GANDiffFace: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations

Pietro Melzi 1 Christian Rathgeb 2,3 Ruben Tolosana 1 Ruben Vera-Rodriguez 1 Dominik Lawatsch 2 Florian Domin 2 Maxim Schaubert 2

1 Biometrics and Data Pattern Analytics Laboratory, Universidad Autonoma de Madrid, Spain
2 secunet Security Networks AG, Essen, Germany
3 Hochschule Darmstadt, Germany

Abstract

Face recognition systems have significantly advanced in recent years, driven by the availability of large-scale datasets. However, several issues have recently come up, including privacy concerns that have led to the discontinuation of well-established public datasets. Synthetic datasets have emerged as a solution, even though current synthesis methods present other drawbacks such as limited intra-class variations, lack of realism, and unfair representation of demographic groups. This study introduces GANDiffFace, a novel framework for the generation of synthetic datasets for face recognition that combines the power of Generative Adversarial Networks (GANs) and Diffusion models to overcome the limitations of existing synthetic datasets. In GANDiffFace, we first propose the use of GANs to synthesize highly realistic identities and meet target demographic distributions. Subsequently, we fine-tune Diffusion models with the images generated with GANs, synthesizing multiple images of the same identity with a variety of accessories, poses, expressions, and contexts. We generate multiple synthetic datasets by changing GANDiffFace settings, and compare their mated and non-mated score distributions with the distributions provided by popular real-world datasets for face recognition, i.e., VGG2 and IJB-C. Our results show the feasibility of the proposed GANDiffFace, in particular the use of Diffusion models to enhance the (limited) intra-class variations provided by GANs towards the level of real-world datasets.

1. Introduction

In recent years, the development of face recognition technology has experienced a significant increase in the use of synthetic datasets. This trend has been facilitated by the proposal of numerous approaches for the generation of synthetic faces [35], resulting in an augmentation and diversification of the datasets for face recognition [47, 26].

Figure 1: Overview of our GANDiffFace framework based on the combination of GAN and Diffusion models. GANDiffFace creates synthetic datasets for face recognition with the properties listed in blue. From each identity synthesized with the GAN-based module, a personalized Diffusion-based module generates images with realistic intra-class variations that, once filtered, will compose the final dataset.

Synthetic datasets provide several advantages compared to real-world datasets [19]. Firstly, they offer a promising solution to some privacy concerns associated with real datasets, which are usually based on the collection of face images of individuals without their knowledge or consent from various online sources [31]. Secondly, synthetic face generators provide potentially infinite data. This is of particular importance because established datasets have been dismissed due to privacy concerns [16], and regulatory frameworks such as the EU-GDPR require the informed consent of individuals prior to the collection and use of personal data [43]. Finally, if the synthesis process is controllable, datasets with desired demographic characteristics (and labels for free) can be easily obtained, unlike real-world datasets that unequally represent diverse demographic groups [30], among other aspects.

Among generative models, Generative Adversarial Networks (GANs) have been widely used to synthesize face images due to their ability to learn complex distributions and generate high-quality images of human faces [20, 39], especially the recent version of StyleGAN3 [22, 2]. However, GANs generate images based on the patterns learned
GAN-based module alone, and we describe related works that use synthetic illumination, characterized by easily controllable intra-class variation, in comparison to real-world datasets. This affects the performance of face recognition models trained with synthetic data and evaluated with real data, as observed in.

More recently, Diffusion models have gained popularity and outperformed GAN models in multiple tasks, including image synthesis. A Diffusion model consists of a Markov chain that gradually adds random noise to data and learns to reverse it, to generate the desired output from noise. Diffusion models can shape their outputs according to text or images, and generate a wider variety of images compared to GAN models. However, unlike GANs, Diffusion models do not learn explicit latent representations of face images, making their demographic attributes and intra-class variations less controllable.

