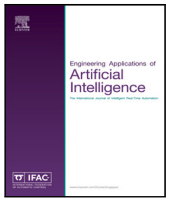
Engineering Applications of Artificial Intelligence 133 (2024) 108495

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research Paper

Generating bimodal privacy-preserving data for face recognition

Darian Tomašević [a,*], Fadi Boutros [b], Naser Damer [b], Peter Peer [a], Vitomir Štruc [c]

[a] University of Ljubljana, Faculty of Computer and Information Sciences, Večna pot 113, Ljubljana, 1000, Slovenia
[b] Fraunhofer Institute for Computer Graphics Research IGD, Fraunhoferstr. 5, Darmstadt, 64283, Germany
[c] University of Ljubljana, Faculty of Electrical Engineering, Tržaška cesta 25, Ljubljana, 1000, Slovenia

ARTICLE INFO

Keywords:
Image synthesis
Face-based biometrics
Privacy-preserving data
Multispectral recognition
Generative adversarial networks

ABSTRACT

The performance of state-of-the-art face recognition systems depends crucially on the availability of large-scale training datasets. However, increasing privacy concerns nowadays accompany the collection and distribution of biometric data, which has already resulted in the retraction of valuable face recognition datasets. The use of synthetic data represents a potential solution, however, the generation of privacy-preserving facial images useful for training recognition models is still an open problem. Generative methods also remain bound to the visible spectrum, despite the benefits that multispectral data can provide. To address these issues, we present a novel identity-conditioned generative framework capable of producing large-scale recognition datasets of visible and near-infrared privacy-preserving face images. The framework relies on a novel identity-conditioned dual-branch style-based generative adversarial network to enable the synthesis of aligned high-quality samples of identities determined by features of a pretrained recognition model. In addition, the framework incorporates a novel filter to prevent samples of privacy-breaching identities from reaching the generated datasets and improve both identity separability and intra-identity diversity. Extensive experiments on six publicly available datasets reveal that our framework achieves competitive synthesis capabilities while preserving the privacy of real-world subjects. The synthesized datasets also facilitate training more powerful recognition models than datasets generated by competing methods or even small-scale real-world datasets. Employing both visible and near-infrared data for training also results in higher recognition accuracy on real-world visible spectrum benchmarks. Therefore, training with multispectral data could potentially improve existing recognition systems that utilize only the visible spectrum, without the need for additional sensors.

1. Introduction

Modern face recognition systems heavily rely on deep learning models and the availability of large-scale training datasets to achieve state-of-the-art performance (Rot et al., 2019; Vitek et al., 2021; Batagelj et al., 2021; Emeršič et al., 2021). In the past, such datasets were commonly acquired from online sources, social media and other web platforms containing facial images captured in various settings. Nowadays, however, the collection, distribution and use of biometric data is accompanied by ever-increasing privacy and ethical concerns (Jasserand, 2018; Meden et al., 2021) and is governed by privacy acts and data-protection legislation, such as the General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019). The consequences of these developments are especially evident when discussing face image datasets collected through web-scraping without proper consent. Upholding proposed regulations for such datasets is not only impractical but near-impossible, which has recently resulted in the retraction of several valuable face recognition datasets in their entirety or in parts (Guo et al., 2016a; Bansal et al., 2017; Cao et al., 2018). Alternatively, manually gathering the required large-scale datasets represents a labor-intensive and time-consuming task (Vitek et al., 2020). Even then, potential use cases of the gathered data must be clearly defined in the consent agreement, which may limit future research.

To improve existing biometric solutions, researchers are also investigating the use of near-infrared and thermal data, which contain cues not present in the commonly utilized visible spectrum (Bourlai, 2016; Chambino et al., 2021; Rose et al., 2022; Martins et al., 2022). Merging the different data sources and extending models to operate on multispectral data therefore holds great potential for enhancing recognition and segmentation performance. However, the availability of large-scale multispectral datasets is rather limited (Sequeira et al., 2017; Panetta et al., 2018). This is particularly true for data concurrently captured across different spectra, as it necessitates custom setups of multiple imaging sensors.

* Corresponding author.
E-mail address: darian.tomasevic@fri.uni-lj.si (D. Tomašević).

https://doi.org/10.1016/j.engappai.2024.108495
Received 15 November 2023; Received in revised form 5 March 2024; Accepted 19 April 2024
Available online 30 April 2024
0952-1976/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

To address the increasing privacy concerns as well as the lack of large-scale biometric datasets, recent research is considering the use of synthetic data (Boutros et al., 2023d,e; Joshi et al., 2024). Such research efforts are primarily fueled by the advances in deep generative models. In particular, Style-based Generative Adversarial Networks (StyleGANs) (Karras et al., 2019, 2020b, 2021) and Diffusion Models (DMs) (Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022), which have recently facilitated the generation of diverse large-scale datasets of photorealistic images. These approaches have also been extended to allow for conditioned data generation based on the desired image features, thus enabling better control over the synthesized images (Shoshan et al., 2021; Zhang et al., 2023a). By exploiting these capabilities, researchers have also demonstrated the possibility of identity conditioned generation of face images (Qiu et al., 2021; Boutros et al., 2022, 2023c,b). The result is the creation of large-scale synthetic datasets that can be used to train highly-accurate face recognition models. Unfortunately, however, there still exists a performance gap compared to models trained on real-world data. Though, it is slowly being bridged by novel approaches that balance two important aspects of recognition datasets, i.e. identity separability and intra-identity diversity (Boutros et al., 2023b).

Despite these incredible advancements, existing research has not yet adequately explored the privacy of synthetic data. Generative models do not necessarily guarantee the preservation of privacy, since the produced samples might still contain identities that match real-world subjects used for training the data generators (Tinsley et al., 2021). Solving such identity leakage is critical for ensuring privacy-preservation in synthetic datasets, but has not been discussed widely so far in the open literature (Singh et al., 2024). Additionally, despite the potential of recent multispectral recognition approaches to enhance the state-of-the-art, the generation of suitable multispectral data has remained rather limited (Tomašević et al., 2022), apart from cross-spectral image translation (Wu et al., 2019; Luo et al., 2022).

In this paper, we address the outlined privacy concerns and the lack of multispectral recognition data by introducing a novel generative framework called ArcBiFaceGAN. Our framework extends the existing StyleGAN-based methods for synthetic data generation (Karras et al., 2020a; Boutros et al., 2022; Tomašević et al., 2022) and enables the simultaneous synthesis of privacy-preserving visible (VIS) and near-infrared (NIR) face images conditioned on identity features of a pretrained recognition model, as seen in Fig. 1. This, in turn, facilitates the creation of large-scale multispectral datasets with diverse and high-quality samples of synthetic identities that can be used to train face recognition models without breaching the privacy of real-world subjects. To this end, ArcBiFaceGAN utilizes a novel identity-conditioned Dual-Branch StyleGAN2 model to generate VIS-NIR image pairs of a given input identity sampled from the latent space of the ArcFace recognition model (Deng et al., 2019a). In addition, the framework relies on an innovative Privacy and Diversity (PD) filter that ensures the removal of privacy breaching identities, with the use of the above recognition model (Deng et al., 2019a). It also improves both identity separability and intra-identity diversity, by rejecting identities that match previously generated identities and removing samples that are too similar to previous samples of the same identity. We compare the synthesis capabilities of our ArcBiFaceGAN framework with the state-of-the-art, in terms of quality, diversity, identity separability, and privacy through a series of experiments on the multispectral Tufts Face Database (Panetta et al., 2018). Furthermore, we employ the produced multispectral synthetic data to train a modern recognition model and evaluate its performance on five state-of-the-art recognition benchmarks. Overall, we showcase that our framework achieves highly competitive synthesis results and enables better recognition performance than even real-world data, while preserving the privacy of real-world subjects.

In summary, this paper makes the following contributions:

- We propose ArcBiFaceGAN, a potent framework for generating large-scale multispectral datasets, suitable for training modern recognition approaches in a privacy-aware manner.
- We present a novel identity-conditioned Dual-Branch StyleGAN2 model, capable of creating diverse and high-quality aligned visible and near-infrared image pairs of synthetic identities, based on a small-scale dataset of poorly aligned training images.
- We introduce a novel Privacy and Diversity (PD) filter that ensures the removal of privacy-breaching synthetic samples while improving both intra-identity diversity and identity separability of samples.
- We demonstrate that recognition models trained on large-scale synthetic datasets can surpass those trained on smaller real-world datasets. In addition, we show that multispectral synthetic data can even be used to improve the recognition performance on only visible spectrum data.

2. Related work

This section provides an overview of existing research on image generation and positions our contributions with respect to the state-of-the-art approaches for generating biometric data suitable for recognition tasks. For an in-depth discussion on synthetic data for face recognition, the reader is referred to some of the recent surveys on this topic (Boutros et al., 2023e; Joshi et al., 2024).

2.1. Image generation

Image synthesis techniques have undergone drastic evolution in the past decade with the emergence of deep generative models. In particular, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) were the first models to enable the creation of convincing and hiqh-quality images. This was achieved by employing two distinct neural networks; a generator, responsible for producing images, and a discriminator, designed to identify synthetic images and in turn guide the generator. In the following years, numerous enhancements to GAN models were proposed, including the use of multiple discriminators (Durugkar et al., 2017) and progressive architectures (Karras et al., 2018) to improve the quality of produced images. Issues with training stability were also addressed with novel regularization methods (Mescheder et al., 2018; Miyato et al., 2018) and custom distribution distance measures (Arjovsky et al., 2017; Gulrajani et al., 2017). In addition, control over the image generation process was explored with the use of class labeled inputs (Mirza and Osindero, 2014). However, despite extensive analysis, the source of various stochastic features in the generated images remained unknown and, hence, challenging to control (Bau et al., 2019).

Overall, the generator continued to function as a black box and the GAN architecture remained fairly unchanged, until the introduction of the Style-based GAN model (i.e. StyleGAN) by Karras et al. (2019). To achieve a level of image quality similar to real-world images and improve control over the synthesis process, the authors proposed separating the generator into a mapping and a synthesis network. The first determines the image style and passes this information to the latter network, which generates the corresponding image. Nevertheless, the images generated by StyleGAN (Karras et al., 2019) still contained noticeable artifacts, e.g. blob-like shapes. To address these issues, Karras et al. (2020b) proposed an improved model (StyleGAN2), which contained redesigned building blocks in the synthesis network and utilized additional skip connections. In addition, the authors resolved the issue of textures sticking to image coordinates rather than underlying surfaces, which enabled the generation of smooth image sequences (Karras et al., 2021). Notably, Karras et al. (2020a) also introduced a novel Adaptive Discriminator Augmentation (ADA) method

Fig. 1. Samples generated by our ArcBiFaceGAN framework. The framework enables the synthesis of diverse high-quality aligned visible (VIS) and near-infrared (NIR) images of new synthetic identities that do not breach the privacy of real-world subjects.

that applies various augmentations to images before they are passed to the discriminator network. This drastically lowered the amount of training data required to learn the model, whilst preventing the augmentations from leaking to the generator and affecting the image quality. These enhancements allowed for a wide variety of new practical use cases in various fields, e.g. image-based biometrics, with scarce large-scale datasets.

Recently, however, the synthesis capabilities of GAN-based approaches have been rivaled by the development of novel diffusion models (Dhariwal and Nichol, 2021). At their core, diffusion models learn to reverse a process that gradually degrades the training data with noise at different scales (Sohl-Dickstein et al., 2015). Initially, these models were applied directly to data in the pixel space, however, their performance was significantly improved, both in terms of speed and quality, by instead utilizing a lower-dimensionality latent space of a pretrained autoencoder (Rombach et al., 2022). Combining diffusion models with external text encoders has also enabled guided image generation, conditioned on text prompts (Saharia et al., 2022), which has garnered incredible success in both industry and research (Croitoru et al., 2023). Control over the generation process has also been improved by conditioning pretrained diffusion models on a variety of spatial-based inputs, e.g. segmentation masks, hand-drawn sketches, depth maps, edge maps and even pose skeletons via an external trainable copy of the model (Zhang et al., 2023b). These capabilities have been expanded to facilitate simultaneous prompting based on encoded text and image features with the use of a decoupled cross-attention mechanism (Ye et al., 2023). In addition, fine-tuning approaches have been introduced that can add new concepts to pretrained diffusion models based on a limited amount of input data. Namely, DreamBooth (Ruiz et al., 2023) has been extensively utilized to generate synthetic images of desired real-world subjects. However, it often struggles with identity consistency when generating new images of real-world identities. Nevertheless, in comparison to GAN-based approaches, current diffusion models still suffer from slow inference speeds and their application in low-data regimes has mostly remained limited to fine-tuning (Moon et al., 2022).

Additionally, most generative models have remained tied solely to data in the visible spectrum (Tomašević et al., 2022), despite the benefits that multispectral data can provide in a variety of tasks and the overall need for larger multispectral datasets (Bourlai, 2016). To address this, we build in this work on the StyleGAN2-ADA (Karras et al., 2020a) model and extend it for the creation of multispectral identity-conditioned facial images.

2.2. Synthesis of biometric recognition datasets

The utilization of synthetic data for training powerful deep learning models has become a focal point in the field of biometrics in recent years. This interest is fueled by ever-increasing privacy and copyright-related concerns regarding the use and sharing of real-world biometric data (Jasserand, 2018). This is especially evident when discussing face recognition datasets, as face images have commonly been collected via web-scraping without suitable consent in the past. Several of these crucial large-scale datasets have recently been retracted or removed from public repositories (Jasserand, 2022), due to the introduction of various privacy acts and data protection regulations, e.g., the GDPR (Hoofnagle et al., 2019). This poses a potential issue for future face recognition research, which heavily relies on the availability of large-scale datasets to achieve state-of-the-art results.

In an effort to address these concerns, researchers are exploring the creation of synthetic biometric data with modern generative approaches, which can then, in turn, be used to train recognition models in a privacy-preserving manner. Specifically, existing research has explored the generation of face recognition datasets, but also data for other biometric modalities (Joshi et al., 2024), by relying on advancements in conditional generative models (Boutros et al., 2023e). The suitability of synthetic face images for biometric recognition was recently analyzed by Zhang et al. (2021). With the use of modern face quality assessment techniques, they showed that images generated by StyleGAN-based (Karras et al., 2019, 2020b) models could compete with the quality and utility of real-world samples. A human study of synthetic face images was also carried out by Shen et al. (2021), who demonstrated that not even human subjects could reliably distinguish synthetic images from real ones.

Deng et al. (2020) proposed a novel StyleGAN-based framework, called DiscoFaceGAN, which utilized 3D face priors and a custom imitative-contrastive learning scheme to achieve a disentangled and interpretable latent space. The framework enabled control over several attributes of the face image generation process, including the identity, expression, pose and lighting. However, Qiu et al. (2021) showcased that training face recognition models with data of DiscoFaceGAN resulted in worse performance in comparison to training with real-world data. They attributed the performance difference to the domain gap between synthetic and real images as well as the low intra-identity variation of synthetic images. To address the observed weaknesses, they extended DiscoFaceGAN with identity and domain mixup of synthetic and real data during training. Recently, Boutros et al. (2022) proposed an alternative face generation method, called SFace, based on the StyleGAN2-ADA model (Karras et al., 2020a). To create face recognition datasets of synthetic identities they conditioned the generative model on identity labels in the form of a one-hot encoded vector. In their experiments, the authors demonstrated that SFace achieved higher intra-identity variation than previous methods, at the cost of lower identity separability. They also observed that a low, but not negligible, cross-identifiability exists between synthetic and real samples, at least with large-scale training datasets. However, the reliance on one-hot encoded vectors presents a limit on the amount of possible identities that can be generated with this approach. In their follow up work (Boutros et al., 2023c), the authors also investigated the possibility of creating recognition data by disentangling identity information from latent spaces of pretrained non-conditional StyleGAN models (Karras et al., 2020a, 2021). They showcased that by determining latent identity directions, they could construct positive and negative class examples without the need for identity-labeled datasets.

More recently, Boutros et al. (2023b) explored the generation of face recognition data with modern latent diffusion models (Rombach et al., 2022). In their work, they proposed the IDiff-Face model, which conditions the denoising U-Net (Ronneberger et al., 2015) network on identity features of a pretrained face recognition model. Additionally, the authors introduced contextual partial dropout during training to prevent overfitting on real-world identities and enable control over the trade-off between identity separability and intra-identity diversity. Throughout their experiments they showcased that IDiff-Face achieved unprecedented image quality and intra-identity diversity, thus enabling the training of better performing recognition models.

However, despite the advancements in identity-conditioned synthesis capabilities, the generation of recognition datasets that do not breach the privacy of real-world subjects has not been adequately discussed in the open literature so far. Furthermore, existing research has only explored the generation of face-image data in the visible spectrum, despite the advantages that multispectral data can provide, especially for improving recognition performance (Bourlai, 2016). This is likely due to the small scale and poor alignment of available multispectral datasets, which present a difficult obstacle for training deep generative models (Tomašević et al., 2022).

Differently from existing works, we focus in this paper specifically on the generation of privacy-preserving multispectral face recognition datasets. To this end, we build on existing GAN-based approaches (Tomašević et al., 2022; Boutros et al., 2022; Karras et al., 2020a) and propose a novel identity-conditioned generative framework, ArcBiFaceGAN, that is capable of generating aligned visible (VIS) and near-infrared (NIR) face images of synthetic identities, even when faced with a limited amount of poorly aligned training data. Additionally, we introduce a novel filtering component, that enables the removal of privacy-breaching identities while ensuring better intra-identity diversity. To enable more precise control over the identity aspect than existing GAN-based approaches, we also condition our model on identity features of a pretrained recognition model.

3. Methodology

The main contribution of this paper is the proposed ArcBiFaceGAN framework that enables the generation of identity-labeled privacy-preserving bimodal face images. This section presents an overview of the framework and provides detailed descriptions of its main components.

3.1. Overview of ArcBiFaceGAN

The proposed ArcBiFaceGAN framework relies on two key components: (i) the identity-conditioned Dual-Branch StyleGAN2 (DB-StyleGAN2) model (Section 3.2) that generates VIS-NIR image pairs of a desired identity, and (ii) the Privacy and Diversity (PD) filter (Section 3.3), which ensures the synthesis of new privacy-preserving identities and diverse intra-identity samples. Together these components facilitate the generation of high-quality bimodal synthetic data, which can be used to train deep face recognition models without breaching the privacy of real-world subjects, as depicted in the right column of Fig. 2.

The image generation process is based on two input vectors, the randomly sampled latent code $z \in \mathbb{Z}$ and the identity code $id \in \mathbb{A}$, sampled from the latent space of a pretrained recognition model (Deng et al., 2019a). The mapping network f combines the inputs into the style information $w \in \mathbb{W}$ that is then passed to the Dual-Branch Synthesis network g to create the corresponding VIS and NIR images (i.e. $\mathbf{x}^{id}_{VIS}$ and $\mathbf{x}^{id}_{NIR}$) of the desired identity id. These samples are then evaluated by the Privacy and Diversity filter PD that removes samples if they (i) do not contain a face, (ii) contain privacy-breaching identities, (iii) contain identities that have already been generated, (iv) do not contain the same identity as previous samples for a given id, or (v) contain samples that are too similar to previously generated samples of a given id, as detailed in Section 3.3. Formally, this synthesis process can be expressed as:

$$\{\mathbf{x}^{id}_{VIS}, \mathbf{x}^{id}_{NIR}\} = PD(\{\mathbf{x}^{id}_{VIS}, \mathbf{x}^{id}_{NIR}\}) = PD(g(w)) = PD(g(f(id, z))). \quad (1)$$

Thus, based on the randomly sampled latent code z and identity code id the proposed ArcBiFaceGAN framework produces a multispectral image pair $\{\mathbf{x}^{id}_{VIS}, \mathbf{x}^{id}_{NIR}\}$ of a privacy-preserving synthetic identity id.

