MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network

Naser Damer[⋆†], Alexandra Moseguí Saladié[⋆], Andreas Braun[⋆†], Arjan Kuijper[⋆†]
[⋆]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[†]Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany
Email: naser.damer@igd.fraunhofer.de

Abstract

Face morphing attacks aim at creating face images that are verifiable to be the face of multiple identities, which can lead to building faulty identity links in operations like border crossing. Research has been focused on creating more accurate attack detection approaches by considering different image properties. However, all the attacks considered so far are based on manipulating facial landmarks localized in the morphed face images. In contrast, this work presents novel face morphing attacks based on image generated by generative adversarial networks. We present the MorGAN structure that considers the representation loss to successfully create realistic morphing attacks. Based on that, we present a novel face morphing attacks database (MorGAN database) that contains 1000 morph images for both, the proposed MorGAN and landmark-based attacks. We present vulnerability analysis of two face recognition approaches facing the proposed attacks. Moreover, the detectability of the proposed MorGAN attacks is studied, in the scenarios where this type of attacks is know and unknown. We concluded with pointing out the challenge of detecting such unknown novel attacks and an analysis of detection performances of different features in detecting such attacks.

1. Introduction

The recent deep-learning driven performance gains in face recognition [46], along with the relatively high social acceptance [5], have pushed automatic face recognition to be a key technology in security sensitive deployments such as identity management (e.g. travel documents) [34].

The critical nature of face biometric applications makes it a target of malicious attacks. One of such attacks is the presentation attack, which is attacking a biometric system by presenting a biometric characteristic to the biometric capture subsystem with the goal of interfering with the operation of the biometric system [22]. These attacks can aim at impersonating a different identity by presenting a copy of their characteristics [9, 23] (spoofing), disabling face detection or recognition using physical objects [10], or moving the face in an intention to avoid identification [8]. From a new perspective, Ferrara et al. [14] presented a new form of attack that aims at presenting one face reference image that is, automatically and by human experts, successfully matched to more than one person, i.e. face morphing attack. If such morphing attacks are used in travel or identity documents, it would allow multiple subjects to verify their identity to the one associated with the document. This faulty subject link to the document identity can lead to a wide range of illegal activities, including financial transactions, illegal immigration, human trafficking, and circumventing legal black identity lists. These morphing attacks have been constructed so far by interpolating facial landmarks between the two face images creating the attack, whether by manual means (e.g. using GNU Image Manipulation Program v2.8 GIMP) [40], or automatic triangulation or averaging [39]. These attacks were accompanied with detection algorithms based on classifying handcrafted features [38], image quality measures [35], and CNN-based representations [40]. However, all the considered attacks were based on located facial landmarks.

Rather than focusing on enhancing the attack detection of the landmark-based morphing attacks, this work foresees future attacks and proposes creating morphing attacks based on automatically generated images by generative adversarial networks. To do that, we specifically designed and trained the MorGAN architecture and successfully created morphing attack images of realistic quality. Based on that, we present a novel database of attacks that contains 1000 attack images for both the proposed MorGAN attacks and the landmark based attacks, with 1500 bona fide references and 1500 bona fide probes. We demonstrate the vulnerability of face recognition algorithms to the proposed attacks by evaluating two deep learning-based face recognition solutions. We also emphasize the challenge in detecting our novel at-

978-1-5386-7180-1/18/$31.00 ©2018 IEEE

tacks, especially if there were not previously known to the attack detector. Therefore, we analyze the detection ability across attack types with two different attack detection approaches based on handcrafted and CNN-based features.

2. Related work

The possibility of creating a morphed face image, out of two images of two subjects, was introduced by Ferrara et al. [14]. They compared morphed images with images of the original subjects using two face recognition solutions, and concluded with the high vulnerability of face recognition to such attacks. This has driven the work on similar attacks on different modalities such as fingerprints [13] and irises [41]. Further studies considered the human expert vulnerability to morphed face images when comparing faces [15, 43]. These works concluded that human experts, to a large extend, can fail in detecting morphed face images.

Different works have been proposed to detect such attacks. Ramachandra et al. [38] were first to propose the automatic detection of morphed face images, by extracting local image descriptors as the Binarised Statistical Image Features (BSIF) that tries to capture textural properties of the image, that are later classified using a Support Vector Machine (SVM). Later works looked into using CNN-based features [40], image quality measures [35], the effect of printing and re-scanning the images [45], and differences between triangulating and averaging the facial landmarks on the detection [39]. A standardized manner to evaluate the vulnerability of biometric systems to morphing attacks was recently proposed by Scherhag et al. [44]. A recent work by Ferrara et al. [16] viewed the morphing attack problem from a different perspective by proposing an approach to revert the morphed face image (demorph) enough to reveal the identity of the legitimate document owner, given a bona fide capture. All the discussed previous works developed and evaluated their approaches based on morphing attacks databases that were created based on facial landmarks.