In this study, we propose a novel framework called GANDiffFace to generate synthetic datasets for face recognition, by combining the advantages of both GAN and Diffusion models. We use StyleGAN3 to generate synthetic identities, and create six different images for each identity by manipulating their pose, expression, and illumination attributes in the latent space. For attribute manipulation, we follow the approach proposed by a previous work that investigates the use of automatically generated synthetic datasets for benchmarking face recognition systems. We observe that such synthetic datasets are not suitable for other tasks, e.g. the training of face recognition systems, because of their limited intra-class variations. Hence, we propose the use of DreamBooth, a recent framework for the “personalization” of Diffusion models, to generate more realistic intra-class variations. Given as input the six images previously generated for a specific subject, DreamBooth fine-tunes a pretrained text-to-image Diffusion model to bind a unique identifier with that subject. The unique identifier allows to synthesize fully-novel photorealistic images of the subject contextualized in different scenes, poses, views, and lighting conditions, by leveraging the semantic prior embedded in the model.

The main contributions of the study are:

- Proposal of GANDiffFace, a novel framework for the generation of synthetic datasets for face recognition.
- GANDiffFace generates photorealistic images of synthetic identities with enhanced intra-class variations. Additionally, specific demographic distributions can be obtained by manipulating the latent space of StyleGAN3 during identity generation.
- Two different datasets with the same synthetic identities are generated at different steps of GANDiffFace: i) with the GAN-based module alone, and ii) with the combination of GAN-based and Diffusion-based modules. We provide a direct comparison (based on the same identities) between the two synthetic datasets, and further compare them to real-world datasets.
- We make available the synthetic dataset generated with GANDiffFace, characterized by easily controllable and realistic intra-class variations. Our dataset represents equally balanced demographic groups, defined in terms of race, age, and gender, and contains labels of several face attributes. Hence, it enables the training/testing of multiple facial analysis applications.

The remainder of this work is organized as follows: in Section 2 we describe related works that use synthetic datasets for face recognition. In Section 3 we describe the modules of our proposed GANDiffFace framework. In Section 4 we provide an evaluation on our synthetic datasets, and in Section 5 we discuss limitations and future works, drawing the conclusions of this work.

### 2. Related works

Numerous technologies have been proposed to generate synthetic datasets for face recognition. The applicability of synthetic datasets to face recognition has been investigated in, to compensate for the lack of publicly available large-scale test datasets, and in, to provide a taxonomy and further discussion. In Table 1, we compare the most relevant synthetic datasets for face recognition proposed in the literature.

StyleGAN2 is used to generate synthetic identities in. With the property of linear separability of StyleGAN2’s latent space, multiple images of the original identities are generated while changing three attributes, i.e. illumination, pose, and expression. Linear separability allows to find a hyperplane in the latent space that separates populations of latent vectors according to different values for a specific attribute. The normal vector to this hyperplane represents the direction along which latent vectors, i.e. the representations of synthetic images in the latent space, can be moved to modify the specific attribute. The approach proposed in presents some limitations addressed by our GANDiffFace, namely the demographic bias inherited from StyleGAN2, and the limited intra-class variations generated.

---

1[https://github.com/PietroMelzi/GANDiffFace]
Table 1: Overview of the synthetic datasets for face recognition applications proposed in the literature.

<table>
<thead>
<tr>
<th>Method</th>
<th>Category</th>
<th>Realism</th>
<th>Controllable demographics</th>
<th>Intra-class variations</th>
<th>Fully synthetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent space [8]</td>
<td>GAN</td>
<td>high</td>
<td>low</td>
<td>low</td>
<td>yes</td>
</tr>
<tr>
<td>SYNFace [33]</td>
<td>GAN</td>
<td>high</td>
<td>low</td>
<td>low</td>
<td>yes</td>
</tr>
<tr>
<td>SFace [4]</td>
<td>GAN</td>
<td>high</td>
<td>low</td>
<td>low</td>
<td>no</td>
</tr>
<tr>
<td>DigiFace-1M [3]</td>
<td>3D model</td>
<td>low</td>
<td>medium</td>
<td>high</td>
<td>yes</td>
</tr>
<tr>
<td>DCFace [24]</td>
<td>Diffusion</td>
<td>medium</td>
<td>low</td>
<td>high</td>
<td>no</td>
</tr>
<tr>
<td>GANDiffFace (ours)</td>
<td>GAN + Diffusion</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>yes</td>
</tr>
</tbody>
</table>

