

# If It's Not Enough, Make It So: Reducing Authentic Data Demand in Face Recognition through Synthetic Faces

Andrea Atzori<sup>1</sup>, Fadi Boutros<sup>2,3</sup>, Naser Damer<sup>2,3</sup>, Gianni Fenu<sup>1</sup>, Mirko Marras<sup>1</sup>

<sup>1</sup> Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy

<sup>2</sup> Fraunhofer Institute for Computer Graphics Research, Germany

<sup>3</sup> Department of Computer Science, TU Darmstadt, Darmstadt, Germany

**Abstract**—Recent advances in deep face recognition have spurred a growing demand for large, diverse, and manually annotated face datasets. Acquiring authentic, high-quality data for face recognition has proven to be a challenge, primarily due to privacy concerns. Large face datasets are primarily sourced from web-based images, lacking explicit user consent. In this paper, we examine whether and how synthetic face data can be used to train effective face recognition models with reduced reliance on authentic images, and thus mitigating large authentic data collection concerns. First, we explored the performance gap among recent state-of-the-art face recognition models, trained only on synthetic data or authentic data. Then, we deepened our analysis by training a state-of-the-art backbone with various combinations of synthetic and authentic data, gaining insights into optimizing the limited use of the latter for verification accuracy. Finally, we assessed the effectiveness of data augmentation approaches on synthetic and authentic data, with the same goal in mind. Our results highlighted the effectiveness of FR trained on combined datasets, particularly when combined with appropriate augmentation techniques.

## I. INTRODUCTION

Face Recognition (FR) stands as one of the predominant biometric technologies in various applications, ranging from logical access control [27], [42], [4] to portable devices [9]. Existing FR models have reached impressive accuracy thanks to the recent advances in deep learning, including margin-penalty-based softmax losses [14], [42], [4], [27] and neural architectures [21], [8]. This development, however, has been made possible by the public availability of large-sized, identity-labeled, FR training databases [11], [20], [48].

Face databases include millions of images often gathered from the Internet without user consent, thereby giving rise to concerns about their legal and ethical use for FR development. Notably, the approval of the General Data Protection Regulation (GDPR) [1] by the European Union (EU) in 2018 has significantly amplified criticism concerning the privacy-related issues associated with the use of publicly available

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research. This work has been also funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

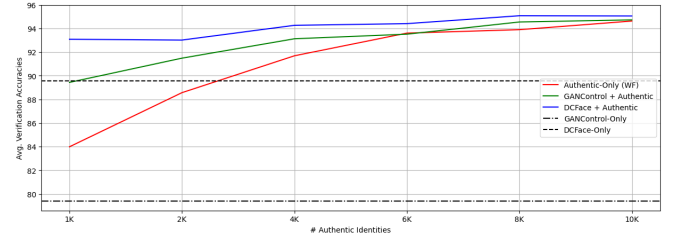


Fig. 1: Average verification accuracy on the testing dataset (y-axis) vs. number of authentic identities from CASIA-WebFace (WF) [44] in the training dataset (x-axis). In all settings, the verification accuracy improved by increasing the number of authentic identities in the training dataset. Also, combining synthetic (black line: DCFace [28], green line: ExFaceGAN(GANControl) [7]) with a limited subset of authentic data improved the FR performance, in comparison to the case where only a limited subset of authentic identities (red line) is used to train FR. The number of synthetic identities in the combined dataset experiments is fixed (10K).

face datasets obtained without adequate consent. The GDPR explicitly affirms individuals' "right to be forgotten" and imposes more stringent regulations concerning the acquisition, distribution, and utilization of biometric data, making adherence to such regulations a key, intricate endeavor when applied to face datasets. Consequently, several databases frequently employed for FR training, including but not limited to [11], [3], [20], have been retracted by their creators, to prevent potential legal issues. These circumstances have raised doubts about the continuity of FR research, as the availability of such fundamental resources has become uncertain.

Synthetic data has recently emerged as a promising alternative to genuine datasets for FR training [10], serving as a viable solution to address the prevailing legal and ethical concerns [36], [6], [8], [2], [29]. This research trajectory has been significantly propelled by the advancement of Deep Generative Models (DGMs), which facilitate the generation of synthetic samples thanks to their ability to capture the probability distribution underlying a face dataset. This generative process can be constrained to various factors, including but not limited to age, facial expression, head pose, and lighting [15], [40], [41], [18]. The majority of DGM techniques employed for the creation of synthetic faces build upon Generative Adversarial Networks (GANs) [19], [15],

[39], [40], [6] or Diffusion Models (DMs) [22], [35], [17].

Deep face recognition methods, such as [30], [8], [36], have increasingly leveraged GANs [15], [24] to generate synthetic data for FR training. However, state-of-the-art (SOTA) FR models trained on synthetic data have often exhibited a drop in verification accuracy, in comparison to FR models trained on authentic data. This discrepancy in performance primarily stems from the training datasets' limited identity discrimination [6] or their minimal intra-class variance [8], [36]. DMs have emerged as a viable alternative to GANs for image synthesis, but this progress comes at the cost of stability and a substantial reduction in training performance. Unfortunately, several questions are still open on the extent to which authentic and synthetic data can be effectively combined to tackle the limitations inherent to both domains.

In this paper, our primary objective is to investigate the suitability of combined (authentic and synthetic) training datasets for model development in FR. This exploration can provide a better understanding of how to address both the performance inadequacies associated with synthetic datasets and the privacy-related concerns that characterize authentic datasets, thereby facilitating the creation of highly accurate FR models, with reduced authentic data demand. As teased in Figure 1, training FR models with a combination of synthetic data and subsets of authentic identities is visibly beneficial in terms of FR performance in comparison to cases where only limited subsets of authentic identities are available, as we assumed that the synthetic data could be generated without any limitation and that small numbers of authentic identities could be collected with proper user consent. Specifically, our novel contribution is threefold:

- We investigated the impact of training models with different ratios of synthetic and authentic identities, under a fixed total number of training images.
- We explored whether a training dataset with a fixed amount of synthetic identities and an increasing number of authentic ones could lead to (at least) comparable accuracy w.r.t. a large, authentic-only dataset.
- We evaluated the influence of training data augmentation on accuracy, by augmenting (i) both the synthetic and authentic subsets and (ii) exclusively the synthetic subset, identifying and discussing the optimal settings.

## II. RELATED WORK

With the ongoing success of deep generative models in generating high-quality and realistic face images [40], [16], [22], [35], several recent works proposed the use of synthetic data in FR development [36], [8], [2], [6], [5], [28], [10]. The synthetic data for FR training can be categorized based on the type of underlying generative model under three categories, GAN-Based [36], [8], [6], [7], digital rendering [2], and diffusion-based [31], [5], [28]

SynFace [36] and SFace [6] were among the earliest works that proposed the use of GAN-generated face images for FR training. To enhance intra-class diversity in synthetic data, SynFace proposed to generate synthetic data using an attribute-conditional GAN model, i.e., DiscoFaceGAN [16],

and perform identity and domain mixup. SFace also analyzed the impact of Style-GAN [26] training under class conditional settings. Furthermore, the extent to which transferring knowledge from the pretrained model on authentic data improves the performance of synthetic-based FR was analyzed, going beyond FR training only on synthetic data. Usynth-Face [8] leveraged unlabelled synthetic data for unsupervised FR training. The work also presented an extensive data augmentation method based on GANs [16] along with color and geometric transformation [12]. Conversely, ExFaceGAN [7] presented a framework to disentangle identity information in learned latent spaces of unconditional GANs, aiming at generating multiple samples of any synthetic identity.

