Deng, Mitigating bias in face recognition using skewness-aware reinforcement learning, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2020. 7, 9, 10, 11, 12, 13, 18, 21, 24, 30, 35, 37, 38

[59] V. Cherepanova, S. Reich, S. Dooley, H. Souri, M. Goldblum, T. Goldstein, A deep dive into dataset imbalance and bias in face identification, arXiv preprint arXiv:2203.08235 (2022). 7

[60] H. Zhang, M. Grimmer, R. Ramachandra, K. Raja, C. Busch, On the applicability of synthetic data for face recognition, in: Proceedings of the IEEE International Workshop on Biometrics and Forensics, 2021. 8

[61] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, Advances in Neural Information Processing Systems 34 (2021) 852–863. 8

[62] M. Falkenberg, A. B. Ottsen, M. Ibsen, C. Rathgeb, Child face recognition at scale: Synthetic data generation and performance benchmark, arXiv preprint arXiv:2304.11685 (2023). 8

[63] L. Colbois, T. de Freitas Pereira, 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, 2021. 8

[64] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training generative adversarial networks with limited data, in: Advances in Neural Information Processing Systems, 2020. 8

[65] T. Karras, S. Laine, T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2019. 9, 10, 12, 13, 17, 19, 33

[66] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, 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, 2017. 9, 10, 11, 12, 13, 18, 35, 37

[67] M. E. Erakιn, U. Demir, H. K. Ekenel, On Recognizing Occluded Faces in the Wild, in: Proceedings of the International Conference of the Biometrics Special Interest Group, 2021. 9, 10, 11, 12, 13, 18, 35, 37

[68] K. Karkkainen, 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, 2021. 10

[69] N. I. of Standards, T. (NIST), Frvt 1:1 verification, https://pages.nist.gov/frvt/html/frvt11.html (2023). 15

[70] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,

in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Vol. 96, 1996, pp. 226–231. 17

[71] M. L. Ngan, P. J. Grother, K. K. Hanaoka, Ongoing Face Recognition Vendor Test (FRVT) Part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms (2020). 17, 18

[72] J. Wang, Y. Liu, Y. Hu, H. Shi, T. Mei, FaceX-Zoo: A PyTorch Toolbox for Face Recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021. 17, 18

[73] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016. 18, 20, 24, 26, 27, 30

[74] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, RetinaFace: Single-shot Multi-level Face Localisation in the Wild, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2020. 19, 21, 22

[75] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2018. 20

[76] J. Deng, J. Guo, X. An, Z. Zhu, S. Zafeiriou, Masked face recognition challenge: The insightface track report, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 21

[77] D. McNeely-White, B. Sattelberg, N. Blanchard, R. Beveridge, Canonical face embeddings, IEEE Transactions on Biometrics, Behavior, and Identity Science 4 (2022) 197–209. 21

[78] D. McNeely-White, B. Sattelberg, N. Blanchard, R. Beveridge, Canonical face embeddings, IEEE Transactions on Biometrics, Behavior, and Identity Science 4 (2) (2022) 197–209. 22

[79] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, CosFace: Large margin cosine loss for deep face recognition, in: Proceedings IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 22, 27, 33

[80] J. Deng, J. Guo, T. Liu, M. Gong, S. Zafeiriou, Sub-center ArcFace: Boosting Face Recognition by Large-scale Noisy Web Faces, in: Proceedings of the European Conference on Computer Vision, 2020. 23

[81] J. Cheng, T. Liu, K. Ramamohanarao, D. Tao, Learning with bounded instance and label-dependent label noise, in: Proceedings of the International Conference on Machine Learning, 2020. 23

[82] M. Wang, Y. Zhang, W. Deng, Meta balanced network for fair face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (11) (2021) 8433–8448. 24

[83] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (10) (2016) 1499–1503. 24

[84] L. Chai, D. Bau, S.-N. Lim, P. Isola, What makes fake images detectable? Understanding properties that generalize, in: Proceedings of the European Conference on Computer Vision, 2020. 25

[85] M. Tan, Q. Le, EfficientNetV2: Smaller Models and Faster Training, in: Proceedings of the International Conference on Machine Learning, 2021. 26, 30

[86] F. Boutros, N. Damer, F. Kirchbuchner, A. Kuijper, ElasticFace: Elastic Margin Loss for Deep Face Recognition, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2022. 26

[87] F. Boutros, M. Klemt, M. Fang, A. Kuijper, N. Damer, Unsupervised face recognition using unlabeled synthetic data, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, 2023. 27

[88] E. D. Cubuk, B. Zoph, J. Shlens, 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, 2020. 27

[89] A. Lumini, L. Nanni, Overview of the combination of biometric matchers, Information Fusion 33 (2017) 71–85. 33