In an analogous way, the linear separability of StyleGAN3’s latent space is exploited to generate a large-scale synthetic dataset of children’s faces, named HDA-SynChildFaces [11]. Compared to the previous work, in HDA-SynChildFaces the latent space is manipulated during identity generation to balance the race distribution of the dataset. The work reveals that children consistently perform worse than adults in various face recognition systems.

SYNFace proposes the use of DiscoFaceGAN for the synthesis of face images, a disentangled learning scheme that enables precise control of targeted face properties such as identity, pose, expression, and illumination [33]. DiscoFaceGAN generates realistic face images by sampling random noise from multiple normal distributions, each one independently controlling a different face attribute. SYNFace identifies in poor intra-class variations the reason of the performance gap existing between face recognition systems trained with synthetic and real datasets. To mitigate it, the intermediate states of two synthetic identities mixed together are considered as novel identities. SYNFace generates mostly frontal-view images, the identity preservation or variation of mixed identities is not evaluated, and a further mix with real images is required to bridge the gap between synthetic and real world data.

In SFace a privacy-friendly synthetically generated face dataset is proposed, based on the training of StyleGAN2-ADA with real datasets, and the setting of identity labels as class labels to create synthetic data [4]. Hence, a 1:1 correspondence can be observed between real and synthetic identities, with the consequent sharing of face attributes (but not identity). SFace provides an unrealistic mated score distribution, shifted towards the non-mated distribution, and unlike other GAN-based methods maintains a tight correspondence between the synthetic identities and the real ones used during training.

DigiFace-1M, a large-scale synthetic dataset obtained by rendering digital faces with a computer graphics pipeline, is proposed for face recognition in [3]. Each identity of DigiFace-1M is defined as the unique combination of facial geometry, texture, eye color, and hair style, while other parameters (i.e. pose, expression, environment, and camera distance) are varied to render multiple images. With aggressive data augmentation, this work significantly reduces the synthetic-to-real domain gap, establishing the new state-of-the-art performance for face recognition models trained on synthetic data. Furthermore, this method does not rely on real data for training the generative model, differently from GAN models. However, we observe some limitations in DigiFace-1M: the textures of the synthetic images appear unrealistic, and the demographic distribution of the synthetic dataset is not analyzed.

More recently, a Diffusion model called DCFace has been proposed for synthetic face recognition [24]. DCFace is composed of: i) a sampling stage for the generation of synthetic identities, and ii) a mixing stage for the generation of face images whose identity comes from the sampling stage and the style is selected from a “style bank” of images. Both components are based on Diffusion models, showing considerable ability to generate unique and diverse identities. Compared to SYNFace and DigiFace-1M, DCFace claims to provide better intra-class variations, but relies on real face images for the “style bank”. While synthetic data could in principle also be used for the “style bank”, this may reduce intra-class variations in the generated dataset. We raise criticism about the use of real data, as newly generated synthetic images contain sharp details from the real images used as style reference and the method is not fully synthetic. Furthermore, with DCFace specific face attributes cannot be manipulated either during the sampling or mixing stages.

As of today, synthetic datasets based on Diffusion models are promising, but still in a primitive stage. Our proposed GANDiffFace framework combines the advantages of GAN models, i.e. generation of highly realistic faces and control of the latent space, with enhanced intra-class variations achieved by recent Diffusion models.

