BiFaceGAN: Bimodal Face Image Synthesis

Darian Tomašević[1*], Peter Peer[1] and Vitomir Štruc[2]

[1]Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia.
[2]Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia.

*Corresponding author(s). E-mail(s): darian.tomasevic@fri.uni-lj.si;
Contributing authors: peter.peer@fri.uni-lj.si; vitomir.struc@fe.uni-lj.si;

Abstract

Modern face recognition and segmentation systems, such as all deep learning approaches, rely on large-scale annotated datasets to achieve competitive performance. However, gathering biometric data often raises privacy concerns and presents a labor-intensive and time-consuming task. Researchers are currently also exploring the use of multispectral data to improve existing solutions, limited to the visible spectrum. Unfortunately, the collection of suitable data is even more difficult, especially if aligned images are required. To address the outlined issues, we present a novel synthesis framework, named BiFaceGAN, capable of producing privacy-preserving large-scale synthetic datasets of photorealistic face images, in the visible and the near-infrared spectrum, along with corresponding ground truth pixel-level annotations. The proposed framework leverages an innovative Dual-Branch Style-based Generative Adversarial Network (DB-StyleGAN2) to generate per-pixel aligned bimodal images, followed by an ArcFace Privacy Filter (APF) that ensures the removal of privacy-breaching images. Furthermore, we also implement a Semantic Mask Generator (SMG) that produces reference ground truth segmentation masks of the synthetic data, based on the latent representations inside the synthesis model and only a handful of manually labeled examples. We evaluate the quality of generated images and annotations through a series of experiments and analyze the benefits of generating bimodal data with a single network. We also show that privacy-preserving data filtering does not notably degrade the

image quality of produced datasets. Finally, we demonstrate that the generated data can be employed to train highly successful deep segmentation models, which can generalize well to other real-world datasets.

Keywords: Image synthesis, Face-based biometric data, Multispectral images, Generative adversarial networks

1 Introduction

Deep learning approaches nowadays present the backbone of many state-of-the-art solutions across a variety of biometric tasks [1–4]. However, to achieve competitive performance such models require large-scale training datasets appropriate for the task at hand. Despite the current availability of online datasets, the collection and sharing of biometric data is becoming increasingly more difficult, due to growing privacy and copyright-related concerns and restrictions [5, 6]. Recently, strict data regulations even led to the retraction of valuable biometric datasets, in their entirety or in parts, due to the use of personal data without clear consent and other ethical issues [7–9]. Alternatively, manually gathering the required amount of biometric data is extremely time-consuming and labor-intensive, especially if the task requires semantic ground truth pixel-level annotations [10, 11]. Here, the consent agreement should also be phrased carefully, to allow for the use of collected data in various potentially-unforeseen research directions.

Meanwhile, novel biometric research is exploring ways of utilizing multispectral data to further improve current deep learning solutions, which are predominantly based on visible spectrum images [12–18]. For instance, many important cues are present in the near-infrared spectrum and not in the visible spectrum, and vice versa [19, 20]. Combining both data sources thus has clear potential in various biometric recognition and segmentation approaches. Unfortunately, large-scale multispectral datasets of aligned images are rather scarce and even more difficult to gather, due to the need for multiple imaging sensors and complex camera setups to enable simultaneous image capturing [21–23].

To address the lack of biometric datasets and the growing privacy concerns that accompany their distribution, researchers are considering the use of synthetic data to train various deep learning approaches [24–26]. This solution has recently gained ground, due to the rapid development of deep generative models, such as Generative Adversarial Networks (GANs) [27, 28]. Modern Style-based GAN approaches have enabled the synthesis of diverse photo-realistic images [29, 30], even when trained on limited reference data [31], thus making them suitable for use on smaller scale datasets. Several procedures have also been developed to produce pixel-level semantic labels of the generated images with minimal human intervention by exploiting the latent information of GANs [32, 33]. Despite these advances, the application of generative methods in multispectral-based biometrics has mostly remained limited

to cross-spectral image translation [34–36], while the actual simultaneous generation of multispectral data has remained poorly discussed [37].

Inspired by the increasing need for large-scale multispectral datasets, we present in this chapter, a novel bimodal generative framework, called BiFaceGAN, which extends our previous BiOcularGAN [37] approach to the more complex face modality with diverse characteristics, e.g. variations in pose, age, ethnicity, and gender. The BiFaceGAN framework enables the synthesis of privacy-preserving photorealistic bimodal face images, in both the visible (VIS) and the near-infrared (NIR) spectrum, accompanied by ground truth pixel-level annotations, as seen in Figure 1. To enable the generation of near-per-pixel aligned bimodal data from a small-scale training dataset of poorly aligned images, we utilize an innovative Dual-Branch StyleGAN2 design (DB-StyleGAN2) [37] and a custom two-phase training regime, designed based on the insights obtained from our previous architecture [37]. We also introduce a novel privacy-preserving filtering component based on the ArcFace model [38], called ArcFace Privacy Filter (APF), which ensures the generation of privacy-preserving datasets, by removing image samples whose identities match the identities of real subjects from the DB-StyleGAN2 training set. The framework also includes an auxiliary Semantic Mask Generator (SMG) component, which exploits semantic-rich latent information within the DB-StyleGAN2 model to produce accurate ground truth segmentation masks. We evaluate the data produced by our proposed framework through a series of experiments. First, we analyze the synthesis capabilities of the framework, in terms of image quality and diversity, with the use of the multispectral Tufts Face dataset [22]. Next, we explore the ability of privacy-based data filtering to produce privacy-preserving datasets and also investigate its effects on the image quality of the generated datasets. Lastly, we examine the utility of the generated ground truth segmentation data, by using it to train segmentation models and evaluate how well they perform on other real-world face datasets.

Overall, we make the following key contributions:

- We introduce a novel framework capable of generating privacy-preserving high-quality labeled bimodal face datasets of convincing and aligned images in both the visible (VIS) and the near-infrared (NIR) spectrum, along with corresponding ground truth pixel-level semantic annotations.
- We extend the existing Dual-Branch StyleGAN2 approach [37] to the more complex face modality and propose a custom training regime that enables stable training even on poorly aligned face datasets.
- We introduce a novel privacy-preserving data filtering step, ArcFace Privacy Filter (APF), which ensures that identities in the synthetically generated datasets do not match the real subjects of the Dual-Branch StyleGAN2 training dataset, whilst retaining image quality.
- We demonstrate that creating ground truth segmentation masks based on the bimodal data generation process results in better pixel-level labels than other unimodal approaches. In addition, we show that segmentation models perform better when utilizing data from both VIS and NIR domains.

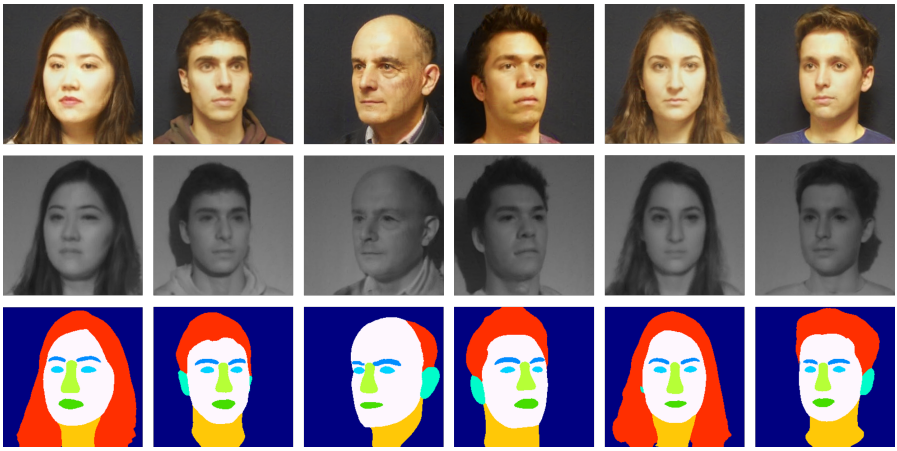

Fig. 1 Samples produced by the proposed BiFaceGAN framework. The framework generates high-quality privacy-preserving near-per-pixel aligned visible (VIS) and near-infrared (NIR) images accompanied by ground truth pixel-level annotations.

2 Related work

In this section, we review work closely related to the proposed BiFaceGAN framework. We position our contribution with respect to existing research on image generation and specifically the generation of face images, as well as the automatic creation of labeled datasets.

2.1 Image generation

The rapid evolution of image synthesis techniques in the past decade was largely enabled by the emergence of deep generative models, most notably Generative Adversarial Networks (GANs) [27]. This machine learning architecture consists of two neural networks, the generator and the discriminator, which compete against each other to facilitate the creation of photorealistic images.

Several works have since then expanded on the GAN model. Methods aimed at further improving the quality and resolution of synthetic images ranged from utilizing multiple discriminators [39] to progressively learning the generation process [40]. To address problems with training instability, researchers developed new regularization methods [41, 42] and utilized custom distance metrics for comparing synthetic and real distributions [43, 44]. Control over features of the generated images was also improved with the use of additional class label inputs [45]. Unfortunately, despite significant progress and extensive analysis, the inner workings of the models and the origin of certain features were not well understood [46].

To alleviate these issues a novel Style-based GAN model, known as StyleGAN, was proposed by Karras *et al.* [28], inspired by advances in style-transfer approaches. To achieve unmatched performance the authors split the generator into the mapping and the synthesis network. The first determines the style

of images within an intermediate latent space, while the latter relies on this style information to generate images. This in turn allows for better control over the image synthesis process.

Karras *et al.* [29] have since then addressed various artifacts in the produced images, e.g. blob-like shapes and location preference of features, by redesigning the building blocks of the generator, in the new approach named StyleGAN2. Later, the authors also focused on enabling smoother interpolation across the feature space by solving artifacts related to textures sticking to image coordinates rather than the underlying generated surfaces [30].

To reduce the immense amount of required data for training successful models, Karras *et al.* [31] also proposed the use of image augmentations during training. They introduced a novel Adaptive Discriminator Augmentation mechanism (ADA), which adaptively applies geometric and color transforms to images before the discriminator so that the augmentations do not leak to the generator. This lowers the required amount of training data by several orders of magnitude, thus enabling new practical use cases, namely in image-based biometrics, due to the small size of datasets.

2.2 Synthesis of biometric data

The application of generative methods in the field of biometrics has attracted attention in recent years, especially due to increasing privacy concerns related to collecting and sharing of biometric data [5, 6]. This interest is further supported by the lack of available large-scale biometric datasets, suitable for training deep recognition and segmentation models [11, 23].

To address these permeating issues, generative methods such as GANs [27] have been applied in a variety of tasks and approaches. Data de-identification methods present a possible solution, however, the anonymized data is still limited by the scale of the original dataset [6]. In comparison, by utilizing modern GAN models to generate completely new synthetic data we are able to create privacy-preserving datasets of high-quality and diversity [25].

Zhang et al. [24] explored the suitability of synthetic StyleGAN-based face images for recognition methods. They observed only minor differences between real and synthetic samples with modern face quality assessment approaches. Shen et al. [47] also performed a human-based study and showcased that even human individuals can be fooled by synthetic face images generated by modern generative methods. Recently, Qiu *et al.* [48] analyzed the performance gap between face recognition models trained on either real data or synthetic data produced by GANs. They identified it was caused by poor intraclass variations and the domain gap between real and synthetic images. To address this they introduced the SynFace GAN-based model which relied on identity and domain mixup between real and synthetic data during its training.

Boutros *et al.* [49] proposed a novel unsupervised approach for training recognition models on synthetic datasets. To achieve this, they utilized GANs to not only generate synthetic images but also induce face-changing augmentations. They improved the traditional training paradigm by maximizing the

similarity between two augmented views of the same image while minimizing the similarity to other images. This enables training without labeled datasets and, in turn, increases the potential value of synthetic datasets.

