Full length article

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 \(^b\), Zhen Lei \(^l,k\), Suvidha Tripathi \(^i\), Mahak Kothari \(^i\), Md Haider Zama \(^i\), Debayan Deb \(^i\), Bernardo Bieseck \(^j,k\), Pedro Vidal \(^i\), Roger Granada \(^i\), Guilherme Fickel \(^i\), Gustavo Führ \(^i\), David Menotti \(^j\), Alexander Unnervik \(^m,n\), Anjith George \(^m\), Christophe Ecabert \(^m\), Hatif 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 \(^q\), Guido Borghi \(^q\), Lorenzo Pellegrini \(^q\), Enrique Mas-Candela \(^s\), Ángela Sánchez-Pérez \(^s\), Andrea Atzori \(^s\), 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
\(^d\) IE, CAS, China
\(^e\) School of Cyber Security, UCAS, China
\(^f\) MAIS, CASIA, China
\(^g\) School of Artificial Intelligence, UCAS, China
\(^h\) CAIR, HRISI, CAS, China
\(^i\) LENS, Inc., US
\(^j\) Federal University of Paraná, Curitiba, PR, Brazil
\(^k\) Federal Institute of Mato Grosso, Pontes e Lacerda, Brazil
\(^l\) unic - idTech, Brazil
\(^m\) Idiap Research Institute, Switzerland
\(^n\) Ecole 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
\(^u\) Fraunhofer IGD, Germany
\(^v\) TU Darmstadt, Germany

A R T I C L E  I N F O

Keywords:
FRCSyn-onGoing
Face recognition
Generative AI
Demographic bias
Benchmark

A B S T R A C T

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

* Correspondence to: Universidad Autonoma de Madrid, C. Francisco Tomás y Valiente, 11, 28049 Madrid, Spain.
E-mail address: ruben.tolosana@uam.es (R. Tolosana).

https://doi.org/10.1016/j.inffus.2024.102322
Received 22 December 2023; Received in revised form 11 January 2024; Accepted 26 February 2024
Available online 5 March 2024
1566-2535/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
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.

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 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, 6, 7]. 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 [8]. Numerous studies indicate that DL models, even when trained on extensive databases, experience notable performance drops when confronted with previously unseen conditions [9, 10]. 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 [11]. Various limitations associated with established databases and benchmarks are extensively discussed in [12]. For instance, LFW [5] is considered to have a limited number of images per subject for SOTA challenges such as illumination, pose, and occlusion invariances.

In recent years, the literature has introduced various approaches for generating synthetic face content [13–15] intended for different applications, including FR [16–18] and digital face manipulations, commonly known as DeepFakes [19–21]. Additionally, synthetic content has been created for other biometric modalities [22–24]. 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 databases. Secondly, synthetic face generators have been created for other biometric modalities [22–24]. 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

---

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 [26] and the observed limitations in FR technology across demographic groups [17, 37] and challenging conditions [8]. 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 [9]. 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 [14] 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: DFFace [16] and GANDiffFace [28]. 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 Fig. 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 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-on-Going, 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 remainder 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-on-Going. 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-on-Going 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-on-Going 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 [4,8,12,39]. 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 [12]. 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 [11]. 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 [12]. 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 [47–49], 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 [54–56], 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 [4,17]. 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 [17].

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 [18], 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 [28,62] 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

Fig. 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.
graphics pipeline [14]. 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 [16] and GANDiffFace [28]. Examples of face images synthesized with both DCFace and GANDiffFace are included in Fig. 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 [17].

Table 1 provides the details of the public databases considered in FRCSyn-onGoing. Id = Identities, Img = Images.