3. Proposed method

The graphical representation of our proposed GANDiffFace framework is provided in Figure 2. GANDiffFace consists of two modules: the first one dedicated to the synthesis of identities, based on StyleGAN3 [22] and transformation in its latent space (Section 3.1), and the second one responsible for the creation of realistic intra-class variations, based on DreamBooth [37] (Section 3.2).
Identity generation. We first generate an initial random set of 256,000 synthetic images with StyleGAN3 (pre-trained with FFHQ dataset [23]), and label them with FairFace, a classifier of demographic attributes (i.e., race, gender, and age) [21]. The distributions of demographic attributes obtained in the random set are reported in Figure 3, highlighting the bias present in StyleGAN3. We remove from the initial set images with poor quality as well as those belonging to young subjects. For quality assessment we use MagFace with backbone iResNet100, a state-of-the-art system that learns feature embeddings whose magnitudes represent face sample quality [29]. We eliminate the 10% of images with the lowest magnitude, that usually contain artifacts, sunglasses, or belong to children. Then, we also eliminate images of people in the age intervals 0-2, 3-9, and 10-19, as we focus only on adult identities.

Face attribute representation. A framework to interpret the disentangled face representation learned by StyleGAN and study the properties of the facial semantics encoded in its latent space was initially proposed in [41]. The framework is based on the training of linear Support Vector Machines (SVMs) in the latent space to separate two distinct populations of latent vectors according to a binary target attribute. The normal vector to the resulting hyperplane boundary of the trained SVM represents the direction to follow in the latent space to edit the target attribute of face images. This approach has proved successful even in the latent space of StyleGAN2 [8] and StyleGAN3 [11].

In this work, we label our synthetic dataset according to pose (yaw and pitch) with 3DDFA_V2 [15, 14], expression (neutral, happy, sad, surprise, disgust, anger, contempt) with DMUE [40], and illumination by comparing the pixel intensity of the right and left half of face images. We also consider the labels provided by FairFace for gender, age, and race. For each attribute of interest, we represent two populations with an equal number of latent vectors, selected...
GAN-based intra-class variations

Demographic-specific transformation: race, age, and gender), an arbitrary large number of the expression of the random identities can be generated to represent target demographic groups. In the following we describe the sequence of operations required to generate identities with target demographic attributes, and provide (limited) intra-class variations with a GAN-based approach. In Figure 4, we also provide graphical examples of these operations for random identities.

Figure 4: Sequence of transformations to generate identities with target demographic attributes and GAN-based intra-class variations. Initial identities with demographic attributes similar to the target ones have been selected.

Equation of the transformation. To neutralize a latent vector $w$ with respect to the attribute $a$, it is possible to project $w$ onto the hyperplane boundary of attribute $a$, as follows:

$$w_{\alpha} = w - (w^T \alpha) \cdot \alpha,$$

(2)

where $w_{\alpha}$ is the resulting neutralized latent vector, in which the attribute $a$ results in a neutral condition. By combining the operations of transformation and neutralization to modify the demographic attributes of face images (i.e. race, age, and gender), an arbitrary large number of identities can be generated to represent target demographic groups. In the following we describe the sequence of operations required to generate identities with target demographic attributes, and provide (limited) intra-class variations with a GAN-based approach. In Figure 4, we also provide graphical examples of these operations for random identities.

1. **Pose neutralization**: the pose of the random identities generated with StyleGAN3 is neutralized, by projecting their latent vectors on the hyperplane boundaries relative to yaw and pitch.

2. **Expression neutralization**: the expression of the random identities is neutralized, by projecting their latent vector on the hyperplane boundary relative to the current expression of each identity, and subsequently moving the resulting latent vectors in the direction of neutral expression (opposite direction with respect to current expression).

