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 came 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].

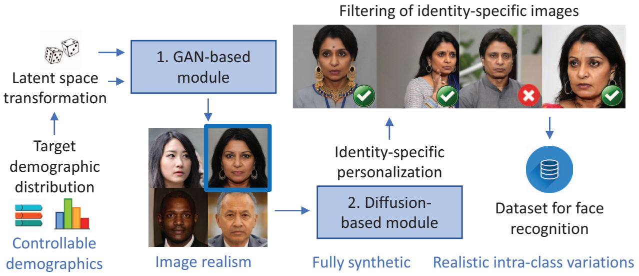

Filtering of identity-specific images
Latent space transformation
1. GAN-based module
Target demographic distribution
Controllable demographics
Image realism
Identity-specific personalization
2. Diffusion-based module
Dataset for face recognition
Fully synthetic
Realistic intra-class variations

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 from the training data, with limited control over the generated features, and possible biases towards certain demographic groups over-represented during training [27]. To tackle this issue, some methods have been proposed in the literature to modify attributes of synthetic face images, *i.e.* pose, illumination, and demographics. Target attributes can be injected into the generative component of GANs [7, 12], or alternatively the latent structure of GANs, *i.e.* their internal representation of face images, can be properly manipulated to meet the target attributes [41, 46]. However, GAN-generated images have been found to exhibit insufficient variability between the images of the same individual (*i.e.* intra-class variation), in comparison to real-world datasets [8]. This affects the performance of face recognition models trained with synthetic data and evaluated with real data, as observed in [33].

More recently, Diffusion models have gained popularity and outperformed GAN models in multiple tasks, including image synthesis [10]. 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 [17]. Diffusion models can shape their outputs according to text or images [44], and generate a wider variety of images compared to GAN models [25]. However, unlike GANs, Diffusion models do not learn explicit latent representations of face images, making their demographic attributes and intra-class variations less controllable [10].

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 (Figure 1). 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 (detailed in Section 3.1) proposed by a previous work that investigates the use of automatically generated synthetic datasets for benchmarking face recognition systems [8]. 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 [37], 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 [37].

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,[1] 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 [47], to compensate for the lack of publicly available large-scale test datasets, and in [5], 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 [8]. 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 [8] 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

Method	**Category**	**Realism**	**Controllable demographics**	**Intra-class variations**	**Fully synthetic**
Latent space [8]	GAN	high	low	low	yes
HDA-SynChildFaces [11]	GAN	high	high	low	yes
SYNFace [33]	GAN	high	low	low	no
SFace [4]	GAN	high	low	low	yes
DigiFace-1M [3]	3D model	low	medium	high	yes
DCFace [24]	Diffusion	medium	low	high	no
GANDiffFace (ours)	**GAN + Diffusion**	**high**	**high**	**high**	**yes**

Table 1: Overview of the synthetic datasets for face recognition applications proposed in the literature.

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.