<table>
<thead>
<tr>
<th>Database</th>
<th>Framework</th>
<th>Use</th>
<th># Id</th>
<th># Img/Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCFace [16]</td>
<td>DCFace</td>
<td>Train</td>
<td>10K</td>
<td>50</td>
</tr>
<tr>
<td>GANDiffFace [28]</td>
<td>GANDiffFace</td>
<td>Train</td>
<td>10K</td>
<td>50</td>
</tr>
<tr>
<td>CASIA-WebFace [54]</td>
<td>Real-world</td>
<td>Train</td>
<td>10.5K</td>
<td>47</td>
</tr>
<tr>
<td>FFHQ [65]</td>
<td>Real-world</td>
<td>Train</td>
<td>70K</td>
<td>1</td>
</tr>
<tr>
<td>BUPF-BalancedFace [58]</td>
<td>Real-world</td>
<td>Eval</td>
<td>24K</td>
<td>45</td>
</tr>
<tr>
<td>AgeDB [66]</td>
<td>Real-world</td>
<td>Eval</td>
<td>570</td>
<td>29</td>
</tr>
<tr>
<td>CPFP-FFP [40]</td>
<td>Real-world</td>
<td>Eval</td>
<td>500</td>
<td>14</td>
</tr>
<tr>
<td>ROF [67]</td>
<td>Real-world</td>
<td>Eval</td>
<td>180</td>
<td>31</td>
</tr>
</tbody>
</table>

Fig. 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. 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) BUPF-BalancedFace [58], (ii) AgeDB [66], (iii) CPA-FPP [40], and (iv) ROF [67]. BUPF-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 BUPF-BalancedFace [58] for facilitating easier and more consistent evaluation. A list of 8,000 random comparison pairs is generated from identities in the BUPF-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], CPA-FPP [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 datasets are considered for training and evaluation, we also intend to analyze 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 CPA-FPP, we exclusively consider the frontal-profile comparisons specified in the original evaluation protocol, excluding all frontal-fronal comparisons. This results in a total of
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.

<table>
<thead>
<tr>
<th>Task 1: synthetic data for demographic bias mitigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: training only with CASIA-WebFace [54] and FFHQ [65];</td>
</tr>
<tr>
<td>Metrics: accuracy, FNMR/FMR = %, AVG, GAP;</td>
</tr>
<tr>
<td>Ranking: AVG (across demographic groups) vs SD of accuracy, see Section 4.3 for more details.</td>
</tr>
</tbody>
</table>

Sub-Task 1.1: training exclusively with synthetic databases
- Train: DCFace [16] and GANDiffFace [28];
- 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 = %, 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.

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

---

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, 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 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-on-Going, we consider a number of comparisons for each demographic group and database in the order of 10^7. 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: GAP = (REAL − SYN)/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 [16], 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: TO = AVG − 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 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-on-going: Description of systems

FRCSyn-on-going 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 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 Fig. 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 [16] and GANDiffFace [28], 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 ResNet-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 BUP-T-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 Fig. 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-S0 [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 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 [14,33,35] 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 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 Fig. 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 \( e \in \mathbb{R}^{12} \) and \( z \in \mathbb{R}^{256} \), respectively, containing the subject and its ethnic group features. This schema is shown on the top of Fig. 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 \( X = x_i, y_i, w_j \), 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

\(^3\) https://github.com/zws98/wacv_frcsyn
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 databases DCFace [16] and GANDiffFace [28] 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 Fig. 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 $e_{i,50}$ 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 $112 \times 112$ 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 [16] and GANDiffFace [28] 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 $e_{IR50}$ and $e_{IR101}$ respectively. The layer was independently trained using FFHQ with both models trained, with $e_{IR101}$ as labels and $e_{IR50}$ 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 $\hat{e}_{50}$, the output of the linear layer, and $e_{101}$. In effect, this linear layer allows for an estimated projection of the embedding from the iResNet-50 embedding space into the iResNet-101 embedding space, allowing both embeddings to be evaluated in a common embedding space. The average of these embeddings $e_{mean}$ 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 Fig. 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 [58,82] 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 $112 \times 112$ 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.\footnote{https://github.com/ndido98/frcsyn}

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 Fig. 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 affect classifier performance [84]. Their baseline model was trained employing the CASIA-WebFace database [54]; the proposed model employed both DCFace [16] and GANDiffFace [28]. 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.\footnote{https://github.com/gsarridis/fair-face-verification-with-synthetic-data}

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 Fig. 8. In 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 $1 \times 1$ 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 [16].

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 Fig. 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.