3. **Demographic-specific transformation**: the latent vectors (neutralized according to pose and expression) are modified by applying transformations in the direction of the boundaries of interest. The pre-selection of random identities with demographic attributes close to the target ones may help to prevent transformations from estimating latent vectors outside of the StyleGAN3 distribution of faces [11]. We consider 70 different demographic groups, obtained by combining the seven

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Number of images</th>
<th>Validation accuracy</th>
<th>Average distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pose: Yaw</td>
<td>100,000</td>
<td>100%</td>
<td>1.39</td>
</tr>
<tr>
<td>Pose: Pitch</td>
<td>100,000</td>
<td>99%</td>
<td>0.98</td>
</tr>
<tr>
<td>Expression: Happy</td>
<td>13,390</td>
<td>100%</td>
<td>1.11</td>
</tr>
<tr>
<td>Expression: Contempt</td>
<td>11,014</td>
<td>92%</td>
<td>0.46</td>
</tr>
<tr>
<td>Expression: Surprise</td>
<td>8,328</td>
<td>89%</td>
<td>0.53</td>
</tr>
<tr>
<td>Expression: Disgust</td>
<td>4,436</td>
<td>95%</td>
<td>0.84</td>
</tr>
<tr>
<td>Expression: Sad</td>
<td>2,606</td>
<td>85%</td>
<td>0.45</td>
</tr>
<tr>
<td>Expression: Anger</td>
<td>2,440</td>
<td>91%</td>
<td>0.74</td>
</tr>
<tr>
<td>Illumination</td>
<td>15,000</td>
<td>72%</td>
<td>0.18</td>
</tr>
<tr>
<td>Gender</td>
<td>100,000</td>
<td>100%</td>
<td>1.33</td>
</tr>
<tr>
<td>Age</td>
<td>37,736</td>
<td>96%</td>
<td>0.85</td>
</tr>
<tr>
<td>Race: White</td>
<td>64,846</td>
<td>100%</td>
<td>1.00</td>
</tr>
<tr>
<td>Race: Latino-Hispanic</td>
<td>33,762</td>
<td>98%</td>
<td>0.92</td>
</tr>
<tr>
<td>Race: East Asian</td>
<td>12,964</td>
<td>100%</td>
<td>1.12</td>
</tr>
<tr>
<td>Race: Middle Eastern</td>
<td>13,112</td>
<td>92%</td>
<td>0.57</td>
</tr>
<tr>
<td>Race: Indian</td>
<td>8,244</td>
<td>100%</td>
<td>1.59</td>
</tr>
<tr>
<td>Race: Southeast Asian</td>
<td>5,180</td>
<td>100%</td>
<td>1.43</td>
</tr>
<tr>
<td>Race: Black</td>
<td>5,356</td>
<td>100%</td>
<td>1.92</td>
</tr>
</tbody>
</table>

Table 2: List of boundaries calculated in this work, with information about each SVM training. Average distance is the distance of latent vectors from the hyperplane boundary. at the two extremes of the score distribution of the target attribute. We train each SVM with a maximum number of 100,000 latent vectors, depending for each attribute on the amount of data available to represent populations. In case of categorical attributes, i.e. expression and race, numerical values are provided respectively by DMUE and FairFace for all the possible categorical attributes. Hence, we train multiple one-vs-one SVMs to separate each expression from the neutral one, and multiple one-vs-all SVMs for each different race. In Table 2 we report all the boundaries calculated in this work, providing additional information about the training of each SVM. High validation accuracy demonstrates the goodness of our boundaries, except for illumination that turns out to be unreliable. The entire training of boundaries is carried out exclusively with synthetic data.

**Latent space transformation.** The approach used to modify face attributes by applying transformations in the latent space has been described in detail in previous works [8, 11]. For clarity, here we only summarize its key points. We can transform a latent vector $w$, that represents a face image in the latent space of StyleGAN3, to modify its attribute $a$ according to the following operation:

$$w_{\alpha} = w + \alpha \cdot n_a,$$

(1)

where $n_a$ is the normal vector to the hyperplane that separates populations according to the attribute $a$, $\alpha$ is the degree of the transformation, and $w_{\alpha}$ is the resulting latent vector, in which the attribute $a$ results modified according to the direction of the transformation. To neutralize a latent vector $w$ with respect to the attribute $a$, it is possible to project $w$ onto the hyperplane boundary of attribute $a$, as follows:

$$w_{\alpha} = w - (w^T n_a) \cdot n_a,$$

(2)

where $w_{\alpha}$ is the resulting neutralized latent vector, in which the attribute $a$ results in a neutral condition. By combining the operations of transformation and neutralization to modify the demographic attributes of face images (i.e. race, age, and gender), an arbitrary large number of identities can be generated to represent target demographic groups. In the following we describe the sequence of operations required to generate identities with target demographic attributes, and provide (limited) intra-class variations with a GAN-based approach. In Figure 4, we also provide graphical examples of these operations for random identities.