Boutros *et al.* [25] also proposed an approach for generating labeled face datasets of synthetic identities for purposes of developing face recognition models. To this end, they relied on the use of the novel StyleGAN2-ADA model [31], which they conditioned on identity class labels. In their work, they analyzed the identity transfer from generator training to the generated data as well as the identifiability of the authentic data in the trained models.

Unfortunately, the generation of multispectral face data has, to the best of our knowledge, not been adequately explored despite the growing research interest in multispectral solutions, due to the valuable cues available in the non-visible light domains [13, 17, 23]. The small scale and scarcity of multispectral face datasets along with the poor alignment of data present a difficult obstacle for training generative methods. These issues are less prominent with the ocular modality, due to the design of custom sensors. In our previous work [37] we investigated the synthesis of multispectral ocular data and introduced the first bimodal StyleGAN-based framework (BiOcularGAN) for generating visible and near-infrared spectrum data along with corresponding segmentation mask. Through a series of experiments, we showcased the clear potential of utilizing synthetic data for training modern deep segmentation approaches.

In this work, we build on previous approaches and extend the application of generative models to poorly aligned multispectral face datasets. In addition, we explore the generation of privacy-preserving data with the use of current recognition technologies and present a novel BiFaceGAN framework for synthesis of privacy-preserving aligned visible and near-infrared face images with accompanying fine-grained ground truth segmentation masks.

2.3 Semi-supervised segmentation

The generation of semantic segmentation masks represents an extensively studied task in computer vision. Prevalent deep models [50, 51] rely on supervised learning to solve this task. However, to achieve competitive performance such methods require domain-specific large-scale annotated datasets, which are labor-intensive and time-consuming to gather. Thus, researchers are exploring semi-supervised ways of utilizing the power of generative models to create segmentation masks and reduce the amount of human involvement.

Several GAN-based approaches have been developed to produce accurate segmentation masks with a limited amount of pixel-level labeled images along with a larger set of weak annotations. Souly *et al.* [52] replaced the discriminator of a conditional GAN with a multi-class classifier, which acted as the segmentation network at test time. Hung *et al.* [53] instead treated the generator as the segmentation network. Based on an input image, it produced a probability map of semantic labels. Recently, Mittal *et al.* [54] introduced one of the first approaches, which relied only on the small set of pixel-level training samples. The approach utilized both a generator-based segmentation network

as well as a separate multi-label classification branch, to address both low-level and high-level segmentation errors.

More recent approaches investigated the simultaneous creation of synthetic image and segmentation mask pairs. To achieve this they relied on the novel StyleGAN [28] architecture. Li *et al.* [32] proposed that the latent representation inside the image-synthesis model could be utilized to create corresponding segmentation masks. They introduced a novel SemanticGAN model, which relied on a separate label synthesis branch and two discriminator networks. The labeling branch was trained separately, due to the limited data regime. The model achieved incredible segmentation performance with only less than a hundred labeled samples. Zhang *et al.* [33] proposed an alternative StyleGAN-based framework, named DatasetGAN, which relied on extracting latent features from the synthesis network and passing them to an ensemble multi-layer perceptron (MLP) classifier. By utilizing a split training process for the StyleGAN and the ensemble classifier, they achieved state-of-the-art results and required only a handful of annotations. Pakhomov *et al.* [55] built on the DatasetGAN framework [33] and presented a method which does not require any labeled training samples. To achieve this they instead applied k-means clustering to the extracted latent feature vectors. Unfortunately, this resulted in overall lower segmentation accuracy and blob-like artifacts.

3 Generation of synthetic face images

In this section, we describe the inner workings of the proposed BiFaceGAN framework. We begin with an overview of the framework and then define in detail its main components.

3.1 Overview of BiFaceGAN

The BiFaceGAN framework relies on three main components to enable the generation of privacy-preserving photo-realistic bimodal images and corresponding segmentation masks, as presented in Fig. 2. The generation of visible (VIS) and near-infrared (NIR) spectrum images is handled by the **Dual-Branch StyleGAN2** generative model (**DB-StyleGAN2**) while associated ground truth segmentation masks are created by a separate **Semantic Mask Generator** (**SMG**) component. To ensure that privacy constraints are met we also utilize a separate **ArcFace Privacy Filter** (**APF**), to remove images that contain real-world identities. The produced data can thus be used for training deep biometric segmentation solutions in a privacy-preserving manner.

The data generation process of the Dual-Branch StyleGAN2 starts with a randomly sampled latent input $\mathbf{z} \in \mathcal{Z}$, which is first mapped to an intermediate latent code $\mathbf{w} \in \mathcal{W}$ by the mapping network f. This representation encodes the style of the images to be created. The code is then passed to the synthesis network g that generates pixel-aligned images in the visible and near-infrared domain $\mathbf{x}_{VIS} \in \mathbb{R}^{W \times H \times 3}$ and $\mathbf{x}_{NIR} \in \mathbb{R}^{W \times H}$. The generated images are then filtered by the ArcFace Privacy Filter (denoted as APF) based on their

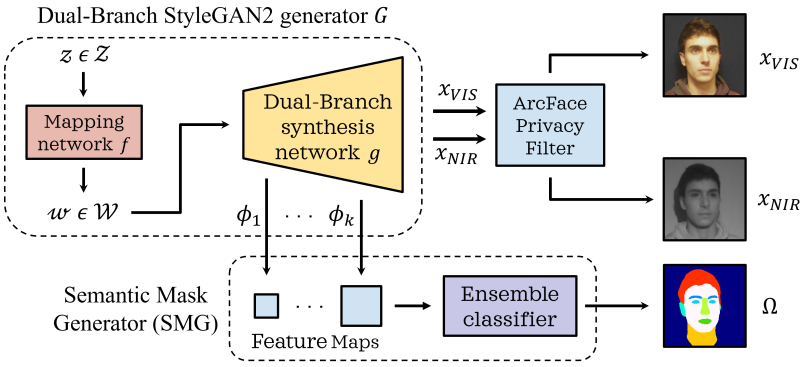

Fig. 2 Overview of the BiFaceGAN data generation process. The Dual-Branch synthesis network generates pairs of VIS and NIR images, whose style is determined by the mapping network. The produced samples are then passed through the ArcFace Privacy Filter (APF), which ensures that the synthetic identities present in the final datasets do not match the identities of the data used for training the Dual-Branch StyleGAN2 model. To create ground truth pixel-level labels of the images, the Semantic Mask Generator (SMG) extracts and exploits internal feature maps of the synthesis network.

identity similarity with the training data of the DB-StyleGAN2 model. Here, only images that are below a predefined threshold τ are kept. This synthesis process can more formally be defined as:

$$\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}\} = APF(\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}\}) = APF(G(\mathbf{z})) = APF(g(f(\mathbf{z}))). \quad (1)$$

During this process, the latent feature maps within the synthesis network are extracted, upsampled, and passed to the Semantic Mask Generator component (SMG), which with the use of an ensemble classifier produces ground truth pixel-level class labels $\Omega \in \mathbb{R}^{W \times H}$, i.e. segmentation masks, as follows:

$$\Omega = SMG(\phi_1(\mathbf{w}), \phi_2(\mathbf{w}), \ldots, \phi_k(\mathbf{w})). \quad (2)$$

Here $\phi_k(\mathbf{w})$ represents the mappings between the input $\mathbf{w}$ to the synthesis network up to a certain layer k in the synthesis network and $SMG(*)$ denotes the mapping of the SMG component from the set of latent features $\Phi = (\phi_1(\mathbf{w}), ..., \phi_k(\mathbf{w}))$ to the mask Ω. Altogether, the BiFaceGAN framework generates the triplet $\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}, \Omega\}$ based on the input latent code $\mathbf{z}$ sampled from a normal distribution.

3.2 Dual-Branch StyleGAN2 generator

The main component of BiFaceGAN is the Dual-Branch StyleGAN2 generator G, which follows the dual-branch generator architecture of our previous BiOcularGAN approach [37]. The model presents an extension of the original StyleGAN2 generator [29, 31] with the addition of separate synthesis branches along the synthesis network, as illustrated in Fig. 3 and Fig. 4.

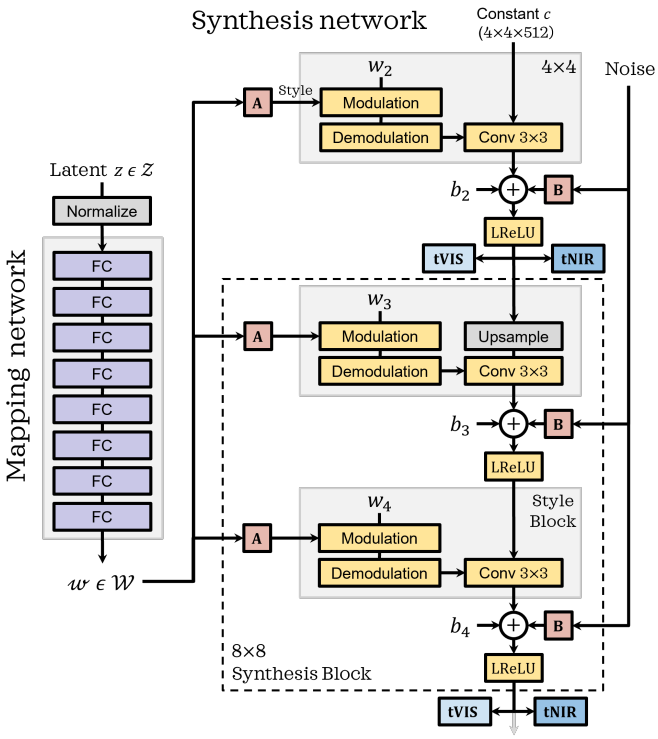

Fig. 3 Components of the Dual-Branch StyleGAN2 generator. The mapping network determines the style of images generated by the synthesis network. The latter consists of synthesis blocks tied to a specific resolution, which produce bimodal images. This is achieved with the tVIS ("toVIS") and tNIR ("toNIR") output layers that map high-dimensional latent representations to image data.

The mapping network f entails 8 fully-connected layers that are responsible for transforming a 512-dimensional input latent code z into an intermediate 512-dimensional latent code $\mathbf{w}$. Meanwhile, the synthesis network consists of a series of synthesis blocks, one for each resolution size, ranging from 4×4 to 256×256. Each of the synthesis blocks consists of two separate style blocks (light gray boxes), corresponding to the auxiliary style inputs received from the mapping network through affine transformations A. This information is then embedded into the convolutional weights of the synthesis network with the use of modulation and demodulation operations [29]. Together, these operations mimic the effects of the Adaptive Instance Normalization (AdaIN) technique of the original StyleGAN [28] while ensuring better signal processing.

The forward pass through the synthesis network begins with a constant input c $(4 \times 4 \times 512)$ which is passed to the first 4×4 style block. Convolution is then applied, with the use of style-infused convolutional weights. After this, bias and noise are added, where the latter is obtained via noise broadcast operations B. The signal is then passed through the Leaky ReLU (LReLU)

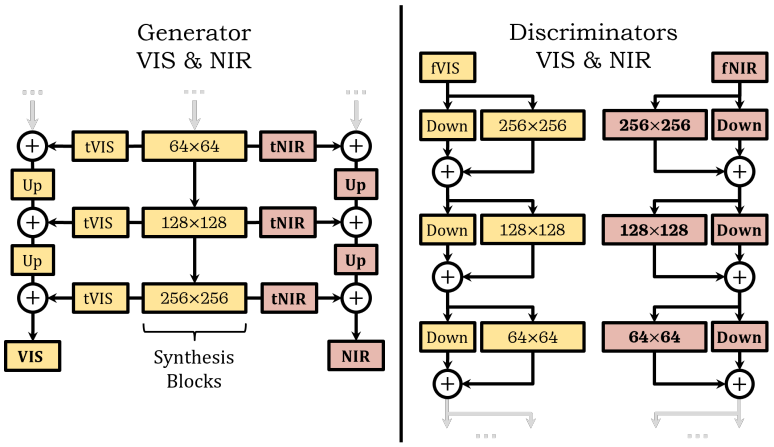

Fig. 4 High-level overview of the Dual-Branch StyleGAN2 generator and discriminator architectures. The generator produces visible and near-infrared images via separate synthesis branches and $tVIS$ and $tNIR$ convolutional layers. For training, the model utilizes two separate discriminators, one for each light spectrum. These receive image data from $fVIS$ and $fNIR$ operations, which transform images back to a high-dimensional representation. Upsampling and downsampling are represented by Up and Down.

activation function [56]. This process is repeated with each style block and the resolution is increased with each new synthesis block up to a final resolution of 256×256. The network thus consists of 7 consecutive synthesis blocks.