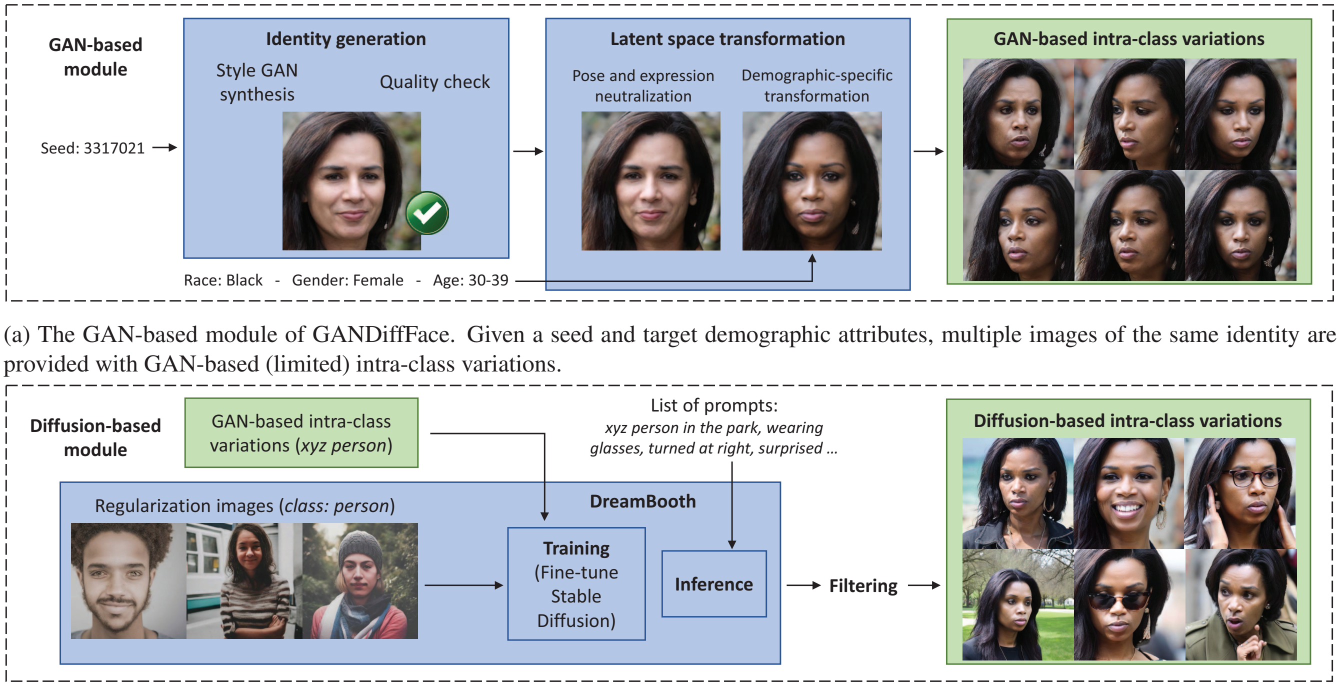
DigiFace-1M, a large-scale synthetic dataset obtained by rendering digital faces with a computer graphics pipeline, is proposed for face recognition in [3]. Each identity of *DigiFace-1M* is defined as the unique combination of facial geometry, texture, eye color, and hair style, while other parameters (*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).


GAN-based module
Seed: 3317021
Identity generation
Style GAN synthesis
Quality check
Race: Black - Gender: Female - Age: 30-39
Latent space transformation
Pose and expression neutralization
Demographic-specific transformation
GAN-based intra-class variations
Diffusion-based module
GAN-based intra-class variations (xyz person)
List of prompts:
xyz person in the park, wearing glasses, turned at right, surprised ...
Regularization images (class: person)
DreamBooth
Training (Fine-tune Stable Diffusion)
Inference
Filtering
Diffusion-based intra-class variations

(a) The GAN-based module of GANDiffFace. Given a seed and target demographic attributes, multiple images of the same identity are provided with GAN-based (limited) intra-class variations.

(b) The Diffusion-based module of GANDiffFace. Given a few images with GAN-based (limited) intra-class variations (a) of the synthetic identity *xyz*, a set of images of the class *person* for regularization (generated by Diffusion model), and a list of prompts, multiple images of the same identity are provided with augmented Diffusion-based intra-class variations.

Figure 2: Graphical representation of the GAN-based (a) and Diffusion-based (b) modules of the GANDiffFace framework.

3.1. GAN-based module

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].

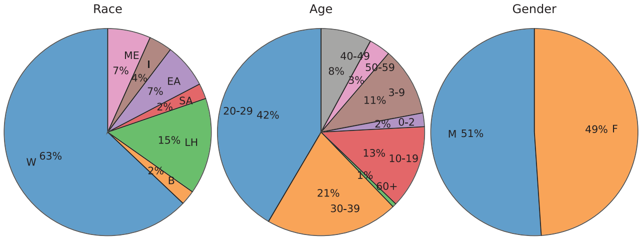


Figure 3: Race (W=White, B=Black, LH=Latino_Hispanic, SA=Southeast Asian, EA=East Asian, ME=Middle Eastern, I=Indian), age, and gender (M=male, F=Female) distributions in the initial random set of 256,000 identities generated with StyleGAN3 [22] and labelled with FairFace [21].