1. **Pose neutralization**: the pose of the random identities generated with StyleGAN3 is neutralized, by projecting their latent vectors on the hyperplane boundaries relative to yaw and pitch.

2. **Expression neutralization**: the expression of the random identities is neutralized, by projecting their latent vector on the hyperplane boundary relative to the current expression of each identity, and subsequently moving the resulting latent vectors in the direction of neutral expression (opposite direction with respect to current expression).

3. **Demographic-specific transformation**: the latent vectors (neutralized according to pose and expression) are modified by applying transformations in the direction of the boundaries of interest. The pre-selection of random identities with demographic attributes close to the target ones may help to prevent transformations from estimating latent vectors outside of the StyleGAN3 distribution of faces [11]. We consider 70 different demographic groups, obtained by combining the seven
The quality of the images generated with text-we filter out images in which the face
Face detection: we extract ArcFace embeddings
4. GAN-based intra-class variations: the latent vectors of demographic-specific identities can be further modiﬁed according to the boundaries of pose, expression, and illumination, to generate (limited) intra-class variations for each synthetic identity.

We observe that kinship ties, multiethnic unions, and population aging can be simulated by applying different demographic transformations to the same original identity.

3.2. Diffusion-based module

Text-to-image models enable high-quality and diverse synthesis of images based on text prompts. They rely on their strong semantic prior, learned from a large collection of image-caption pairs, to bind a word with various images in different poses and contexts [34, 38]. However, these models lack the ability to preserve the identity of a subject consistently across synthesized images. To overcome this issue we consider Dreambooth, a novel framework that ﬁnetunes text-to-image models (in this case Stable Diffusion [36]) to bind new words with speciﬁc subjects, and synthesize novel renditions of subjects in different contexts while maintaining their distinctive features [37].

Training. We use the images generated by the GAN-based module of GANDiffFace to ﬁne-tune Stable Diffusion, a state-of-the-art Diffusion text-to-image model [36]. We apply Dreambooth to bind a unique token (we use $xyz$) with a speciﬁc synthetic identity, and implant it into the output domain of Stable Diffusion. To refer to the identity, we use text prompts containing the token $xyz$ followed by the class name of the identity, in our case person. Hence, the minimum text prompt to refer to the identity is: “$xyz$ person”. The class name (i.e. person) enables the model to use its prior knowledge of the class, and an additional class-speciﬁc prior preservation loss helps to prevent the model to associate the class with the speciﬁc identity. These components serve as regularization, as they alleviate overﬁtting and encourage diversity in the resulting images [37].

Previous studies highlighted the importance of parameter settings to ﬁne-tune the Stable Diffusion model, especially in case of the person class [1]. We ﬁne-tune Stable Diffusion with 6 input images for each synthetic identity, 200 images of the class person for regularization (generated by Stable Diffusion itself), and for 1,000 epochs, also allowing the ﬁne-tuning of the text encoder. Given the high number of identities in our dataset and possible interferences between tokens in the vocabulary, we ﬁne-tune a specific Stable Diffusion model for each synthetic identity.