After each synthesis block, the network produces two outputs with the use of separate 1×1 convolutional layers, labeled as $tVIS$ ("toVIS") and $tNIR$ ("toNIR"). These layers interpret the latent representation inside the network as visible and near-infrared spectrum data respectively. These outputs are then used to form separate VIS and NIR synthesis branches, which upsample and merge the data to form the final bimodal images. The shallow depth of external branches allows the latent representation inside the network to carry most of the semantic-related image information. This is not only favorable in terms of training but also enables the synthesis of aligned images in two spectra from a single latent code. The rationale behind such a design is that it allows the images in both modalities to retain a similar structure of the face while the high-level appearance, tied to the different imaging sensors, is modeled in through the shallow output branches, which act as a sort of renderer. The described bimodal generation process is visualized in Figure 4.

3.3 Dual discriminator networks

The Dual-Branch StyleGAN2 model utilizes two separate discriminators (D_{VIS} and D_{NIR}) corresponding to images of the two light spectra, as depicted in Fig. 4. These evaluate whether the produced synthetic images are real or fake. This assessment allows us to train a model capable of generating photorealistic images in both domains with characteristics representative of the training distributions. The two discriminators share the same ResNet-like [57]

downsampling design, following the work of Karras *et al.* [29]. Furthermore, we rely on the Adaptive Discriminator Augmentation (ADA) [31] mechanism, which enables training of GANs in a low-data regime. This is crucial for enabling the use of generative models on small-scale biometric datasets.

During training, the images produced by the generator are first augmented with various geometric and color transforms based on an adaptive probability p following the ADA technique [31]. The altered images are then introduced to the discriminator through 1×1 convolutional layers denoted as $fVIS$ ("from VIS") and $fNIR$ ("from NIR"). The data is then passed through a series of downsampling resolution blocks that contain two convolutional layers along with a separate residual connection. The final block of the discriminator contains only a single convolutional layer, which transforms the latent representation of the discriminator into a binary decision. Similarly to the generator, the discriminator architecture consists of 7 resolution blocks in total, each of which downsamples the resolution by a power of two.

3.4 Dual-Branch StyleGAN2 training

The proposed model deviates from the original StyleGAN2 model [29] with the dual synthesis branch design, which simultaneously produces matching bimodal images in the two different light spectra. To train the model we thus utilize two separate discriminators and a corresponding multi-task adversarial learning objective. This facilitates the generation of photorealistic images in both domains while allowing the model to produce semantically similar bimodal images. The learning process is visualized in Fig. 5.

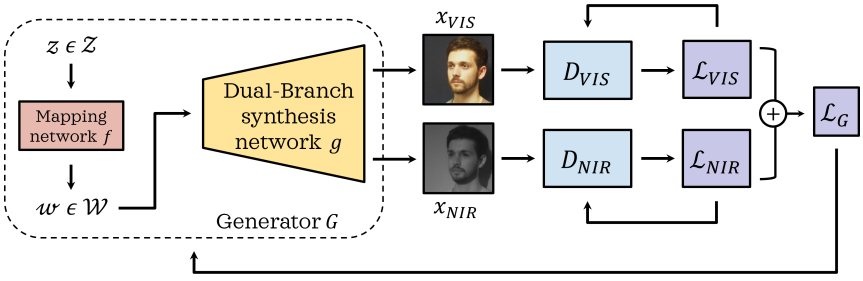

Fig. 5 Training procedure of the Dual-Branch StyleGAN2 component. The model is trained using two separate discriminators (D_{VIS} and D_{NIR}), one for each light spectrum, and the corresponding losses ($\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$). These are combined into $\mathcal{L}_G$ to train the generator G.

To define the adversarial learning objectives with a two-discriminator environment, we build on established unimodal learning strategies of StyleGAN-based approaches. We adopt the Non-Saturating Logistic loss [27] as well as two regularization methods (R_1 [42] and path length regularization [29]). As is standard practice, we implement the loss function with the soft-plus operation $sp(x) = \log(1 + \exp(x))$. Additionally, we apply regularization in a lazy manner, i.e. only every 16 mini-batches.

To train the two discriminators of bimodal images (D_{VIS} and D_{NIR}) we introduce two separate learning objectives, $\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$. These can formally be defined as:

$$\mathcal{L}_b = sp(-D_b(\mathbf{y}_b)) + sp(D_b(\mathbf{x}_b)) + \frac{\gamma_{R_1}}{2}\,\mathbb{E}\left[\nabla D_b(\mathbf{y}_b)^2\right];\ b \in \{VIS, NIR\}. \quad (3)$$

Here $\mathbf{y}_b$ denotes real (training) images from a given light spectrum b, while $\mathbf{x}_b$ represents synthetic images produced with the dual-branch generator G, following the process defined in Eq. (1). The hyperparameter of the R_1 regularization method is denoted with γ_{R_1} and is heuristically computed using the resolution res and batch size bs as $\gamma_{R_1} = 10^{-4}(2res^2/bs)$, as proposed by Karras *et al.* [29].

The learning objective of the generator $\mathcal{L}_G$ combines the two discriminator learning objectives $\mathcal{L}_{VIS}$ and $\mathcal{L}_{NIR}$. However, it only utilizes the terms tied to synthetic images, i.e. terms involving $\mathbf{x}_b$. In addition, it relies on the use of path length regularization $\gamma_{R_{PL}}$ [29]. Formally, $\mathcal{L}_G$ can be expressed as follows:

$$\mathcal{L}_G = \sum_{b \in B} w_b \cdot sp(-D_b(\mathbf{x}_b)) + \gamma_{R_{PL}}\,\mathbb{E}\Big(\Big\|\sum_{b \in B} \nabla(\mathbf{x}_b q_b)\Big\| - a\Big)^2. \quad (4)$$

Here, notations of the previous equation apply. In addition, q represents an image with normally distributed pixel intensities, a denotes the average of the computed norm, and weight w controls the influence of each spectrum on the final loss value. Meanwhile, the path length regularization hyperparameter $\gamma_{R_{PL}}$ is determined with $\gamma_{R_{PL}} = \ln 2/(res^2(\ln res - \ln 2)$, following the procedure of Karras *et al.* [29].

3.5 ArcFace Privacy Filter (APF)

We extend the Dual-Branch StyleGAN2 model with an additional privacy-preserving filtering step, to address the alarming similarity between identities of real and synthetic images, observed when training on small-scale biometric datasets. With this auxiliary step, we are able to prevent privacy-breaching synthetic images from reaching the final public synthetic datasets. We achieve this by removing images that surpass a certain similarity threshold with identities of the DB-StlyeGAN2 training set.

To filter the images our method relies on the recently introduced and widely-adopted ArcFace recognition model [38], which utilizes the iResNet-101 architecture [57, 58] pretrained on the MS1MV3 dataset [59]. In our work, we utilize the ArcFace model to construct a novel filtering component (ArcFace Privacy Filter) for our BiFaceGAN framework.

Once a bimodal image pair has been synthesized by DB-StyleGAN2 we pass the visible spectrum sample through the ArcFace model to obtain its 512-dimensional feature vector representation $AF(\mathbf{x})$, which contains valuable identity information. We then compare the obtained feature vector with

the feature representations of images from the DB-StyleGAN2 training set to determine the identity similarity and in turn the overall privacy of the synthetic image. To compare the feature vectors of two images we rely on the use of the cosine similarity metric [60], which has been extensively used in face recognition and verification research [38, 61, 62]. The similarity between real images $\mathbf{y}$ and synthetic images $\mathbf{x}$ is determined as follows:

$$similarity(\mathbf{x}_{VIS}, \mathbf{y}_{VIS}) = \frac{AF(\mathbf{x}_{VIS}) \cdot AF(\mathbf{y}_{VIS})}{\|AF(\mathbf{x}_{VIS})\| \|AF(\mathbf{y}_{VIS})\|}. \tag{5}$$

The similarity score is then compared to the desired privacy threshold τ, which determines whether the identities of synthetic images match those of real images used for training the DB-StyleGAN2 model. To speed up the similarity evaluation process, we first only compare the generated image with a subset of representative real images, i.e. frontal facing examples, one for each identity. In the next step, the synthetic sample is compared to all real images of the identity which achieved the highest score in the previous step. In comparison, assessing the similarity with all training images during the data generation process would be incredibly inefficient at least from a practical sense, especially since datasets often contain several images of each subject.

For our experiments, we select $\tau = 0.6$ as the privacy threshold, based on the mean and standard deviations of inter-class similarity as well as qualitative observations when comparing real and synthetic images. Nevertheless, the threshold can be adapted to achieve the desired privacy of synthetic data.

3.6 Semantic Mask Generator (SMG)

Additionally, we extend the Dual-Branch StyleGAN2 model to also generate ground truth semantic masks Ω corresponding to the created synthetic images. To achieve this, we exploit semantic-rich latent feature maps present within the synthesis network of the generator, inspired by Zhang *et al.* [33]. To implement the mapping $SMG(*)$ in Eq. (2) between latent feature maps and pixel-level labels, we utilize an ensemble classifier that operates on a per-pixel level. This process forms the last component of our model, i.e. the Semantic Mask Generator (SMG), which is depicted in Fig. 6.

As with our previous BiOcularGAN approach [37], we extract latent feature maps $\phi_k(\mathbf{w})$ directly after each Leaky ReLU (LReLU) [56] activation function in the synthesis network. Here the consecutive number of the LReLU operation is represented with k while $\mathbf{w}$ denotes the intermediate latent code input to the synthesis network. With this, we obtain semantic-filled latent information shared by the bimodal face images before they are rendered in a given light spectrum. This process differs from the DatasetGAN approach [33], which instead extracts features from the AdaIN operation of the initial StyleGAN [28] model that no longer exists in StyleGAN2 [31].

The extracted feature maps are then upsampled to the desired image resolution and merged to create a $W \times H \times d$ semantic representation of the

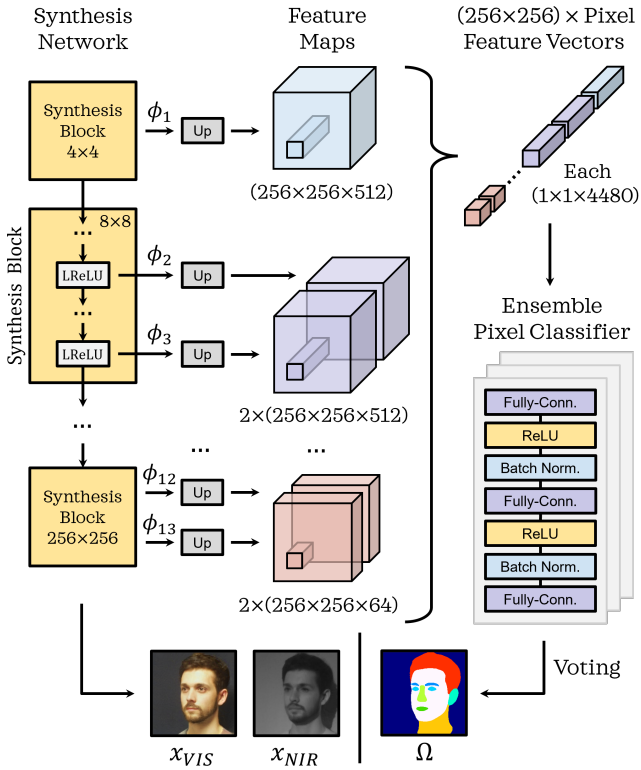

Fig. 6 Overview of the Semantic Mask Generator (SMG) component. The component exploits high-dimensional latent features (ϕ) extracted from LReLU operations of the DB-StyleGAN2 synthesis network to produce ground truth segmentation masks (Ω) corresponding to the synthetic images.

produced image. Here W and H dimensions are linked to the image resolution (in this case 256×256) and d represents the length of the combined feature maps, where the length of each is determined by the resolution of the origin synthesis block. Higher resolutions yield shorter feature maps, as defined by the following resolution-length mappings: $\{4 : 512, 8 : 512, 16 : 512, 32 : 512, 64 : 256, 128 : 128, 256 : 64\}$. For images of resolution 256×256 we extract 13 feature maps, one from each style block. Altogether this forms a tensor of $256 \times 256 \times 4480$. Thus, each of the image pixels is linked to a 4480-dimensional semantic-rich feature vector.