3.2. Identity-conditioned dual-branch StyleGAN2

The synthesis capabilities of the proposed ArcBiFaceGAN framework are rooted in the identity-conditioned Dual-Branch StyleGAN2 (DB-StyleGAN2). The proposed model extends the dual-branch architecture of the BiOcularGAN approach (Tomašević et al., 2022) to enable the synthesis of identity-specific samples. Differently from existing identity-conditioned GAN-based approaches, our method does not condition the generative model on class labels (e.g. one-hot encoded vectors as with SFace (Boutros et al., 2022)) but rather on identity features from the latent space of a pretrained ArcFace recognition model (Deng et al., 2019a). This, in turn, facilitates more flexible and detailed control over the identity sampling and bypasses potential limits of identity capacity. In total, the identity-conditioned DB-StyleGAN2 consists of two key components, i.e., (i) the generator that produces data and (ii) the discriminators that are responsible for training, as depicted in Fig. 2. Details on the two components are given below.

3.2.1. Identity-conditioned DB-StyleGAN2 generator

The generator network of our identity-conditioned DB-StyleGAN2 is presented in Fig. 3. As can be seen, it is split into two connected networks, where the mapping network f determines the style of the image that the synthesis network g generates. The mapping network receives as input two 512-dimensional latent codes, the random input z and the identity condition id, that jointly guide the generative process. The identity condition id is first passed through an initial fully-connected layer that reinterprets the identity information but retains the feature dimensions. Both the z and the reinterpreted id feature codes are then normalized separately and concatenated into a single latent code. The combined code is then passed through 8 fully-connected layers, which map the input information to an intermediate 512-dimensional latent code w, that defines the style of the image to be created. Here, the initial identity-based fully-connected layer is crucial for enabling better control over the image generation process during inference. In particular, more diverse identities can be generated by multiplying the sampled identity condition id, as demonstrated in Section 4.4.4. Importantly, this can be done without affecting the quality of produced samples because the multiplication effect is limited to the initial fully-connected layer and thus does not overwhelm the entire mapping network and consequently the synthesis network.

The synthesis network g incorporates seven consecutive synthesis blocks that increase in powers of two from 4×4 to 256×256. Each block consists of two style blocks, except the first, which contains only one style block. As input, the first block receives a constant $4 \times 4 \times 512$ feature c, which is then passed to a convolutional 3×3 layer. Meanwhile, the style information w from before is introduced into the convolutional weights in order to influence the generative process. This is done by utilizing the modulation and demodulation operations (Karras et al., 2020b), which simulate the Adaptive Instance Normalization (AdaIN) technique (Karras et al., 2019). After each convolutional layer, additional bias and noise are also incorporated into the signal, before the Leaky ReLU activation function (Maas et al., 2013) is applied. The signal is then passed to the next style block and the process repeats. If this entails crossing to the next synthesis block then the signal is also upsampled to the correct size.

After each respective synthesis block two output feature maps are created, one for the VIS and one for the NIR spectrum, by passing the

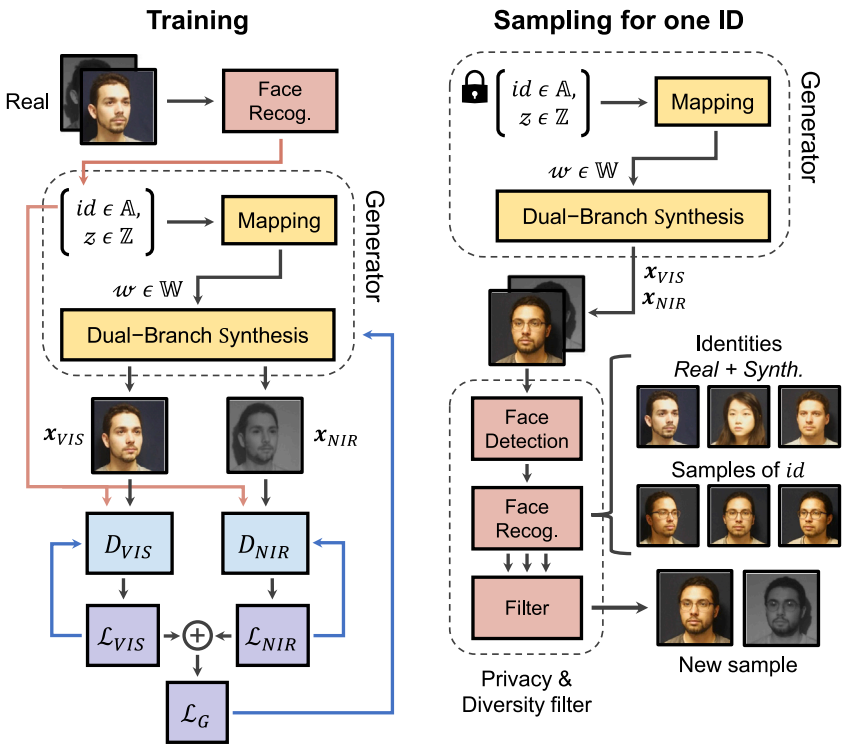


Fig. 2. High-level overview of the proposed ArcBiFaceGAN framework. The framework conditions a Dual-Branch StyleGAN2 model (Tomašević et al., 2022) on identity features id of a pretrained recognition model (Deng et al., 2019a) to generate visible (VIS) and near-infrared (NIR) images of a desired identity. The model is trained using separate discriminators (D_{VIS} and D_{NIR}) for each spectrum, which compute the loss based on input images and the id feature. During the data generation process, we first sample and lock the id vector, and then randomly sample input latent codes z to generate multiple samples of the target identity. The created samples are then passed to the Privacy and Diversity filter that: (i) removes samples whose identities match any real-world subject or previously generated identity, and (ii) ensures intra-identity diversity by removing images that are too similar to previous samples of the same identity.

current latent signal through separate 1×1 convolutional layers. These layers act as a sort of renderer that applies the correct spectrum-related appearance onto the shared feature map. On the outside, the synthesis network forms two branches that upsample and merge the spectrum-specific outputs of the synthesis blocks. The result is the creation of VIS and NIR image pairs, which are well-aligned due to the shared semantic information.

3.2.2. Identity-conditioned discriminators

To train the multispectral generator we rely on two identity-conditioned discriminator networks, one for each light spectrum (D_{VIS} and D_{NIR}). Their role is to determine the authenticity of the synthesized images, i.e. decide if the images are synthetic or genuine. The produced feedback is then used to improve the generator and facilitate the synthesis of VIS-NIR image pairs that compete in quality with the training distribution. Both discriminator networks utilize the same ResNet-like (He et al., 2016) downsampling architecture, similar to previous StyleGAN-based approaches (Karras et al., 2020b,a; Tomašević et al., 2022), as illustrated in Fig. 4.

Each discriminator receives as input either a synthetic or a training sample in the corresponding spectrum as well as the identity feature id. The image samples are first passed through a 1×1 convolutional layer before reaching a series of seven resolution blocks, each consisting of two convolutional layers connected by an auxiliary skip connection. These downsample the image resolution from 256×256 back to 4×4, each step by a power of two. At the end, the discriminator incorporates a single convolutional layer and a fully-connected layer that takes into account the obtained latent representation of the image as well as the identity information to decide on the authenticity of the presented sample. Before that, however, the original id feature is passed through a separate mapping network akin to the network of the generator but with different weights and without the latent code z, to enable better decision making.

3.2.3. Identity-conditioned training

As noted above, the identity-conditioned DB-StyleGAN2 utilizes two discriminator networks to train and improve its synthesis capabilities. The feedback of both discriminators is combined in a multi-task adversarial learning objective, which augments existing multispectral-based objectives (Tomašević et al., 2022) with the additional identity information, as seen in the left column of Fig. 2. Differently from existing methods (Boutros et al., 2022), we rely on identity features that are extracted from face images with a pretrained ArcFace recognition model (Deng et al., 2019a).

At the core of the learning objective lies the Non-Saturating Logistic loss (Goodfellow et al., 2014) that is implemented with the soft-plus operation, i.e. $sp(x) = \log(1 + \exp(x))$, as is standard practice. The loss function is accompanied by two crucial regularization methods, including R_1 regularization (Mescheder et al., 2018) and path length regularization (R_{PL}) (Karras et al., 2020b), that stabilize and improve the training process. These are applied only every 16 mini-batches so as not to negatively impact the speed of training.

The aim of both discriminator networks is to improve their ability to distinguish between authentic and synthetic samples. Thus, their corresponding learning objectives $\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$ can be defined as:

$$\mathcal{L}_b = sp(-D_b(id, \mathbf{y}_b)) + sp(D_b(id, \mathbf{x}_b)) + \frac{\gamma_{R_1}}{2} \mathbb{E}\left[\|\nabla D_b(id, \mathbf{y}_b)\|^2\right], \tag{2}$$

where b represents either the visible (VIS) or the near-infrared (NIR) spectrum. Here, real (training) images and produced synthetic images are denoted with $\mathbf{y}_b$ and $\mathbf{x}_b$ respectively, while id represents the identity feature. Lastly, the regularization hyperparameter γ_{R_1} is determined as $\gamma_{R_1} = 10^{-4}(2res^2/bs)$ based on the batch size bs and resolution res of the images (Karras et al., 2020b).

The two discriminator learning objectives $\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$ are then combined to form the objective of the generator network $\mathcal{L}_G$, as visualized in Fig. 2. However, the generator is tied solely to the production

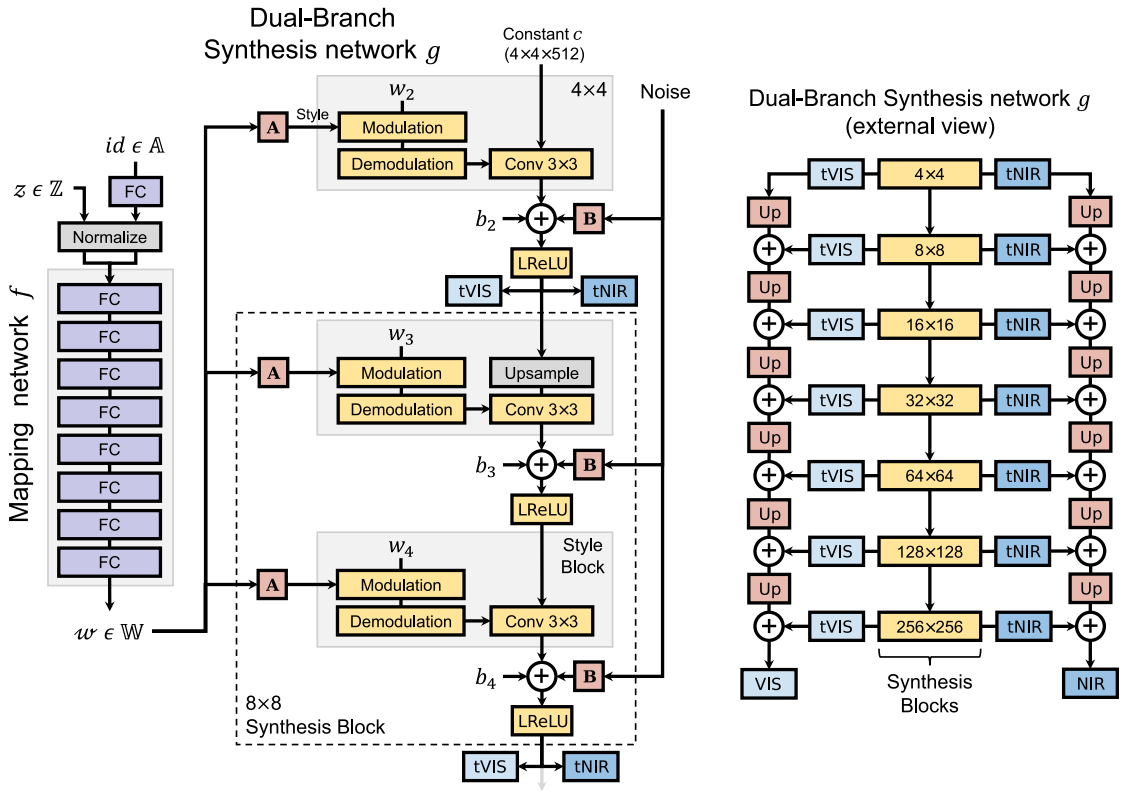


Fig. 3. Illustration of the identity-conditioned DB-StyleGAN2 generator architecture. The generator consists of a mapping network g and a synthesis network f. The synthesis network f features a main branch that encodes the facial semantics and a pair of output branches after each processing block that render the images in the VIS or NIR spectrum, denoted as tVIS and tNIR. These outputs are then upsampled and merged to form the final generated VIS and NIR images. The networks rely on fully-connected layers (FC), convolutional layers (Conv), and upsampling (Up).

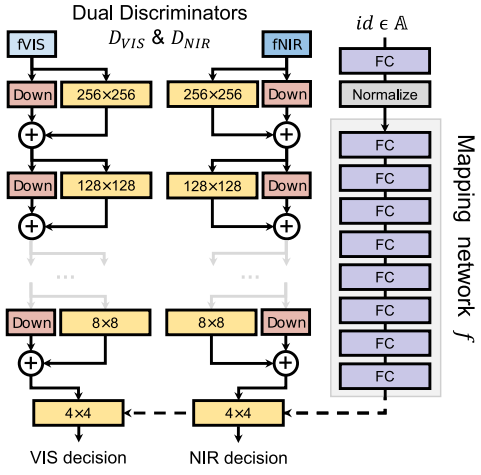


Fig. 4. Illustration of the identity-conditioned discriminator networks. To train the identity-conditioned DB-StyleGAN2 we utilize two separate discriminator networks (D_{VIS} and D_{NIR}), one for each light spectrum. These determine the authenticity of generated images, which is used to improve the synthesis capabilities. The discriminators operate on the high-dimensional representations of images, which are obtained from fVIS and fNIR blocks. In the image, Down denotes downsampling.

of synthetic images, and never encounters real-world samples. This is reflected in the merged learning objective, as it incorporates only terms from the discriminator objective that involve synthetic images $\mathbf{x}_b$. Thus, the generator learning objective can be formally defined as:

$$\mathcal{L}_G = \sum_{b \in B} w_b \cdot sp(-D_b(id, \mathbf{x}_b)) + \gamma_{R_{PL}} \mathbb{E}\Big(\Big\| \sum_{b \in B} \nabla(\mathbf{x}_b q_b) \Big\| - a \Big)^2, \tag{3}$$

where the notations defined before also apply. Furthermore, w represents a weight parameter that determines the effect of a given spectrum b on the final loss value, q denotes images with normally distributed pixel intensities, and a is the average of the computed norm. Here, the regularization hyperparameter $\gamma_{R_{PL}}$ is computed based on the resolution res as $\gamma_{R_{PL}} = \ln 2/(res^2(\ln res - \ln 2))$ (Karras et al., 2020b).

Importantly, due to the small-scale of available multispectral face datasets, we also critically depend on the Adaptive Discriminator Augmentation (ADA) technique (Karras et al., 2020a) to enable more stable training in a low-data regime. This process entails augmenting synthetic or authentic images, before they are passed to the discriminator networks, with geometric and color transforms based on an adaptive probability.

3.3. Privacy and diversity filter

In our work, we propose a novel Privacy and Diversity (PD) filter, which can be used as an auxiliary component during data generation to ensure the creation of privacy-preserving datasets and improve both identity separability and the intra-identity diversity of samples. This is crucial for creating large-scale high-quality recognition datasets that can be used to train biometric systems in a way that does not breach the privacy of real-world subjects.

The proposed PD filter relies on identity features which are extracted from images with a pretrained face recognition model, namely ArcFace (Deng et al., 2019a). In addition, the filter utilizes the pretrained Multi-Task Cascaded Convolutional Neural Network (MTCNN) (Zhang et al., 2016) face detector, to determine the presence of faces in the image. To compare the similarity of different identities, our filter incorporates the cosine similarity function (Nguyen and Bai, 2010), following existing research (Deng et al., 2019a; Wang et al., 2018;

Boutros et al., 2023a). Formally, the similarity sim between two identity features a and b can be computed as:

$$sim(a, b) = \frac{a \cdot b}{\|a\| \, \|b\|}. \quad (4)$$

To determine which images are problematic, the filter utilizes predetermined identity similarity thresholds τ. These are based on the characteristics of the training dataset. Specifically, we rely on the mean μ and standard deviation σ values of intra-identity similarity of real images to determine the thresholds. The specific thresholds used in our experiments are presented below, when explaining the filtering process. However, these thresholds can be changed to attain the desired privacy and diversity levels of synthetic data, suitable for the task at hand.

During the data generation process, the proposed PD filter follows the following procedure:

- **Step 1 - Identity Initialization:** When creating the first sample of a new synthetic identity, the filter uses the pretrained face detector to check if the synthetic sample even contains a face. Based on the obtained landmarks, it also determines if the face is in a frontal pose. This is done by ensuring that the distance between the nose and each eye landmark is similar, considering a predetermined threshold based on the distance between they two eyes. The filter then removes the image, if it does not suit this criterion. This initial step is performed to ensure that all subsequent samples of the same identity have a decent base for identity comparison and forces the model to synthesize a frontal image during initialization.
- **Step 2 - Privacy Filter:** Next, for each of the generated samples the filter extracts a 512-dimensional identity feature vector using the pretrained face recognition model. Here, the computed feature vector is first compared to the feature vectors of the real-world identities of the training dataset. Note that we utilize only one feature vector per identity that is most representative for each real-world subject to (drastically) speed up the sampling procedure. This representative feature vector is extracted from the sample that is most similar to all other samples of the same identity, typically the one with a front facing subject. If the synthetic identity in question is too similar, in terms of cosine similarity, to any of the real-world identities, based on a predetermined threshold $\tau_{ID} = \mu - 2\sigma$, it is discarded. In turn, this ensures the removal of privacy-breaching synthetic samples. A similar check is then performed but with previously generated synthetic identities. If the new synthetic identity is too similar to an existing synthetic identity with the same threshold τ_{ID}, then the sample is again discarded. Thus the proposed PD filter not only ensures privacy, but also improves the identity separability of the generated synthetic identities.
- **Step 3 - Consistency Check:** If the generated sample in question is not the first sample of given synthetic identity, then the proposed filter also checks that the generated identity is consistent with the first generated sample of a given id. Again, the similarity of identity features is measured and compared with the predetermined threshold τ_{ID}. Furthermore, the sample is compared to all previously generated samples of the same identity to ensure that it is also not too similar, so as to avoid duplicate samples with the same pose. For this, the filter utilizes a high identity similarity threshold $\tau_{High} = \mu + \sigma$. This process thus results in improved intra-identity diversity of synthetic samples. If the generated sample passes all the listed criteria then it is allowed into the final synthetic dataset.

It should also be noted that the decision to mostly rely on representative identity features was made to address the otherwise inefficient sampling process, whilst achieving almost the same results. Sampling speed becomes an issue especially in cases where datasets contain large quantities of samples for each subject. However, if required, the identity comparison could be made more extensive, e.g. in a real-world application with extremely sensitive data.