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

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

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:

$$w_a = w + \alpha \cdot n_a, \tag{1}$$

where n_a is the normal vector to the hyperplane that separates populations according to the attribute a, α is the degree of the transformation, and w_a 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:

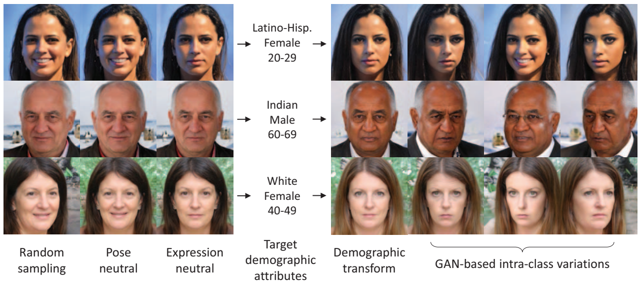


Figure 4: Sequence of transformations to generate identities with target demographic attributes and GAN-based intra-class variations. Initial identities with demographic attributes similar to the target ones have been selected.

$$w_{n_a} = w - (w^T n_a) \cdot n_a, \tag{2}$$

where w_{n_a} 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

races, five adult age intervals, and two genders reported in Figure 3. In total, we generate at this step 700 different identities (10 identities for each of the 70 demographic groups).

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 fine-tunes 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.

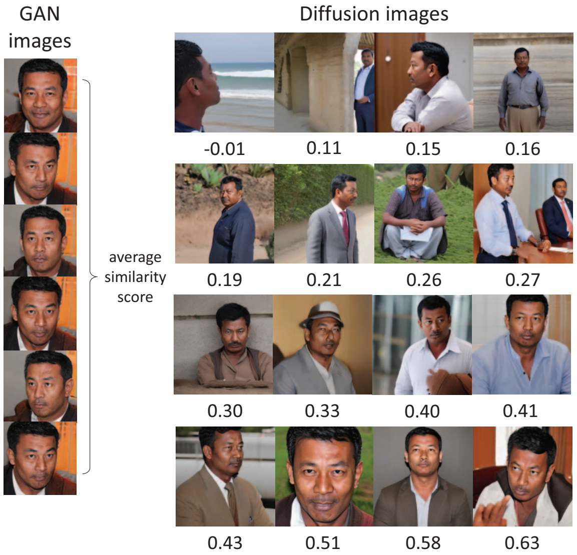


Figure 5: Identity preservation scores, obtained for each image on the right by averaging their similarity scores with the GAN images on the left. Threshold t_{ip} regulates the intra-class variations provided by GANDiffFace.

Inference. Once fine-tuned with DreamBooth, the Stable Diffusion model can generate images of the specific synthetic identity in multiple contexts, according to the provided prompts. To generate synthetic images with realistic intra-class variations, we evaluate different categories of prompts: accessorization, advanced poses, advanced expressions, and recontextualization. Examples of prompts are: "xyz person wearing scarf", "close photo of xyz person at the beach", "skeptical xyz person", and "full body xyz person with accurate details of face in an indoor place".

We observe that Stable Diffusion allows to specify negative prompts, to prevent the generation of undesired images. Given the large variety of datasets used to train Stable Diffusion, for the entire inference process we specify the following negative prompt: "photo with the style of painting, comics, drawing, or containing text".

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 VGFace2). 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-