To determine the semantic class of each image pixel based on the corresponding feature vector we utilize an ensemble of classifiers, namely an ensemble of 10 three-layer Multi-Layer Perceptrons (MLPs) but we also experiment with a DeepLab-based classifier [63]. As part of the ensemble, we also utilize the majority voting strategy to minimize the noisy artifacts in the final semantic masks. The main advantage of this mask generation process is that we require only a small set of (at least 1) manually annotated images to train

an extremely accurate ensemble classifier since each pixel is its own training sample. More formally, each classifier C in the ensemble is trained using backpropagation to approximate the mapping $C : \mathcal{F} \mapsto \mathcal{M}$. Here $\mathcal{M} = (\Omega_1^*, ..., \Omega_i^*)$ defines the set of manually annotated segmentation masks Ω^* of the i synthetic image pairs x_{VIS} and x_{NIR}. Meanwhile, $\mathcal{F} = (\Phi_1, ..., \Phi_i)$ represents the sets of extracted feature maps belonging to the image pairs, where each set $\Phi = (\phi_1, ..., \phi_{13})$ includes all 13 feature maps extracted during the synthesis of a single image pair.

The Semantic Mask Generator (SMG) represents the last component in our BiFaceGAN framework, alongside the Dual-Branch StyleGAN2 image synthesis model and the ArcFace Privacy Filter. Altogether these components enable the generation of endless privacy-preserving pixel-aligned face images in both the visible and the near-infrared spectrum as well as corresponding ground truth segmentation masks. Thus, a single forward pass through the framework yields a data triplet $\{\mathbf{x}_{VIS}, \mathbf{x}_{NIR}, \Omega\}$.

4 Experiments and results

This section is dedicated to the evaluation of the presented BiFaceGAN framework. The following experiments are divided into two core parts, tied to the assessment of image synthesis capabilities and the evaluation of the generated ground truth segmentation masks. However, we will first introduce the utilized face datasets and preprocessing steps, followed by details of the model implementation and the proposed custom training procedure, as well as the metrics used for evaluation.

4.1 Experimental setup

4.1.1 Datasets and data preparation

Multispectral biometric datasets are currently still scarce and of small scale [22], despite recent advances in multispectral biometric research [13, 14]. The majority of these datasets are aimed at multispectral recognition, which does not necessarily require images in different spectra to match. Thus, datasets often contain images captured in various environments and at different time points. However, training our proposed BiFaceGAN framework necessitates well-matching visible and near-infrared image pairs. Unfortunately, none of the multispectral datasets contain simultaneously captured data. In our work, we utilized the most suitable of these datasets, this being the multispectral Tufts Face dataset [22], to train and evaluate the image synthesis component of the BiFaceGAN framework. For evaluating segmentation models trained on the synthetic data we use the CelebA-Mask-HQ dataset [10], which includes visible spectrum data along with ground truth segmentation masks. Characteristics of the two datasets and the preprocessed subsets are presented in Table 1. Detailed descriptions of the datasets and the preprocessing steps are provided below.

Table 1 Overview of face datasets used in experiments. BiFaceGAN is trained and validated on the multispectral Tufts Face dataset [22], while the visible spectrum CelebAMask-HQ [10] dataset is used for evaluating the quality of produced data, based on the performance of an auxiliary segmentation model.

Dataset	#Images	#IDs	Resolution	Modality†	Purpose‡
Tufts Face [22]	> 10,000	113	3280 × 2464	VIS/NIR	–
Tufts Face* [22]	2207	104	256 × 256	VIS/NIR	TR/SV
CelebA [10]	202,599	10,177	1024 × 1024	VIS	–
CelebAMask-HQ [10]	30,000	–	1024 × 1024	VIS	–
CelebAMask-HQ* [10]	9357	–	256 × 256	VIS	SE

* – processed subset, †VIS – visible spectrum, NIR – near-infrared spectrum
‡TR – training, SV – synthesis validation, SE – segmentation experiments

The **Tufts Face dataset** [22] represents one of the first datasets which contain a variety of heterogeneous data captured on the same individuals. The dataset includes images from three different spectra, i.e. visible light, thermal and infrared, along with a collection of 3D data, video data as well as a set of computerized sketches. In total, the database contains over 10,000 images of 113 participants. However, for the purposes of our research, we rely only on the visible (VIS) and the near-infrared (NIR) image subsets. These were captured in a semi-circle around sitting individuals with a custom quad-camera setup, under either diffused light for VIS images or under 850 nm Infrared 96 for NIR images. From the subsets, we removed images captured in the two maximum angle positions, i.e. side-profile images, to ensure that all faces contained key features, namely two eyes. In addition, we cropped the images to only contain the face modality, as seen in Fig. 7, and removed blurry images in which subjects moved. The final subset thus includes only 2207 VIS-NIR image pairs of 104 subjects.

Unfortunately, despite the custom camera setup, the visible and near-infrared images were not taken simultaneously. This resulted in significant VIS-NIR image pair misalignment, which makes training the proposed Dual-Branch StyleGAN2 model near impossible. Thus, we further processed the cropped images to ensure better alignment. First, we utilized a standard face detection procedure, based on the Histogram of Oriented Gradients (HOG) [64] and a Linear Support Vector Machine (SVM) [65], to obtain facial landmarks for each image, namely the eye positions. Then we computed the angle between the centroids of the eyes based on their image coordinates with the use of the arc-tangent function. We also computed a custom scaling factor, based on the initial distance between eye centroids and their desired distance, to ensure a similar face size across images of different identities and in turn a form of inter-class alignment. The obtained angle, scale, and the midpoint between the eye centroids are then used to define an affine transform that produces the final image. This process allows us to obtain slightly better aligned visible and near-infrared image pairs, as seen in Fig. 7, and also reduce additional complexity of training data e.g. face position. The final preprocessing steps resize

the images to a resolution of 256×256 and split the dataset is in a $9:1$ ratio into the training and holdout sets. Here the training set is used to train the Dual-Branch StyleGAN2 model, while the holdout set is used to evaluate the image synthesis quality of the model.

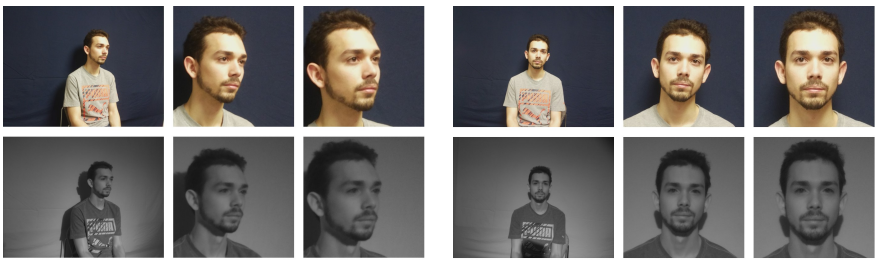

Fig. 7 Preparation steps of the Tufts Face dataset [22]. The original images are first cropped to an intermediate resolution. Eye positions are used to define an affine transform to rotate and scale the image to ensure better-aligned visible light and near-infrared image pairs and less irrelevant variance in the training data.

CelebAMask-HQ [10] is a large-scale visible spectrum dataset that contains hand-annotated semantic labels of a high-quality subset of images from the CelebA dataset [40, 66]. The initial CelebaA dataset [66] consists of $202,599$ images of $10,177$ identities, while the smaller high-quality CelebaA-HQ subset [40] contains only $30,000$ images. CelebAMask-HQ [10] extends on this with the addition of pixel-level semantic labels of 19 semantic classes. For the purposes of our experiments, we simplify the dataset to only 10 key semantic face regions, i.e. eyes, nose, lips, eyebrows, ears, neck, hair, face skin as well as glasses and background. To achieve this, we remove images which contain classes that are not represented in images of the Tufts Face dataset [22], i.e. hats, earrings, and teeth. In addition, we merge certain classes such as the lower and upper lip, or left and right eye. Lastly, we resize the images to a resolution of 256×256. The final subset thus contains 9357 annotated images. It is used to evaluate the performance of segmentation models trained on the synthetic data produced by our model.

4.1.2 Implementation and experimental details.

The entirety of the proposed BiFaceGAN framework is implemented in PyTorch [67] and is made publicly available[1]. The main component of the framework, i.e. the Dual-Branch StyleGAN2 model, was constructed upon the SytleGAN2-ADA implementation [31]. The output resolution (256×256) was chosen based on several factors, including the quality of available datasets as well as the training time and system requirements. Nevertheless, the proposed

[1]Implementation available at: https://github.com/dariant/BiFaceGAN

framework can also be easily expanded to produce higher-resolution images by increasing the depth of the synthesis and discriminator networks.

The proposed Dual-Branch StyleGAN2 model utilizes similar training parameters and procedures as the original StyleGAN2 implementation [31]. To train the model, we first initialize its core part with weights pretrained on the FFHQ [28] dataset. This excludes the near-infrared synthesis branch. Training is then performed with the multi-task adversarial learning objectives (defined in Eq. (4)) in batch sizes of 16. To update the model the Adam optimizer [68] is chosen, with a learning rate of 0.0025 as well as $\beta_1 = 0$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$, selected based on previous research [31].

To allow for a fair comparison of models, we train all models up to a limit of 2500 *kimgs* (thousand images) and use the best performing model, in terms of validation Fréchet Inception Distance (FID) [69]. The models are trained until training diverges or up to 2000 *kimgs*. To improve the stability of training on small-scale face datasets (in our case only 2207 images) we rely on the Adaptive Discriminator Augmentation (ADA) technique [31] to artificially increase the number of training samples.

This approach ensures stable training in a unimodal environment, e.g. training on only visible spectrum images. However, we observe that this is not the case for training the bimodal model, as demonstrated in Fig. 8 by the orange learning curve that depicts the generator loss $\mathcal{L}_G$ from Eq. (4). In comparison with the unimodal version (simply denoted as StyleGAN2), the bimodal version (DB-StyleGAN2) experiences clear training divergence. The loss rapidly increases and the quality of images drops significantly. This transpires early during training, so even the best models up to that point do not produce satisfactory image quality. We believe that this issue most likely occurs due to the poor alignment of visible and near-infrared image pairs in the training dataset since we did not encounter similar issues with bimodal ocular image generation, where the datasets were near-perfectly aligned [37].

Bimodal training improvements. To address this, we propose separating the training of the model into two learning phases, which utilize differently weighted visible light (VIS) and near-infrared (NIR) components of the generator loss function. The proposed approach is represented by the green learning curve in Fig. 8. The first phase is aimed at producing high-quality VIS images. For this purpose, we set the NIR-based and VIS-based loss weights to $w_{NIR} = 0.1$ and $w_{VIS} = 1.0$. This stabilizes the training process, while also allowing the model to still learn a low-quality estimation of the NIR images. While the loss does slightly increase, this is linked to the introduction of artifacts and noise in the NIR images. The first phase of training thus results in high-quality VIS images along with low-quality and less detailed NIR images.