DigiFace-1M [2] employed a digital rendering pipeline, leveraging facial geometry models, a diverse array of textures, hairstyles, and 3D accessories, along with robust data augmentation techniques during training. However, it comes at a considerable computational cost during the rendering process. DigiFace-1M also proposed to combine synthetic with authentic data during FR training to improve the verification accuracy of synthetic-based FR using a small and fixed number of authentic identities.

Very recently, IDiff-Face [5] and DCFace [28] adopted diffusion models to generate synthetic data for FR training, achieving SOTA verification accuracy for synthetic-based FR. Specifically, IDiff-Face included fuzziness in the identity condition to induce variations in the generated data. DCFace proposed a two-stage generative framework. In the first stage, an image of a novel identity using an unconditional diffusion model is generated and an image style from the style bank is selected. In the second stage, the generated images and style are mixed using a dual conditional diffusion model. GANDiffFace [31] proposed to combine GAN and diffusion models to generate synthetic data for FR training that exhibit certain attributes such as controllable demographic distribution.

Although these approaches [28], [5], [2], [6], [36] achieved relatively high verification accuracy when they were evaluated on authentic datasets, their performance is lower than the one obtained by FR models trained on authentic data [14]. To overcome this challenge, SFace [6] applied knowledge distillation on the embedding level between the model trained on the authentic data and the one trained on synthetic data. DigiFace [2] and DCFace [28] combined their synthetic datasets with a subset of the authentic CASIA-WebFace dataset [44], aiming at improving the verification accuracy of synthetic-based FR. Both methods demonstrated that synthetic-based verification accuracy can be improved using a limited set of authentic data. Wang et al. [43] proposed to combine synthetic data from SynFace [36] with a small and limited subset of authentic data. The work proposed to utilize the same backbone to extract features from synthetic and authentic identity and then learn multi-class classification through two classification layers for synthetic and authentic classes. However, these works [2], [28], [43] only utilized a subset of authentic data, with a fixed number of authentic identities, and they did not investigate the performance gain

when different ratios of synthetic and authentic identities combined in the FR training. Recently, several challenges and competitions have been organized in conjunction with top venues, aiming at promoting privacy-friendly synthetic-based FR development. FRCSyn competition [32] was organized at WACV 2024, aiming to explore the use of synthetic data in FR training and to attract the development of solutions for synthetic-based FR. The challenge considered two main tasks, training FR only with synthetic data and training FR with both synthetic and authentic data. The achieved results of the top-performing solutions from FRCSyn [32] competition are further investigated and reported in [33]. A second edition of FRCSyn [13] was organized in conjunction with CVPR 2024. Also, very recently SDFR [38] competition was organized in conjunction with FG 2024, attracting developing and promoting solutions for synthetic-based FR.

Unlike previous research that combined fixed, (in terms of numbers of images and identities) sets of synthetic and authentic data [2], [28], [43], [32], we investigate in detail FR performances using different combinations of authentic and synthetic data. We first fixed the total number of identities in the training dataset and investigated the performance of different ratios of synthetic and authentic identities. Then, we investigated the FR performance using a fixed number of synthetic identities with different subsets of authentic data. Finally, we investigated the impact of data augmentation on synthetic and authentic subsets.

### III. METHODOLOGY

This work aims, at first, to investigate the performance gaps between FR models trained on synthetic datasets and the ones trained on privacy-sensitive authentic datasets. Then, we explored the performance gains of synthetic-based FR by including a limited set of authentic identities in the training dataset. Finally, we investigate the impact of data augmentation on synthetic-based FR verification accuracy.

#### A. Datasets Preparation

The presented investigations in this paper are performed using four training datasets, two authentic and two synthetic.

As the authentic data, we adopted the well-known CASIA-WebFace [44] and MS1MV2 [14] datasets. CASIA-WebFace [44] consists of 0.5M images of 10K identities. MS1MV2 is the refined version of the MS-Celeb-1M dataset [20]. MS1MV2 contains 5.8M images of 85K identities. To provide comparable results across FR trained on authentic-only data and on synthetic and authentic combinations, we utilized a subset of 10K identities of MS1MV2 in the conducted experiments. This subset will be noted as M2-S and contains the first 10K identities from MS1MV2.

The synthetic datasets come from ExFaceGAN [7] and DCFace [28]. Each of these datasets contains 0.5M images of 10K identities, i.e., 50 images per identity. The ExFaceGAN dataset is generated using an identity disentanglement approach [7] on pretrained GAN-Control [40] which is trained on the FFHQ dataset [25]. From now on, the ExFaceGAN(GAN-Control) dataset will be noted as GC.

The second synthetic dataset is generated via DCFace [28]. The latter is based on two-stage diffusion models. In the first stage, a high-quality face image of a novel identity is generated using unconditional diffusion models [22] trained on FFHQ [25], with an image style randomly selected from a style bank. In the second stage, the generated images and styles from stage 1 are mixed using a dual conditional diffusion model [22] trained on CASIA-WebFace [44] to generate an image with a specific identity and style. The dataset generated by DCFace will be noted as DC. All images are aligned and cropped to  $112 \times 112$  using landmarks obtained by the Multi-task Cascaded Convolutional Networks (MTCNN) [45], following [14]. All the training and testing images are normalized to have values between -1 and 1.

#### B. Data Mixing

Given a synthetic dataset  $S$  with a number of identities equal to  $N$  and an authentic dataset  $A$  with a number of identities equal to  $M$ , we combined these two datasets into one dataset  $C = s \cup a$ , where  $s \subseteq S$  and  $a \subseteq A$  are subsets of  $S$  and  $A$ , respectively, with a total number of identities equal to  $H$ , as follows. In the first evaluation setting, we set  $H$  to be equal to  $M$  and sampled  $m < M$  from the authentic dataset and  $M - m < N$  from the synthetic dataset. We evaluated several  $m \in [1K, 2K, 3K, 4K, 5K, 6K, 7K, 8K, 9K]$ . In this stage, we aim to boost the verification accuracy of synthetic-based FR by combining synthetic with authentic face images into one dataset. We fixed the total number of identities, in this case, to provide a fair comparison across the considered settings. In the second stage, we considered the whole synthetic dataset  $S$  and combined it with a subset  $a \subseteq A$ . The sampled  $a$  contains a number of identities equal to 1K, 2K, 4K, 6K, 8K or 10K. In this scenario, we assumed that the synthetic data could be generated without any limitation and that a small number of authentic identities could be collected with proper user consent.