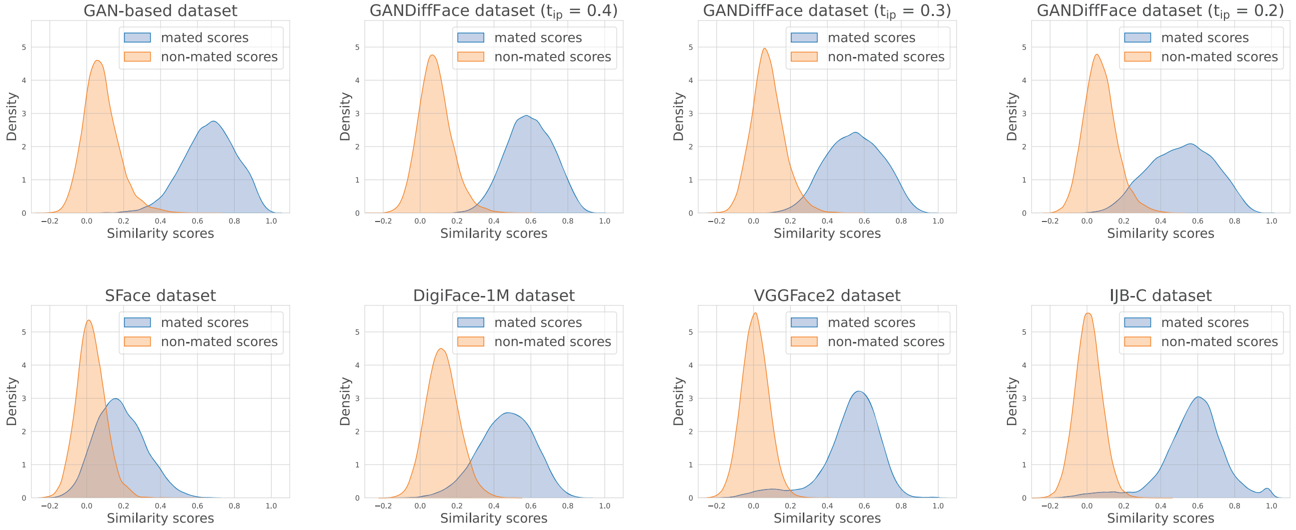


Figure 6: Similarity score distributions obtained from mated and non-mated comparisons randomly selected from our synthetic datasets (first row), and other synthetic, SFace and DigiFace-1M, and real-world, VGGFace2 and IJB-C, datasets (second row).

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

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

Dataset	Mated scores		Non-mated scores		EER
	VGG2	IJB-C	VGG2	IJB-C	
GAN-based	0.69	0.28	0.48	0.42	1.49%
GANDiffFace ($t_{ip} = 0.4$)	**0.16**	**0.09**	0.52	0.46	1.25%
GANDiffFace ($t_{ip} = 0.3$)	**0.16**	0.16	0.48	0.43	2.74%
GANDiffFace ($t_{ip} = 0.2$)	0.23	0.28	0.42	0.37	5.11%
SFace	1.72	2.13	**0.18**	**0.11**	22.53%
DigiFace-1M	0.21	0.41	1.05	1.02	7.92%
VGGFace2	-	0.11	-	0.01	4.51%
IJB-C	0.15	-	0.01	-	3.22%

Table 4: KL divergences of the distributions of each dataset from the real ones provided by VGGFace2 and IJB-C.

thetic SFace and DigiFace-1M datasets compared to the ones provided by GANDiffFace, with Equal Error Rates (EERs) about twice the real ones. This is reflected in higher KL divergences for mated (in case of SFace) and non-mated (in case of DigiFace-1M) score distributions, showing that these datasets fail to reproduce realistic intra and inter-class variations. However, SFace provides the best non-mated score distribution for synthetic datasets, with $KL = 0.18$ from VGGFace2 and $KL = 0.11$ from IJB-C.

5. Conclusion

This study has proposed GANDiffFace, a novel framework that combines the advantages of GAN and Diffusion models to generate synthetic datasets for face recognition with some desired properties. The use of a GAN model for identity generation, *i.e.* StyleGAN3, allows to synthesize images of human faces with high realism, and manipulate the latent space to provide a fair representation of 70 demographic groups. The addition of a Diffusion model, *i.e.* Stable Diffusion, personalized for specific identities 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

[1] Training Stable Diffusion with DreamBooth using Diffusers. https://huggingface.co/blog/dreambooth. Accessed: 2023-05-03. 6

[2] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? Image and video editing with StyleGAN3. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 204–220. Springer, 2023. 1

[3] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. DigiFace-1M: 1 Million Digital Face Images for Face Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023. 3, 7

[4] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *Proceedings of the 2022 IEEE International Joint Conference on Biometrics*, pages 1–11. IEEE, 2022. 3, 7