Once the desired quality is achieved in the visible spectrum, in terms of FID scores [69] and qualitative image assessment, the model is switched to the next phase. The goal of the second phase is to improve the quality of NIR images while retaining the VIS quality from the previous phase. To achieve this, we use equal learning weights $w_{NIR} = 1.0$ and $w_{VIS} = 1.0$. When the

phase change occurs the generator loss also increases, due to the low quality of current NIR images. However, the loss rapidly drops as various artifacts are removed from the produced NIR images. The proposed two-phase learning regime allows us to avoid the explosion of loss observed in the single-phase approach and facilitates more stable model training. In turn, this enables the synthesis of aligned high-quality visible and near-infrared spectrum image pairs even if the training samples are poorly aligned.

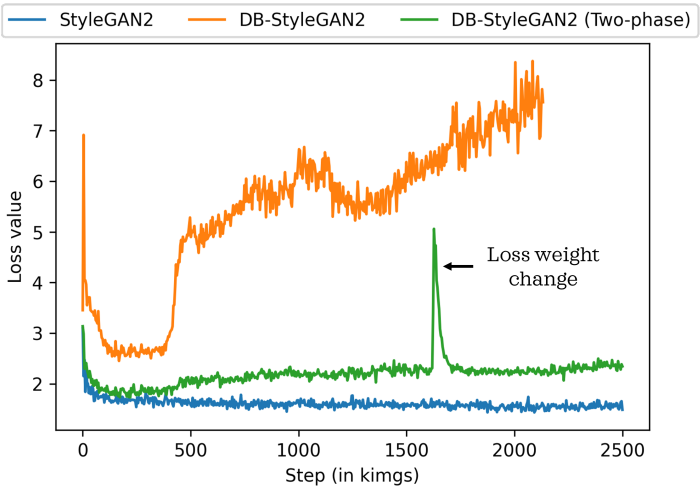

Fig. 8 Generator loss $\mathcal{L}_G$ comparison of different training procedures. The proposed two-phase procedure allows more stable training of the DB-StyleGAN2 model on poorly aligned data.

To train the **Semantic Mask Generator (SMG)** component of the BiFaceGAN framework we utilize the Cross-Entropy learning objective and the Adam optimizer [68], with a learning rate of 10^{-3}. For training, we manually annotate 8 synthetic images with 10 key face regions. To train each classifier of the ensemble we randomly sample image pixels from the data in batches of 64. Training of each classifier is stopped once performance on the synthetic validation set does not improve in 50 batches.

Synthetic masks produced by the SMG component are evaluated in a series of **segmentation experiments**, where we use them to train a state-of-the-art DeepLab-V3 model [70]. To train the segmentation model we again rely on the Cross-Entropy learning objective and the Adam optimizer [68] and train the model with a batch size of 8 and an initial learning rate of 10^{-4}. We reduce the learning rate by a factor of 10 when no validation loss improvement is observed in 5 epochs. Training is then stopped once the learning rate reaches 10^{-8} or if no improvements are observed in 10 consecutive epochs.

Experimental Hardware. The training and evaluation of the models is 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 video RAM.

4.1.3 Evaluation metrics

We evaluate the proposed approach both in terms of the quality of generated images and the accuracy of produced ground truth segmentation masks. For this purpose, we rely on a variety of evaluation metrics, discussed below.

Synthetic image quality. To evaluate the overall quality of synthetic images produced by the proposed model we rely on several metrics which are based on features extracted from deep learning approaches, i.e.:

- **Fréchet Inception Distance (FID)** [69], which estimates the difference between real and synthetic distributions based on image features extracted from the Inception-v3 model [71]. The mean and the covariance matrix of real and synthetic feature vectors are then used to determine the similarity of distributions. This metric has become widespread in image synthesis research [28, 30, 31] because the extracted features mimic the human perception of images well. However, it has been demonstrated, that large sample sizes are required to avoid potential score over-estimation [72]. This can be problematic due to the small scale of our datasets.
- **Learned Perceptual Image Patch Similarity (LPIPS)** [73], an alternative method that uses latent image features from the VGG network [74] pretrained on ImageNet [75] to determine the similarity of image pairs. To compare the entire datasets we randomly sample real and synthetic images and use the mean and standard deviation values of obtained results.
- **Certainty Ratio Face Image Quality Assessment (CR-FIQA)**, a state-of-the-art method that is specifically designed for evaluating face image quality. It utilizes a ResNet-101 backbone [57] to predict the relative classifiability of a face image and in turn, its quality. The model is trained on a classification task on the CASIA-WebFace dataset [76] with the use of the ArcFace loss [38] and a custom Certainty Ratio loss. The latter is based on cosine similarity [60] between the image sample and the corresponding class center as well as the nearest negative class center. Once the model is trained it can be used to estimate the quality of unseen samples.
- ***t*-distributed Stochastic Neighbor Embedding (*t*-SNE)** [77] data dimensionality method, which enables comprehensible visual comparison of the real and synthetic distributions. To allow for comparison, images are first expressed as feature vectors with a ResNet-101 model [57] pretrained on ImageNet [75]. From the distribution of feature vectors, the method then constructs a representative distribution in a lower-dimensional space. To estimate the proximity of distributions the method utilizes the **Kullback-Leibler divergence (KL-divergence)** [78] measure. This measure is then minimized within the *t*-SNE approach to obtain a low-dimensionality embedding that can be visualized in a 2D space.

Segmentation metrics. We conduct our segmentation experiments with the use of several established metrics for evaluating the accuracy of ground truth segmentation masks [2, 79–81]. These metrics being $\boldsymbol{F}_1$ **score**, **Intersection over Union (IoU)** and total **Pixel Error (P.E.)**, where the first two

can both be expressed in terms of Precision and Recall scores. The F_1 score represents the performance as a harmonic mean of the two scores, whereas IoU results are closer to the worst-case performance. In comparison, Pixel Error simply measures the percentage of misclassified image pixels and is as such affected by class imbalance.

4.2 Bimodal synthesis evaluation

In the first set of experiments, we explore the image generation capabilities of our BiFaceGAN framework. To this end, we compare synthetic images produced by our framework with real images. To obtain the synthetic images we train the Dual-Branch StyleGAN2 (DB-StyleGAN2) component on the preprocessed images of the Tufts Face dataset [22] at a resolution of 256×256. In addition, we utilize the proposed ArcFace Privacy Filter (APF) to produce privacy-preserving images. Hereafter, the combination of both components will be denoted as DB-StyleGAN2-APF.

4.2.1 Visual evaluation

In this section, we perform qualitative analysis of the generated face images to showcase the potential of our proposed BiFaceGAN framework. In Fig. 9 we depict both real training samples of the Tufts Face dataset [22] and filtered synthetic samples produced by the DB-StyleGAN2-APF.

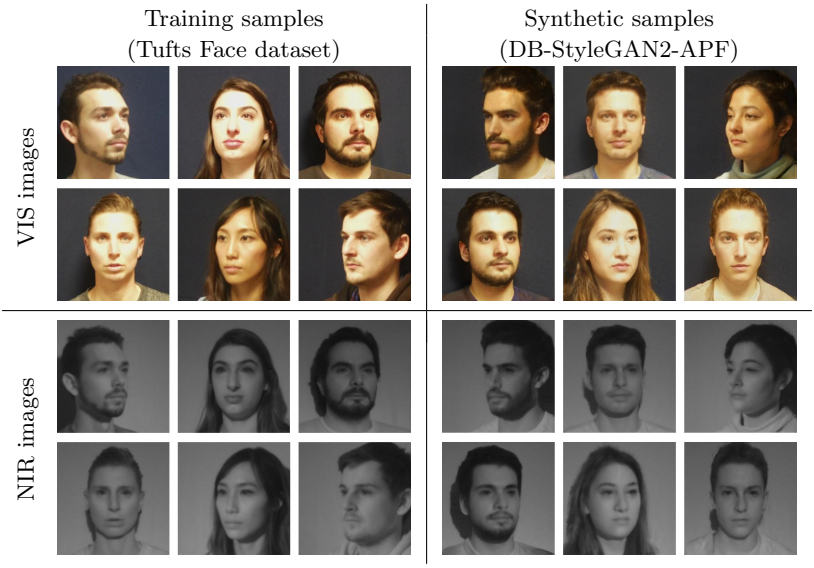

Fig. 9 Real and synthetic image samples in the visible (VIS) and the near-infrared (NIR) spectrum. DB-StyleGAN2-APF is capable of producing privacy-preserving photorealistic images, aligned in both imaging domains.

As can be observed, our DB-StyleGAN2-APF model is able to generate extremely realistic synthetic face image samples, in both the visible (VIS) and the near-infrared (NIR) spectrum. The generated images are of high quality and variety, and the portrayed faces have characteristics that closely resemble the training data. Moreover, the model is capable of accurately reproducing convincing details, such as specular reflections in the eyes, as well as stochastic features like hair placement and texture. The model also generates matching bimodal image pairs which are well-aligned.

Importantly, the high quality of images is achieved despite the small scale of the training dataset. This capability is extremely valuable, due to the current state of multispectral biometric datasets as well as the vital insight and additional cues that multispectral data can provide to various deep learning solutions. In addition, the produced synthetic images preserve the privacy of training subjects, due to the auxiliary ArcFace-based filtering process, which removes synthetic images whose identities match the ones in the DB-StyleGAN2 training set. With this, our model also addresses the growing privacy-related concerns.

4.2.2 VIS-NIR pair alignment

Next, we further investigate the alignment of VIS and NIR face images by analyzing composite images in the YCbCr color space. These are created by replacing the luma channel (Y) of the VIS image in the YCbCr color space with the NIR image. Images obtained with this procedure exhibit changes in the overall color characteristics, as seen in Fig. 10. Importantly, they also clearly illustrate image pair misalignment in the form of color artifacts.

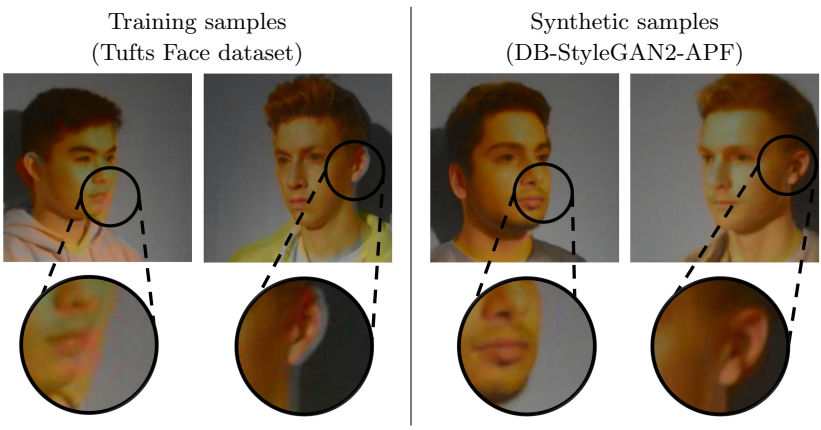

Fig. 10 VIS-NIR alignment in real and synthetic images. Shown are composite images, where the luma channel of the YCbCr representation of the VIS image was replaced by the NIR image. Note that DB-StyleGAN2 generates per-pixel aligned image pairs.

Prominent misalignment artifacts can be observed in the training samples of the Tufts Face dataset [22], despite the intricate preprocessing steps described in Sec. 4.1.1. The artifacts are most clear near the edges of the faces, as seen in the zoomed-in examples in Fig. 10, as well as in discolored clothing. These are mostly caused by the movement of subjects during the capturing process since VIS and NIR images were not gathered simultaneously.