#### C. Data Augmentation

Considering the constrained range of visual diversity in synthetic face data, in comparison to large and realistic appearance variations in authentic data [10], [8], [28], we evaluated in this work the impact of augmenting the FR training dataset on FR verification accuracy, as follows.

a) *Baseline (no augmentation)*: We refer to procedures commonly used for training FR models, which involve augmenting the training dataset solely through the application of a random horizontal flip, with a probability of 0.5.

b) *Random augmentation*: We evaluated the impact of enriching the FR training dataset on the FR verification accuracy by employing random augmentation techniques [12], [8], denoted as RandAug. The augmentation space of RandAug includes the following color and geometric transformations [8]: horizontal-flipping, rotation, x-axis translation, y-axis translation, x-axis shear, y-axis shear, sharpness adjusting, auto-contrasting, adjusting contrast, solarizing, posterizing, equalizing, adjusting color, adjusting brightness, performing a resized crop, and converting to

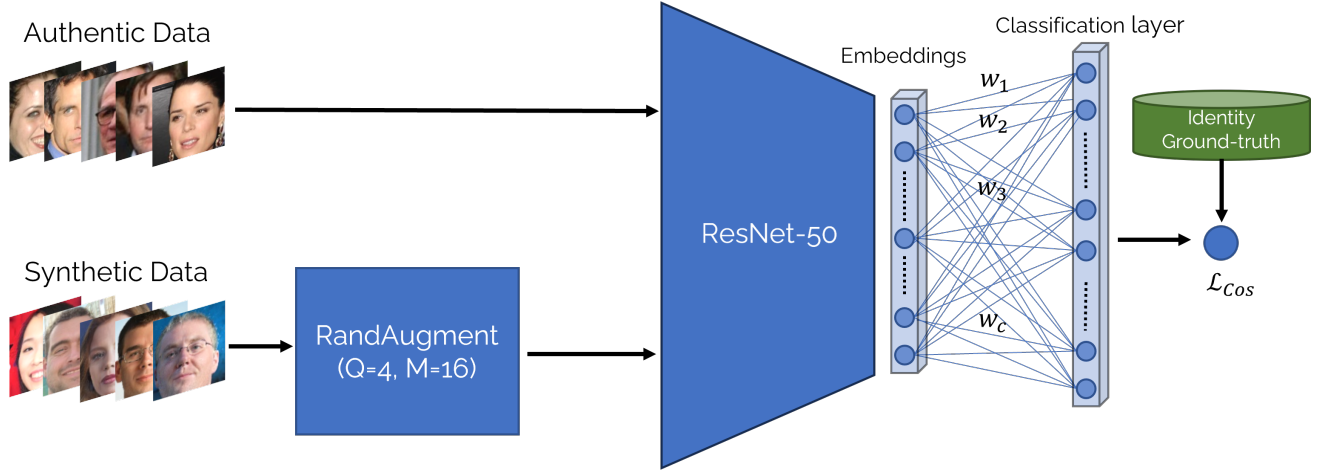


Fig. 2: **FR Training Paradigm Overview.** Subsets of authentic and synthetic data are combined to form the training dataset. During the training phase, only synthetic data is augmented with RandAugment, as well discussed in Section IV. the utilized network architecture in all settings is ResNet50 [21] trained with CosFace [42] loss.

grayscale. RandAug includes two hyper-parameters,  $Q$  and  $M$ , to select the number of operations  $Q$  and the magnitude  $M$  of each transformation, respectively. We set  $M$  to 16 and  $Q$  to 4, following [8]. In our analysis, we first experimented with the application of a random augmentation on both subsets of the training dataset, i.e., authentic and synthetic. In a later stage, we only augmented the synthetic subset (this approach is visually depicted in Figure 2).

#### D. Model Training

We relied on FR models having the well-known ResNet50 architecture [21] as the backbone and CosFace [42] as the loss function; the latter is defined as:

$$L_{CosFace} = \frac{1}{N} \sum_{i \in N} -\log \frac{e^{s(\cos(\theta_{y_i}) - m)}}{e^{s(\cos(\theta_{y_i}) - m)} + \sum_{j=1, j \neq y_i}^c e^{s(\cos(\theta_j))}} \quad (1)$$

where  $c$  is the number of classes (identities),  $N$  is the batch size,  $m$  is the margin penalty applied on the cosine angle  $\cos(\theta_{y_i})$  between the feature representation  $x_i$  of the sample  $i$  and its class center  $y_i$ ,  $s$  is scale parameters. In all conducted experiments, the margin  $m$  is set to 0.35 and the scale parameter  $s$  to 64, following [42].

During the training, Stochastic Gradient Descent (SGD) is employed as an optimizer with an initial learning rate of 0.1. The learning rate is divided by 10 at 22, 30, and 40 training epochs [6]. The models have been trained for 40 epochs with a total batch size of 512.

#### E. Evaluation Benchmarks

We evaluated the trained FR models on the well-known benchmark verification pairs that accompany the following datasets: Labeled Faces in the Wild (LFW) [23], AgeDB-30 [34], Celebrities in Frontal to Profile in the Wild (CFP-FP) [37], Cross-Age LFW [47], and Cross-Pose LFW [46]. Results for all benchmarks are reported as verification accuracy in %, thus adhering to their official evaluation protocol.

## IV. EXPERIMENTAL RESULTS

Our experiments analyzed, at first, the performance gaps between FR models trained on authentic-only and synthetic-only data created with SOTA generative methods (Section IV-A). Then, we explored how combining different ratios of authentic and synthetic data can influence FR performance (Section IV-B) and whether the combination of authentic and synthetic data can actually overcome the limitations of both data domains (Section IV-C). Finally, we investigate the impact of augmentation on synthetic FR performance (Section IV-D). Our experiments were conducted on the GC and DC synthetic datasets.

#### A. Performance gap between FR trained on an authentic dataset and ones trained on synthetic dataset

Table I presents the verification accuracies on five benchmarks achieved by models trained on GC, DC and WF datasets. The table also presents the achieved accuracies of FR models trained on different subsets of authentic WF. As expected, the model trained on authentic WF (10K) outperformed the models trained on synthetic GC (10K) and DC (10K) [28]. The average verification accuracies achieved by the model trained on WF (10K) is 94.63% and the achieved ones by the models trained on GC and DC are 79.38% and 89.56%, respectively. It can be also observed that the model trained on a subset of 3K identities of WF (average accuracies of 90.83%) outperformed models trained on 10K identities of synthetic DC and GC datasets. This large performance gap can be reduced by combining synthetic with a small set of authentic identities as will be presented in the next section.

#### B. Impact of combining synthetic data with a small subset of authentic data on FR performance

The third and fourth groups of results in Table I present the achieved FR verification accuracies of FR models trained on different datasets that are built on by combining subsets