4. Experiments and results

This section covers the evaluation of the proposed ArcBiFaceGAN framework. First, we introduce the utilized face datasets and performance indicators, as well as provide additional implementation details. Next, we present a variety of experiments through which we compare our solution to the state-of-the-art.

4.1. Experimental setup

4.1.1. Training data

Throughout the experiments, we rely on the multispectral **Tufts Face Database** (Panetta et al., 2018) to train and evaluate the different generative models. The dataset, summarized in Table 1, contains heterogeneous facial data of 113 individuals, including over 10,000 images in various spectra. This includes face images in both the visible (VIS) and near-infrared (NIR) spectrum, which were captured with a custom quad-camera setup in a semi-circle around the individuals. VIS images were taken by four visible field cameras under constant diffused light, while four night vision cameras and a 850 nm Infrared 96 light system were used for NIR imaging. Importantly, however, the VIS and NIR images were not captured simultaneously, resulting in notable image pair misalignment.

To make the data more suitable for training our proposed bimodal generative model, we utilized the following preprocessing steps. First, we removed blurry images, in which subjects moved, and side-profile images, which lacked crucial facial features (e.g. two eyes) and then cropped the images to focus on the region of interest, i.e., the face. Next, we defined an affine transform, based on the angle between the centroids of the eyes and the distance between them, with the use of the pretrained Multi-Task Cascaded Convolutional Neural Network (MTCNN) (Zhang et al., 2016) face and landmark detector. Using this approach, we considerably improved the alignment of VIS-NIR image pairs and also ensured a similar face size across the entire dataset. The final preprocessed dataset thus included 2113 VIS-NIR image pairs of 105 individuals, resized to a resolution of 256×256. The data was then split in an approximate $9:1$ ratio into the training and the holdout set. These two sets included images of 95 and 10 identities, respectively, or a total of 1970 training and 143 holdout images.

4.1.2. Evaluation methodology

To evaluate the quality and diversity of images produced by different generative models, we rely on the following standard performance measures that mimic the human perception of images.

- We use the **Fréchet Inception Distance (FID)** (Heusel et al., 2017) to estimate the overall quality of images in comparison to the real-world data. To compare the images, the FID-score calculation procedure first extracts features of images with a pretrained Inception-v3 model (Szegedy et al., 2016) and then estimates the difference between the feature distributions of real and synthetic images. Lower scores imply better correspondence.
- As an alternative, we use the **Learned Perceptual Image Patch Similarity (LPIPS)** (Zhang et al., 2018), which relies on latent image features from a pretrained VGG network (Simonyan and Zisserman, 2014). However, instead of comparing distributions, it estimates the similarity of a given image pair, in our case a real and synthetic image. Thus, to obtain the similarity of two datasets, we measure the LPIPS score between randomly sampled real and synthetic images and then report the mean and standard deviation values. Lower scores again imply better performance.
- We also utilize the **Certainty Ratio Face Image Quality Assessment (CR-FIQA)** (Boutros et al., 2023a) measure that is designed specifically for face images. It measures the quality via the relative classifiability of a given face image with a pretrained ResNet-101 backbone (He et al., 2016). However, because the

Table 1
Overview of face recognition datasets used in our research. The proposed ArcBiFaceGAN is trained and validated with the multispectral Tufts Face Database (Panetta et al., 2018). The other datasets are then used to evaluate the performance of recognition models trained on synthetic data of different generative models.

Dataset	#Images	#IDs	Resolution	Modality	Purpose
Tufts Face Database (Panetta et al., 2018)	> 10,000	113	3280 × 2464	VIS & NIR	–
Tufts Face Database* (Panetta et al., 2018)	2213	105	256 × 256	VIS & NIR	TR & SV
LFW (Huang et al., 2007)	13,233	5749	250 × 250	VIS	REC
CA-LFW (Zheng et al., 2017)	7156	2996	250 × 250	VIS	REC
CP-LFW (Zheng and Deng, 2018)	5984	2296	250 × 250	VIS	REC
AgeDB-30 (Moschoglou et al., 2017)	16,488	568	Var.	VIS	REC
CFP-FP (Sengupta et al., 2016)	7000	500	Var.	VIS	REC

(*) – Preprocessed subset; (TR) – Training; (SV) – Synthesis Validation; (REC) – Recognition experiments; (Var.) – Various resolutions.

model was trained on close-ups of faces, we first use a pretrained Multi-task Cascaded Convolutional Network (MTCNN) face detector (Zhang et al., 2016) to crop the generated images and then evaluate their quality with CR-FIQA (Boutros et al., 2023a).

- To provide a comprehensible visual comparison of real and synthetic distributions, we rely on the *t*-**distributed Stochastic Neighbor Embedding** (*t*-**SNE**) (Van der Maaten and Hinton, 2008) method. Image distributions are first represented with feature vectors obtained from a ResNet-101 model (He et al., 2016) that was pretrained on the ImageNet dataset (Deng et al., 2009). The *t*-SNE method then utilizes the **Kullback–Leibler divergence (KL-divergence)** (Joyce, 2011) to construct a lower-dimensional distribution representative of the initial distribution that can be visualized in a 2D space. To achieve this, the method minimizes the KL-divergence between the low-dimensionality and the original distribution.
- We also analyze the pose of faces in the images in order to determine the intra-identity diversity of samples. For this, we rely on predictions of the state-of-the-art **6DRepNet** (Hempel et al., 2022) head pose estimator that is pretrained on the 300W-LP dataset (Zhu et al., 2016). Specifically, for the evaluation we utilize the yaw predictions, as the diversity of the training Tufts Face Database (Panetta et al., 2018) and the produced synthetic data is highly limited in terms of the pitch and roll of the faces.

Furthermore, we also investigate the suitability of the generated synthetic datasets to train recognition models. To this end, we rely on genuine and imposter score distribution plots, obtained with a pretrained face recognition model (Deng et al., 2019a), and corresponding verification measures. Alongside the mean and standard deviation values of distributions, we measure the **Equal Error Rate (EER)** (Maio et al., 2002), i.e., the operating point on the Receiver Operating Characteristics (ROC) curve, where the False Match Rate (FMR) equals the False Non-Match Rate (FNMR), as well as the **Fisher Discriminant Ratio (FDR)** (Poh and Bengio, 2004), which quantifies the separability of genuine and imposter scores. In addition, we report the lowest FNMR for a FMR that is lower or equal than 1.0% or 0.1%, denoted as **FMR100** and **FMR1000** respectively. To showcase the utility of the produced data in a real-world scenario, we also use it to train a CosFace face recognition model (Wang et al., 2018) that utilizes a ResNet-18 network as the backbone and class-margin criterion as the learning objective (He et al., 2016). The performance of the model is then evaluated in terms of verification accuracy on five popular verification benchmarks, whose main characteristics are listed in Table 1. A more detailed description of the utilized benchmarks is presented below:

- **Labeled Faces in the Wild (LFW)** (Huang et al., 2007) represents an unconstrained verification dataset of 13,233 face images of 5749 identities that was collected from the web.
- **Cross-Age Labeled Faces in the Wild (CA-LFW)** (Zheng et al., 2017) is a subset of the LFW (Huang et al., 2007) dataset, containing 7156 images of 2996 identities, aimed at evaluating recognition performance across a given age gap.
- **Cross-Pose Labeled Faces in the Wild (CP-LFW)** (Zheng and Deng, 2018) represents a subset of the LFW (Huang et al., 2007) dataset that is suited for evaluating cross-pose verification performance. The dataset includes 5984 face images of 2296 identities captured in a variety of poses.
- **AgeDB-30** (Moschoglou et al., 2017) is a dataset of in-the-wild face images, targeted specifically at the evaluation of verification performance across a 30 year age gap. In total, the dataset contains 16,488 images of 568 identities.
- **Celebrities in Frontal-Profile in the Wild (CFP-FP)** (Sengupta et al., 2016) is a verification dataset that is aimed at evaluating cross-pose performance, specifically frontal and profile poses. The dataset entails 500 identities, each with 10 frontal and 4 profile images, i.e. 7000 images in total.

For the purposes of our experiments, all datasets are used to construct 3000 genuine comparison pairs and an equal amount of imposter pairs. For the cross-age and cross-pose datasets the imposter comparison pairs are sampled from the same race and gender, in order to limit the influence of other attributes on the recognition model. In addition, all images are rescaled to a resolution of 112 × 112 to be in-line with the image resolution of the trained recognition model.

It should be noted, however, that these datasets only include image data in the visible (VIS) spectrum. Thus, to evaluate the performance of recognition models that are trained on VIS-NIR image pairs, we utilize a grayscale representation of the available VIS images to mimic the corresponding NIR spectrum. Furthermore, to address the lack of multispectral verification benchmarks, we construct our own benchmark from the holdout set of the preprocessed Tufts Face Database (Panetta et al., 2018). The holdout set includes only 143 VIS-NIR image pairs of 10 identities, due to the already small-scale of the initial dataset. Following the structure of existing verification benchmarks, we use this data to create 1145 genuine comparison pairs and 9008 imposter comparison pairs.

4.1.3. Implementation and experimental details

The ArcBiFaceGAN framework is implemented in PyTorch (Paszke et al., 2019) and is made publicly available.[1] The identity-conditioned Dual-Branch StyleGAN2 model is built upon the SytleGAN2-ADA implementation (Karras et al., 2020a) and outputs images with a resolution of 256 × 256. This resolution was selected based on the resolution requirements of face recognition models used in related works (Boutros et al., 2022, 2023b), however, it can also be adapted to facilitate the generation of higher resolution images. Differently from existing GAN-based approaches (Boutros et al., 2022), our model is conditioned on identity features obtained from a pretrained ArcFace recognition model (Deng et al., 2019a). Specifically, we rely on the publicly available model that is based on the iResnet-101 architecture (He et al., 2016; Duta et al., 2021) and pretrained on the MS1MV3 dataset (Guo et al., 2016b).

[1] https://github.com/dariant/ArcBiFaceGAN

Table 2
Summary of the training procedure for each generative architecture and the recognition experiments. The table includes parameters used for training the different Dual-Branch StyleGAN2 models and the diffusion-based IDiff-Face (Boutros et al., 2023b) model. Parameters used throughout the recognition experiments for training the recognition models are also reported.

DB-StyleGAN2	Parameter value
Batch size	12
Optimizer	Adam (Kingma and Ba, 2015)
Learning rate	25×10^{-4}
β_1, β_2, ϵ	0, 0.99, 10^{-8}
Loss functions	Eq. (2) & Eq. (3)
w_{VIS} & w_{NIR}	1.0 & 0.1 (0.5)
γ_{R_1} & $\gamma_{R_{PL}}$	1.09 & 2.2×10^{-6}
Maximum steps	$2500 kimgs$
ADA target (Karras et al., 2020a)	0.6

IDiff-Face (Boutros et al., 2023b)	Parameter value
Batch size	64
Optimizer	Adam (Kingma and Ba, 2015)
Learning rate	10^{-4}
β_1, β_2, ϵ	0.9, 0.99, 10^{-8}
Loss function	MSE
Maximum steps	20,000
EMA factor	0.75
Horizontal flip chance	0.5
Identity dropout ratio	0.25

Recognition experiments	Parameter value
Batch size	16
Optimizer	SGD
Learning rate	0.1
Momentum	0.9
Weight decay	5×10^{-4}
Loss function	CosFace (Wang et al., 2018)
Scale & Margin	64 & 0.35
Augmentation N & M	4 & 16
Dropout ratio	0.4

ArcBiFaceGAN training parameters. Prior to training, the model is initialized with random weights to avoid breaching the privacy of any subjects used for training any publicly available pretrained version of StyleGAN2 (Karras et al., 2020b). The model is then trained with the multi-task learning objective (Eqs. (2) and (3)) and the Adam optimizer (Kingma and Ba, 2015) in batches of 12 images. The learning rate is set to 25×10^{-4}, $\beta_1 = 0$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$, based on the original implementation (Karras et al., 2020a). During training, we condition the model on normalized identity features of the pretrained ArcFace recognition model (Deng et al., 2019a). To limit the effect of the condition only to the identity aspect and not limit other attributes such as pose, we train the model only with the most representative feature for each real-world identity, i.e. the one that is most similar to features of all other samples of the same identity. Effects of this decision are presented in more detail in Section 4.4.1, where we compare the capabilities of the model trained with representative features for each identity or separate features for each image.

To improve the stability of training on poorly aligned VIS-NIR image pairs, we perform training in two training phases. This entails manipulating the weight that the NIR discriminator loss $\mathcal{L}_{NIR}$ has in the final learning objective $\mathcal{L}_G$. We first begin training with loss weights of $w_{NIR} = 0.1$ and $w_{VIS} = 1.0$, which facilitates the production of high-quality VIS images alongside a low-quality estimation of images in the NIR spectrum. This phase is performed until we achieve convergence and the desired quality of VIS images, in terms of FID (Heusel et al., 2017) scores. The second phase of training is then aimed at improving the quality of NIR images, which we accomplish by increasing the NIR weight to $w_{NIR} = 0.5$. This phase lasts until a desired quality of NIR images is achieved or at maximum up to $2500 kimgs$ (thousand images). Overall, this enables the model to generate more detailed NIR images, while not reducing its synthesis capabilities in the VIS spectrum, and results in the creation of aligned VIS-NIR image pairs. Furthermore, to achieve more stable training on small-scale datasets (in our case 1970 images), we also heavily rely on the Adaptive Discriminator Augmentation (ADA) (Karras et al., 2020a) technique with a target of 0.6 to augment images before they are passed to the discriminators, thus achieving more variety in the training images. This includes augmentations related to rotation, scale, brightness, contrast, hue, saturation and also horizontal flips. The described training process is also summarized in Table 2.

ArcBiFaceGAN data generation procedure. To create data suitable for recognition tasks, we must generate multiple identities each with multiple synthetic images. We begin our data generation process by randomly sampling the latent feature id that determines the identity to be created. To generate multiple images of the same identity, we lock the id feature, and then focus on sampling latent input codes z. For each sample, the locked id and a random z feature are first preprocessed and then passed to the mapping network of our identity-condition DB-StyleGAN2 that generates a pair of visible and near-infrared images $\{\mathbf{x}^{id}_{VIS}, \mathbf{x}^{id}_{NIR}\}$. Preprocessing includes truncating the z feature with a factor of $\psi = 0.7$ to reduces the amount of anomalies (Karras et al., 2020a), and applying a multiplication factor of 4 to the identity feature, to improve the diversity of sampled identities, as demonstrated in Section 4.4.4. The identity feature is then reinterpreted by a fully-connected layer, after which the feature is normalized, thus limiting the effects of the multiplication to the initial layer. The z feature is also normalized separately and then truncated with the identity feature to form the combined input for the rest of model. Once the images $\{\mathbf{x}^{id}_{VIS}, \mathbf{x}^{id}_{NIR}\}$ are generated they are sent to the proposed PD filter for evaluation. The filter removes potentially problematic images, following predetermined similarity thresholds. If this occurs with the first sample of a certain id, then we reset the generation process and sample a new identity features id. However, if this occurs with any later sample, then the id is retained and we simply sample a new random latent code z to continue the generation process. To address potential time issues, we also limit the amount of synthesis attempts for creating samples of a certain identity to 2500. If this limit is reached, the generation process moves on to the next potential synthetic identity.

Privacy and Diversity filter. During the generative process our ArcBiFaceGAN framework utilizes the proposed Privacy and Diversity (PD) filter. To evaluate the similarity of identities in the generated samples, the filter relies on the cosine similarity (Nguyen and Bai, 2010) of identity features obtained with a pretrained recognition model. Specifically, the filter utilizes the ArcFace recognition model (Nguyen and Bai, 2010) with an iResNet-101 (He et al., 2016; Duta et al., 2021) backbone that was pretrained on the MS1MV3 dataset (Deng et al., 2019b). The computed similarity can then be used to perform filtering based on several constraints presented in Section 3.3. The thresholds that are used to distinguish between identities are based on the intra-identity similarity of samples in the training set of the Tufts Face Database (Panetta et al., 2018). In particular, we utilize the mean $\mu = 0.776$ and standard deviation $\sigma = 0.077$ of intra-identity similarity values to determine the threshold for the samples that contain similar identities, i.e. $\tau_{ID} = \mu - 2\sigma = 0.622$, which can be used to remove samples with potentially privacy-breaching identities as well as ensure better identity separability. In addition, we define a threshold for samples that are too similar to existing samples of the same identity, i.e. $\tau_{High} = \mu + \sigma = 0.853$, which can be used to improve intra-identity diversity. These values, are however, selected solely for the purposes of our experiments and can be adapted to suit the desired privacy and diversity needs of a given application.

Recognition experiments. The utility of synthetic data generated by the proposed ArcBiFaceGAN is evaluated in a series of recognition experiments, where we use it to train a small-scale ResNet-18 (He et al., 2016) recognition model. For the purposes of the experiments, we use

the different generative methods to create datasets of 95, 500 and 1000 identities, with 32 images per identity. To prepare the images for the recognition model, we crop and align them with the MTCNN face detector (Zhang et al., 2016), thus resulting in face images with a resolution of 112×112. Training is performed in batches of 16 images with the CosFace loss function (Wang et al., 2018) and the Stochastic Gradient Descent (SGD) optimizer with 0.9 momentum and a weight decay of 5×10^{-4}. The initial learning rate is set to 0.1 and is lowered by a factor of 10 after the 22nd, the 30th, and the 35th epoch. During training the model also relies on random augmentations (Cubuk et al., 2020) with 4 random operations (N) and a magnitude (M) of 16 to add more variety to the training images. A dropout ratio of 0.4 is also utilized to prevent overfitting. Training is then stopped once no improvements in 5 epochs are observed, in terms of verification accuracy, on the LFW (Huang et al., 2007) benchmark. Table 2 contains a summary of the described training process and the used parameters.

In our experiments, we analyze the performance of two variants of the presented recognition model. The first operates on the VIS spectrum, while the second is based on both VIS and NIR data. To this end, the latter model is adapted to receive an input with 4 channels. In the end, the performance of the trained recognition models is evaluated on five state-of-the-art verification benchmarks (Huang et al., 2007; Zheng et al., 2017; Zheng and Deng, 2018; Moschoglou et al., 2017; Sengupta et al., 2016). However, these only contain data in the visible spectrum. Thus to evaluate the performance of the multispectral recognition model, we adapt the existing benchmarks by using the grayscale representation of VIS images as a replacement for the NIR spectrum. Furthermore, we also perform verification on the holdout data of the Tufts Face Database (Panetta et al., 2018) by constructing genuine and imposter comparison pairs among all combinations of samples.

Implementation of state-of-the-art baselines. We rely on the SFace (Boutros et al., 2022) and IDiff-Face (Boutros et al., 2023b) generative models to evaluate the performance of our proposed ArcBiFaceGAN framework. To allow for a fair comparison, we adapt the SFace (Boutros et al., 2022) model to also utilize the same underlying Dual-Branch StyleGAN2 architecture (Tomašević et al., 2022) as our ArcBiFaceGAN. This means that we also rely on the same training procedure and parameters. Thus, the only difference between the two is the identity condition that is used. The SFace (Boutros et al., 2022) model utilizes a one-hot encoded vector as a form of class label to determine the identity that is to be generated. Here, the length of the vector is determined by the amount of identities in the training dataset, i.e. 95 in the case of the Tufts Face Database (Panetta et al., 2018). However, this in turn limits the amount of potential identities that can be achieved with the original SFace (Boutros et al., 2022) model. To address this issue, we also extend the sampling procedure of the existing SFace (Boutros et al., 2022) and propose the so called Mix-SFace approach. To enable the generation of more identities, that are not tied to only the main indices of the one-hot encoded vector, we create new random identities by combining two to five randomly sampled identities of the vector. These vectors are then interpreted by a fully-connected layer at the start of the mapping network (Boutros et al., 2022).