Filtering. The quality of the images generated with text-to-image Diffusion models highly depends on the correct speciﬁcation of text prompts [45]. In our inference phase, we consider some prompts that may work well for most but not all the identities, to enhance the intra-class variations resulting in our dataset. For this reason, an important component of our GANDiffFace framework is the ﬁltering of the generated images, which is carried out in three stages:

1. Face detection: we ﬁlter out images in which the face detector SCRFD-10G [13] detects no face.
2. Identity preservation: we extract ArcFace embeddings (backbone iResNet100) [9] from synthetic images generated with both the GAN-based module only and the
entire GANDiffFace framework. Then, for each synthetic image generated with GANDiffFace, we calculate the average of its cosine similarity with the 6 GAN images that represent the same identity (previously used to fine-tune the Stable Diffusion model). We filter out images if the similarity score is below a threshold $t_{ip} = 0.3$. We note that similarity scores are computed between images of different domains, allowing the removal of outliers images generated with the Diffusion-based module, and no comparison is carried out between images of the same domain. In Figure 5, we include examples of the average similarity scores between GAN and Diffusion images for a random identity.

3. **Gender preservation:** we label the remaining images by gender with FairFace, and filter out images with a gender different from the corresponding GAN images.

4. **Evaluation**

This section analyzes the similarity scores obtained with four versions of our synthetic dataset, in order to provide a comparison with the score distributions of existing synthetic and real-world datasets used for face recognition.

4.1. **Synthetic datasets**

We generate four datasets selecting different settings of our GANDiffFace framework, all of them containing the same 700 synthetic identities. We only use the GAN-based module of GANDiffFace to generate a synthetic dataset provided with GAN-based (limited) intra-class variations. We use the entire GANDiffFace with identity preservation filter $t_{ip} = 0.3$ (default) to evaluate the impact of the Diffusion model on intra-class variations. Then, we use the entire GANDiffFace framework with $t_{ip} = 0.2$ and $t_{ip} = 0.4$ to evaluate different intra-class variations. Additionally, we consider subsets of two synthetic datasets, i.e. SFace [4] and DigiFace-1M [3]. With the latter, we provide a comparison with a dataset based on 3D model.

4.2. **Real-world datasets**

We consider two real-world datasets widely used for face recognition, VGGFace2 [6] and IJB-C [28]. VGGFace2 is a large-scale dataset containing images from the web of around 9,000 identities, with large variations in pose, age, illumination, ethnicity and profession. IJB-C contains around 3,000 identities, with focus on occlusions and diversity of ethnicity and profession. According to IJB-C annotations, we remove multiple images taken from the same video and images with small faces. Both datasets have been discontinued, underlining the necessity of the generation of synthetic datasets with realistic intra-class variations.

For a fair comparison with our synthetic datasets, we filter out real images with a MagFace quality lower than 24.45 [29]. This is the threshold used to eliminate the 10% of images with the lowest magnitude during GANDiffFace identity generation (Section 3.1). We are interested in the generation of datasets with high quality images. Datasets with low quality images can be obtained with data augmentation, and their evaluation is out of the scope of this work.

4.3. **Similarity score distributions**

For each identity in real or synthetic datasets, we randomly select 10 images and generate 20 mated and 20 non-mated comparisons, and calculate the cosine similarity of their ArcFace (backbone iResNet100) embeddings [9] (Figure 6). We use ArcFace as it is open source and widely used for face recognition. Then, we measure the diversity between synthetic and real score distributions, the latter as reference, with Kullback–Leibler (KL) divergence.

In Table 3 we report the mean and standard deviation of mated and non-mated comparisons in the different datasets, as well as the number of identities. The reason behind the limited size of our GANDiffFace datasets is the high computational cost required for generation, but larger datasets can be produced. Analyzing the results, the use of a Diffusion model reduces the mean of mated scores from 0.67 (obtained with GAN-based module only) to 0.51, for $t_{ip} = 0.2$. This value is closer to the means of real datasets, i.e. 0.52 for VGGFace2 and 0.57 for IJB-C. The IJB-C mean is affected by a peak in the score distribution for values close to 1, due to the comparison of images taken from the same video and not detected in the annotation file. According to Table 4, GANDiffFace with $t_{ip} \geq 0.3$ reproduces mated distributions similar to the real ones ($KL = 0.16$ from VGGFace2). We observe that the mated distribution of the GAN-based dataset is very far from the VGGFace2 one ($KL = 0.69$), while the mated comparisons with high score in IJB-C help to reduce KL divergence to 0.28 for GAN-based and 0.09 for GANDiffFace with $t_{ip} = 0.4$.