Train Data	Id/Img.	LFW (↑)	AgeDB-30 (↑)	CFP-FP (↑)	CA-LFW (↑)	CP-LFW (↑)	Avg (↑)
GC (10K)	10K / 50	93.22	77.60	73.03	82.25	70.80	79.38
DC (10K)	10K / 50	<b>98.55</b>	<b>89.70</b>	<b>85.33</b>	<b>91.60</b>	<b>82.62</b>	<b>89.56</b>
WF (1K)	1K / 45	95.60	81.97	80.81	84.78	76.83	83.99
WF (2K)	2K / 45	97.23	86.30	87.88	89.36	82.07	88.57
WF (3K)	3K / 45	98.62	90.32	90.04	90.83	84.30	90.82
WF (4K)	4K / 45	97.76	91.32	91.70	91.72	85.98	91.69
WF (5K)	5K / 45	99.15	92.75	93.24	92.40	87.38	92.98
WF (6K)	6K / 46	99.25	93.80	93.36	92.85	88.77	93.61
WF (7K)	7K / 46	99.37	93.58	94.07	93.13	89.05	93.84
WF (8K)	8K / 46	99.35	<b>94.71</b>	95.01	93.18	89.25	93.90
WF (9K)	9K / 46	99.42	94.32	95.03	93.73	89.70	94.44
WF (10K)	10K / 46	<b>99.55</b>	94.55	<b>95.31</b>	<b>93.78</b>	<b>89.95</b>	<b>94.63</b>
GC (1K) $\cup$ WF (9K)	10K / 46	<b>99.43</b>	<b>94.62</b>	95.10	<b>93.12</b>	<b>90.02</b>	<b>94.46</b>
GC (2K) $\cup$ WF (8K)	10K / 47	99.35	94.07	<b>95.19</b>	93.05	89.78	94.29
GC (3K) $\cup$ WF (7K)	10K / 47	99.35	93.58	94.90	92.98	89.73	94.11
GC (4K) $\cup$ WF (6K)	10K / 48	99.38	93.68	94.10	92.67	89.43	93.85
GC (5K) $\cup$ WF (5K)	10K / 48	99.30	93.42	93.34	92.60	89.07	93.55
GC (6K) $\cup$ WF (4K)	10K / 48	99.17	92.18	92.87	92.45	88.05	92.94
GC (7K) $\cup$ WF (3K)	10K / 49	98.93	91.93	91.97	91.82	86.98	92.33
GC (8K) $\cup$ WF (2K)	10K / 49	98.62	90.18	89.60	90.45	85.68	90.91
GC (9K) $\cup$ WF (1K)	10K / 50	98.17	89.15	87.49	88.78	83.30	89.38
DC (1K) $\cup$ WF (9K)	10K / 46	<b>99.50</b>	94.38	95.43	<b>93.63</b>	90.10	94.61
DC (2K) $\cup$ WF (8K)	10K / 47	99.45	94.65	<b>95.56</b>	93.55	<b>90.18</b>	<b>94.68</b>
DC (3K) $\cup$ WF (7K)	10K / 47	99.33	<b>94.75</b>	95.14	93.52	90.13	94.57
DC (4K) $\cup$ WF (6K)	10K / 48	99.43	94.28	95.01	93.42	90.45	94.52
DC (5K) $\cup$ WF (5K)	10K / 48	99.37	94.25	94.90	93.33	89.57	94.28
DC (6K) $\cup$ WF (4K)	10K / 48	99.27	94.10	94.41	93.33	89.52	94.13
DC (7K) $\cup$ WF (3K)	10K / 49	99.35	93.55	94.27	92.87	89.03	93.81
DC (8K) $\cup$ WF (2K)	10K / 49	99.17	93.27	93.77	92.68	88.10	93.40
DC (9K) $\cup$ WF (1K)	10K / 50	99.03	93.20	92.54	92.23	87.32	92.86

TABLE I: Verification accuracy on size-fixed training sets (authentic dataset: WF). First rows group: accuracy on full synthetic datasets. Second rows group: accuracy on different subsets of authentic WF. Third and fourth rows group: accuracy on different subsets of authentic WF and synthetic GC and DC datasets, respectively. It can be noticed that the accuracy of synthetic-based models improved by adding limited subsets of authentic data. Highest accuracy scores for each row group in **bold**.

of authentic and synthetic identities. The total number of identities in each case is fixed to be equal to 10K, while the ratio of synthetic and authentic identities varies.

Combining only 1K authentic identities with 9K identities from DC enhanced the FR model accuracy w.r.t. the DC synthetic baseline of 0.49% on LFW, 3.90% on AgeDB-30, 8.45% on CFP-FP, 0.69% on CA-LFW and 5.69% on CP-LFW, with an overall average improvement of 3.68% across datasets. Combining the same 1K authentic identities with 9K synthetic identities from GC resulted in even more noticeable improvements over the respective baseline: 5.31% on LFW, 14.88% on AgeDB-30, 20.35% on CFP-FP, 7.94% on CA-LFW and 17.66% on CP-LFW, with an overall 12.60% average improvement.

As expected, a consistent improvement was observed on all benchmarks as the ratio of involved authentic identities increased. Combining subsets of identities from GC and WF (third group of rows) steadily improved the average verification accuracy from 89.38% (9K synthetic and 1K authentic identities) to 94.46% (1K synthetic and 9K authentic identities). The same trend was also noted while combining DC subsets with the same authentic identities (fourth group of rows), with an increase in the average accuracy from 92.86% to 94.68%.

### C. Impact of combining fixed numbers of synthetic identities with different subsets of authentic identities on FR performance

Tables II and III present the FR verification accuracy achieved by models trained on a fixed number of synthetic identities (10K) combined with a subset of authentic identities

Train Data	Id/Img.	LFW (↑)	AgeDB-30 (↑)	CFP-FP (↑)	CA-LFW (↑)	CP-LFW (↑)	Avg (↑)
GC (10K)	10K / 50	93.22	77.60	73.03	82.25	70.80	79.38
DC (10K)	10K / 50	98.55	89.70	85.33	91.60	82.62	89.56
WF (1K)	1K / 45	95.60	81.97	80.81	84.78	76.83	83.99
GC (10K) $\cup$ WF (1K)	11K / 49	97.95	88.80	87.24	89.68	83.57	89.44
DC (10K) $\cup$ WF (1K)	11K / 49	99.13	93.45	92.83	92.43	87.63	<b>93.09</b>
WF (2K)	2K / 45	97.23	86.30	87.88	89.36	82.07	88.57
GC (10K) $\cup$ WF (2K)	12K / 49	98.67	90.45	90.90	91.30	85.92	91.49
DC (10K) $\cup$ WF (2K)	12K / 49	99.15	93.03	92.61	92.52	87.78	<b>93.02</b>
WF (4K)	4K / 45	97.76	91.32	91.70	91.72	85.98	91.69
GC (10K) $\cup$ WF (4K)	14K / 49	99.27	92.57	93.06	92.32	88.45	93.13
DC (10K) $\cup$ WF (4K)	14K / 49	99.45	94.47	94.64	93.20	88.58	<b>94.27</b>
WF (6K)	6K / 46	99.25	93.80	93.36	92.85	88.77	93.61
GC (10K) $\cup$ WF (6K)	16K / 49	99.38	93.30	94.39	91.12	89.43	93.52
DC (10K) $\cup$ WF (6K)	16K / 49	99.42	94.35	94.59	93.73	89.97	<b>94.41</b>
WF (8K)	8K / 46	99.35	94.71	95.01	93.18	89.25	93.90
GC (10K) $\cup$ WF (8K)	18K / 48	99.46	94.23	95.23	93.37	90.47	94.55
DC (10K) $\cup$ WF (8K)	18K / 49	99.55	95.15	95.88	93.83	90.98	<b>95.08</b>
WF (10K)	10K / 46	99.55	94.55	95.31	93.78	89.95	94.63
GC (10K) $\cup$ WF (10K)	20K / 48	99.52	94.53	95.26	93.82	90.53	94.73
DC (10K) $\cup$ WF (10K)	20K / 49	99.50	94.92	95.71	94.08	90.98	<b>95.06</b>

TABLE II: Verification accuracy of models trained on synthetic datasets combined with different subsets authentic WF. First group: accuracy of models trained on synthetic GC and DC datasets, each with 10K identities. The first row in each group: accuracy of models trained by using different subsets of authentic WF only. Second and third rows in each group: accuracy of models obtained by combining different subsets of authentic WF with 10K synthetic identities from GC and DC datasets, respectively. It can be noticed that the accuracy of synthetic-based models improved by adding a limited subset of authentic data. Moreover, combining synthetic with authentic datasets (e.g., GC (10K)  $\cup$  WF (10K)) led to higher accuracy, in comparison to the case where FR is only trained on authentic WF dataset/subsets. For each row group, the highest accuracy score is highlighted in **bold**.

ties (1K, 2K, 4K, 6K, 8K, or 10K) from the WF and MS-2 authentic datasets, respectively.