Conversely, we rely solely on the original IDiff-Face (Boutros et al., 2023b) model for the purposes of our experiments. This inherently gives it an edge over the other GAN-based approaches, as it is limited only to the generation of visible spectrum data. The IDiff-Face (Boutros et al., 2023b) model follows the architecture of recently introduced latent diffusion models (Rombach et al., 2022) and is based on the denoising U-Net (Ronneberger et al., 2015) model. This network contains four resolution levels, each with two consecutive residual blocks. To enable identity-conditioned image generation the model introduces identity features of a pretrained face recognition model (Deng et al., 2019a) into the generative process with the cross-attention mechanism. Attention blocks are applied in all residual blocks apart from the first resolution level. This denoising model then operates within the latent space of a pretrained autoencoder network (Rombach et al., 2022). Based on results of the proposed work, we also rely on contextual partial dropout that prevents overfitting on real-world identities by partially dropping out the identity context during training with a probability of 25%, which is especially important due to the small scale of the utilized dataset. In total, the diffusion process entails 1000 time steps and is controlled by the linear diffusion variance schedule. The denoising model is trained to generate 128×128 images with the Mean Squared Error (MSE) loss function and the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. The initial learning rate is set to 10^{-4} and is controlled by an annealing cosine learning rate schedule with warm restarts. Training is performed in batches of 64 images across 20,000 steps, which is adjusted for the smaller scale of the utilized dataset in comparison to the original work (Boutros et al., 2023b). Here, the learning rate is first adjusted after 2500 steps and then in phases that are twice as long as before. Throughout training the weights of the model are also adjusted with the Exponential Moving Average (EMA) technique and a negative exponential factor of 0.75. Data augmentation in the form of horizontal flips is also applied with a probability of 50%. The described training parameters are also summarized in Table 2.

Experimental hardware. All experiments are conducted on a Desktop PC with an AMD Ryzen 7 5800X CPU with 32 GB of RAM and two Nvidia RTX 3060 GPU cards, each with 12 GB of VRAM.

4.2. Face image synthesis capabilities

In the first series of experiments, we evaluate the quality and diversity of the data produced by our proposed ArcBiFaceGAN framework. To this end, we compare the synthesis capabilities of our method to two state-of-the-art face image generative frameworks, i.e. the GAN-based SFace approach (Boutros et al., 2022) and the diffusion-based IDiff-Face approach (Boutros et al., 2023b). To allow for a fair comparison, we use each framework to generate 95 identities, with 32 samples per identity. This corresponds to the number of identities in the training subset of the Tufts Face Database (Panetta et al., 2018). We note again that for our experiments, the SFace approach (Boutros et al., 2022) has been adapted to also generate bimodal data, i.e. it relies on the same underlying Dual-Branch StyleGAN2 architecture that ArcBiFaceGAN uses, whereas the IDiff-Face approach (Boutros et al., 2023b) remains unchanged, and thus only generates images in the VIS spectrum.

4.2.1. Quality and diversity evaluation

We begin our experiments by analyzing the overall quality of images produced by the different generative frameworks. For this purpose, we measure the Fréchet Inception Distance (FID) (Heusel et al., 2017) and the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) between the generated synthetic datasets and both the training (T) and holdout (H) real-world set. We report these values in Table 3, along with results of the Certainty Ratio Face Image Quality Assessment (CR-FIQA) (Boutros et al., 2023a) measure that determines the suitability of images for training face recognition approaches. For context, we also report the FID and LPIPS scores between the training and the holdout set, as well as the CR-FIQA score of the training set. These scores are also separately measured for the visible and the near-infrared spectrum. Furthermore, we provide image samples of different generative frameworks in the left column of Fig. 5. Here, each row contains images that best match the real-world identity depicted in the first row, in terms of cosine similarity of identity embeddings from a pretrained ArcFace recognition model (Deng et al., 2019a).

We first investigate how our initial ArcBiFaceGAN, without filtering, performs in comparison to the original SFace (Boutros et al., 2022) model. As can be observed in Table 3, our approach achieves slightly higher quality with regards to LPIPS and CR-FIQA, but scores notably lower in terms of FID, across both spectra. However, the main difference between the models is actually found in the number of possible

Table 3
Evaluation of image quality in terms of FID (Heusel et al., 2017), **LPIPS** (Zhang et al., 2018) **and CR-FIQA** (Boutros et al., 2023a) **scores.** The reported FID and LPIPS scores are obtained by comparing synthetic samples from datasets of 95 identities with either the training (T) or holdout (H) set of the Tufts Face Database (Panetta et al., 2018). Here FID compares two distributions, while LPIPS compares each image pair. Differently, CR-FIQA only evaluates each synthetic sample in terms of face image quality, i.e., utility for face recognition.

Sp.	Dataset	PD	FID ↓ – T (H)	LPIPS ↓ – T (H)	CR-FIQA ↑
VIS	Tufts (H vs. T) (Panetta et al., 2018)	–	52.324	0.489 ± 0.106	1.743 ± 0.259
	StyleGAN2 (Karras et al., 2020a)	–	19.772 (53.143)	0.477 ± 0.114 (0.474 ± 0.113)	1.711 ± 0.222
	DB-StyleGAN2 (Tomašević et al., 2022)	–	22.701 (56.370)	0.470 ± 0.115 (0.481 ± 0.109)	1.699 ± 0.251
	SFace (Boutros et al., 2022)	–	26.189 (58.739)	0.443 ± 0.095 (0.436 ± 0.100)	1.779 ± 0.193
	Mix-SFace (Boutros et al., 2022)	–	34.793 (62.697)	0.431 ± 0.097 (0.444 ± 0.090)	1.792 ± 0.164
		✓	28.898 (56.706)	0.564 ± 0.071 (0.566 ± 0.070)	1.824 ± 0.149
	ArcBiFaceGAN (Ours)	–	38.873 (62.034)	0.428 ± 0.098 (0.425 ± 0.087)	1.841 ± 0.108
		✓	29.920 (55.666)	0.453 ± 0.113 (0.454 ± 0.104)	1.835 ± 0.140
	IDiff-Face (N) (Boutros et al., 2023b)	–	28.602 (69.832)	0.520 ± 0.101 (0.528 ± 0.096)	1.559 ± 0.332
	IDiff-Face (Boutros et al., 2023b)	–	46.789 (80.327)	0.526 ± 0.123 (0.536 ± 0.111)	1.843 ± 0.175
NIR	Tufts (H vs. T) (Panetta et al., 2018)	–	43.497	0.404 ± 0.090	1.575 ± 0.338
	StyleGAN2 (Karras et al., 2020a)	–	23.599 (72.028)	0.591 ± 0.070 (0.586 ± 0.066)	1.498 ± 0.296
	DB-StyleGAN2 (Tomašević et al., 2022)	–	22.813 (65.676)	0.380 ± 0.101 (0.402 ± 0.096)	1.477 ± 0.310
	SFace (Boutros et al., 2022)	–	28.157 (53.069)	0.366 ± 0.085 (0.375 ± 0.082)	1.553 ± 0.286
	Mix-SFace (Boutros et al., 2022)	–	36.056 (56.701)	0.362 ± 0.088 (0.380 ± 0.078)	1.557 ± 0.239
		✓	28.373 (50.457)	0.364 ± 0.096 (0.376 ± 0.096)	1.642 ± 0.242
	ArcBiFaceGAN (Ours)	–	37.952 (55.528)	0.347 ± 0.086 (0.364 ± 0.082)	1.666 ± 0.206
		✓	27.589 (49.012)	0.368 ± 0.100 (0.381 ± 0.101)	1.680 ± 0.199

(T) – Training set; (H) – Holdout validation set; (↓) – Lower is better; (↑) – Higher is better.
(Sp.) – Spectrum of light; (PD) – Privacy and Diversity filter.

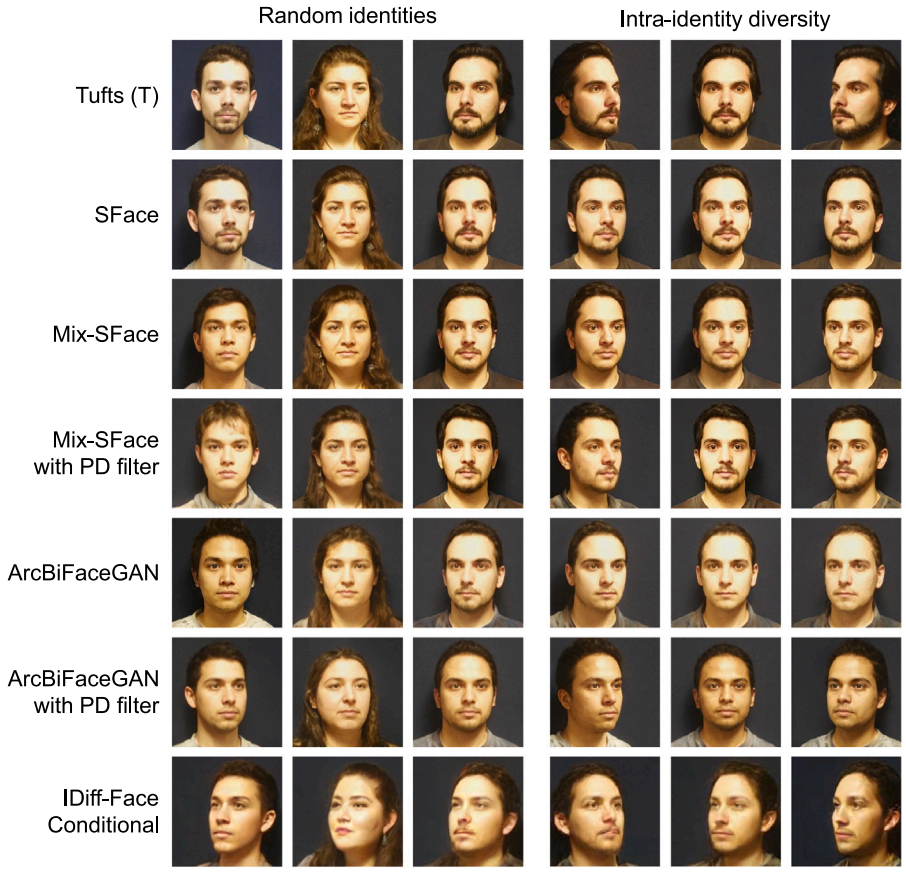


Fig. 5. Comparison of real-world and synthetic visible spectrum samples. The first row contains images from the training dataset. Images in the following rows are produced by a different generative framework. The left column contains samples that best match training identities, while images in the right column showcase intra-identity diversity.

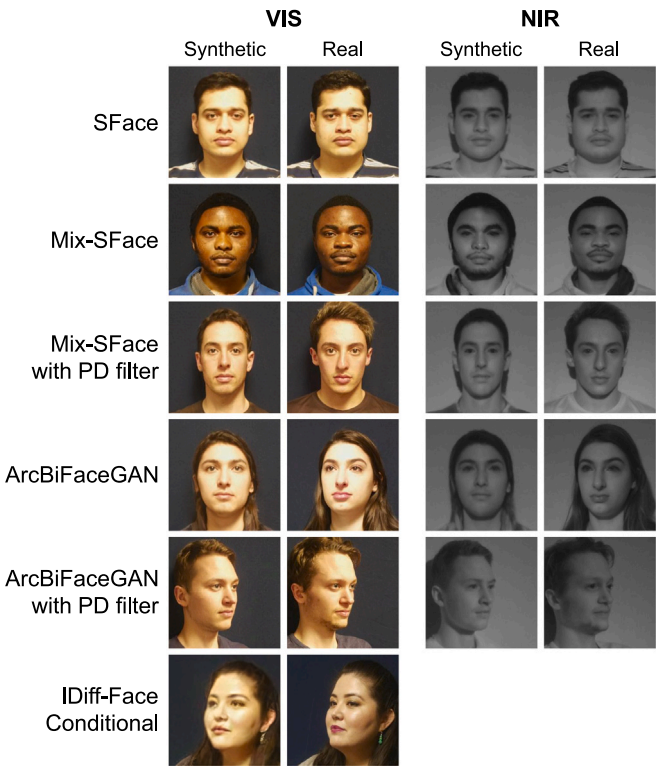


Fig. 6. Privacy-preservation comparison of different generative approaches. Displayed are synthetic samples that have the highest identity similarity with samples from the training set of the Tufts Face Database (Panetta et al., 2018). The most similar samples are determined via cosine similarity of features obtained with a pretrained ArcFace recognition model (Deng et al., 2019a).

unique identities. The SFace (Boutros et al., 2022) approach relies on a one-hot encoded vector to determine the identity. Thus, its original implementation is only capable of sampling as many identities as there are identities in the training set. Even then, however, the produced synthetic identities highly resemble those found in the training set, as can be seen in Fig. 5. Conversely, our approach utilizes an identity feature vector, obtained from a pretrained face recognition model, as the identity condition that is passed to the Dual-Branch StyleGAN2 model. This inherently avoids the identity capacity limit, as we can simply randomly sample new random feature vectors during inference to create new identities.

To improve upon the original SFace (Boutros et al., 2022) method and allow for a more fair comparison with our ArcBiFaceGAN, we propose to mix different one-hot encoded vectors during inference. For the purposes of these experiments, we sum two to five randomly selected one-hot encoded vectors during sampling to create a new identity. This creation of mixed identities bypasses the limited number of possible identities and reduces the overall resemblance to real-world identities. However, similarities still remain, which is problematic from a privacy point of view. In terms of quality indicators, the performance of the so called Mix-SFace approach lies between performance of the original SFace (Boutros et al., 2022) and our ArcBiFaceGAN approach.

To address the privacy-breaching issues that affect both the Mix-SFace approach and our ArcBiFaceGAN (to a lesser extent), we propose the use of an auxiliary Privacy and Diversity (PD) filter during the creation of synthetic datasets, as detailed in Section 3.3. In short, this filter applies a set of restrictions that determine the acceptable similarity between new synthetic identities and both training identities as well as existing samples. The drastic effect of the proposed PD filter can be observed in the identity changes in Fig. 6. Overall, the filtering process ensures that the produced samples do not breach the privacy of subjects from the training set, whilst substantially improving the quality of both Mix-SFace and ArcBiFaceGAN approaches in terms of FID with respect to both training and holdout samples across both spectra. Notable is also the improvement in CR-FIQA scores of the Mix-SFace approach. However, the LPIPS scores are affected slightly negatively, most notably with the Mix-SFace approach. Interestingly, we also observe that all generative approaches achieve higher CR-FIQA scores than the training set and that the produced data is more similar to the training set, in terms of FID, than the training is to the holdout set. In addition, we analyze the effect of the identity condition on the generative process. We observe that the capability for controlling identities can have its cost, as the image quality of discussed GAN-based approaches is lower than that of their non-conditional variants, i.e. the single spectrum StyleGAN2 and the multispectral Dual-Branch StyleGAN2 model, at least in terms of FID scores.

Lastly, we compare our approach with the diffusion-based visible spectrum IDiff-Face (Boutros et al., 2023b) model. The non-conditional variant, i.e. IDiff-Face (N) (Boutros et al., 2023b), achieves better FID scores with regards to the training distribution. However, this is not the case when comparing the produced data with the holdout set, where the results point to a substantially lower quality than all previous methods. The model also performs worse than our proposed ArcBiFaceGAN in terms of LPIPS scores, with regards to both the training and the holdout set, as well as CR-FIQA scores, where we observe not only a lower mean value but also a notable increase in standard deviation. It should be noted, however, that these results belong to the baseline non-conditional version IDiff-Face (N) (Boutros et al., 2023b), which is not capable of generating multiple samples of the same identity.

In comparison, the identity conditioned version IDiff-Face (Boutros et al., 2023b), performs drastically worse in terms of FID and even slightly worse with regards to LPIPS scores. Interestingly, however, the model achieves a substantially higher CR-FIQA score, that competes with the quality of our ArcBiFaceGAN approach. This suggests a more suitable quality of images for training recognition models, despite the lower FID and LPIPS scores. However, when examining the visualized samples in Fig. 5 we observe that the quality of images is indeed rather poor. In comparison with the GAN-based approaches, the images produced by IDiff-Face (Boutros et al., 2023b) contain noticeable unnatural artifacts and are more blurry, even though the model is tasked with the synthesis of only a single spectrum, thus avoiding issues with alignment of VIS-NIR pairs. These issues point to possible issues with training diffusion models on the small scale multispectral Tufts Face Database (Panetta et al., 2018).

The presented results showcase the suitability of our ArcBiFaceGAN framework for creating high-quality face images. Overall, our solution outperforms the original SFace (Boutros et al., 2022) model, in terms of CR-FIQA scores and FID scores with regards to the holdout set, whilst enabling the production of less privacy breaching datasets. Furthermore, it also achieves higher quality images across all measures in comparison to the proposed Mix-SFace approach, likely due to the feature-based identity condition that provides more information to the generative process than a one-hot encoded vector. Finally, our framework is also more suited for training in a low data regime than the IDiff-Face (Boutros et al., 2023b), which is crucial for producing new multispectral synthetic datasets or augmenting existing ones.

4.2.2. T-SNE analysis

To obtain further insight into the differences between synthetic and real-world image distributions of the methods tested in the previous section, we visually compare the distributions with the t-distributed Stochastic Neighbor Embedding (t-SNE) technique (Van der Maaten and Hinton, 2008). In Fig. 7, we depict the t-SNE for each spectrum, featuring image distributions of different generative frameworks alongside the training and holdout sets. For clarity we only plot 250 random samples from each dataset. Furthermore, we provide the Kullback–Leibler divergence (KL-divergence) values (Joyce, 2011) employed in the creation of the t-SNE plots in Table 4.

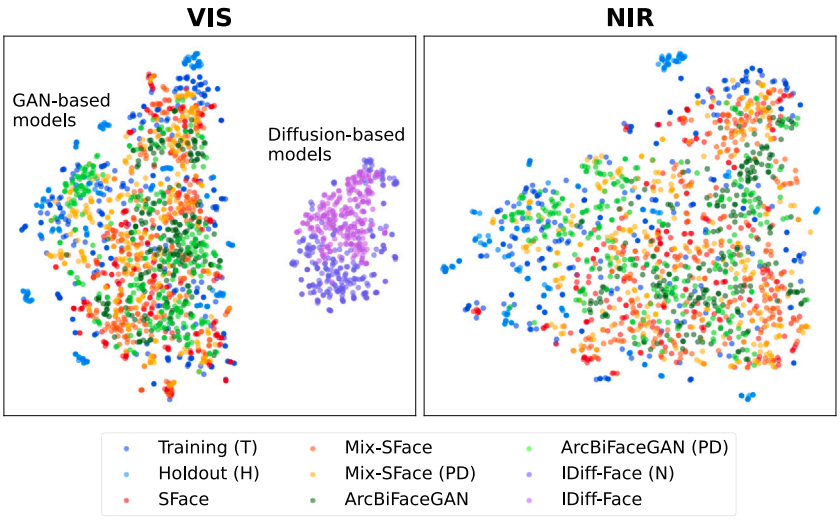


Fig. 7. Synthetic and training distributions presented with 2D t-SNE plots. For clarity the plots are generated with 250 image samples randomly selected from datasets with 95 identities or similarly sized datasets of non-conditional methods. The plot includes samples created by different generative frameworks as well as samples from the training and holdout sets.