Image generation has, in the recent years, opened new possibilities due to the latest advancements in deep learning techniques combined with generative models [37]. There are different deep learning-based approaches of generative models, however, the two most dominant are Variational Autoencoders (VAE) [28] and Generative Adversarial Networks (GANs) [19]. VAE-based techniques learn the parameters of a probability distribution representing the data by projecting into a learned latent space and reconstruct samples from that space. VAEs have stable training dynamics, but tend to produce very blurry outputs discarding high-frequency details [17]. Additionally, VAE uses a pixel-wise reconstruction loss function that is not completely aligned with the goal of a generative model. By contrast, GAN-based approaches are explicitly trained to produce the most plausible and realistic images combining two networks playing in an adversarial game. Even if GANs have unstable and often oscillatory training dynamics, they generate sharper and higher-quality samples than the best VAE techniques [49].

Recently, several face image generation approaches have been suggested based on GANs. Age-cGAN [3] aimed to change the age of the face image by implementing a conditional GAN and an encoder in the same pipeline. Their main contribution is a latent vector optimization step that allows to preserve the identity of the person by comparing the feature vector of the original and the reconstructed image. Moreover, Karras et al. [24] introduced a new GAN architecture for generating high-resolution face images starting from lower resolution ones and progressively grow the generator and discriminator architecture. However, this approach requires high computational power and time. Within the scope of attacking face recognition systems, Sharif et al. proposed to automatically add eyeglasses to face images, so that it could succeed in targeted (impersonation) or untargeted (dodging) attacks [47]. They create an attack network called adversarial generative nets (AGNs) that crates attacks on two pre-trained DNN-based face recognition networks VGG [36] and OpenFace [2]. Although they show realistic glasses and prove to successfully attack the mentioned recognition systems, the attack need to be planned for each recognition network individually, which is not always possible in a real-case scenario.

3. MorGAN

The general idea of GANs is to establish a game between two players: generator and discriminator. The first player, the generator *G*, produces samples from a distribution, ideally indistinguishable from the training distribution (produce an image $\tilde{x}$ from a noise vector z). The discriminator *D* is trained to do the assessment of the samples and determine whether they are coming from the training set (real) or from the generator (fake). The discriminator learns in a supervised manner to divide the inputs into two different classes, real or fake, while the generator is trained to fool the discriminator. Therefore, the generative model is set against an adversary, a discriminative model. The discriminator architecture is similar to a standard convolutional neural network (CNN), while the generator consists of consist of deconvolutional layers. Once the GAN model is trained, one can discard the discriminator and use the generator to generate images from the approximated data distribution p_g.

In order to train a model to reconstruct an input image without losing the identity we adapted the vanilla GAN formulation [19]. The original generative model is setup to produce images by sampling from a set of vectors of the latent space, which describes the image appearance. Therefore, there is no way of figuring out which initial noise vector will produce specific features in the image. Consequently, we need to reverse the mapping of the generator and project images back into the latent space. However, since mapping from image space *X* into latent space *Z* is non-trivial and it requires the inversion of the generator *G*.

MorGAN is inspired by the work of Dumoulin *et al.* [12] by expanding the work of Donahue *et al.* [11] by a stochastic encoder. In [11], a GAN in which a third network is added to explicitly learn the reverse mapping was added. They propose to include an encoder *E*, which takes an image sample $x \in X$ and maps it into a vector $z \in Z$, such that when *z* is passed through the generator it produces an image $\tilde{x}$ close to the original image *x*. Here, the encoder *E* is trained at the same time as the discriminator *D* and the generator *G*. Because of the simultaneous training, the discriminator *D* needs to discriminate not only in data space, synthetic samples from real data, but jointly in data and latent space, distinguishing between two joint distributions. Several methods have been proposed to achieve joint distribution matching, including [26, 6, 50, 18]. Nevertheless, they all target a different objective, the unsupervised training of unpaired data. Our approach will split the generator *G* into two different networks, encoder G_z (the new added model), and decoder G_x, referred previously as the generator, with two probability distributions over *x* and *z*:

- Encoder joint distribution: $q(x, z) = q(x)q(z|x)$.
- Decoder joint distribution: $p(x, z) = p(z)p(x|z)$.