Training FR models on subsets of authentic identities combined with synthetic data consistently improved the verification accuracy as the number of authentic identities in the training dataset increased. Models trained on a subset of 10K identities from WF combined with 10K identities from GC achieved average verification accuracy improvements from 89.94% (1K authentic identities) to 94.73% (10K authentic identities). Similarly, training on a subset of 10K identities from WF combined with DC improved the average verification accuracy from 93.09% (1K authentic identities) to 95.06% (10K authentic identities). Notably, the performance obtained by training the model on DC combined with 10K authentic identities (95.06%) was slightly worse than the one obtained by combining it with 8K authentic identities (95.08%) from WF.

Training FR models on M2-S combined with either GC or DC led to similar observations, reporting verification accuracy improvements from 83.66% to 95.08% and from 90.72% to 92.75%, respectively. As visually emerging from Figure 3, all the settings in which the models were trained on combined datasets achieved higher verification accuracy than their respective synthetic baseline, with best improvements of about 30%.

As visually illustrated in Figure 3, increasing the number of authentic identities did improve, as expected, the FR performances with the best performance achieved by using 10K authentic identities, i.e. utilizing the whole authentic dataset. However, similar performances, or even higher, can



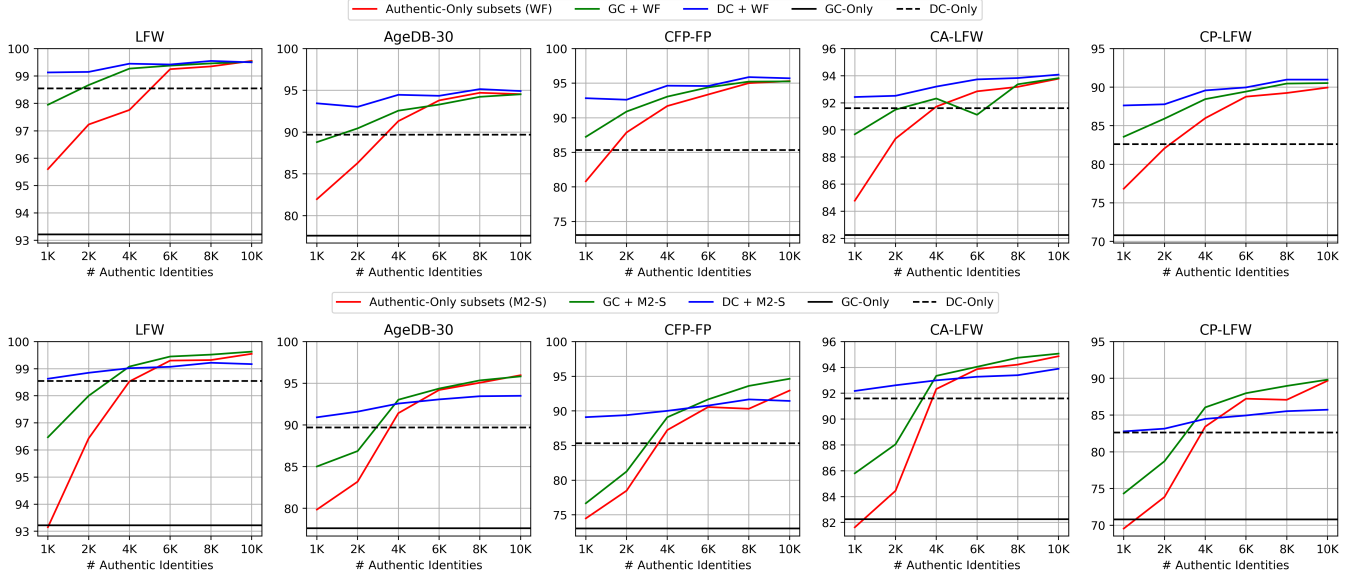


Fig. 3: Average verification accuracy on testing data (y-axis) vs. number of authentic identities in training data (x-axis). The top row figures refer to authentic data sampled from WF, second row figures refer to authentic data sampled from M2-S. The verification accuracy improved by increasing the number of authentic identities in the training dataset. Also, combining synthetic with a limited subset of authentic datasets (black and green lines) improved the FR performance in comparison to the case where only a limited subset of authentic identities (red lines) is used to train FR. The number of synthetic identities in the combined dataset experiments is fixed (10K). The results in these plots correspond to the ones reported in Table II Table III.

Train Data	Id/Img.	LFW (↑)	AgeDB-30 (↑)	CFP-PP (↑)	CA-LFW (↑)	CP-LFW (↑)	Avg (↑)
GC (10K)	10K / 50	93.22	77.60	73.03	82.25	70.80	79.38
DC (10K)	10K / 50	98.55	89.70	85.33	91.60	82.62	89.56
M2-S (10K)	10K / 45	93.15	79.83	74.49	81.62	69.55	79.73
GC (10K) ∪ M2-S (1K)	11K / 49	96.47	85.02	76.67	85.80	74.33	83.66
DC (10K) ∪ M2-S (1K)	11K / 49	98.63	90.92	89.10	92.18	82.78	<b>90.72</b>
M2-S (2K)	2K / 45	96.43	83.20	78.49	84.45	73.85	83.28
GC (10K) ∪ M2-S (2K)	12K / 49	98.00	86.87	81.26	88.05	78.72	86.58
DC (10K) ∪ M2-S (2K)	12K / 49	98.85	91.60	89.39	92.62	83.15	<b>91.12</b>
M2-S (4K)	4K / 45	98.53	91.42	87.24	92.33	83.43	90.59
GC (10K) ∪ M2-S (4K)	14K / 49	99.08	93.03	89.09	93.35	86.05	<b>92.12</b>
DC (10K) ∪ M2-S (4K)	14K / 49	99.02	92.57	90.01	93.00	84.48	91.82
M2-S (6K)	6K / 46	99.30	94.20	90.56	93.87	87.23	93.03
GC (10K) ∪ M2-S (6K)	16K / 49	99.45	94.37	91.66	94.05	87.98	<b>93.50</b>
DC (10K) ∪ M2-S (6K)	16K / 49	99.07	93.07	90.77	93.28	84.95	92.23
M2-S (8K)	8K / 46	99.32	95.07	90.31	94.22	87.08	93.20
GC (10K) ∪ M2-S (8K)	18K / 48	99.52	95.35	93.61	94.75	88.97	<b>94.44</b>
DC (10K) ∪ M2-S (8K)	18K / 49	99.22	93.45	91.66	93.40	85.53	92.65
M2-S (10K)	10K / 46	99.55	95.97	92.94	94.87	89.67	94.60
GC (10K) ∪ M2-S (10K)	20K / 48	99.63	95.83	94.64	95.07	89.82	<b>95.00</b>
DC (10K) ∪ M2-S (10K)	20K / 49	99.17	93.50	91.44	93.90	85.73	92.75

TABLE III: Verification accuracy of models trained on synthetic datasets combined with different subsets authentic M2-S. The first group: accuracy of models trained on synthetic GC and DC datasets, each with 10K identities. The first row in each group: accuracy of models trained by using different subsets of authentic M2-S only. Second and third rows in each group: accuracy of models obtained by combining different subsets of authentic M2-S with 10K synthetic identities from GC and DC datasets, respectively. It can be noticed that the accuracy of synthetic-based models improved by adding a limited subset of authentic data. Moreover, combining synthetic with authentic datasets (e.g., GC (10K) ∪ M2-S (10K)) led to higher accuracy, in comparison to the case where FR is only trained on authentic M2-S dataset/subsets. For each row group, the highest accuracy score is highlighted in **bold**.

be achieved by utilizing a small set of authentic identities, e.g. 4K authentic identities, combined with 10K synthetic identities from either GC or DC.