Table 4
Comparison of Kullback–Leibler (KL) divergence values of the t-SNE plots in Fig. 7. Reported values are obtained by comparing synthetic images of a given generative framework to the real-world training and holdout samples. Lower values imply more similarity between synthetic and real distributions and thus better performance.

KL-divergence		vs. (T)		vs. (H)	
Dataset	PD	VIS	NIR	VIS	NIR
Tufts Face (H) (Panetta et al., 2018)	–	6.644	6.661	–	–
SFace (Boutros et al., 2022)	–	3.341	3.071	6.009	5.186
Mix-SFace (Boutros et al., 2022)	–	3.056	2.958	4.671	4.798
	✓	**2.902**	2.602	4.539	4.335
ArcBiFaceGAN (Ours)	–	3.831	2.930	5.219	4.873
	✓	3.061	**2.179**	**4.289**	**3.950**
IDiff-Face (N) (Boutros et al., 2023b)	–	9.640	–	8.980	–
IDiff-Face (Boutros et al., 2023b)	–	8.834	–	8.193	–

The plots in Fig. 7 reveal that the distributions of most generative frameworks overlap rather well with the real-world distributions. The main outliers are again the distributions of the two diffusion-based IDiff-Face (Boutros et al., 2023b) models in the visible spectrum. These can easily be separated from the rest, including the training and holdout distributions, as also indicated by the high KL-divergence values in Table 4. The two distributions are also substantially less spread out, in comparison to other distributions, which points to a possible low diversity of samples. Overall, this suggests that these models lack the capability of producing images that would match the quality and diversity of real world samples.

From the plots, we can also discern that the GAN-based approaches that utilize the proposed Privacy and Diversity (PD) filter overlap slightly better with both the training and the holdout distributions, as they cover a larger area. An example of this can be seen on the left side of both plots, where more samples reach the ends of the real-world distributions. This is supported by the lower KL-divergence scores achieved by both Mix-SFace and ArcBiFaceGAN when combined with the PD filter, as seen in Table 4 across both spectra. This suggests that the addition of the PD filter actually increases the diversity of created images and also makes sense, when we consider the alternative. Without it we often sample identities that are most common in the latent space, which results in the creation of similar identities.

It should also be noted that the shapes of the training and holdout distributions match fairly well, despite containing different identities. This implies that the distance between different samples does not necessarily correspond to a difference in identity. Thus, we are not incentivized to search for distributions with the highest divergence from real-world samples. Interestingly, the GAN-based approaches overlap with the training distribution better than the training does with the holdout, similarly to previous FID results. However, the overlap is also better with the holdout distribution, so this does not indicate a case of overfitting. Overall, our proposed ArcBiFaceGAN with the PD filter achieves the lowest KL-divergence values across most scenarios. The PD filter also has a drastically larger positive impact on KL-divergence values when applied to ArcBiFaceGAN rather than to Mix-SFace.

4.2.3. *Evaluating intra-identity diversity*

Next, we investigate the diversity of intra-identity samples produced by the different generative frameworks. This aspect of data generation is critically important, especially when discussing synthetic data for training recognition approaches. To achieve high recognition performance the generated dataset should contain diverse identities, each with a large number of diverse samples. Unfortunately, the multispectral Tufts Face Database (Panetta et al., 2018) contains rather low diversity between samples. Subjects are captured from different angles but this is done in a controlled environment with the same background. Thus, the diversity is mainly limited to the varying face pose, along with slight lighting changes. This drastically limits the amount of diversity that any generative framework can learn.

Due to these characteristics, we focus our analysis on the pose diversity that can be found in the training set, namely the yaw rotation of faces. To this end, we utilize a pretrained state-of-the-art head pose estimator 6DRepNet (Hempel et al., 2022) to obtain the yaw rotation of each sample. We then analyze the yaw rotation distributions across

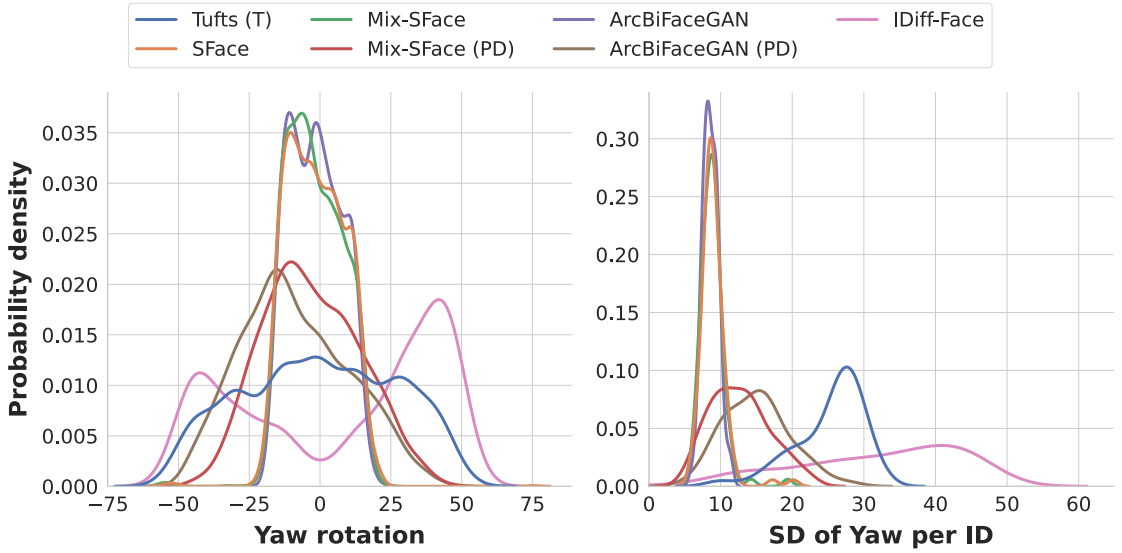


Fig. 8. Comparison of image diversity through yaw rotation of faces. The left plot displays the distribution of yaw rotations across all samples of the datasets. Differently, the right plot contains distributions of Standard Deviation (SD) yaw rotation values obtained for each identity separately, thus highlighting intra-identity diversity of the datasets. The yaw rotation of faces is determined with a pretrained head pose estimator (Hempel et al., 2022).

Table 5
Comparison of image diversity based on yaw rotations of faces. Reported are the mean and standard deviation values of the different distributions from Fig. 8. The left column describes the distribution of yaw rotations across all samples. The right includes results of Standard Deviation (SD) distributions obtained across samples for each identity. The yaw rotation of samples is determined with a pretrained head pose estimator (Hempel et al., 2022).

Dataset	PD	Yaw rotation	SD of Yaw per ID
Tufts Face (T) (Panetta et al., 2018)	–	-0.265 ± 26.591	25.262 ± 4.740
SFace (Boutros et al., 2022)	–	-1.600 ± 9.995	8.983 ± 1.841
Mix-SFace (Boutros et al., 2022)	–	-1.955 ± 9.762	8.852 ± 1.694
	✓	-2.948 ± 17.127	12.704 ± 4.045
ArcBiFaceGAN (Ours)	–	-1.563 ± 9.038	8.588 ± 1.061
	✓	-8.766 ± 19.173	15.036 ± 4.568
IDiff-Face (Boutros et al., 2023b)	–	7.534 ± 34.918	32.025 ± 11.199

(T) – Training set; (PD) – Privacy and Diversity filter.

all samples as well as for each identity separately. Specifically, to determine the intra-identity diversity, we compute the standard deviation of samples for each identity and then form distributions based on these values. We provide plots of the dataset-wide distribution and the intra-identity standard deviation distributions in Fig. 8. We also report the mean and standard deviation of the different distributions in Table 5. In addition, we depict, in the right column of Fig. 5, three samples with different face poses from the right-most identity shown in the left column.

From Fig. 8 and Table 5 we can discern that all GAN-based methods without the PD filter generate data with similarly low levels of both dataset-wide diversity and intra-identity diversity. To address this, our proposed PD filter not only removes privacy-breaching samples, but also removes samples if they are too similar to previously generated samples for a given identity. The effect of the PD filter on the data diversity of both Mix-SFace (Boutros et al., 2022) and ArcBiFaceGAN can be clearly observed by the more spread out distributions with lower peaks. Importantly, the distributions of intra-identity diversity are also moved further to the right. As seen in Fig. 5 the produced samples cover a larger range of face poses. Interestingly, this effect is more pronounced with our ArcBiFaceGAN approach as it achieves a notably higher diversity scores than the Mix-SFace (Boutros et al., 2022), despite similar performance without the PD filter. These observations are also reflected in the intra-diversity plot, as the distribution of our ArcBiFaceGAN framework with filtering best matches the real-world distribution.

In comparison to the GAN-based approaches, the diffusion-based IDiff-Face (Boutros et al., 2023b) model scores the highest in terms of intra-identity diversity, with values even higher than the training dataset, as seen in Table 5. A similar observation can be made for the dataset-wide distribution, but only in terms of standard deviation. Despite the promising scores, the bimodal shape of the dataset-wide distribution in Fig. 8 reveals a notable problem, i.e. a lack of front facing samples. In contrast the real-world or GAN-based datasets form uniform distributions with peaks near the middle. This problem can also be observed in samples of both columns in Fig. 5, where most samples take on a more profile-based view. Overall, the distributions of the IDiff-Face (Boutros et al., 2023b) model simply do not correspond well to the real-world samples.

4.2.4. Exploration of the latent space

In the above experiments we demonstrated the suitability of our proposed framework for identity-conditioned image generation, where the identity of a subject is determined by the *id* input while other aspects of the image, e.g. the pose, are based on the regular z input of StyleGAN-based architectures (Karras et al., 2019, 2020b). In this section, we specifically explore the control offered by the separate inputs and their potential entanglement, as well as the linearity of their latent spaces. To this end, we generate images by linearly interpolating between two randomly sampled points in the identity space the identity while keeping the z input fixed to a randomly selected point and vice versa. In Fig. 9 we depict samples generated by this process, with three

Fig. 9. Interpolation results generated with ArcBiFaceGAN. The patterns are generated by linear interpolation between two randomly selected points in either identity or latent **Z** space, while the other input remains fixed. The leftmost and rightmost images correspond to sampled points, while the images in between represent the results of interpolation.

examples for each interpolated input. From the resulting images, we can discern that interpolating between two identities does not affect the overall pose of the face. Meanwhile interpolating over the z input enables drastic changes to the pose of the face with minimal influence on the identity. All interpolations also display a smooth gradual transition either between two identities or two poses, confirming the linearity of the latent spaces. Overall, these results highlight the disentanglement that can be achieved with our ArcBiFaceGAN framework, especially when training on most representative identity features for each real-world subjects to limit the effect on the pose, as is further explored in Section 4.4.1.

4.3. Recognition experiments

The second set of experiments revolves around the suitability of synthetic data for training face recognition approaches. This includes exploring the identity separability of generated datasets and investigating how synthetic (multispectral) data can be used to improve the performance of modern recognition models. Similarly to before, we compare our proposed ArcBiFaceGAN with the state-of-the-art conditional generative frameworks, i.e. SFace (Boutros et al., 2022) and IDiff-Face (Boutros et al., 2023b), as well as the presented Mix-SFace (Boutros et al., 2022) variant. In addition, we investigate how the introduced Privacy and Diversity (PD) filter affects the final generated datasets. For the purposes of the experiments, we use the different generative frameworks to produce multiple datasets with increasing amounts of identities, i.e. 95, 500 and 1000, with 32 samples per identity in order to analyze their full potential.

4.3.1. Identity separability

We begin our face recognition experiments by evaluating the separability of identities generated by our proposed framework and comparing it to the state-of-the-art. To analyze this capability in a quantitative manner, we utilize a pretrained ArcFace recognition model (Deng et al., 2019a) to extract identity embeddings and use them to construct genuine and imposter distribution plots in Fig. 10, which are based on all possible genuine pairs and an equal amount of randomly sampled imposter pairs. We also report the corresponding verification performance in Table 6, including the Equal Error Rate (EER) (Maio et al., 2002), the False Non-Match Rate (FNMR) at or below a False Match Rate (FMR) of 1.0% and 0.1%, denoted as FMR100 and FMR1000, as well as the mean and standard deviation of each distribution along with the Fisher Discriminant Ratio (FDR) (Poh and Bengio, 2004). For brevity, the plots only contain datasets with 95 identities, as increases in scale did not alter the overall distribution shapes. Numerical results for the different scales are, however, still provided in Table 6.

Comparing real-world and synthetic identity separability. First let us address the unusual lack of overlap between the genuine and imposter distributions in the Tufts Face Database (Panetta et al., 2018). This is typically not observed in large-scale real-world datasets, as can be seen in related works (Boutros et al., 2022, 2023b). Likely, this is caused by the small-scale of the multispectral dataset and its incredibly limited intra-identity diversity, apart from face pose changes. Apart from these factors, perfect identity separability is also not indicative of a good recognition training dataset, due to the ease of separating the identities. To train successful face recognition models we require a delicate balance of identity separability and intra-identity diversity (Boutros et al., 2023b).

In comparison to the Tufts Face Database (Panetta et al., 2018), all produced synthetic datasets showcase some form of overlap and verification errors. The lowest EER and FMR100 scores are obtained by the original SFace (Boutros et al., 2022) approach, along with the highest overall distribution separability in terms of FDR. This is rather expected, considering the nature of the model, as it utilizes a one-hot encoder vector as the identity condition. While this choice does limit the amount of possible identities that can be created, these identities should be separable, given a properly trained model. Interestingly, however, the model performs drastically worse in terms of FMR1000 than other generative models, i.e. results in more false non-matches when a FMR of 1-in-1000 is desired. This drastic increase between

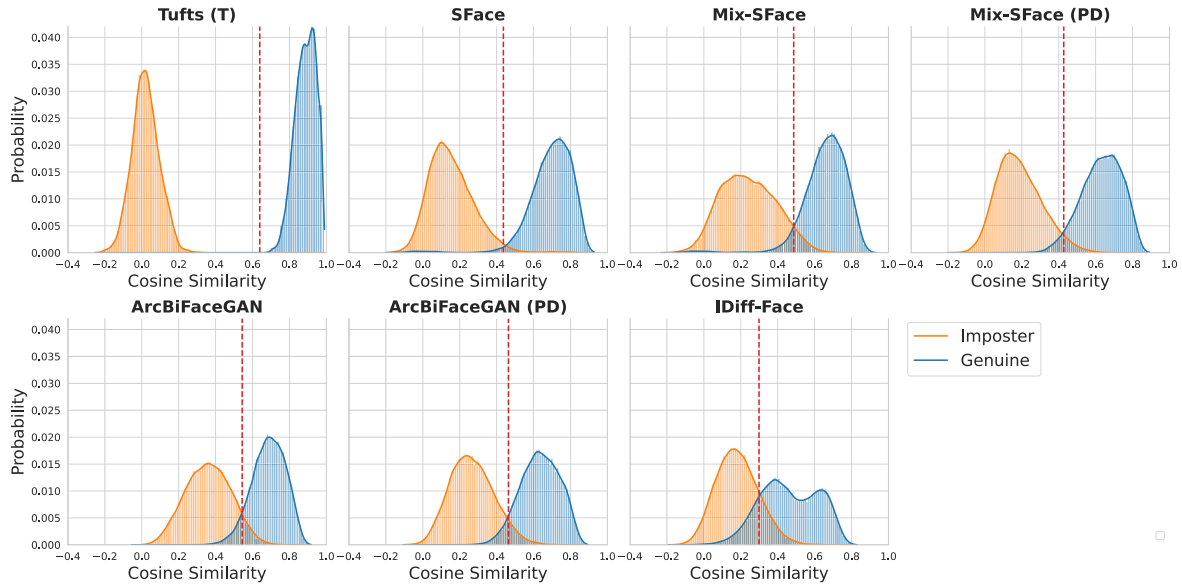


Fig. 10. Identity separability comparison of different datasets with plots of genuine and imposter distributions. Distributions are obtained by computing the identity similarity of a pair of visible spectrum face images, either of the same or different identities (i.e. genuine and imposter pairs). For each dataset, all possible genuine pairs are considered, along with an equal amount of randomly sampled imposter pairs. Identity similarity is determined by the cosine similarity of identity features, obtained for each image with a pretrained ArcFace recognition model (Deng et al., 2019a). To allow a comparison with the real-world dataset, the figure only contains plots of datasets with 95 identities.

Table 6
Quantitative identity separability analysis of different datasets. Reported are verification-based results, i.e. Equal Error Rate (EER) and the false non-match rate at a False Match Rate of 1% (FMR100) or 0.1% (FMR1000), of the distributions of visible spectrum face images in Fig. 10. Included are also the mean and standard deviation of the distributions along with their overall separability in terms of Fisher Discriminant Ratio (FDR). The best results of synthetic datasets are marked in bold.

Dataset	PD	#IDs	EER ↓	FMR100 ↓	FMR1000 ↓	Gen. μ (σ)	Imp. μ (σ)	FDR ↑
Tufts Face (T) (Panetta et al., 2018)	–	95	0.000	0.000	0.000	0.887 (0.056)	0.021 (0.074)	87.069
SFace (Boutros et al., 2022)	–	95	**0.020**	**0.041**	0.797	0.694 (0.122)	0.155 (0.124)	**9.663**
Mix-SFace (Boutros et al., 2022)	–	95	0.056	0.199	0.519	0.659 (0.121)	0.242 (0.148)	4.754
	✓	95	0.040	0.140	0.392	0.633 (0.109)	0.188 (0.123)	7.332
	✓	500	0.031	0.101	0.382	0.643 (0.109)	0.162 (0.124)	8.523
	✓	1000	0.030	0.088	**0.337**	0.647 (0.114)	0.150 (0.125)	8.594
ArcBiFaceGAN (Ours)	–	95	0.066	0.270	0.540	0.687 (0.091)	0.360 (0.122)	4.638
	✓	95	0.058	0.253	0.573	0.635 (0.105)	0.267 (0.119)	5.360
	✓	500	0.043	0.167	0.475	0.648 (0.107)	0.230 (0.122)	6.643
	✓	1000	0.037	0.130	0.418	0.657 (0.107)	0.214 (0.121)	7.465
IDiff-Face (Boutros et al., 2023b)	–	95	0.157	0.523	0.700	0.463 (0.156)	0.181 (0.114)	2.115
	–	500	0.151	0.506	0.677	0.474 (0.152)	0.186 (0.116)	2.267
	–	1000	0.152	0.507	0.681	0.475 (0.152)	0.188 (0.115)	2.257

(PD) – Privacy and Diversity filter, (Gen.) – Genuine distribution; (Imp.) – Imposter distribution. (#IDs) – Number of identities; (↓) – Lower is better; (↑) – Higher is better.

FMR100 and FMR1000 scores points to an imposter distribution with a long right tail, meaning that some generated identities might be extremely similar. Differently from slight similarities, which provide a valid challenge for face recognition training, such high similarity anomalies might not be desirable.