Analyzing both probabilities, one can observe that the marginals $q(x)$ and $p(z)$ are terms mentioned in the vanilla GAN description, $q(x)$ corresponds to the data distribution, defined before as p_{data}, and $p(z)$ is a known distribution p_z, such as the standard Normal distribution. Therefore, one can easily sample from both of them. The training procedure is performed as described by Gan *et al.* [18]. After a sample *x* is sampled from $q(x)$, the encoder generates a sample $\hat{z}$ from the conditional distribution $q(z|x)$. Hence, the data pair $(x, \hat{z})$ is a sample from the encoder joint distribution. On the other hand, a data pair $(\tilde{x}, z)$ can be sampled from the decoder by first sampling z from $p(z)$ and then drawing $\tilde{x}$ from $p(x|z)$.

MorGAN main objective is to match both joint distributions playing an adversarial game. In particular, the method should learn the bidirectional mapping given by the conditionals distributions. In other words, we want to match the conditional distribution $q(z|x)$, defined by the encoder G_z, with the conditional distribution in the opposite direction $p(x|z)$, defined by the decoder G_x. If this is accomplished, we can guarantee that the tuples $(x, G_z(x))$ and $(G_x(z), z)$ are indistinguishable from each other. The MorGAN training objective is defined as a minimax game

$$\min_{G_x, G_z} \max_{D} v(\theta_{G_x}, \theta_{G_z}, \theta_D), \tag{1}$$

where

$$\begin{aligned} v(\theta_{G_x}, \theta_{G_z}, \theta_D) = \mathbb{E}_{x \sim q(x)} & \underbrace{[\mathbb{E}_{\hat{z} \sim q(z|x)} \, log(D(x, z))]}_{log D(x, G_z(x))} + \\ & + \mathbb{E}_{z \sim p(z)} \underbrace{[\mathbb{E}_{\tilde{x} \sim p(x|z)} \, log(1 - (D(x, z)))]}_{log(1 - D(G_x(z), z))} . \end{aligned} \tag{2}$$

In this adversarial game, the model learns only by marginals samples, it is not required to compute conditional densities, but only that gradient back-propagation is achievable along all the model. This property eases the task of training this complex configuration and reduces the differences between vanilla GAN and our model. To summarize the tow introduced improvements, first, the generator has two components, the encoder $G_z(x)$ and the decoder $G_x(z)$, complementary inverse to each other. Second, the discriminator is trained to distinguish between joint pairs: samples from the encoder $(x, \hat{z})$ and samples from the decoder $(\tilde{x}, z)$. Such as in previous works, the generator minimizes the Jensen-Shannon divergence [30] between joint distributions under the assumption of optimal discriminator.

The MorGAN framework does not use simple autoencoders but variational autoencoder, to avoid the possibilities of non-continuous latent space leading to unrealistic output because of interpolation between encoded vectors. Variational Autoencoders (VAEs), on the other hand, learn the parameters of a probability distribution representing the data. Their latent spaces are forced to be continuous and this allows an easy random sampling and interpolation. Thus, instead of mapping the input to a fixed vector, the input will be mapped into a distribution. The original encoded vector *z* is split into two vectors, one representing the mean μ of the distribution and the other one representing the standard deviation σ. Now, the vector *z* needed to reconstruct the input image will be sampled from the learned distribution.

The introduction of this sampling raises a new problem, the model require to compute back-propagation through all the network and it is not possible to push gradients to a sampling node. Therefore, re-parametrization trick [28, 4, 42] is used. Instead of sampling directly from the desired distribution, the latent vector *z* is expressed as:

$$z = \mu(x) + \sigma(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1), \tag{3}$$

where μ and σ are learned parameters, and ϵ is the stochastic part from which the latent vector is sampled from.

From the image morphing perspective, interpolating two images in the pixel space gives an overlapping of two images without blending into a single object. On the contrary, doing the interpolation in the latent space produces a smooth transition between images. Therefore, we linearly interpolate two faces in the latent space to generate the morphing face. Based on that, the face morphing is going to be performed in three steps given two face images x_1 and x_2:

1. Encode x_1 and x_2 into the latent space:

$$\hat{z}_1 = G_z(x_1), \quad \hat{z}_2 = G_z(x_2). \tag{4}$$

2. Linearly interpolate z_1 and z_2 with a factor $\beta = 0.5$:

$$\hat{z} = (\beta - 1)\hat{z}_1 + \beta\hat{z}_2. \tag{5}$$

3. Decode the morphed latent vector $\hat{z}$ into the image space:

$$\tilde{x} = G_x(\hat{z}). \tag{6}$$

Equation 5 is a linear interpolation, where α determines how much information the morphed image is taking from each original image x_1 and x_2. In our case, both images should give the same amount of information in order to preserve both identities, therefore the alpha equals 0.5.