Training the models on GC combined with authentic identity subsets from WF led to 6.49%, 3.30%, 1.57%, −0.10%, 0.69% and 0.11% improvements in terms of average verification accuracy. Similarly, training on DC combined WF subsets led to 10.83%, 5.02%, 2.81%, 0.85%, 1.26% and 0.45% improvements in terms of average verification accuracy. Each comparison was made by taking into consideration an authentic subset involving 1K, 2K, 4K, 6K, 8K, and 10K identities, respectively. It also emerges that, under a given amount of authentic identities, the impact of DC is more evident than that of GC.

Following the same protocol, training FR models on DC (GC) combined with authentic identity subsets from M2-S led to average verification accuracy improvements of 13.78% (4.93%), 9.41% (3.96%), 1.36% (1.69%), −0.86% (0.51%), −0.59% (1.33%), and −1.96% (0.42%). These results further confirm that training on a large synthetic dataset combined with small authentic identity subsets results in a substantial increase in verification accuracy and that DC identities are more effective than those from GC. However, differently from what we observed for WF, the performance achieved by training a FR model on DC combined with 6K or more authentic identities from M2-S not only achieved lower performances compared to combining with GC identities but also adversely affected performance with respect to training only on authentic data. This observation sheds light on

Train Data	Method	LFW (↑)	AgeDB-30 (↑)	CFP-PP (↑)	CA-LFW (↑)	CP-LFW (↑)	Avg (↑)
GC (10K) ∪ WF (1K)	-	97.95	88.80	87.24	89.68	83.57	89.44
GC (10K) ∪ WF (1K)	RAug-All	98.27	88.87	89.31	89.35	83.82	<b>89.92</b>
GC (10K) ∪ WF (1K)	RAug-Synt	98.25	88.47	88.76	89.68	83.47	89.73
GC (10K) ∪ WF (2K)	-	98.67	90.45	90.90	91.50	85.92	<b>91.49</b>
GC (10K) ∪ WF (2K)	RAug-All	98.80	89.88	90.52	90.67	85.65	91.10
GC (10K) ∪ WF (2K)	RAug-Synt	98.83	90.61	89.60	91.30	85.75	91.22
GC (10K) ∪ WF (4K)	-	99.27	92.57	93.06	92.32	88.45	93.13
GC (10K) ∪ WF (4K)	RAug-All	99.02	92.23	92.69	92.07	87.67	92.74
GC (10K) ∪ WF (4K)	RAug-Synt	99.10	92.28	93.03	92.45	88.92	<b>93.16</b>
GC (10K) ∪ WF (6K)	-	99.38	93.30	94.39	91.12	89.43	93.52
GC (10K) ∪ WF (6K)	RAug-All	99.37	92.80	94.33	92.78	88.81	93.62
GC (10K) ∪ WF (6K)	RAug-Synt	99.35	93.13	94.77	92.75	89.97	<b>93.79</b>
GC (10K) ∪ WF (8K)	-	99.46	94.23	95.23	93.37	90.47	<b>94.55</b>
GC (10K) ∪ WF (8K)	RAug-All	99.32	93.25	94.47	92.35	89.10	93.70
GC (10K) ∪ WF (8K)	RAug-Synt	99.48	94.25	95.13	93.48	90.40	<b>94.55</b>
GC (10K) ∪ WF (10K)	-	99.52	94.53	95.26	93.82	90.53	<b>94.73</b>
GC (10K) ∪ WF (10K)	RAug-All	99.48	93.97	95.27	93.53	90.38	94.53
GC (10K) ∪ WF (10K)	RAug-Synt	99.42	94.37	95.40	93.57	90.60	<b>94.62</b>

TABLE IV: Impact of augmenting the training dataset on FR verification accuracies (authentic dataset: WF; synthetic dataset: GC). For different subsets of the authentic dataset combined with the synthetic dataset, we reported the results under three settings: no augmentation, both synthetic and authentic samples are randomly augmented (RAug-All) and only synthetic samples are augmented (RAug-Synt). For each group of experimental settings, the best average accuracy is in **bold**. The authentic identities are sampled from WF and the synthetic ones are from GC.

the importance of devising future work about methods for creating synthetic data functional to improve model accuracy, in an end-to-end manner.

#### D. Impact of data augmentation

Tables IV and V present the achieved verification accuracies by FR models trained on different subsets of authentic and synthetic data and by augmenting the training dataset with RandAug presented in Section III-C.

The reported results demonstrated that augmenting both authentic and synthetic subsets (RAug-All settings) used to train FR models leads to overall lower average verification accuracy w.r.t. the case in which no augmentation is applied to the training data. On average, the verification accuracy of models trained on a fully augmented training dataset (with subsets from WF as authentic identities), was impacted by  $-0.01\%$ ,  $+0.23\%$ ,  $-0.11\%$ ,  $+0.21\%$ ,  $-0.18\%$  and  $-0.14\%$  while including synthetic identities from DC and by  $+0.54\%$ ,  $-0.43\%$ ,  $-0.43\%$ ,  $+0.10\%$ ,  $-0.92\%$  and  $-0.22\%$  while including synthetic identities from GC. Each comparison was made by taking into consideration combined and authentic training sets involving 1K, 2K, 4K, 6K, 8K, and 10K authentic identities, respectively.

A slight but consistent improvement, w.r.t. the baselines trained on data with no augmentation, is achieved by augmenting only the synthetic training subset (RAug-Synt settings). In this scenario, FR models trained on authentic identity subsets from WF combined with augmented synthetic samples from DC led to  $-0.01\%$ ,  $+0.49\%$ ,  $+0.10\%$ ,  $+0.27\%$ ,  $-0.06\%$  and  $-0.06\%$  changes in average verification accuracy. Similarly, training FR models on authentic identity subsets from WF combined with augmented synthetic samples from GC led to  $+0.31\%$ ,  $-0.30\%$ ,  $+0.03\%$ ,  $+0.52\%$ ,  $0\%$ ,  $+0.17\%$  improvements in term of average verification accuracy.

Train Data	Method	LFW (↑)	AgeDB-30 (↑)	CFP-PP (↑)	CA-LFW (↑)	CP-LFW (↑)	Avg (↑)
DC (10K) ∪ WF (1K)	-	99.13	93.45	92.83	92.43	87.63	<b>93.09</b>
DC (10K) ∪ WF (1K)	RAug-All	99.15	93.23	92.81	92.40	87.82	93.08
DC (10K) ∪ WF (1K)	RAug-Synt	99.13	93.43	92.80	92.38	87.70	<b>93.09</b>
DC (10K) ∪ WF (2K)	-	99.15	93.03	92.61	92.52	87.78	93.02
DC (10K) ∪ WF (2K)	RAug-All	99.18	93.02	93.57	92.57	87.83	93.23
DC (10K) ∪ WF (2K)	RAug-Synt	99.25	93.25	93.48	92.86	88.46	<b>93.46</b>
DC (10K) ∪ WF (4K)	-	99.45	94.47	94.64	93.20	89.58	94.27
DC (10K) ∪ WF (4K)	RAug-All	99.33	94.32	94.86	92.93	89.37	94.16
DC (10K) ∪ WF (4K)	RAug-Synt	99.37	94.33	94.60	93.67	89.83	<b>94.36</b>
DC (10K) ∪ WF (6K)	-	99.42	94.35	94.59	93.73	89.97	94.41
DC (10K) ∪ WF (6K)	RAug-All	99.40	94.42	95.40	93.53	90.28	94.61
DC (10K) ∪ WF (6K)	RAug-Synt	99.50	94.32	95.06	93.95	90.48	<b>94.66</b>
DC (10K) ∪ WF (8K)	-	99.55	95.15	95.88	93.83	90.98	<b>95.08</b>
DC (10K) ∪ WF (8K)	RAug-All	99.55	95.02	95.69	93.78	90.53	94.91
DC (10K) ∪ WF (8K)	RAug-Synt	99.62	94.77	95.73	94.02	90.98	<b>95.02</b>
DC (10K) ∪ WF (10K)	-	99.50	94.92	95.71	94.08	90.98	95.04
DC (10K) ∪ WF (10K)	RAug-All	99.55	94.72	95.71	94.07	90.48	94.91
DC (10K) ∪ WF (10K)	RAug-Synt	99.55	94.85	95.71	94.10	91.11	<b>95.06</b>