Despite the poor alignment of the training data, our Dual Branch StyleGAN2 model (DB-StyleGAN2) is still able to generate well-aligned VIS-NIR image pairs, which display virtually no alignment artifacts, Even in the worst examples, only faint misalignment artifacts can be detected near the edges of the face, as seen in the rightmost image of Fig. 10. These results reveal that our model is able to overcome a certain amount of training data misalignment and still produce near-per-pixel aligned synthetic image pair samples. This synthesis capability is attributed to the semantic-rich latent representation inside the core DB-StyleGAN2 architecture that is shared by the two shallow synthesis branches, which generate images in the VIS and NIR light spectrum.

4.2.3 Privacy-preserving data filtering

In the following section, we explore similarities between the synthetic and training data, as well as analyze the effects of the ArcFace Privacy Filter (APF) component on the produced data. Specifically, we compare the identities of the DB-StyleGAN2 training samples with identities of the unfiltered and filtered synthetic datasets. For this comparison, we utilize features obtained from the ArcFace model [38] and evaluate their likeness with the cosine similarity score [60]. We compute this similarity between each generated image and all training samples.

In Fig. 11 we present two synthetic samples from the unfiltered datasets with the highest similarity score (middle columns) along with their most similar training samples (left columns). As can be seen, the DB-StyleGAN2 model is able to produce data that is distinct from the training set, despite sharing many visual characteristics. Unfortunately, the identities of the produced images resemble the identities of the DB-StyleGAN2 training data, which is also reflected in the high similarity scores (0.892 and 0.897). Thus, the individuals from the training set could be identified with the use of our synthetic data, despite the overall data being distinct. In turn, this raises privacy concerns regarding the use of the produced synthetic data.

To address this, we propose the use of an auxiliary ArcFace Privacy Filter (APF) component during the data generation process, as described in Sec. 3.5. This component filters the data produced by the DB-StyleGAN2 model based on the computed similarity scores and a privacy threshold hyperparameter τ, which is set to 0.6 for the purposes of these experiments. To evaluate the effect of the proposed privacy filter, we also present in Fig. 11 synthetic samples from the filtered set (right columns) which are most similar to the selected training samples. We can observe that the identities from the resulting filtered dataset differ drastically from the identities of the real samples

used to train the DB-StyleGAN2 model. This can be seen in shape differences of many face components, e.g. differences in the jawline, nose, ears, eyes, and lips. Interestingly, changes can also be observed in the clothing. These observations are also supported by the lower similarity scores (0.598 and 0.589). Most importantly, we note that the presented images remain realistic and convincing, despite their lower similarity score. These observations suggest that we are able to filter the data based on identity similarity without lowering the overall quality of images, at least from a qualitative perspective, and in turn, allow the model to generate high-quality privacy-preserving synthetic data.

Fig. 11 Identity similarity with and without privacy-preserving filtering. Training samples of the Tufts Face dataset [22] (left) are accompanied by the most similar synthetic samples from either the dataset obtained with the proposed ArcFace Privacy Filter (right) or from the unfiltered one (middle). The cosine similarity scores between ArcFace features of training and synthetic images are provided below the images.

4.2.4 State-of-the-art comparison and ablation study

In the following section, we perform an in-depth analysis of the image generation process. We conduct an ablation study of the Dual-Branch (DB) synthesis design (Sec. 3.2) and the ArcFace Privacy Filter (APF) component (Sec. 3.5) and analyze their effects on the overall quality of produced images. To this end, we train two additional single-spectrum StyleGAN2 models, one for each light spectrum. We also utilize synthetic data produced by the models introduced in the previous section, i.e. the bimodal Dual-Branch StyleGAN2 (DB-StyleGAN2) and the bimodal version with the ArcFace Privacy Filter component (denoted as DB-StyleGAN2-APF). In total, each dataset used in the analysis consists of 5000 synthetic samples. Metrics used throughout the experiments are presented in Sec. 4.1.3.

We first evaluate the overall quality and realism of the produced synthetic images with two metrics prominently used in the field of image generation, **Fréchet Inception Distance (FID)** [69] and **Learned Perceptual Image Patch Similarity (LPIPS)** [73]. To compare the performance of different generative model designs, we compute the mentioned metrics between the

synthetic datasets, produced by the models, and the training and validation sets of the Tufts Face dataset [22]. We report the obtained results in Tab. 2. To provide additional context to the values, we also present scores acquired between the training and the holdout sets.

Table 2 Comparison of image quality in terms of FID and LPIPS scores. Evaluation is performed between 5000 synthetic samples produced by each generative model design (StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF) and the training and holdout sets. Lower scores are better.

Data from	LS	FID (T)	FID (H)	LPIPS (T)	LPIPS (H)
StyleGAN2	VIS	14.259	31.342	0.485 ± 0.114	0.484 ± 0.112
StyleGAN2	NIR	28.554	36.189	0.599 ± 0.069	0.598 ± 0.069
DB-StyleGAN2	VIS	**13.452**	**30.492**	$\mathbf{0.483 \pm 0.113}$	0.482 ± 0.111
	NIR	**17.206**	**28.554**	$\mathbf{0.386 \pm 0.098}$	$\mathbf{0.387 \pm 0.099}$
DB-StyleGAN2-APF	VIS	16.322	32.686	**0.483** ± 0.117	$\mathbf{0.479 \pm 0.115}$
	NIR	18.532	29.144	**0.386** ± 0.101	0.397 ± 0.097
(H) vs. (T)	VIS	23.229		0.493 ± 0.113	
	NIR	16.918		0.397 ± 0.097	

(LS) – light spectrum; (T) – training set; (H) – holdout validation set

From the results, we can observe that our Dual-Branch design is able to compete with the image quality of the original single-spectrum StyleGAN2 approach, whilst generating two matching images in the VIS and NIR spectrum at once. Moreover, our DB-StyleGAN2 actually achieves better results in terms of both metrics and both light spectra. This difference is most noticeable in the FID results, especially for the NIR domain, where the original StyleGAN2 achieves a score of 28.554 and the Dual-Branch version a score of 17.206. A considerable difference between the two also exists in terms of the LPIPS metric, even when considering the large standard deviation values. The cause of this drastic performance difference, in terms of NIR data generation, likely lies in the quality difference between VIS and NIR training samples. The training VIS images contain more detail than their NIR counterpart. Thus, we suspect that utilizing both VIS and NIR images provides more valuable semantic information to the bimodal model during training and inference, which is reflected in the generation of higher-quality NIR images. Meanwhile, the original single-spectrum model is trained only on the lower-quality NIR samples, resulting in lower-quality synthetic data. In comparison, when discussing the VIS generation task, only a slight difference is present between the two model designs, both in terms of FID and LPIPS scores. However, the bimodal model still achieves better image quality, likely due to the additional cues present in the NIR spectrum.

Interestingly, we observe that the addition of the ArcFace Privacy Filter (APF), which acts as a privacy-preserving filtering step, only slightly lowers the quality of synthetic images produced by the DB-StyleGAN2 model.

This quality reduction is only observed with the FID metric and even then, the scores on the NIR domain still drastically surpass scores of the original StyleGAN2 model. Meanwhile, LPIPS results demonstrate that the effect of removing images with high identity similarity does not affect the quality of images. When comparing images with the training set, the only difference is noted in the standard deviation of scores. Surprisingly, a slight improvement in LPIPS scores is actually observed with the addition of the APF component on the holdout set and the VIS domain. Overall, the obtained results showcase that our proposed approach is able to produce privacy-preserving data without negatively affecting the quality of the synthetic images in a meaningful way.

We also note that all generative models achieve better FID and LPIPS scores on the training set (columns denoted with (T)) than the holdout set (last row), at least in the VIS domain, though results are comparable also in the NIR domain. This suggests that the synthetic models are capable of generating data that shares more similarities with the training data than the training data does with the holdout set. In comparison, scores between the synthetic and the holdout set (columns denoted with (H)) are lower, as is to be expected. Nevertheless, the finding still demonstrates the potential of the proposed models for generating realistic images.

To obtain a more comprehensive understanding of synthetic image distributions we utilize the ***t*-distributed Stochastic Neighbor Embedding (*t*-SNE)** method [77] to visualize the distributions in a lower-dimensionality space. In Fig. 12 we present *t*-SNE plots for each light spectrum which include 200 randomly sampled images from each of the synthetic datasets and the training set. In addition, we report in Tab. 3 the **Kullback-Leibler divergence (KL-divergence)** values [78] used for the plots.

Table 3 Kullback-Leibler (KL) divergence values of the *t*-SNE plots in Fig. 12. Divergence values are computed between training sample images and the synthetic images produced by the discussed generative models (i.e., StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF). Lower values are better.

Kullback-Leibler divergence from the training set		
Data from	VIS	NIR
StyleGAN2 (unimodal)	1.785	2.988
DB-StyleGAN2 (bimodal)	**1.150**	1.417
DB-StyleGAN2-APF (bimodal)	1.392	**1.331**

From the plots, we can observe that the synthetic distributions of all presented model variations overlap fairly well with the training set in the visible spectrum, at least from a qualitative perspective. Kullback-Leibler divergence values, which are the base for the plots, reveal that both bimodal DB-StyleGAN2 approaches achieve better scores and in turn, better overlap with the training set than the single-spectrum StyleGAN2 model. Similarly to previous observations, the addition of the privacy filter, in the form of the

APF component, reduces the overlap. Despite this, the approach still outperforms the single-spectrum one. Conversely, in the near-infrared spectrum plot, we can observe a clear separation between the single-spectrum distribution and the rest. This qualitative lack of overlap is also further supported by the Kullback-Leibler divergence values. Interestingly, we also note that the addition of the APF component actually reduces the divergence scores, pointing to a possible increase in similarity with the training set.

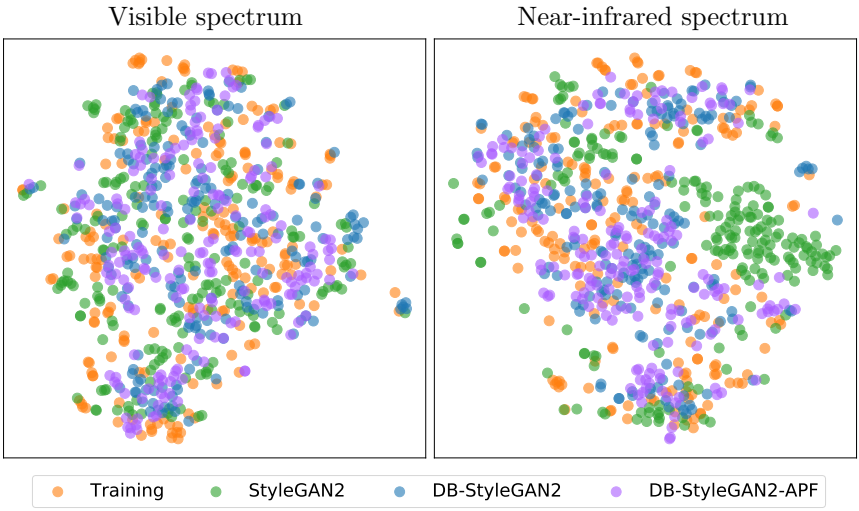

Fig. 12 Comparison of synthetic and training image samples with *t*-SNE plots (in 2D). Plots are generated with 200 randomly sampled images from synthetic sets produced by the discussed generative models (StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF) and the training set.

Last but not least, we also evaluate the quality of produced images with a state-of-the-art biometric approach, known as **CR-FIQA** [62], which is specifically designed for assessing the quality of face images. We compute the CR-FIQA score for each image in the VIS domain and report the score distributions of each synthetic and real dataset in Fig. 13 along with their mean and standard deviation values.