For non-mated comparisons, we observe that synthetic datasets present distributions skewed towards positive values, differently from real datasets (Figure 6). KL divergence is generally bigger for synthetic non-mated distributions, showing difficulty to reproduce realistic inter-class variations (Table 4). While no significant difference can be observed between the non-mated distributions of GAN-based and GANDiffFace ($t_{ip} = 0.3$) datasets, KL divergences are slightly higher for $t_{ip} = 0.4$, and decrease with $t_{ip} = 0.2$: from 0.48 to 0.42 with regard to VGGFace2, and from 0.43 to 0.37 with regard to IJB-C. This may be due to the inclusion in the synthetic dataset of images less similar to the GAN-based ones, which showed a slightly positive mean for non-mated comparison scores.

Finally, we observe worse score distributions for the syn-
Figure 6: Similarity score distributions obtained from mated and non-mated comparisons randomly selected from our synthetic datasets (first row), and other synthetic, SFace and DigiFace-1M, and real-world, VGGFace2 and IJB-C, datasets (second row).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Id.</th>
<th>Mated scores</th>
<th>Non-mated scores</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAN-based</td>
<td>Syn</td>
<td>700</td>
<td>0.67 ± 0.14</td>
<td>0.08 ± 0.10</td>
<td></td>
</tr>
<tr>
<td>GANDiffFace (t_{tp} = 0.4)</td>
<td>Syn</td>
<td>700</td>
<td>0.59 ± 0.12</td>
<td>0.08 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>GANDiffFace (t_{tp} = 0.3)</td>
<td>Syn</td>
<td>700</td>
<td>0.55 ± 0.15</td>
<td>0.08 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>GANDiffFace (t_{tp} = 0.2)</td>
<td>Syn</td>
<td>700</td>
<td>0.51 ± 0.17</td>
<td>0.07 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>SFace</td>
<td>Syn</td>
<td>411</td>
<td>0.18 ± 0.13</td>
<td>0.02 ± 0.08</td>
<td></td>
</tr>
<tr>
<td>DigiFace-1M</td>
<td>Syn</td>
<td>2,000</td>
<td>0.47 ± 0.15</td>
<td>0.12 ± 0.09</td>
<td></td>
</tr>
<tr>
<td>VGGFace2</td>
<td>Real</td>
<td>8,515</td>
<td>0.52 ± 0.16</td>
<td>0.01 ± 0.07</td>
<td></td>
</tr>
<tr>
<td>IJB-C</td>
<td>Real</td>
<td>2,557</td>
<td>0.57 ± 0.17</td>
<td>0.01 ± 0.07</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Number of identities and means of mated/non-mated score distributions of synthetic and real datasets.

with DreamBooth, allows the fully synthetic generation of a dataset with realistic intra-class variations.

A limitation of GANDiffFace consists in the high computational cost required to fine-tune identity-specific Diffusion models. This was the main reason for the generation of 700 identities, but many more can be generated. Also, Diffusion images present some artifacts observable at human level. Nevertheless, they usually affect parts of human bodies such as hands that are cut out for face recognition.

In future works, we plan to use the synthetic dataset generated with GANDiffFace to deploy face recognition systems, given its desired properties of realistic intra-class variations and fair representation of multiple demographic groups. Also, future works can focus on i) reducing the KL divergence from non-mated score distributions of real datasets, to reproduce more accurately real-world inter-class variations, and ii) improving the quality of DeepFakes [35, 42] and attacks [18, 32] through GANDiffFace.
Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860813 - TReSPAsS-ETN. This study is supported by the project INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER). This research is based upon work supported by the Hessian Ministry of the Interior and Sport in the course of the Bio4ensics project and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[16] Jules, Harvey, Adam. LaPlace. Exposing.ai, 2021. 1