TABLE V: Impact of augmenting the training dataset on FR verification accuracies (authentic dataset: WF; synthetic dataset: DC). For different subsets of the authentic (WF) dataset combined with the synthetic dataset (DC), we reported the results under three settings: no augmentation, both synthetic and authentic samples are randomly augmented (RAug-All) and only synthetic samples are augmented (RAug-Synt). For each group of experimental settings, the best average accuracy is in **bold**. Notably, augmenting only the synthetic samples led to higher verification accuracies in most of the settings, in comparison to cases where both authentic and synthetic samples are augmented.

Train Data	Method	LFW (↑)	AgeDB-30 (↑)	CFP-PP (↑)	CA-LFW (↑)	CP-LFW (↑)	Avg (↑)
GC (10K) ∪ M2-S (1K)	-	96.47	83.02	76.67	85.80	74.33	83.66
GC (10K) ∪ M2-S (1K)	RAug-Synt	96.27	82.15	77.64	85.22	75.70	83.40
GC (10K) ∪ M2-S (2K)	-	98.00	86.87	81.26	88.05	78.72	<b>86.58</b>
GC (10K) ∪ M2-S (2K)	RAug-Synt	96.33	83.05	78.74	86.10	77.73	84.43
GC (10K) ∪ M2-S (4K)	-	99.08	93.03	89.09	93.35	86.05	<b>92.12</b>
GC (10K) ∪ M2-S (4K)	RAug-Synt	97.98	86.60	83.00	89.35	81.72	87.53
GC (10K) ∪ M2-S (6K)	-	99.45	94.37	91.66	94.05	87.98	<b>93.50</b>
GC (10K) ∪ M2-S (6K)	RAug-Synt	98.63	88.70	84.80	91.03	85.40	89.71
GC (10K) ∪ M2-S (8K)	-	99.52	95.35	93.61	94.75	88.97	<b>94.44</b>
GC (10K) ∪ M2-S (8K)	RAug-Synt	98.95	91.25	86.46	92.38	87.30	91.27
GC (10K) ∪ M2-S (10K)	-	99.63	95.83	94.64	95.07	89.82	<b>95.00</b>
GC (10K) ∪ M2-S (10K)	RAug-Synt	99.33	87.77	88.00	91.28	87.70	90.82

TABLE VI: Impact of augmenting the training dataset on FR verification accuracies (authentic dataset: M2-S; synthetic dataset: GC). Verification accuracy comparison between baselines trained with combined authentic and synthetic identities (-) and the same with data augmentation applied only on synthetic samples (RAug-Synt). For each group of results, best average accuracies are highlighted in **bold**. The authentic identities are sampled from M2-S, the synthetic ones from GC.

Given the consistent enhancements observed by augmenting only the synthetic samples in our combined training datasets, we investigated whether this behavior is confirmed across different sources of authentic identities. To address this, we extended our investigation to training FR models on synthetic data combined with authentic identities from M2-S. Results are reported in Table VI and VII. In these settings, it can be observed that FR models trained on augmented GC data combined with authentic identities from M2-S constantly led to slightly lower verification performances than the baselines, which were impacted by  $+0.23\%$ ,  $-2.49\%$ ,  $-4.83\%$ ,  $-4.09\%$ ,  $-3.38\%$  and  $-4.42\%$  while involving 1K, 2K, 4K, 6K, 8K and 10K authentic identities, respectively. FR models trained with augmented DC samples achieved 1.32%, 1.00%, 1.11%, 0.94%, 0.68% and 0.93% improvements. The consistency of our results while train-

Train Data	Method	LFW (↑)	AgeDB-30 (↑)	CFP-PP (↑)	CA-LFW (↑)	CP-LFW (↑)	Avg (↑)
DC (10K) ∪ M2-S (1K)	-	98.63	90.92	89.10	92.18	82.78	90.72
DC (10K) ∪ M2-S (1K)	RAug-Synt	98.85	91.33	90.73	92.40	85.80	<b>91.86</b>
DC (10K) ∪ M2-S (2K)	-	98.85	91.60	89.39	92.62	83.15	91.12
DC (10K) ∪ M2-S (2K)	RAug-Synt	98.82	92.03	90.89	92.30	85.85	<b>91.98</b>
DC (10K) ∪ M2-S (4K)	-	99.02	92.57	90.01	93.00	84.48	91.82
DC (10K) ∪ M2-S (4K)	RAug-Synt	99.13	93.05	92.03	92.90	86.82	<b>92.79</b>
DC (10K) ∪ M2-S (6K)	-	99.07	93.07	90.77	93.28	84.95	92.23
DC (10K) ∪ M2-S (6K)	RAug-Synt	99.13	93.28	92.43	93.22	87.18	<b>93.05</b>
DC (10K) ∪ M2-S (8K)	-	99.22	93.45	91.66	93.40	85.53	92.65
DC (10K) ∪ M2-S (8K)	RAug-Synt	99.27	93.70	92.74	93.35	87.20	<b>93.25</b>
DC (10K) ∪ M2-S (10K)	-	99.17	93.50	91.44	93.90	85.73	92.75
DC (10K) ∪ M2-S (10K)	RAug-Synt	99.38	93.92	93.17	93.42	87.95	<b>93.57</b>

TABLE VII: Impact of augmenting the training dataset on FR verification accuracies (authentic dataset: M2-S; synthetic dataset: DC). Verification accuracy comparison between baselines trained with combined authentic and synthetic identities (-) and the same with data augmentation applied only on synthetic samples (RAug-Synt). For each group of results, best average accuracies are highlighted in **bold**. The authentic identities are sampled from M2-S, the synthetic ones from DC.

ing FR models on augmented DC samples combined with authentic identities further reinforces our previous considerations. This points out that augmenting synthetic data is a promising direction for improving the FR performances, in case the FR model is trained on combining subsets of synthetic and authentic data.

## V. CONCLUSIONS AND FUTURE WORK

To conclude, we connect the insights coming from individual analysis and discuss the results and future research. Our initial analysis aimed to comprehend the performance differences between FR models exclusively trained on authentic data and those trained solely on synthetic data generated through SOTA generative methods. This investigation led to a better understanding of the impact of data source variation on model performance, under an equal number of identities across datasets. Our findings align consistently with prior research [44], [28], confirming that FR models trained on the large, full authentic dataset exhibit the highest performance, achieving an average accuracy of 94.63% across the benchmarks in our experiments.