In contrast, both proposed approaches that resolve the unique identity limit of the original SFace (Boutros et al., 2022), i.e. the Mix-SFace variant and our ArcBiFaceGAN, achieve worse overall identity separability across most reported results in Table 6. This is also reflected in larger overlap between the genuine and imposter distributions, caused by the drastic shift of the imposter distributions to the right and their overall flatter shape. Worse results are however not surprising, due to the sampling approach taken by both proposed frameworks. Since we either randomly combine one-hot encoded vectors (Mix-SFace) or randomly sample features from a latent space of a pretrained recognition model (ArcBiFaceGAN), we are likely to generate similar identities at some point. This is especially true when considering the low amount of training data, which limits the possible identity diversity learnt by the generative models. Nevertheless, both approaches actually attain notably lower FMR1000 scores than SFace (Boutros et al., 2022), despite larger FMR100 values. While this does indicate the presence of beneficial similar identities (e.g. identities of cosine similarity near the EER threshold), it also showcases a lower amount of potentially problematic extremely similar identities.

The reported results on identity separability reveal additional issues with training the IDiff-Face (Boutros et al., 2023b) model on the limited amount of training images. The genuine and imposter distributions in Fig. 10 showcase a considerable amount of overlap and poor identity separability. Additionally, the genuine distribution follows an unusual bimodal shape, further indicating a potential lack of intra-identity diversity, i.e. coverage across the different face poses. The results in Table 6 support these observations, as the model scores drastically worse than other generative models across all measures, apart from FNMR1000 where it still outperforms the SFace approach (Boutros et al., 2022).

Influencing identity separability with the PD filter. Lastly, we explore the effect of our proposed Privacy and Diversity (PD) filter on the identity separability of datasets generated by both the Mix-SFace and the ArcBiFaceGAN framework. Fig. 10 and Table 6 reveal that the introduction of the PD filter drastically shifts the imposter distributions to the left and increasing their peaks. This improves scores across the majority of reported measures, most noticeably in terms of EER

and FDR, results in an overall increased separability among generated identities, thus addressing the issues of sampling similar identities. Interestingly, when generating larger datasets with increased amounts of identities (e.g. 500 and 1000), our PD filter actually ensures the increase of identity separability. This can be observed in substantial improvements in terms of all results with the increasing number of identities and leads to scores that are closer to those of the SFace approach (Boutros et al., 2022), whilst achieving a drastically lower FMR1000 value.

4.3.2. Face recognition training on synthetic datasets

The experiments decribed above focused on analyzing the synthesis capabilities of the proposed ArcBiFaceGAN framework to determine the overall quality of the generated data. However, in order to evaluate the utility and applicability of the created synthetic data in a real-world scenario, we need to investigate how well it performs in the task of training recognition models. To this end, we train a modern CosFace recognition model (Wang et al., 2018), based on the ResNet-18 network architecture (He et al., 2016), with the synthetic data produced by different generative frameworks. This includes the state-of-the-art SFace (Boutros et al., 2022) and IDiff-Face (Boutros et al., 2023b) models along with the introduced Mix-SFace variant (Boutros et al., 2022) and our ArcBiFaceGAN. In addition, we explore how the generation of privacy-preserving data with the proposed Privacy and Diversity (PD) filter affects the suitability of the generated datasets. We evaluate the overall performance and genaralizability of the different synthetic-based recognition models with the use of five widespread face recognition benchmarks. This includes the unconstrained LFW (Huang et al., 2007) dataset, the cross-age AgeDB-30 (Moschoglou et al., 2017) and CA-LFW (Zheng et al., 2017) datasets, as well as the cross-pose CFP-FP (Sengupta et al., 2016) and CP-LFW (Zheng and Deng, 2018) datasets. To investigate how the models perform on real-world multispectral data, we also perform evaluation on the small holdout set of the Tufts Face Database (Panetta et al., 2018). The verification accuracies obtained by the different synthetic-based face recognition models are reported in Table 7. Here it should be noted, that differently from existing works (Boutros et al., 2022, 2023b), we validate the recognition model during training only on the LFW (Huang et al., 2007) benchmark, instead of across all benchmarks. In this section, we focus on the analysis of the results tied to solely the visible spectrum, while the use of multispectral data is explored in the following section.

Real-world and synthetic-based recognition. First, let us discuss the recognition performance obtained with different datasets of 95 identities, to allow for a fair comparison with the real-world Tufts Face Database (Panetta et al., 2018). When considering the results with the datasets generated without the proposed PD filter, we observe that the original SFace (Boutros et al., 2022) approach results in the highest verification accuracies on the LFW and CA-LFW (Zheng et al., 2017) benchmarks. However, on the other four benchmarks, we achieve drastically better performance when utilizing data produced by our ArcBiFaceGAN approach. Meanwhile, the results of the Mix-SFace and the IDiff-Face (Boutros et al., 2023b) model present a middle ground between the performance of SFace (Boutros et al., 2022) and our ArcBiFaceGAN. In comparison, training a recognition model on real-world data delivers overall better performance on most benchmarks apart from AgeDB-30 (Moschoglou et al., 2017) and CA-LFW (Zheng et al., 2017), where synthetic-based models prevail. Importantly, all above discussed solutions share the same crucial downside, i.e. they all rely on potentially privacy-breaching images.

Suitability of privacy-preserving data for recognition. The results in Table 7 reveal that utilizing PD filtered datasets instead does not drastically affect the overall accuracy of the trained recognition models. The performance achieved with our Mix-SFace approach does slightly decrease on all benchmarks. However, we can observe that it actually improves with ArcBiFaceGAN on the LFW (Huang et al., 2007) and CA-LFW (Zheng et al., 2017) benchmarks. Most notably, the SFace (Boutros et al., 2022)-based approach achieves higher accuracies in all verification benchmarks, when using the PD filter during the creation of the dataset. This showcases that it is possible to train face recognition models without breaching the privacy of real-world subjects. Nevertheless, it should be noted, that the recognition performance is in certain instances close to random (e.g. on AgeDB-30 (Moschoglou et al., 2017)), with either real-world or synthetic datasets. This poor generalizability to real-world benchmarks is likely caused by the small size of the datasets (only 95 identities) and the extremely limited diversity of samples, caused by the controlled environment in which the real-world images were captured.

Scalability of synthetic recognition datasets. In order to explore possible solutions, we investigate the scalability of our proposed approaches. To this end, we create datasets with a increasing amounts of unique identities (500 and 1000) with the different generative approaches that support this. From Table 7, we discern a substantial improvement in accuracies when increasing the amount of identities with Mix-SFace and ArcBiFaceGAN. Creating 500 identities with Mix-SFace leads to better performance across most benchmarks apart from AgeDB-30 (Moschoglou et al., 2017). In comparison, results based on 500 identities created by ArcBiFaceGAN, showcase substantial improvements in all benchmarks, where the approach also outperforms the Mix-SFace variant. On the LFW (Huang et al., 2007) and CA-LFW (Zheng et al., 2017) benchmarks, this configuration even achieves better results than when utilizing real-world data. We observe similar performance increases with the IDiff-Face (Boutros et al., 2023b) model, which yields better results on AgeDB-30 (Moschoglou et al., 2017) than our ArcBiFaceGAN approach, but lower among other benchmarks. Lastly, utilizing synthetic datasets of 1000 identities again results in improved performance in the majority of experimental settings. With the Mix-SFace approach we observe notably higher accuracies on the AgeDB-30 (Moschoglou et al., 2017) and CFP-FP (Sengupta et al., 2016) benchmarks along with better verification on the holdout set of the Tufts Face Database (Panetta et al., 2018). However, this is accompanied by slightly worse performance on other benchmarks. Differently, increasing the dataset scale of IDiff-Face (Boutros et al., 2023b) heavily negatively impacts the recognition performance on AgeDB-30 (Moschoglou et al., 2017), CFP-FP (Sengupta et al., 2016) and CP-LFW (Zheng and Deng, 2018), whilst not providing noticeable performance boosts on other benchmarks. These scalability issues are likely caused by the low identity separability investigated in Section 4.3.1. In comparison, our proposed ArcBiFaceGAN showcases the best scalability, as we observe improved performance on all benchmarks, apart from CP-LFW (Zheng and Deng, 2018), when expanding the synthetic dataset to 1000 identities. Most notable here is the substantial accuracy increase on the previously problematic AgeDB-30 (Moschoglou et al., 2017) benchmark.

Suitability of ArcBiFaceGAN. Among all generative approaches, the proposed ArcBiFaceGAN model leads to overall highest scores across the different verification benchmarks in terms of all dataset scales, with only few exceptions that are based on minimal accuracy differences (e.g. on CFP-FP (Sengupta et al., 2016)). Most importantly, by utilizing the synthetic dataset of 1000 identities generated by our proposed ArcBiFaceGAN and the presented PD filter, we achieve drastically better verification performance than with the real-world dataset. This can be observed in all benchmarks apart from CP-LFW (Zheng and Deng, 2018), but even there the results of our method remain competitive. Overall, the presented results showcase the value of synthetic data for training biometric recognition models, especially when faced with small-scale real-world datasets with limited diversity. Furthermore, the performance achieved with our approach demonstrates that high identity separability of training data does not always translate to better recognition models. Compared to ArcBiFaceGAN, both SFace (Boutros et al., 2022) and Mix-SFace attained better identity separability but substantially worse intra-identity diversity. Thus, when creating state-of-the-art synthetic datasets, we should always consider the trade-off between identity separability and intra-identity diversity, and seek to balance both aspects.

Table 7
Verification performance of recognition models trained with synthetic data of different generative frameworks. Reported is the accuracy across 5 state-of-the-art verification benchmarks and all image pairs of the holdout set of the Tufts Face Database (Panetta et al., 2018). During training, the models are validated (Val.) on the LFW (Huang et al., 2007) benchmark to prevent overfitting. Overall, the table is split into two sections according to the type of data used for training, i.e. visible spectrum (VIS) or multispectral, as indicated in the first column. Since verification benchmarks do not include near-infrared (NIR) data, we instead use a grayscale version of VIS images in its place, when evaluating multispectral recognition models. The best performance on each benchmark for each spectral section is marked in bold. The best scores that are based on 95 identities are also underlined.

Training setting				Val. ↑	Verification accuracy on benchmarks ↑				
Sp.	Dataset	PD	#IDs	LFW	AgeDB-30	CA-LFW	CFP-FP	CP-LFW	Tufts (H)
VIS	Tufts Face (T) (Panetta et al., 2018)	–	95	0.674	0.511	0.532	0.595	0.548	0.973
	SFace (Boutros et al., 2022)	–	95	0.655	0.493	0.537	0.580	0.526	0.945
		✓	95	0.644	0.506	0.541	0.599	0.537	0.949
	Mix-SFace (Boutros et al., 2022)	–	95	0.647	0.522	0.535	0.581	0.537	0.940
		✓	95	0.634	0.511	0.526	0.559	0.519	0.939
		✓	500	0.676	0.504	0.552	0.579	0.543	0.966
		✓	1000	0.667	0.537	0.550	**0.600**	0.541	**0.981**
	ArcBiFaceGAN (Ours)	–	95	0.640	0.529	0.523	0.593	0.540	0.946
		✓	95	0.645	0.506	0.541	0.571	0.531	0.940
		✓	500	0.676	0.517	0.552	0.593	**0.547**	0.976
		✓	1000	**0.692**	**0.549**	**0.562**	0.598	0.545	**0.981**
	IDiff-Face (Boutros et al., 2023b)	–	95	0.648	0.504	0.531	0.567	0.534	0.948
		–	500	0.668	0.522	0.547	0.593	0.540	0.954
		–	1000	0.670	0.497	0.548	0.574	0.535	0.960
VIS & NIR	Tufts Face (T) (Panetta et al., 2018)	–	95	0.699	0.531	0.561	0.616	0.548	0.978
	SFace (Boutros et al., 2022)	–	95	0.656	0.538	0.542	0.587	0.522	0.948
		✓	95	0.666	0.513	0.535	0.582	0.531	0.950
	Mix-SFace (Boutros et al., 2022)	–	95	0.665	0.506	0.534	0.591	0.530	0.952
		✓	95	0.627	0.508	0.534	0.564	0.525	0.945
		✓	500	0.703	0.521	0.578	0.579	0.547	0.960
		✓	1000	0.721	**0.561**	**0.580**	0.600	0.540	0.958
	ArcBiFaceGAN (Ours)	–	95	0.652	0.511	0.535	0.584	0.530	0.948
		✓	95	0.670	0.496	0.557	0.587	0.539	0.946
		✓	500	0.715	0.496	0.579	0.608	0.545	0.979
		✓	1000	**0.727**	0.554	**0.580**	**0.619**	**0.560**	**0.979**

(Sp.) – Spectrum of light; (↑) – Higher is better; (T/H) – Training/Holdout set; (PD) – Privacy and Diversity filter.

4.3.3. Recognition with multispectral data

So far, we demonstrated the possibility of using synthetic visible spectrum data to train successful recognition models. In this section, we extend our recognition-based analysis to the bimodal capabilities of the proposed ArcBiFaceGAN model to investigate the value and utility of generated datasets that include both visible (VIS) and near-infrared (NIR) images. To this end, we repeat our previous recognition experiments but limit our comparison to only GAN-based approaches that we have adapted to rely on the same Dual-Branch StyleGAN2 architecture (Tomašević et al., 2022). Additionally, to allow recognition based on VIS-NIR image pairs, we alter the ResNet-18 architecture (He et al., 2016) of the CosFace recognition model (Wang et al., 2018), to accept a four-channel image as input, with the additional channel dedicated to the NIR data. To compare the performance obtained with datasets from different generative frameworks, we must also modify the existing recognition benchmarks, due to the lack of benchmarks with aligned VIS-NIR image pairs. We therefore propose to create new VIS-NIR benchmarks by simply using the grayscale version of existing VIS images to mimic the required NIR spectrum. This in turn allows for a fair comparison with previous experiments, as we can run the new four-channel recognition models on the same benchmarks as before. To allow for easier comparison, we report the results of these experiments in the second half of Table 7.

Real-world and synthetic-based multispectral recognition. We observe that training the recognition model on real-world VIS-NIR image pairs of the Tufts Face Database (Panetta et al., 2018) results in drastically increased performance than when relying on VIS data. This indicates the potential of the proposed rudimentary solution that mimics NIR data for further improving the real-world accuracy of modern recognition models without requiring the setup of additional NIR sensors. Performance improvements when utilizing VIS-NIR data are also seen with the non-filtered synthetic dataset produced by SFace (Boutros et al., 2022), as it leads to higher accuracies on all benchmarks, apart from CP-LFW (Zheng and Deng, 2018). In comparison, the performance increases are not as pronounced with datasets created by Mix-SFace and our ArcBiFaceGAN. Both achieve noticeably better accuracy on the validation LFW (Huang et al., 2007) benchmark, and CPF-FP (Sengupta et al., 2016) and CA-LFW (Zheng et al., 2017) respectively, but the performance on other benchmarks is similar or lower.

Privacy-preserving multispectral recognition. With the addition of the proposed PD filter to the sampling process, we observe an overall slight decrease in benchmark scores, when generating datasets of 95 identities. The only major improvements are seen with SFace (Boutros et al., 2022) on CP-LFW (Zheng and Deng, 2018) and with ArcBiFaceGAN on LFW-based benchmarks (Huang et al., 2007; Zheng et al., 2017; Zheng and Deng, 2018). When compared to only relying on the VIS spectrum, the recognition performance improvements are quite noticeable. This is especially true with our ArcBiFaceGAN on the LFW-based benchmarks (Huang et al., 2007; Zheng et al., 2017; Zheng and Deng, 2018) as well as the CFP-FP (Sengupta et al., 2016) and the holdout Tufts Face (Panetta et al., 2018) benchmarks. Similar accuracy increases are also observed with the proposed Mix-SFace approach, on CA-LFW (Zheng et al., 2017) and on Tufts Face holdout (Panetta et al., 2018), as well as both cross-pose benchmarks (Zheng and Deng, 2018; Sengupta et al., 2016). With SFace (Boutros et al., 2022), however, the improvements are limited to the LFW (Huang et al., 2007) and AgeDB-30 (Moschoglou et al., 2017) benchmarks. Meanwhile, the accuracy on other benchmarks decreases, likely due to the lower CR-FIQA quality of generated NIR images, observed in Section 4.2.1. However, despite the effect on performance, the proposed PD filter facilitates the generation of privacy-preserving dataset.

Scalability of synthetic VIS-NIR datasets. When increasing the scale of the synthetic datasets, we again notice drastic performance boosts, consistent with observations from the single spectrum experiments. The Mix-SFace dataset of 500 identities displays improvements

on all benchmarks, with the largest differences being on the LFW-related benchmarks (Huang et al., 2007; Zheng et al., 2017; Zheng and Deng, 2018). With 1000 identities these improvements continue and are most notable on the AgeDB-30 (Moschoglou et al., 2017) benchmark. However, performance does slightly decrease on the CP-LFW (Zheng and Deng, 2018) and holdout Tufts Face (Panetta et al., 2018) benchmarks. Compared to results on the VIS spectrum, the accuracy does substantially increase on LFW (Huang et al., 2007), AgeDB-30 (Moschoglou et al., 2017) and CA-LFW (Zheng et al., 2017) but remains similar on the CFP-FP (Sengupta et al., 2016) and CP-LFW (Zheng and Deng, 2018) benchmarks. However, worse performance is achieved on the holdout set of the Tufts Face Database (Panetta et al., 2018), possibly due to the misalignment found between the VIS-NIR image pairs. In comparison, our proposed ArcBiFaceGAN approach displays even better scalability than our Mix-SFace variant. This can be discerned from higher accuracies across all benchmarks, apart from AgeDB-30 (Moschoglou et al., 2017), when increasing the dataset size to 500 identities. Performance again substantially increases with 1000 identities, with the most drastic improvement observed specifically on the AgeDB-30 (Moschoglou et al., 2017) benchmark. When compared to accuracies obtained on the VIS spectrum, we notice drastic improvements in the large majority of experiments. The only performance decrease is seen on AgeDB-30 (Moschoglou et al., 2017) when using a dataset of 500 identities. Minimally lower performance is also observed on the holdout set of the Tufts Face Database (Panetta et al., 2018), again likely due to VIS-NIR image pair misalignment.

Suitability of ArcBiFaceGAN. Overall, the best performing recognition model is based on the 1000 identity dataset of the filtered ArcBiFaceGAN approach. It performs better than all other synthetic-based models on all benchmarks, apart from AgeDB-30 (Moschoglou et al., 2017), and also outperforms the model that was trained on real-world data of the Tufts Face Database (Panetta et al., 2018) in all verification aspects. Likely, the performance with our proposed ArcBiFaceGAN can be explained by the better balance in terms of identity separability and intra-identity of generated samples, observed in Sections 4.2.3 and 4.2.1. The obtained results reveal the incredible potential of utilizing multispectral data for training recognition models instead of only data in the visible spectrum. In addition, they support the notion of possibly training modern recognition models on pairs of VIS-NIR images, in order to achieve better recognition performance on visible spectrum data, thus bypassing the need for additional sensors in real-world scenarios. Furthermore, these results showcase that our proposed ArcBiFaceGAN framework is more suitable for creating large-scale privacy-preserving synthetic datasets that lead to better performing recognition models on unseen real-world data, than existing state-of-the-art generative approaches.

4.3.4. Data augmentation with synthetic samples

The main goal of this work is to explore the possibility of replacing real-world datasets with synthetic ones. However, to further analyze the utility of the generated datasets, we also briefly investigate how they can be used to augment existing small-scale multispectral datasets. Thus, we repeat the recognition experiments with combined data of the training set of the Tufts Face Database (Panetta et al., 2018) and datasets of different generative approaches. To allow for a fair comparison and retain brevity, we limit the experiments to datasets with 95 identities.