[5] Fadi Boutros, Vitomir Struc, Julian Fierrez, and Naser Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, page 104688, 2023. 2

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2018. 7

[7] Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. SSCGAN: Facial attribute editing via style skip connections. In *Proceedings of the European Conference on Computer Vision*, pages 414–429. Springer, 2020. 2

[8] Laurent Colbois, Tiago de Freitas Pereira, and Sébastien Marcel. On the use of automatically generated synthetic image datasets for benchmarking face recognition. In *Proceedings of the IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE, 2021. 2, 3, 4, 5

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 6, 7

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021. 2

[11] Magnus Falkenberg, Anders Bensen Ottsen, Mathias Ibsen, and Christian Rathgeb. Child face recognition at scale: Synthetic data generation and performance benchmark. *arXiv preprint arXiv:2304.11685*, 2023. 3, 4, 5

[12] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, 129(7):2288–2307, 2021. 2

[13] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection, 2021. 6

[14] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3DDFA. https://github.com/cleardusk/3DDFA, 2018. 4

[15] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision*, 2020. 4

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

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[18] Mathias Ibsen, Christian Rathgeb, Fabian Brechtel, Ruben Klepp, K Pöppelmann, A George, S Marcel, and C Busch. Attacking Face Recognition with T-shirts: Database, Vulnerability Assessment and Detection. *IEEE Access*, 2023. 8

[19] Indu Joshi, Marcel Grimmer, Christian Rathgeb, Christoph Busch, Francois Bremond, and Antitza Dantcheva. Synthetic data in human analysis: A survey. *arXiv preprint arXiv:2208.09191*, 2022. 1

[20] Amina Kammoun, Rim Slama, Hedi Tabia, Tarek Ouni, and Mohmed Abid. Generative adversarial networks for face generation: A survey. *ACM Computing Surveys*, 55(5):1–37, 2022. 1

[21] Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 4

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 1, 3, 4

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[24] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. DCFace: Synthetic Face Generation with Dual Condition Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[25] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[26] Adam Kortylewski, Andreas Schneider, Thomas Gerig, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Training deep face recognition systems with synthetic data. *arXiv preprint arXiv:1802.05891*, 2018. 1

[27] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. Studying Bias in GANs Through the Lens of Race. In *Proceedings of the European Conference on Computer Vision*, pages 344–360. Springer, 2022. 2

[28] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. IARPA Janus Benchmark-C: Face Dataset and Protocol. In *Proceedings of the International Conference on Biometrics*, pages 158–165. IEEE, 2018. 7

[29] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 4, 7

[30] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. SensitiveNets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020. 1

[31] Madhumita Murgia and Max Harlow. Who's using your face? The ugly truth about facial recognition. *Financial Times*, 19, 2019. 1

[32] Joao C Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020. 8

[33] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. SynFace: Face Recognition With Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021. 2, 3

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 6

[35] Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch. *Handbook of digital face manipulation and detection: from DeepFakes to morphing attacks*. Springer Nature, 2022. 1, 8

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 6

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 6

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 6

[39] Divya Saxena and Jiannong Cao. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys*, 54(3):1–42, 2021. 1

[40] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6248–6257, 2021. 4

[41] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 2, 4

[42] Ruben Tolosana, Sergio Romero-Tapiador, Ruben Vera-Rodriguez, Ester Gonzalez-Sosa, and Julian Fierrez. DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. *Engineering Applications of Artificial Intelligence*, 110:104673, 2022. 8

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

[44] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 2

[45] Yutong Xie, Zhaoying Pan, Jinge Ma, Jie Luo, and Qiaozhu Mei. A prompt log analysis of text-to-image generation systems. *arXiv preprint arXiv:2303.04587*, 2023. 6

[46] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2M-GAN: Learning to Manipulate Latent Space Semantics for Facial Attribute Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2951–2960, 2021. 2

[47] Haoyu Zhang, Marcel Grimmer, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. On the applicability of synthetic data for face recognition. In *Proceedings of the IEEE International Workshop on Biometrics and Forensics*, pages 1–6. IEEE, 2021. 1, 2