Similarly to previous results, the CR-FIQA distributions of all three generative approaches in the VIS domain share the same distribution shape and display only slight differences overall. In comparison, the training and validation distributions are skewed more to the right, have a drastically lower peak, and have notably higher standard deviation values. The highest mean value is observed with the training distribution. Meanwhile, the validation set actually includes more outliers that form the tail on the left, which results in a lower mean value. When comparing the original StyleGAN2 model with its bimodal Dual-Branch alternative, we note that the distribution of the DB-StyleGAN2 approach has a slightly lower peak and is slightly more skewed to the left. Despite these differences, the mean value remains near identical,

while the standard deviation interestingly slightly drops. The filtering process of component APF skews the distribution slightly more to the left, which results in a lower mean value but also a lower standard deviation. Nevertheless, the privacy filter does not have a drastic influence on the overall distribution of quality scores, thus showcasing its suitability and potential for future privacy-preserving synthesis approaches.

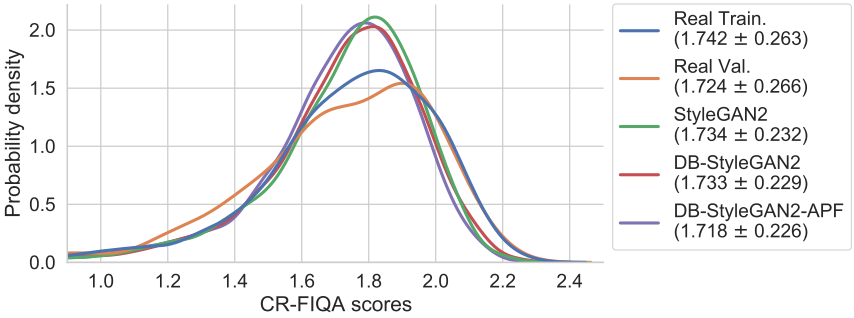

Fig. 13 Image quality comparison in terms of CR-FIQA score distributions. Distributions of the Tufts Face dataset [22] (training and holdout set) are reported alongside distributions of 5000 synthetic samples produced by the discussed generative models (StyleGAN2, DB-StyleGAN2, and DB-StyleGAN2-APF). The mean and standard deviation values are also reported in the legend. Higher scores are better.

4.3 Segmentation evaluation

The second set of experiments is tied to the evaluation of the semi-supervised label generation process, enabled by the Semantic Mask Generator (SMG) component of the proposed BiFaceGAN framework. Specifically, we investigate how the generation of bimodal data as well as privacy-preserving filtering affects the quality of corresponding masks. We also explore how the performance of current deep segmentation models can be improved with the use of bimodal data. Lastly, we analyze the choice of pixel classifiers used in the SMG component of BiFaceGAN.

Throughout the following segmentation experiments, we utilize 8 synthetic face images annotated with 10 semantic face regions, i.e. eyes, nose, lips, eyebrows, ears, neck, hair, face skin as well as glasses, and background. These are labeled based on the combined information provided from the VIS and NIR imaging domains, due to the different cues that they provide. The annotated synthetic samples are then used to train the SMG component that is tied to a different image generation model depending on the experiment. The BiFaceGAN framework, which entails these components, is then used to generate synthetic samples with corresponding ground truth segmentation masks, which are then used to train a state-of-the-art DeepLab-V3 [70] segmentation model. Here 5000 synthetic samples are used for training and 500 for validation. The trained segmentation model is then tested on the visible spectrum

CelebAMask-HQ dataset [10] since the bimodal Tufts Face dataset [22] does not contain any ground truth annotations. The models are evaluated in terms of Intersection Over Union (IoU), F_1 score, and overall Pixel Error (P.E.). The performance of the trained model is then used as a proxy to determine the quality of synthetic data used for training.

4.3.1 State-of-the-art segmentation comparison

In the following section, we compare the mask generation process with its initial implementation, the state-of-the-art dataset generation framework, known as DatasetGAN [33]. However, to allow for a fair comparison, we update DatasetGAN to also utilize the single-spectrum StyleGAN2 model (and not StyleGAN1). This ensures that the main difference between the approaches is the bimodal synthesis ability, which in turn influences the latent representation inside the DB-StyleGAN2 model that is used to create masks of the synthetic data. It should also be noted, that the DeepLab-V3 segmentation model [70], used for evaluation, is trained only using the VIS spectrum data and corresponding masks, due to the spectral limitations of the visible light evaluation dataset CelebAMask-HQ [10].

Table 4 Cross-dataset segmentation performance. DeepLab-V3 is trained on synthetic Tufts Face data produced by different StyleGAN2 models and evaluated on the CelebAMask-HQ dataset [10]. Synthetic data of our BiFaceGAN framework enables better segmentation performance than the data of the state-of-the-art DatasetGAN framework.

DeepLab-V3 trained on synthetic data, evaluated on CelebAMask-HQ					
Framework	**Images from**	**Classifier**	**IoU ↑**	**F_1↑**	**P.E. ↓**
DatasetGAN	StyleGAN2	eMLP	0.573 ± 0.094	0.679 ± 0.094	0.166 ± 0.085
BiFaceGAN	DB-StyleGAN2	eMLP	$\mathbf{0.587 \pm 0.098}$	$\mathbf{0.689 \pm 0.096}$	0.158 ± 0.094
BiFaceGAN	DB-StyleGAN2-APF	eMLP	0.584 ± 0.098	0.686 ± 0.096	0.156 ± 0.091
BiFaceGAN	DB-StyleGAN2-APF	eDL	0.578 ± 0.098	0.681 ± 0.096	$\mathbf{0.154 \pm 0.091}$

(eMLP) – ensemble of Multi-Layer Perceptrons; (eDL) – ensemble of DeepLab-V3 models
(↑) – Higher is better; (↓) – Lower is better

Segmentation results on the evaluation dataset with differently trained DeepLab-V3 segmentation models [70] are presented in Tab. 4. From the results of the first two rows, we can discern that the DeepLab-V3 model trained on the data produced by our BiFaceGAN framework achieves better segmentation performance in terms of all metrics. This is also supported by score distribution plots in Fig. 14. Here, we can see that the BiFaceGAN distributions of F_1 and IoU scores are skewed more to the right and in turn also have a lower peak than the DatasetGAN distributions. Meanwhile, the pixel error distribution is skewed more to the left, indicating lower overall error scores. Sample segmentation results in Fig. 15 reveal that the segmentation model trained on the synthetic data of our BiFaceGAN is able to better classify the hair and neck semantic classes. The produced masks also have more rounded edges that better fit the different face features. In addition, regions with shadows are better classified, e.g. regions under the eyes or chin.

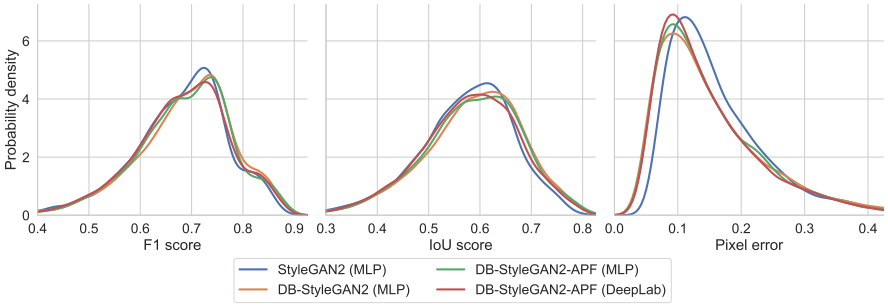

Fig. 14 Cross-dataset segmentation performance comparison. Distributions of segmentation performance scores corresponding to results in Tab. 4. Results were obtained with the DeepLab-V3 model trained on the synthetic data produced by different StyleGAN2-based models and evaluated on the CelebAMask-HQ dataset [10].

Fig. 15 Qualitative segmentation comparison of DeepLab-V3 trained with either data produced by DatasetGAN or BiFaceGAN. The models are evaluated on samples of the CelebAMask-HQ dataset [10]. Results correspond to the first two rows of Tab. 4.

Overall these results suggest that the bimodal nature of the DB-StlyeGAN2 model leads to a more semantic-rich latent representation within the generative model, based on which the SMG component is able to generate higher-quality training segmentation samples. This in turn enables the training of better performing biometric segmentation models, which can distinguish between different semantic regions of the face.

4.3.2 Training with privacy-preserving data

Next, we investigate how the addition of privacy-preserving filtering with the ArcFace Privacy Filter (APF) influences the quality of masks generated by the SMG component. To this end, we create an additional synthetic dataset with DB-StyleGAN2, however, this time we also utilize the APF component to filter out privacy-breaching images. The obtained data is then again used to train the DeepLab-V3 segmentation model as was done in the previous section.

From segmentation performance scores presented in Tab. 4 we can observe that filtering the training synthetic data hardly influences the accuracy of the final segmentation model. Even with a slight difference, the model still outperforms the model trained on data of the DatasetGAN framework. This

is also reflected in the distribution plots of Fig. 14. Here the distributions (in green) closely follow the distributions of the unfiltered approach. The main difference is a slight drop in the distribution peak that is redistributed to the left side in the F_1 and IoU plots. Interestingly, the filtering process actually positively influences the pixel error score, in terms of both mean and standard deviation values as well as the overall distribution shape. This could be caused by slight improvements in the background semantic class since this metric is most influenced by the majority class.

The obtained results suggest that privacy-preserving synthetic data can be utilized instead of the initial privacy-breaching synthetic data to train deep biometric solutions, such as segmentation models, without any notable difference or detrimental effects on the performance of the models.

4.3.3 Choice of pixel classifiers

As part of our work we also experiment with replacing the ensemble MLP classifier of the Semantic Mask Generator (SMG) component with the decoder network of the DeepLab-V3 model [70]. To evaluate this approach we used the privacy-preserving BiFaceGAN framework and generated an additional synthetic dataset with the updated SMG component. Then an auxiliary DeepLab-V3 segmentation model was trained on the produced data and compared to the previously obtained models in Tab. 4 and Fig. 14.

As can be seen, the proposed change negatively impacts the segmentation accuracy in terms of IoU and F_1 scores, which both display lower mean values and distributions skewed more to the left. Interestingly, the change does however lower the overall pixel error, as can be seen by the lower mean value and a higher distribution peak. This is again possibly caused by the improved segmentation of the background class. Unfortunately, this, in turn, likely reduces the segmentation accuracy of other smaller but more important classes, at least based on the negative effect on the first two metrics. The obtained results thus further support the choice of the initial pixel classifier, i.e. the ensemble MLP proposed by Zhang *et al.* [33].

4.3.4 Segmentation with VIS and NIR data

To showcase the potential of multispectral approaches, we also investigate the effect of utilizing both the visible (VIS) and the near-infrared (NIR) spectrum data for the purposes of segmentation. To this end, we train two DeepLab-V3 segmentation models on the privacy-preserving synthetic dataset created by the BiFaceGAN framework with the DB-StyleGAN2-APF method. Here, one segmentation model is trained only on the synthetic VIS images and the corresponding masks, whilst the other is trained on both the synthetic VIS and NIR images, along with the ground truth segmentation masks. Notably, the latter DeepLab-V3 model was also adapted to accept input images with a channel size of 4. Unfortunately, due to the lack of suitable annotated multispectral datasets, we must rely solely on qualitative analysis to evaluate the difference between the approaches. Thus, we use the two trained segmentation models

on the validation set of the Tufts Face dataset [22] and display the obtained segmentation results in Fig. 16.

As can be observed, the use of both VIS and NIR data allows the segmentation model to better distinguish between certain semantic regions, such as the neck region with facial hair and the normal hair region, e.g. second and last image. Furthermore, it also enables better segmentation of the glasses region, as can be seen in the first image, as well as the segmentation of hair or lack thereof, e.g. third and fifth image. This is likely due to the additional semantic information that is present within the NIR images that enables the model to make better segmentation decisions. Overall, the results showcase the potential of utilizing multispectral data to improve biometric segmentation solutions.

Fig. 16 Segmentation improvements when utilizing bimodal (VIS & NIR) instead of unimodal (VIS) data. Two DeepLab-V3 models are trained on privacy-preserving synthetic data of BiFaceGAN, either on the VIS data or on the entire VIS & NIR data pair. Performance is then evaluated on the validation part of the Tufts Face dataset [22].