Verification scores reported in Table 8 showcase that augmenting the training datasets with synthetic data of all generative methods results in performance increases on most verification benchmarks. In addition, we observe that utilizing data of either Mix-SFace (Boutros et al., 2022) or ArcBiFaceGAN results in better verification performance across all benchmarks than when using data of the original SFace (Boutros et al., 2022) method. This is likely due to the higher similarity between synthetic identities of SFace (Boutros et al., 2022) and real-world identities.

Interestingly, without the PD filter the performance of our ArcBiFaceGAN-based augmentation is actually worse than that of the non-filtered Mix-SFace approach (Boutros et al., 2022). However, the addition of the proposed PD filter substantially improves the performance of the ArcBiFaceGAN-based augmentation, but it often does not positively impact the Mix-SFace (Boutros et al., 2022) method. Thus in total, our ArcBiFaceGAN with the PD filter achieves the highest performance on the majority of benchmarks. Overall, the drastic performance gains obtained by augmenting the Tufts Face Database (Panetta et al., 2018) with only a small set of synthetic identities showcase the incredible potential and applicability of synthetic data in a variety of scenarios.

4.4. Ablation studies

The following section is dedicated to the evaluation of the implementation details and training configurations of the proposed ArcBiFaceGAN framework.

4.4.1. Choice of training identity conditions

In the discussed experiments, we trained our proposed ArcBiFaceGAN approach by conditioning it on representative identity features for each real-world subject rather than utilizing separate identity features for each sample. With this we limit the impact of the identity condition on other aspects of the generated images, e.g. the pose, as described in Section 4.1.3. This decision was made based on the analysis of samples generated with both training configurations. As before we generated datasets with 95 identities and 32 samples per identity. During the generation process, we also skipped the Privacy and Diversity (PD) filter, to ensure that the observations were solely focused on the capabilities of the proposed identity-conditioned DB-StyleGAN2 model. To investigate possible undesired effects of the identity condition, we evaluate the diversity of generated samples in terms of the yaw rotation, as defined in Section 4.2.3. The obtained dataset-wide distribution, i.e. distribution of yaw rotation across all samples, and the intra-identity distribution, i.e. distribution of standard deviation values of yaw rotation for each identity, are displayed in Fig. 11.

When observing the dataset-wide distributions (left plot), we note that the configuration that utilizes features of each sample results in a distributions with more pronounced valleys than when training with only representative identity features. This shows that certain yaw rotations are underrepresented in the generated samples. However, an even larger difference between the two configurations can be discerned from the intra-identity distributions (right plot). Relying on different identity features for each sample results in a distribution of standard deviation values that is heavily skewed to the left and thus exhibits a lack of yaw rotation diversity across samples of most identities. These results highlight the adverse effect that utilizing separate identity features for each sample during training can have on the capabilities of the trained model, specifically the production of diverse samples in terms of pose. The disentanglement of these features can, however, be addressed by relying on a single representative identity feature for each training subject, as can be seen by the drastic improvement in the intra-identity distribution.

Scores reported in Table 9 also showcase that the chosen training configuration does not negatively affect the quality of produced samples across both light spectra. We also utilize the generated samples to form genuine and imposter distributions, as described in Section 4.3.1. Results reported in Table 10 also display a notable improvement in identity separability across all verification measures, when training the model with representative identity features.

Table 8
Verification performance of recognition models trained with combinations of real and synthetic datasets with 95 identities. Reported is the accuracy across 5 state-of-the-art verification benchmarks and all image pairs of the holdout set of the Tufts Face Database (Panetta et al., 2018). During training, the models are validated (Val.) on the LFW (Huang et al., 2007) benchmark to prevent overfitting. Overall, the table is split into two sections according to the type of data used for training. The best performance on each benchmark for each spectral section is marked in bold.

Training setting			Val. ↑	Verification benchmarks ↑				
Sp.	Dataset	PD	LFW	AgeDB-30	CA-LFW	CFP-FP	CP-LFW	Tufts (H)
VIS	Tufts Face (T) (Panetta et al., 2018)	–	0.674	0.511	0.532	0.595	0.548	0.973
	(T) + SFace (Boutros et al., 2022)	–	0.660	0.523	0.545	0.592	0.548	0.965
	(T) + Mix-SFace (Boutros et al., 2022)	–	0.682	**0.529**	0.549	0.607	0.551	0.964
		✓	0.688	0.520	**0.559**	0.575	**0.549**	0.970
	(T) + ArcBiFaceGAN (Ours)	–	0.645	0.511	0.526	0.563	0.533	0.948
		✓	**0.689**	0.522	**0.559**	**0.594**	0.532	**0.980**
	(T) + IDiff-Face (Boutros et al., 2023b)	–	0.688	0.523	0.550	0.588	0.540	0.969
VIS & NIR	Tufts Face (T) (Panetta et al., 2018)	–	0.699	0.531	0.561	0.616	0.548	0.978
	(T) + SFace (Boutros et al., 2022)	–	0.716	0.568	0.575	0.598	0.544	0.946
	(T) + Mix-SFace (Boutros et al., 2022)	–	0.713	0.575	**0.584**	0.599	**0.556**	0.935
		✓	0.700	0.572	0.568	0.584	0.543	0.946
	(T) + ArcBiFaceGAN (Ours)	–	0.706	0.593	0.568	0.598	0.544	**0.950**
		✓	**0.723**	**0.599**	0.576	**0.601**	0.535	0.935

(Sp.) – Spectrum of light; (↑) – Higher is better; (T/H) – Training/Holdout set; (PD) – Privacy and Diversity filter.

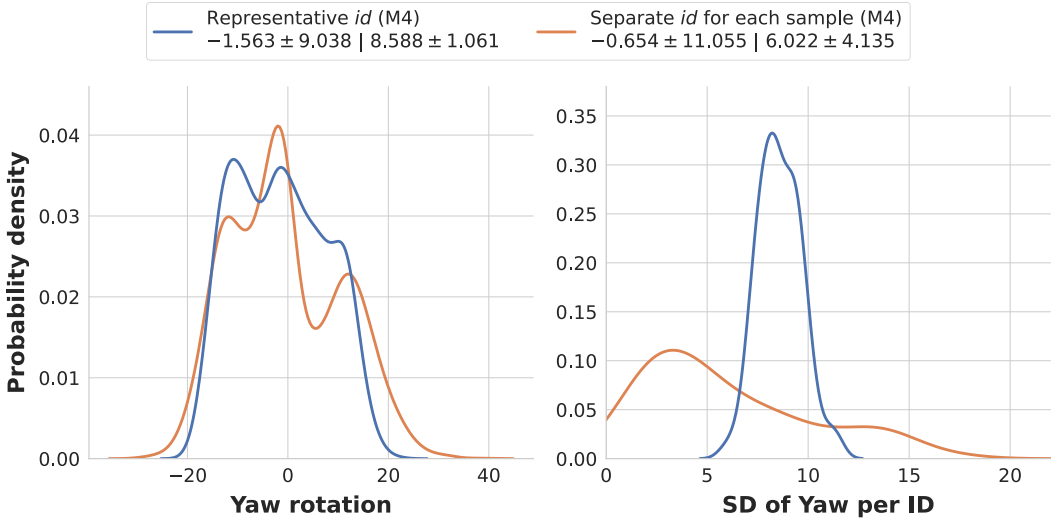


Fig. 11. Image diversity through yaw rotation of faces with different identity conditions. The left plot displays the distribution of yaw rotations, obtained with a pretrained head pose estimator (Hempel et al., 2022), across all samples of the datasets. Differently, the right plot contains distributions of Standard Deviation (SD) yaw rotation values obtained for each identity separately. The mean and standard deviation for each distribution are also reported in the legend.

4.4.2. Amount of training data

As part of our ablation studies, we also investigate how the amount of training and holdout data influences the synthesis capabilities of our proposed ArcBiFaceGAN framework. To this end, we split the preprocessed Tufts Face Database (Panetta et al., 2018) into two similarly sized parts, with the training set containing 1118 VIS-NIR image pairs of 53 identities, while the remaining 995 image pairs of 52 identities are part of the holdout set. We then use the former set to train our proposed ArcBiFaceGAN model, following the training procedure described in Section 4.1.3. The trained model is then used to generate two synthetic datasets, one with 53 identities, corresponding to the smaller scale of the training dataset, and another with 95 identities, to allow for a more fair comparison with previous methods. As before, we generate 32 samples for each identity without the use of the proposed PD filter, to highlight the synthesis capabilities of the model. These configurations are denoted as *Tufts Face Half* in Tables 9 and 10. Here it should also be noted that the reported FID (Heusel et al., 2017) and LPIPS (Zhang et al., 2018) measures for these configurations also utilize the smaller training and larger holdout set for comparison.

With both configurations we can discern a substantial drop in quality across all measures reported in Table 3, when compared to the original configuration (in bold) that is trained on the larger training set. This drop is also more notable for the NIR spectrum. We also observe that training on the smaller dataset results in worse identity separability, as seen in Table 10. The overlap between identities is especially notable in terms of the EER and FDR measures. Overall, these results highlight the difficulty of training the multispectral ArcBiFaceGAN framework in such a low data regime. In this setting worse synthesis capabilities are expected, especially when considering the already small size of the initial multispectral Tufts Face Database (Panetta et al., 2018). However, despite the lower quality and identity separability, the proposed framework is still able to produce synthetic face image samples in both the VIS and the NIR spectrum.

4.4.3. Training regularization methods

During training, our ArcBiFaceGAN framework relies on two existing regularization methods, R_1 regularization (Mescheder et al., 2018) and path length (R_{PL}) regularization (Karras et al., 2020b). Their weights are determined by heuristic formulas based on image resolution and batch size, following their initial implementations (Karras et al., 2020b,a; Mescheder et al., 2018), as described in Section 3.2.3.

Table 9
Evaluation of image quality obtained with different ArcBiFaceGAN configurations, in terms of FID (Heusel et al., 2017), **LPIPS** (Zhang et al., 2018) **and CR-FIQA** (Boutros et al., 2023a) **scores.** The reported FID and LPIPS scores are obtained by comparing synthetic samples with either the training (T) or the holdout (H) set of the Tufts Face Database (Panetta et al., 2018). Differently, CR-FIQA evaluates each synthetic samples in terms of face image quality, i.e. its utility for face recognition, without the need for a real-world reference. The baseline configuration that is used throughout the paper is written in bold.

Sp.	Configuration of ArcBiFaceGAN	FID ↓ – T (H)	LPIPS ↓ – T (H)	CR-FIQA ↑
VIS	Separate *id* for each sample (M4)	49.930 (71.436)	0.437 ± 0.102 (0.440 ± 0.091)	1.886 ± 0.113
	Representative *id* (M4)	**38.873 (62.034)**	$\mathbf{0.428 \pm 0.098}$ ($\mathbf{0.425 \pm 0.087}$)	$\mathbf{1.841 \pm 0.108}$
	Representative *id* (M1)	64.983 (78.281)	0.436 ± 0.097 (0.440 ± 0.090)	1.798 ± 0.099
	Representative *id* (M8)	40.604 (64.843)	0.433 ± 0.097 (0.433 ± 0.099)	1.850 ± 0.117
	No regularization (M4)	288.106 (326.667)	0.701 ± 0.041 (0.700 ± 0.033)	0.646 ± 0.156
	Tufts Face Half 53 (M4)	56.171 (63.510)	0.573 ± 0.077 (0.569 ± 0.078)	1.649 ± 0.422
	Tufts Face Half 95 (M4)	54.095 (61.423)	0.568 ± 0.077 (0.569 ± 0.076)	1.679 ± 0.430
NIR	Separate *id* for each sample (M4)	45.172 (61.397)	0.350 ± 0.093 (0.366 ± 0.088)	1.676 ± 0.168
	Representative *id* (M4)	**37.952 (55.528)**	$\mathbf{0.347 \pm 0.086}$ ($\mathbf{0.364 \pm 0.082}$)	$\mathbf{1.666 \pm 0.206}$
	Representative *id* (M1)	46.921 (59.769)	0.346 ± 0.090 (0.365 ± 0.083)	1.497 ± 0.172
	Representative *id* (M8)	39.160 (56.920)	0.349 ± 0.088 (0.366 ± 0.080)	1.689 ± 0.194
	No regularization (M4)	215.014 (241.046)	0.576 ± 0.078 (0.579 ± 0.073)	0.433 ± 0.204
	Tufts Face Half 53 (M4)	75.828 (85.471)	0.376 ± 0.099 (0.381 ± 0.095)	1.435 ± 0.415
	Tufts Face Half 95 (M4)	72.002 (81.439)	0.372 ± 0.097 (0.382 ± 0.097)	1.472 ± 0.417

(Sp.) – Spectrum of light; (T/H) – Training/Holdout set; (M#) – Identity multiplication factor; (↓/↑) – Lower/Higher is better.

Table 10
Identity separability analysis of samples produced by different ArcBiFaceGAN configurations. Reported are verification-based results of the genuine and imposter distributions of visible spectrum images generated without the PD filter. This includes the Equal Error Rate (EER) (Maio et al., 2002), the false non-match rate at a False Match Rate of 1% (FMR100) or 0.1% (FMR1000) as well as the mean and standard deviation of the distributions along with their overall separability in terms of Fisher Discriminant Ratio (FDR) (Poh and Bengio, 2004). The baseline configuration used throughout the experiments is marked in bold.

Configuration of ArcBiFaceGAN	EER ↓	FMR100 ↓	FMR1000 ↓	Gen. μ (σ)	Imp. μ (σ)	FDR ↑
Separate *id* for each sample (M4)	0.101	0.333	0.591	0.751 (0.091)	0.473 (0.118)	3.477
Representative *id* (M4)	**0.066**	**0.270**	**0.540**	**0.687 (0.091)**	**0.360 (0.122)**	**4.638**
Representative *id* (M1)	0.286	0.889	0.982	0.718 (0.081)	0.607 (0.108)	0.677
Representative *id* (M8)	0.052	0.226	0.576	0.690 (0.096)	0.317 (0.130)	5.346
Tufts Face Half 53 (M4)	0.156	0.412	0.682	0.630 (0.264)	0.291 (0.209)	1.014
Tufts Face Half 95 (M4)	0.156	0.439	0.730	0.637 (0.259)	0.299 (0.211)	1.018

(M#) – Identity multiplication factor; (↓/↑) – Lower/Higher is better; (Gen./Imp.) – Genuine and imposter distribution.

To assess the effect of these regularization methods, we also trained the underlying identity-conditioned DB-StyleGAN2 model with and without them. The configuration without regularization exhibited extremely unstable training that resulted in mode collapse, a common problem of GAN-based architectures (Brock et al., 2018). As a result, the images generated with the best performing model showcase incredibly low quality and diversity, and often do not even contain discernible faces. Quantitative results, denoted as *No regularization* in Table 9, confirm these observation with a drastic drop across all reported measures in both the visible and the near-infrared spectrum. This showcases the importance of utilizing the R_1 and R_{PL} regularization methods (Karras et al., 2020b,a; Mescheder et al., 2018) to facilitate stable training, especially on small-scale datasets such as the multispectral Tufts Face Database (Panetta et al., 2018).

4.4.4. Effect of the multiplication parameter during inference

As part of our last ablation study, we investigate the effect of the proposed multiplication factor that is applied to the sampled identity code *id* during data generation to improve the diversity of sampled identities. For the purposes of the experiments, the choice of this factor was determined by analyzing the identity separability of data generated with either a factor of 1, i.e. no multiplication, 4 or 8. Each configuration was used to generate 95 identities with 32 samples without the Privacy and Diversity (PD) filter. Genuine and imposter distributions were then constructed based on the identity features extracted from these samples with the pretrained ArcFace recognition model (Deng et al., 2019a), following the procedure described in Section 4.3.1.

Table 10 includes the corresponding verification results, where the different multiplication parameters are denoted as $M1$, $M4$ and $M8$. The reported scores reveal a large overlap between the genuine and imposter distributions when utilizing the $M1$ configuration, which signifies a low identity separability between samples of different identities and an overall lack of identity variety. This is particularly problematic when also utilizing the proposed PD filter, as the generation of synthetic identities that the filter considers new is almost impossible, which stalls the data generation process. However, as observed in the Table 10, this issue can be addressed with a higher identity multiplication factor during sampling. The scores improve drastically across all measures with the $M4$ configuration, showcasing more diversity of sampled identities and, in turn, better identity separability. Importantly, more likely sampling of new distinct identities, also enables more efficient generation of privacy-preserving data with the proposed PD filter. Interestingly, a further increase in the multiplication factor with the $M8$ configuration only leads to minor improvements over the $M4$ configuration, with the FMR1000 score actually being worse. The quality-based results reported in Table 9 also showcase that a multiplication factor above 1 drastically improves the quality of samples. Out of the three configurations, the $M4$ configuration performs the best in terms of FID and LPIPS scores, but is slightly surpassed in CR-FIQA score by the $M8$ configuration.

Overall, these results demonstrates the advantages of utilizing the proposed identity multiplication approach during the data generation process. Considering the drastic improvements between the $M1$ and $M4$ configurations and the slight negative effect on quality with the $M8$ configuration, we select a multiplication factor of 4 as the baseline for our ArcBiFaceGAN framework.

Table 11
Comparison of real-world time requirements of different generative methods. Reported is the time required for training to converge and the time required to generate datasets with either 95, 500 or 1000 identities, each with 32 samples.

Generative model	PD	Training time[a]	Generation time [min]		
			95	500	1000
StyleGAN2 (Karras et al., 2020a)	–	≈ 50 h	1.7	8.8	17.6
DB-StyleGAN2 (Tomašević et al., 2022)	–	≈ 87 h	2.1	10.8	21.5
SFace (Boutros et al., 2022)	–	≈ 65 h	6.9	–	–
Mix-SFace (Boutros et al., 2022)	–	≈ 65 h	7.1	42.7	88.7
	✓	–	15.3	389.6	2258.8
ArcBiFaceGAN (Ours)	–	≈ 82 h	6.4	35.3	67.4
	✓	–	26.0	751.5	6731.1
IDiff-Face (N) (Boutros et al., 2023b)	–	≈ 7 h	75.7	401.6	804.9
IDiff-Face (Boutros et al., 2023b)	–	≈ 11 h	105.7	562.3	1128.3

[a] Approximate estimate.

4.5. Real-world training and generation requirements

The last set of experiments is focused on the real-world requirements for employing the generative models. This includes the time and the video memory required for both training and data generation.

4.5.1. Real-world time analysis

We first compare our proposed ArcBiFaceGAN approach with the state-of-the-art in terms of training speed and dataset generation times. To this end, we report in Table 11 the time required for the different generative methods models to converge during training, along with the time required to generate the datasets with or without the proposed Privacy and Diversity (PD) filter. Additionally, for each approach we report the time required to create the datasets with either 95, 500 or 1000 identities with 32 image samples per identity. All reported scores are obtained with the hardware described in Section 4.1.

We begin by analyzing the effects of identity-conditioning on the training speed of GAN-based approaches. The results in Table 11 reveal that the additional conditioning actually speeds up training convergence, when compared to training the unconditional DB-StyleGAN2 (Tomašević et al., 2022). However, all these approaches require longer training than the single spectrum StyleGAN2, due to the additional complexity of training on poorly aligned VIS-NIR image pairs. This is also reflected in more time required to generate the datasets, where the approaches based on the dual-branch architecture have to produce two images at once. However this is still faster than running the single spectrum StyleGAN2 twice.