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 [16], 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 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 Fig. 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 Fig. 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 [16] and GANDiffFace [28] synthetic data to reduce bias in current FR technology. As shown in Fig. 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
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

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

### Sub-Task 1.2 (Bias mitigation): Synthetic + Real data

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

### Sub-Task 2.1 (Overall improvement): Synthetic data

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

### Sub-Task 2.2 (Overall improvement): Synthetic + Real data

<table>
<thead>
<tr>
<th>Pos.</th>
<th>Team</th>
<th>AVG [%]</th>
<th>FNMR@FMR = %</th>
<th>AUC [%]</th>
<th>GAP [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CBSR</td>
<td>94.95</td>
<td>10.82 (1)</td>
<td>97.92 (1)</td>
<td>−3.69</td>
</tr>
<tr>
<td>2</td>
<td>LENS</td>
<td>92.40</td>
<td>17.67 (4)</td>
<td>96.58 (5)</td>
<td>−1.63</td>
</tr>
<tr>
<td>3</td>
<td>Idiap</td>
<td>91.74</td>
<td>23.27 (5)</td>
<td>96.87 (4)</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>BOVIFOCR</td>
<td>91.34</td>
<td>16.51 (2)</td>
<td>97.03 (3)</td>
<td>1.77</td>
</tr>
<tr>
<td>5</td>
<td>MeVer</td>
<td>87.66</td>
<td>17.10 (3)</td>
<td>97.40 (2)</td>
<td>−1.57</td>
</tr>
<tr>
<td>6</td>
<td>UNICA</td>
<td>84.86</td>
<td>29.35 (6)</td>
<td>91.46 (6)</td>
<td>−27.43</td>
</tr>
</tbody>
</table>

![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.](image)

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] 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 exclusively used DCFace [16], the majority of teams employed both DCFace [16] and GANDiffFace [28], 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 (Figs. 11 and 12, associated with Sub-Tasks 1.1 and 1.2, respectively) or database (Figs. 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 Fig. 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 (Fig. 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...
P. Melzi et al.

Fig. 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.

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 (Figs. 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 (Fig. 13), the winning team BOVIFOCR-UFRJ 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 (Fig. 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 is 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 generated with GAN-DiffFace [17,28]. 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 = %, 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 = % in both Sub-Tasks 2.1 and 2.2 (–22.59%). This confirms that using DCFace [16] and GANDiff-Face [28] 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-PP [40], and ROF [67]) and across all metrics (i.e., accuracy, FNMR@FMR = %, 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 substantially benefits from the training of FR systems with the proposed synthetic databases, i.e., DCFace and GDANDiff-Face. 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.

CRediT authorship contribution statement

Pietro Melzi: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing original draft, Writing – review & editing. Ruben Tolsana: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – original draft, Writing – review & editing. Christian Rathgeb: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Aythami Morales: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Ivan DeAndres-Tame: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Julián Fierrez: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Javier Ortega-García: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Weisong Zhao: Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Zheyu Yan: Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Xiao-Yu Zhang: Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Jinlin Wu: Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Zhen Lei: Formal analysis, Investigation, Software, Supervision, Validation, Visualization, Roles/Writing – review & editing. Suvidha Tripathi: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Mahak Kothari: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Md Haider Zama: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Debanj Deb: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Bernardo Biessack: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Pedro Vidal: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Roger Granada: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Guillerme Fickel: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Alexander Unnervik: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Anjith George: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Christophe Ecabet: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Hafet Ottoshi Shahreza: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Parsa Rahimi: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Sébastien Marcel: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Ioannis Sarridis: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Christos Koutlis: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Georgia Balsou: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Symeon Papadopoulos: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Fadi Boutros: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Naser Damer: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Gianni Fenu: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Andrea Atzori: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing. Mirko Marra: Formal analysis, Investigation, Software, Validation, Visualization, Roles/Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have included the link to my data/code in the manuscript.
Acknowledgments

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 TReSPAS-ETN (No 860813) and is supported by INTER-ACTION, Spain (PID2021-126521OB-100 MICINN/FEDER), Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2), Spain R&D Agreement DGCC/UAM/PUAM for Biometrics and Cybersecurity, Spain, and Autonomous Community of Madrid (ELLIS Unit Madrid). 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, Germany within their joint support of the National Research Center for Applied Cybersecurity ATHENE. BioLab acknowledges Andrej Pilzer from the NVIDIA AI Technology Center, EMEA, Italy, for his support. MeVer was supported by the EU Horizon Europe project MAMMoth (Grant Agreement 101070285).

References