4.4 Real-world training and inference time comparison

Lastly, we analyze the training and inference times of the different BiFaceGAN components and compare them to the current state-of-the-art. In Tab. 5 we report both the time required to reach convergence of models during training as well as the time required to produce a single sample, averaged over 1000 samples. Results are obtained on the hardware described in Sec. 4.1.2.

From the first two rows of Tab. 5, we can discern that our bimodal Dual-Branch StyleGAN2 design takes longer to train than the original unimodal StyleGAN2. However, it should be noted, that the bimodal version is trained on two times the number of images (VIS and NIR) and in a longer two-phase regime, which ensures training stability. Thus, an increase in overall training time is expected. Interestingly, during inference, our model is able to compete with the speed of the unimodal approach, as it requires only around 22% longer (14.898ms instead of 12.114ms on average) to generate two images in the two different imaging domains. Meanwhile, generating two images with the unimodal model would necessitate two forward passes with two separate spectrum-specific models.

Table 5 Training and inference time of each component. Our DB-StyleGAN2 model produces two images whilst matching the speed of the unimodal StyleGAN2 model. Privacy-preserving filtering slows the generation process, however, the production of corresponding ground truth segmentation masks still takes the longest.

Components	**Training time**†	**Inference time [ms]**
StyleGAN2	$\sim$ 50 hours	12.114 ± 1.541
DB-StyleGAN2	$\sim$ 87 hours	14.898 ± 2.092
ArcFace Privacy Filter	/	80.976 ± 3.996
SMG (MLP)	$\sim$ 12 minutes	982.755 ± 12.010
SMG (DeepLab-V3)	$\sim$ 1 minute	5227.822 ± 67.051

†Approximate estimate

When also utilizing the proposed ArcFace Privacy Filter (APF) component to ensure privacy-preserving data synthesis, we can observe that this additional filtering step slows the data generation process by 80.976ms on average. In comparison with the speed of the above-discussed image synthesis models, this takes significantly longer. Nevertheless, the perks of producing privacy-preserving data far outweigh the increase in inference times, as the data generation process is still relatively fast. It should also be noted that the speed of the proposed privacy-preserving filtering step is highly dependent on the size and structure of the real-world dataset.

In our proposed BiFaceGAN framework, the Semantic Mask Generator (SMG) component takes the longest during inference, by a large margin. The synthesis of privacy-preserving images takes less than 100ms, while the production of corresponding ground truth segmentation masks with the use of an ensemble MLP classifier takes around 1000ms. The alternative approach, which utilizes the decoder of the DeepLab-V3 model, performs even worse in terms of speed, while not surpassing the quality of masks produced by the ensemble MLP. It does, however, take less time to train. Overall, the reported results reveal, that improvements to the SMG component would benefit the speed of the data generation process the most.

5 Conclusion

In this chapter, we investigated the generation of synthetic multispectral face data to address the data requirements of various biometric deep learning solutions and the increasing privacy concerns connected to biometric data. To this end, we presented a novel generative framework, called BiFaceGAN, which is capable of producing high-quality privacy-preserving synthetic face images in the visible and the near-infrared spectrum along with corresponding ground truth pixel-level annotations. To produce visually convincing and near-per-pixel aligned bimodal images, the framework relies on a Dual-Branch StyleGAN2 model, which uses a custom training regime to combat training instability, caused by poorly aligned real-world datasets. During inference,

the model utilizes an additional filtering step, implemented with the ArcFace Privacy Filter (APF) component, which ensures privacy-preserving image synthesis whilst retaining image quality. These two components are accompanied by an auxiliary Semantic Mask Generator (SMG) that exploits latent features of DB-StyleGAN2 to produce accurate fine-grained ground truth segmentation masks. Through a series of experiments, we showcased that our framework is capable of competing with current unimodal synthesis approaches in terms of image quality while producing privacy-preserving images in two different light domains at once. Furthermore, we demonstrated the utility of the produced synthetic data, by training both visible spectrum and multispectral-based segmentation models that could generalize well to real-world data. As part of our future work, we plan to investigate the generation of specific synthetic identities found in the latent space to enable training of recognition approaches in a privacy-preserving manner. We also plan to explore domain transferring possibilities between the visible and the near-infrared spectrum to allow for the generation of new multispectral datasets.

Acknowledgments. This research was supported in parts by the Slovenian Research Agency (ARRS) through the ARRS Research Programmes P2-0214 "Computer Vision" and P2-0250 "Metrology and Biometric Systems" as well as the ARRS Project J2-2501 "DeepBeauty" and the ARRS Junior Researcher Program.

References

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

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

[3] Batagelj, B., Peer, P., Štruc, V., Dobrišek, S.: How to correctly detect face-masks for COVID-19 from visual information? MDPI Applied Sciences **11**(5), 2070 (2021)

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

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

[6] Meden, B., Rot, P., Terhörst, P., Damer, N., Kuijper, A., Scheirer, W.J., Ross, A., Peer, P., Štruc, V.: Privacy-enhancing face biometrics: A comprehensive survey. IEEE Transactions on Information Forensics and Security (TIFS) **16**, 4147–4183 (2021)

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

[8] Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The MegaFace benchmark: 1 million faces for recognition at scale. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4873–4882 (2016)

[9] Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In: ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 547–558 (2020)

[10] Lee, C.-H., Liu, Z., Wu, L., Luo, P.: MaskGAN: Towards diverse and interactive facial image manipulation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558 (2020)

[11] Vitek, M., Rot, P., Štruc, V., Peer, P.: A comprehensive investigation into sclera biometrics: A novel dataset and performance study. Neural Computing & Applications **32**(24), 17941–17955 (2020)

[12] Bourlai, T., Cukic, B.: Multi-spectral face recognition: Identification of people in difficult environments. In: IEEE International Conference on Intelligence and Security Informatics, pp. 196–201 (2012)

[13] Bourlai, T.: Face Recognition Across the Imaging Spectrum. Springer, Cham (2016)

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

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

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

[17] Bourlai, T., Hornak, L.A.: Face recognition outside the visible spectrum. Image and Vision Computing **55**, 14–17 (2016)

[18] Bourlai, T., Kalka, N., Cao, D., Decann, B., Jafri, Z., Nicolo, F., Whitelam, C., Zuo, J., Adjeroh, D., Cukic, B., et al.: Ascertaining human identity in night environments. Distributed Video Sensor Networks, 451–467 (2011)

[19] Chambino, L.L., Silva, J.S., Bernardino, A.: Multispectral facial recognition: a review. IEEE Access **8**, 207871–207883 (2020)

[20] Bourlai, T., Whitelam, C., Kakadiaris, I.: Pupil detection under lighting and pose variations in the visible and active infrared bands. In: IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2011)

[21] 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.*: Cross-eyed 2017: Cross-spectral iris/periocular recognition competition. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 725–732 (2017)

[22] Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S.P., Kaszowska, A., Taylor, H.A., Samani, A., *et al.*: A comprehensive database for benchmarking imaging systems. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **42**(3), 509–520 (2018)

[23] Peri, N., Gleason, J., Castillo, C.D., Bourlai, T., Patel, V.M., Chellappa, R.: A synthesis-based approach for thermal-to-visible face verification. In: IEEE International Conference on Automatic Face and Gesture Recognition (F&G), pp. 01–08 (2021)

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

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

[26] Boutros, F., Damer, N., Raja, K., Ramachandra, R., Kirchbuchner, F., Kuijper, A.: Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. Image and Vision Computing **104**, 104007 (2020)

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

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

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

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

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

[32] Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8300–8311 (2021)

[33] Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.-F., Barriuso, A., Torralba, A., Fidler, S.: DatasetGAN: Efficient labeled data factory with minimal human effort. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10145–10155 (2021)

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

[35] Luo, Y., Pi, D., Pan, Y., Xie, L., Yu, W., Liu, Y.: Clawgan: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light. Expert Systems with Applications **191**, 116269 (2022)

[36] Mokalla, S.R., Bourlai, T.: Robust LWIR-based eye center detection through thermal to visible image synthesis. In: IEEE International Conference on Automatic Face and Gesture Recognition (F&G), pp. 1–8 (2021)

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

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

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

(2017)

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

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

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

[43] Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning (ICML), pp. 214–223 (2017)

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

[45] Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

[46] Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: Visualizing and understanding generative adversarial networks. In: International Conference on Learning Representations (ICLR), pp. 1–4 (2019)

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

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

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

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

[51] Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)

[52] Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: IEEE International Conference on Computer Vision (ICCV), pp. 5688–5696 (2017)

[53] Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.-Y., Yang, M.H.: Adversarial learning for semi-supervised semantic segmentation. In: British Machine Vision Conference (BMVC), pp. 1–17 (2018)

[54] Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high-and low-level consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **43**(4), 1369–1379 (2019)

[55] Pakhomov, D., Hira, S., Wagle, N., Green, K.E., Navab, N.: Segmentation in style: Unsupervised semantic image segmentation with StyleGAN and CLIP. arXiv preprint arXiv:2107.12518 (2021)

[56] Maas, A.L., Hannun, A.Y., Ng, A.Y., *et al.*: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning

(ICML), pp. 1–3 (2013)
[57] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
[58] Duta, I.C., Liu, L., Zhu, F., Shao, L.: Improved residual networks for image and video recognition. In: IEEE International Conference on Pattern Recognition (ICPR), pp. 9415–9422 (2021)
[59] Deng, J., Guo, J., Zhang, D., Deng, Y., Lu, X., Shi, S.: Lightweight face recognition challenge. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 0–0 (2019)
[60] Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: Asian Conference on Computer Vision, pp. 709–720 (2010). Springer
[61] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: CosFace: Large margin cosine loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)
[62] Boutros, F., Fang, M., Klemt, M., Fu, B., Damer, N.: CR-FIQA: face image quality assessment by learning sample relative classifiability. arXiv preprint arXiv:2112.06592 (2021)
[63] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(4), 834–848 (2017)
[64] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 886–893 (2005)
[65] Suthaharan, S.: Support vector machine. In: Machine Learning Models and Algorithms for Big Data Classification, pp. 207–235. Springer, New York (2016)
[66] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3730–3738 (2015)
[67] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., *et al.*: PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 8026–8037 (2019)
[68] Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR), pp. 1–5 (2015)
[69] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: 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 (2017)
[70] Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arxiv:1706.05587 (2017)
[71] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)
[72] Borji, A.: Pros and cons of GAN evaluation measures: New developments. Computer Vision and Image Understanding (CVIU) **215**, 103329 (2022)
[73] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018)

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

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

[76] Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)

[77] Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research (JMLR) **9**(86), 2579–2605 (2008)

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

[79] Lozej, J., Meden, B., Štruc, V., Peer, P.: End-to-end iris segmentation using U-Net. In: IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 1–6 (2018)

[80] Rot, P., Emeršič, Ž., Štruc, V., Peer, P.: Deep multi-class eye segmentation for ocular biometrics. In: IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 1–8 (2018)

[81] Vitek, M., Das, A., Pourcenoux, Y., Missler, A., Paumier, C., Das, S., Ghosh, I.D., Lucio, D.R., Jr., L.A.Z., Menotti, D., Boutros, F., Damer, N., Grebe, J.H., Kuijper, A., Hu, J., He, Y., Wang, C., Liu, H., Wang, Y., Sun, Z., Osorio-Roig, D., Rathgeb, C., Busch, C., Tapia, J., Valenzuela, A., Zampoukis, G., Tsochatzidis, L., Pratikakis, I., Nathan, S., Suganya, R., Mehta, V., Dhall, A., Raja, K., Gupta, G., Khiarak, J.N., Akbari-Shahper, M., Jaryani, F., Asgari-Chenaghlu, M., Vyas, R., Dakshit, S., Dakshit, S., Peer, P., Pal, U., Štruc, V.: SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment. In: International Joint Conference on Biometrics (IJCB), pp. 1–10 (2020)