Subsequently, we investigated whether introducing a restricted number of authentic identities led to significant gains in accuracy, compared to models exclusively trained on synthetic identities. As anticipated, our study revealed a consistent improvement across all benchmarks as the number of involved authentic identities increased. FR models trained on subsets comprising both authentic and synthetic data demonstrated remarkably competitive results when compared to models exclusively trained on the entire authentic dataset. Notably, certain combinations that included DC identities even outperformed the full authentic-only baseline.

In our third analysis, we assessed whether a training dataset with a fixed number of synthetic identities and a progressively increasing number of authentic ones could attain accuracy levels, at the very least, comparable to a larger dataset comprising solely authentic identities. For both authentic datasets, all settings involving combined datasets yielded superior outcomes compared to the synthetic-only

baselines, showcasing improvements of approximately 30% in several cases. Moreover, training FR models on synthetic data combined with subsets of authentic identities consistently led to improvements compared to models trained exclusively on authentic subsets. As another observation, models trained on combined datasets that included synthetic data from DC consistently outperformed those involving synthetic data from GC, especially in cross-pose benchmarks.

In a last exploration, we delved into the impact of data augmentation on verification accuracy concerning the combined baselines. Throughout our experiments, we applied augmentation to two distinct sets of images: exclusively on synthetic ones and on the entire combined datasets. Notably, augmenting both synthetic and authentic samples in the combined training datasets generally resulted in inferior performance. Conversely, employing the same augmentation solely on synthetic samples led to remarkable gains in accuracy, indicating this approach as promising for enhancing accuracy when training FR models with combined datasets.

Building upon this work, our next steps involve exploring the performance of a wider range of FR backbones and incorporating additional augmentation, sampling, and domain generalization techniques. We will also explore techniques to optimize the generation of synthetic data specifically designed for training more effective FR models when faced with (very) limited authentic data, in an end-to-end manner. Additionally, in the immediate future, we plan to extend this study with a sensitivity analysis to determine an optimal parameterization for this task, encompassing both domain generalization and data augmentation techniques. This research also suggests that augmenting limited (and likely less demographically inclusive) data with more diverse synthetic data can not only improve overall performance but also enhance fairness across demographic groups. Investigating this aspect further will be one of our next research objectives.

## REFERENCES

- [1] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. 2016.
- [2] G. Bae, M. de La Gorce, T. Baltrusaitis, C. Hewitt, D. Chen, J. P. C. Valentin, R. Cipolla, and J. Shen. Digiface-1m: 1 million digital face images for face recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 3515–3524. IEEE, 2023.
- [3] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*, pages 464–473. IEEE, 2017.
- [4] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1577–1586. IEEE, 2022.
- [5] F. Boutros, J. H. Grebe, A. Kuijper, and N. Damer. Idiff-face: Synthetic-based face recognition through fuzzy identity-conditioned diffusion model. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 2-6, 2023*, October 2023.
- [6] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *IEEE International Joint Conference on Biometrics, IJCB 2022*,



- Abu Dhabi, United Arab Emirates, October 10-13, 2022, pages 1–11. IEEE, 2022.
- [7] F. Boutros, M. Klemm, M. Fang, A. Kuijper, and N. Damer. Exfacegan: Exploring identity directions in gan’s learned latent space for synthetic identity generation. In *IEEE International Joint Conference on Biometrics, IJCB 2023*, September 2023.
  - [8] F. Boutros, M. Klemm, M. Fang, A. Kuijper, and N. Damer. Unsupervised face recognition using unlabeled synthetic data. In *17th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2023, Waikoloa Beach, HI, USA, January 5-8, 2023*, pages 1–8. IEEE, 2023.
  - [9] F. Boutros, P. Siebke, M. Klemm, N. Damer, F. Kirchbuchner, and A. Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.
  - [10] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image Vis. Comput.*, 135:104688, 2023.
  - [11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi’an, China, May 15-19, 2018*, pages 67–74. IEEE Computer Society, 2018.
  - [12] E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020.
  - [13] I. DeAndres-Tame, R. Tolosana, P. Melzi, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, A. Morales, J. Fierrez, J. Ortega-Garcia, et al. Frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data. *arXiv preprint arXiv:2404.10378*, 2024.
  - [14] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022.
  - [15] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5153–5162. Computer Vision Foundation / IEEE, 2020.
  - [16] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5153–5162. Computer Vision Foundation / IEEE, 2020.
  - [17] P. Dhariwal and A. Q. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021.
  - [18] P. Ghosh, P. S. Gupta, R. Uziel, A. Ranjan, M. J. Black, and T. Bolkart. GIF: generative interpretable faces. In *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pages 868–878. IEEE, 2020.
  - [19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.
  - [20] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016.
  - [21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.
  - [22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
  - [23] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in ‘Real-Life’ Images: detection, alignment, and recognition*, 2008.
  - [24] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
  - [25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019.
  - [26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. Computer Vision Foundation / IEEE, 2020.
  - [27] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18729–18738. IEEE, 2022.
  - [28] M. Kim, F. Liu, A. K. Jain, and X. Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12715–12725. IEEE, 2023.
  - [29] J. N. Kolf, T. Rieber, J. Elliesen, F. Boutros, A. Kuijper, and N. Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 806–816. IEEE, 2023.
  - [30] R. T. Marriott, S. Madiouni, S. Romdhani, S. Gentric, and L. Chen. An assessment of gans for identity-related applications. In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020*, pages 1–10. IEEE, 2020.
  - [31] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin, and M. Schaubert. Gandifface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 3078–3087. IEEE, 2023.
  - [32] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X.-Y. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M. H. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, H. O. Shahreza, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltsou, S. Papadopoulos, C. Diou, N. Di Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, A. Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, and M. Marras. Frcsyn challenge at wacv 2024: Face recognition challenge in the era of synthetic data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 892–901, January 2024.
  - [33] P. Melzi, R. Tolosana, R. Vera-Rodriguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fierrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X.-Y. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M. H. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, H. O. Shahreza, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltsou, S. Papadopoulos, C. Diou, N. D. Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, A. Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, and M. Marras. Frcsyn-ongoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems. *Information Fusion*, 107:102322, 2024.
  - [34] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1997–2005. IEEE Computer Society, 2017.
  - [35] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 2021.
  - [36] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. Synface: Face recognition with synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10860–10870. IEEE, 2021.
  - [37] S. Sengupta, J. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV*

- 2016, Lake Placid, NY, USA, March 7-10, 2016, pages 1–9. IEEE Computer Society, 2016.
- [38] H. O. Shahreza, C. Ecabert, A. George, A. Unnervik, S. Marcel, N. Di Domenico, G. Borghi, D. Maltoni, F. Boutros, J. Vogel, et al. Sdfr: Synthetic data for face recognition competition. *arXiv preprint arXiv:2404.04580*, 2024.
- [39] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9240–9249. Computer Vision Foundation / IEEE, 2020.
- [40] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. G. Medioni. Gan-control: Explicitly controllable gans. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14063–14073. IEEE, 2021.
- [41] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H. Seidel, P. Pérez, M. Zollhöfer, and C. Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6141–6150. Computer Vision Foundation / IEEE, 2020.
- [42] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. Computer Vision Foundation / IEEE Computer Society, 2018.
- [43] W. Wang, L. Zhang, C.-M. Pun, and J.-C. Xie. Boosting face recognition performance with synthetic data and limited real data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [44] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.
- [45] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.
- [46] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep.*, 5(7), 2018.
- [47] T. Zheng, W. Deng, and J. Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.
- [48] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10492–10502. Computer Vision Foundation / IEEE, 2021.