In comparison, the diffusion-based IDiff-Face (Boutros et al., 2023b) model is slowed noticeably by the identity-conditioning. In terms of data generation speeds, the model requires roughly 40% more time, while the difference in training times is even larger. Nevertheless, convergence during training is achieved much quicker on the small-scale dataset, in comparison to other GAN-based approaches. On the other hand, the generation times are drastically larger than with other solutions due to the multiple denoising steps required for each image. Here, it should also be noted that the discussed IDiff-Face (Boutros et al., 2023b) only generates images in the visible spectrum. A multispectral variant would likely be even slower. When comparing the identity-conditioned SFace (Boutros et al., 2022) to the proposed Mix-SFace variant, we observe only a slight difference in the time required to generate datasets, caused by the sampling of random vector combinations. Interestingly, our ArcBiFaceGAN achieves slightly faster generation times, however this is likely due to minimal implementation differences.

Importantly, we also explore how our proposed Privacy and Diversity (PD) filter affects dataset generation speeds. We observe a substantial slowdown with the use of the PD filter with both Mix-SFace and ArcBiFaceGAN. The generation speed also gets exponentially slower with increasing amounts of synthetic identities, due to also comparing new synthetic identities to previously generated ones. Here, the generation speed of ArcBiFaceGAN is impacted the most, since it is more difficult to find new identities in a latent space via random sampling, than it is to generate unique one-hot encoded vectors. We could avoid this slow down, by only ensuring the generation of privacy-preserving data. However, this would come at a cost of important benefits related to improving the intra-identity diversity and identity separability. With the use of the PD filter the generation speed of 1000 identities with ArcBiFaceGAN actually becomes substantially slower than the speed of IDiff-Face (Boutros et al., 2023b). Despite this, the above mentioned perks and benefits of generating privacy-preserving data far outweigh the increase in inference times. Furthermore, adding the PD filter to the IDiff-Face (Boutros et al., 2023b) to enable privacy-preserving data generation would be impractical, due to its already slow synthesis speed.

4.5.2. Real-world video memory footprint

Lastly, we compare the different generative models used throughout the experiments in terms of the number of parameters and the video memory (VRAM) required for performing training and inference, as reported in Table 12. We begin by evaluating the differences between our proposed ArcBiFaceGAN framework and the SFace (Boutros et al., 2022) and Mix-SFace (Boutros et al., 2022) approaches. As in the experiments above, the latter models are adapted to also produce data in both the VIS and the NIR spectrum by relying on the DB-StyleGAN2 (Tomašević et al., 2022) architecture that is utilized by ArcBiFaceGAN. Overall, we observe only minor difference between the models across all results, which are tied to the dimensions of the different identity conditions. Our ArcBiFaceGAN utilizes a 512 dimensional identity feature vector, while the dimension of the identity one-hot vector used by SFace (Boutros et al., 2022) and Mix-SFace (Boutros et al., 2022) is determined by the amount of identities in the training dataset (95 in the case of the Tufts Face Database (Panetta et al., 2018)).

From Table 12 we can also observe a difference between conditional (SFace (Boutros et al., 2022), Mix-SFace (Boutros et al., 2022) and ArcBiFaceGAN) and non-conditional (DB-StyleGAN2 (Tomašević et al., 2022)) multispectral models. This difference is caused by the need for an additional mapping network for the discriminators, described in Section 3.2.2, as well as the additional fully-connected layer at the start of the mapping network that interprets the input identity feature, as presented in Section 3.2.1. A drastic decrease in the parameters and the VRAM is also seen with the original single spectrum StyleGAN2 (Karras et al., 2020a), due to the need for only a single output branch in the synthesis network and only a single discriminator network during training. Overall, we also observe a notable decrease of parameters and VRAM usage during data generation with all StyleGAN-based models in comparison to the training process, because the discriminator are not required during inference. Furthermore, the training process of these models utilizes a batch size of 12, while the batch size is set to 1 during data generation to enable the use of the proposed PD filter. When also utilizing the PD filter during data generation, with

Table 12
Comparison of generative models in terms of their footprint. Reported are the number of parameters of each model and the VRAM required for training and data generation. Configurations that utilize the proposed PD filter are also included.

Generative model	PD	Training		Data generation	
		Parameters	VRAM	Parameters	VRAM
StyleGAN2 (Karras et al., 2020a)	–	48,768,547	6367 MiB	24,767,458	1357 MiB
DB-StyleGAN2 (Tomašević et al., 2022)	–	74,052,459	9282 MiB	26,050,409	1383 MiB
SFace (Boutros et al., 2022)	–	76,776,298	9410 MiB	26,361,705	1409 MiB
Mix-SFace (Boutros et al., 2022)	–	76,776,298	9410 MiB	26,361,705	1409 MiB
	✓	–	–	92,013,715	1679 MiB
ArcBiFaceGAN (Ours)	–	77,203,306	9534 MiB	26,575,209	1413 MiB
	✓	–	–	92,227,219	1683 MiB
IDiff-Face (Boutros et al., 2023b)	–	271,663,969	8813 MiB	249,326,546	1309 MiB

the Mix-SFace (Boutros et al., 2022) and ArcBiFaceGAN approaches, we note a drastic increase in the number of parameters, due to the use of the ArcFace (Deng et al., 2019a) recognition model and the MTCNN face detector (Zhang et al., 2016). However, since these models are pretrained and implemented in an efficient manner, the VRAM usage only increases slightly.

In comparison, the diffusion-based IDiff-Face model (Boutros et al., 2023b) utilizes a drastically larger amount of parameters than all StyleGAN-based models. However, due to various optimizations (e.g. use of the half-precision floating-point format) and the selected lower image resolution, this is not reflected in the VRAM usage. Overall, the memory footprint is slightly smaller than that of conditional StyleGAN-based approaches, both during training and data generation. Nevertheless, the IDiff-Face model (Boutros et al., 2023b) is still notably slower in practice than StyleGAN-based approaches, as demonstrated in Section 4.5.1, since it requires multiple passes through the network to generate a single sample.

5. Conclusion

In this paper, we presented ArcBiFaceGAN, a new generative framework that facilitates the synthesis of privacy-preserving multispectral face recognition datasets. At its core, the framework relies on a novel identity-conditioned Dual-Branch StyleGAN2 model capable of generating aligned high-quality visible (VIS) and near-infrared (NIR) face images of synthetic identities, determined by identity features from a pretrained face recognition model. To enable the creation of privacy-preserving datasets the framework also utilizes a novel Privacy and Diversity (PD) filter, which removes synthetic samples with privacy-breaching identities while ensuring better identity separability and intra-identity diversity.

Throughout the experiments we demonstrated that the proposed ArcBiFaceGAN framework is able to compete with synthesis capabilities of state-of-the-art generative methods while producing data that is privacy-preserving. Importantly, we showed that recognition models trained on synthetic data of ArcBiFaceGAN achieve higher verification accuracy on multiple real-world benchmarks than models trained on data of existing generative methods. The generation of more synthetic identities also led to better verification performance than with other methods or even with real-world data, thus exhibiting the possibility of replacing real-world datasets. In addition, we observed that training recognition models on both VIS and NIR data results in higher accuracy even on benchmarks that only contain visible spectrum images and their grayscale representation. This suggests that we could potentially improve the performance of existing recognition systems without requiring any additional sensors.

Overall, our ArcBiFaceGAN framework offers a potent solution for addressing the increasing privacy concerns and the need for multispectral recognition data. However, our work also sheds light on the issues that all generative models face when trained on small-scale multispectral datasets. Future work could extend this research by adding more intricate control over the generative process or by exploring domain transferring possibilities between the different spectra. The use of diffusion-based models for multispectral data generation should also be explored, especially in a low-data regime.

CRediT authorship contribution statement

Darian Tomašević: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft, Formal analysis. **Fadi Boutros:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Supervision, Validation, Writing – review & editing. **Naser Damer:** Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Peter Peer:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Vitomir Štruc:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

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

Data availability

In our research we used publicly available datasets.

Acknowledgments

Supported in parts by the Slovenian Research and Innovation Agency (ARIS) through the ARIS Research Programmes P2-0250 “Metrology and Biometric Systems” and P2-0214 “Computer Vision”, the ARIS Project J2-50065 “DeepFake DAD”, and the ARIS Young Researcher programme.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: International Conference on Machine Learning. ICML, pp. 214–223.
Bansal, A., Nanduri, A., Castillo, C.D., Ranjan, R., Chellappa, R., 2017. UMDFaces: An annotated face dataset for training deep networks. In: IEEE International Joint Conference on Biometrics. IJCB, pp. 464–473.
Batagelj, B., Peer, P., Štruc, V., Dobrišek, S., 2021. How to correctly detect face-masks for COVID-19 from visual information? MDPI Applied Sciences 11 (5), 2070.
Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A., 2019. Visualizing and understanding generative adversarial networks. In: International Conference on Learning Representations. ICLR, pp. 1–4.
Bourlai, T., 2016. Face Recognition Across the Imaging Spectrum. Springer, Cham.
Boutros, F., Fang, M., Klemt, M., Fu, B., Damer, N., 2023a. CR-FIQA: Face image quality assessment by learning sample relative classifiability. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5836–5845.
Boutros, F., Grebe, J.H., Kuijper, A., Damer, N., 2023b. IDiff-Face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In: IEEE/CVF International Conference on Computer Vision. ICCV, pp. 19650–19661.

Boutros, F., Huber, M., Siebke, P., Rieber, T., Damer, N., 2022. SFace: Privacy-friendly and accurate face recognition using synthetic data. In: IEEE International Joint Conference on Biometrics. IJCB, pp. 1–11.

Boutros, F., Klemt, M., Fang, M., Kuijper, A., Damer, N., 2023c. ExFaceGAN: Exploring identity directions in GAN's learned latent. In: IEEE International Joint Conference on Biometrics. IJCB, pp. 1–10.

Boutros, F., Klemt, M., Fang, M., Kuijper, A., Damer, N., 2023d. Unsupervised face recognition using unlabeled synthetic data. In: IEEE International Conference on Automatic Face and Gesture Recognition. FG, pp. 1–8.

Boutros, F., Struc, V., Fierrez, J., Damer, N., 2023e. Synthetic data for face recognition: Current state and future prospects. Image Vis. Comput. 104688.

Brock, A., Donahue, J., Simonyan, K., 2018. Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations. ICLR, pp. 1–35.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2018. VGGFace2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition. FG, pp. 67–74.

Chambino, L.L., Silva, J.S., Bernardino, A., 2021. Multispectral face recognition using transfer learning with adaptation of domain specific units. MDPI Sensors 21 (13), 4520.

Croitoru, F.-A., Hondru, V., Ionescu, R.T., Shah, M., 2023. Diffusion models in vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI).

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2020. Randaugment: Practical automated data augmentation with a reduced search space. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, pp. 702–703.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 248–255.

Deng, J., Guo, J., Xue, N., Zafeiriou, S., 2019a. ArcFace: Additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4690–4699.

Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S., 2019b. Lightweight face recognition challenge. In: IEEE/CVF International Conference on Computer Vision Workshops. ICCVW.

Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X., 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5154–5163.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems. NeurIPS, vol. 34, pp. 8780–8794.

Durugkar, I., Gemp, I., Mahadevan, S., 2017. Generative multi-adversarial networks. In: International Conference on Learning Representations. ICLR, pp. 1–14.

Duta, I.C., Liu, L., Zhu, F., Shao, L., 2021. Improved residual networks for image and video recognition. In: IEEE International Conference on Pattern Recognition. ICPR, pp. 9415–9422.

Emeršič, Ž., Sušanj, D., Meden, B., Peer, P., Štruc, V., 2021. ContexedNet: Context–aware ear detection in unconstrained settings. IEEE Access 9, 145175–145190.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems. NeurIPS, pp. 2672–2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems. NeurIPS, pp. 5769–5779.

Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J., 2016a. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In: Springer European Conference on Computer Vision. ECCV, pp. 87–102.

Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J., 2016b. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In: Springer European Conference on Computer Vision. ECCV, pp. 87–102.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778.

Hempel, T., Abdelrahman, A.A., Al-Hamadi, A., 2022. 6D rotation representation for unconstrained head pose estimation. In: IEEE International Conference on Image Processing. ICIP, pp. 2496–2500.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems. NeurIPS, pp. 6626–6637.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. NeurIPS, vol. 33, pp. 6840–6851.

Hoofnagle, C.J., Van Der Sloot, B., Borgesius, F.Z., 2019. The European union general data protection regulation: What it is and what it means. Inform. Commun. Technol. Law 28 (1), 65–98.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Tech. Rep. 07–49, University of Massachusetts, Amherst.

Jasserand, C., 2018. Massive facial databases and the GDPR: The new data protection rules applicable to research. In: Data Protection and Privacy: The Internet of Bodies. Bloomsbury Publishing, Oxford, pp. 169–188.

Jasserand, C., 2022. Research, the GDPR, and mega biometric training datasets: Opening the pandora box. In: International Conference of the Biometrics Special Interest Group. BIOSIG, pp. 1–6.

Joshi, I., Grimmer, M., Rathgeb, C., Busch, C., Bremond, F., Dantcheva, A., 2024. Synthetic data in human analysis: A survey. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI).

Joyce, J.M., 2011. Kullback-Leibler divergence. In: International Encyclopedia of Statistical Science. Springer, Berlin, pp. 720–722.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations. ICLR, pp. 1–26.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020a. Training generative adversarial networks with limited data. In: Advances in Neural Information Processing Systems. NeurIPS, pp. 12104–12114.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T., 2021. Alias-free generative adversarial networks. In: Advances in Neural Information Processing Systems. NeurIPS, pp. 852–863.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020b. Analyzing and improving the image quality of StyleGAN. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 8110–8119.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations. ICLR, pp. 1–5.

Luo, Y., Pi, D., Pan, Y., Xie, L., Yu, W., Liu, Y., 2022. ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light. Expert Syst. Appl. 191, 116269.

Maas, A.L., Hannun, A.Y., Ng, A.Y., et al., 2013. Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning. ICML, pp. 1–3.

Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.K., 2002. FVC2000: Fingerprint verification competition. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 24 (3), 402–412.

Martins, P., Silva, J.S., Bernardino, A., 2022. Multispectral facial recognition in the wild. MDPI Sensors 22 (11), 4219.

Meden, B., Rot, P., Terhörst, P., Damer, N., Kuijper, A., Scheirer, W.J., Ross, A., Peer, P., Štruc, V., 2021. Privacy-enhancing face biometrics: A comprehensive survey. IEEE Trans. Inform. Forensics Secur. (TIFS) 16, 4147–4183.

Mescheder, L., Geiger, A., Nowozin, S., 2018. Which training methods for GANs do actually converge? In: International Conference on Machine Learning. ICML, pp. 3481–3490.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations. ICLR, pp. 1–26.

Moon, T., Choi, M., Lee, G., Ha, J.-W., Lee, J., 2022. Fine-tuning diffusion models with limited data. In: NeurIPS Workshop on Score-Based Methods. pp. 1–14.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S., 2017. AgeDB: The first manually collected, in-the-wild age database. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, pp. 51–59.

Nguyen, H.V., Bai, L., 2010. Cosine similarity metric learning for face verification. In: Asian Conference on Computer Vision. Springer, pp. 709–720.

Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S.P., Kaszowska, A., Taylor, H.A., Samani, A., et al., 2018. A comprehensive database for benchmarking imaging systems. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 42 (3), 509–520.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. NeurIPS, pp. 8026–8037.

Poh, N., Bengio, S., 2004. A Study of the Effects of Score Normalisation Prior to Fusion in Biometric Authentication Tasks. Tech. rep., IDIAP.

Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., Tao, D., 2021. SynFace: Face recognition with synthetic data. In: IEEE/CVF International Conference on Computer Vision. ICCV, pp. 10880–10890.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10684–10695.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. MICCAI, pp. 234–241.

Rose, J., Liu, H., Bourlai, T., 2022. Multispectral face mask compliance classification during a pandemic. In: Disease Control Through Social Network Surveillance. Springer, Cham, pp. 189–206.

Rot, P., Vitek, M., Meden, B., Emeršič, Ž., Peer, P., 2019. Deep periocular recognition: A case study. In: IEEE International Work Conference on Bioinspired Intelligence. IWOBI, pp. 21–26.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K., 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 22500–22510.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al., 2022. Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems. NeurIPS, vol. 35, pp. 36479–36494.

Sengupta, S., Chen, J.-C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W., 2016. Frontal to profile face verification in the wild. In: IEEE Winter Conference on Applications of Computer Vision. WACV, pp. 1–9.

Sequeira, A.F., Chen, L., Ferryman, J., Wild, P., Alonso-Fernandez, F., Bigun, J., Raja, K.B., Raghavendra, R., Busch, C., de Freitas Pereira, T., et al., 2017. Cross-eyed 2017: Cross-spectral iris/periocular recognition competition. In: IEEE International Joint Conference on Biometrics. IJCB, pp. 725–732.

Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., Scheirer, W.J., 2021. A study of the human perception of synthetic faces. In: IEEE International Conference on Automatic Face and Gesture Recognition. FG, pp. 1–8.

Shoshan, A., Bhonker, N., Kviatkovsky, I., Medioni, G., 2021. Gan-control: Explicitly controllable GANs. In: IEEE/CVF International Conference on Computer Vision. ICCV, pp. 14083–14093.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singh, J., Bhatia, H., Vatsa, M., Singh, R., Bharati, A., 2024. SynthProv: Interpretable framework for profiling identity leakage. In: IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 4746–4756.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. ICML, pp. 2256–2265.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2818–2826.

Tinsley, P., Czajka, A., Flynn, P., 2021. This face does not exist... but it might be yours! Identity leakage in generative models. In: IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 1320–1328.

Tomašević, D., Peer, P., Štruc, V., 2022. BiOcularGAN: Bimodal synthesis and annotation of ocular images. In: IEEE International Joint Conference on Biometrics. IJCB, pp. 1–10.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. (JMLR) 9 (86), 2579–2605.

Vitek, M., Hafner, A., Peer, P., Jaklič, A., 2021. Evaluation of deep approaches to sclera segmentation. In: International Convention on Information, Communication and Electronic Technology. MIPRO, pp. 1097–1102.

Vitek, M., Rot, P., Štruc, V., Peer, P., 2020. A comprehensive investigation into sclera biometrics: A novel dataset and performance study. Neural Computing and Applications (NCA) 32, 17941–17955.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W., 2018. CosFace: Large margin cosine loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5265–5274.

Wu, F., You, W., Smith, J.S., Lu, W., Zhang, B., 2019. Image-image translation to enhance near infrared face recognition. In: IEEE International Conference on Image Processing. ICIP, pp. 3442–3446.

Ye, H., Zhang, J., Liu, S., Han, X., Yang, W., 2023. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721.

Zhang, H., Grimmer, M., Ramachandra, R., Raja, K., Busch, C., 2021. On the applicability of synthetic data for face recognition. In: IEEE International Workshop on Biometrics and Forensics. IWBF, pp. 1–6.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 586–595.

Zhang, L., Rao, A., Agrawala, M., 2023a. Adding conditional control to text-to-image diffusion models. In: IEEE/CVF International Conference on Computer Vision. ICCV, pp. 3836–3847.

Zhang, L., Rao, A., Agrawala, M., 2023b. Adding conditional control to text-to-image diffusion models. In: IEEE/CVF International Conference on Computer Vision. ICCV, pp. 3836–3847.

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23 (10), 1499–1503.

Zheng, T., Deng, W., 2018. Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Vol. 5, no. 7. Tech. Rep., Beijing University of Posts and Telecommunications.

Zheng, T., Deng, W., Hu, J., 2017. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z., 2016. Face alignment across large poses: A 3d solution. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 146–155.