A major challenge now is that the original identity is often lost in the generated image. Antipov *et al.* showed in their study [3], that even though the latent vector $\hat{z}$ produced by the encoder results in visually plausible face reconstructions, the identity of the original person is lost in about 50% of cases. Therefore, a modification is proposed for MorGAN. Some studies already contemplated the possibility of identity preservation using GANs [3, 31, 20]. The common suggestion among all of them was to adapt the generator's loss. Our proposal is to combine prior knowledge from data distribution (adversarial loss) with specific knowledge of the subject's face (reconstruction loss). The adversarial loss is a standard cross-entropy cost that evaluates the performance of the generator and the discriminator in an adversarial game. In the MorGAN framework, these losses are defined as follows:

$$\begin{aligned}\mathcal{L}_{GAN-D} = &-\frac{1}{M}\sum_{i=1}^{M}\log(D(x^{(i)}, \hat{z}^{(i)})) \\ &-\frac{1}{M}\sum_{j=1}^{M}\log(1 - D(\tilde{x}^{(j)}, z^{(j)})),\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{GAN-G} = &-\frac{1}{M}\sum_{i=1}^{M}\log(1 - D(x^{(i)}, \hat{z}^{(i)})) \\ &-\frac{1}{M}\sum_{j=1}^{M}\log(D(\tilde{x}^{(j)}, z^{(j)})),\end{aligned} \tag{8}$$

where M is the number of processed samples at each batch. Both losses plays a crucial role in the adversarial game, consequently, they need to be preserved. However, we combine the defined adversarial loss with a reconstruction loss (pixel-wise loss). This synthesis will take place in the

$$\mathcal{L}_{syn} = \mathcal{L}_{GAN-G} + \alpha\mathcal{L}_{pixel}, \tag{9}$$

where α is a weighting parameter to define the importance of $\mathcal{L}_{recon}$ in the generator's loss (here, α = 0.3).

Figure 1 presents the MorGAN structure. There are two simultaneous inputs in the generator, the first one being a sample from the training dataset x that enters to the generator G_x and outputs a latent vector $\hat{z}$. Similarly, a latent vector z is sampled from the Normal distribution and entered to the generator G_z which outputs a generated image $\tilde{x}$. Our modifications are in the pixel loss $\mathcal{L}_{pixel}$ computation (marked in orange in Figure 1). There is no connection between the two generators, therefore, the discriminator will never obtain the direct reconstruction of an image sample. However, with the introduction of the new pixel loss $\mathcal{L}_{pixel}$ (referred before as reconstruction loss), the output from the generator G_x, $\hat{z} = G_x(x)$, is sent to the generator G_z to obtain the reconstructed image defined as:

$$\tilde{x}_{recon} = G_z(G_x(x)). \tag{10}$$

Consequently, the pixel loss that we propose will compare the original image x with the reconstructed image $\tilde{x}_{recon}$ by calculating the pixel-wise L1 loss between both samples:

$$\mathcal{L}_{pixel} = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\left|x^{(i,j)} - \tilde{x}_{recon}^{(i,j)}\right|, \tag{11}$$

where W And H are the width and height of the images.

Due to this modification, the generators parameters θ_{G} are optimized by minimizing this synthesis loss $\mathcal{L}_{syn}$. This adaptation preserves the most prominent facial structure and improve the preservation of the identity. Algorithm 1 introduces the MorGAN in details.

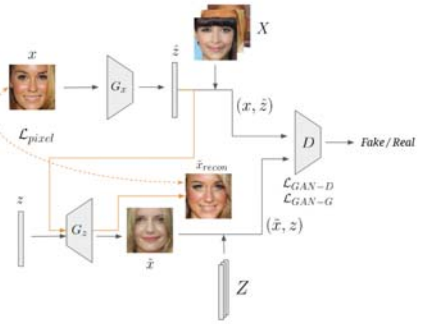

Figure 1: The MorGAN structure.

MorGAN implementation is based on the Theano [48] implementation of ALI's model [12], which also trains simultaneously a decoder and encoder in an adversarial game. This network, at the same time, is based on Deep Convolutional GAN (DCGAN) [37]. During the training, the generator G_x samples a noise vector z of size 512 from a Normal Distribution $z \sim \mathcal{N}(0, 1)$ that is upsampled with several transpose convolutions, as suggested in [37], to finally generate an image $\tilde{x}$ of size 64x64x3 (width, height and channels respectively). Meanwhile, the generator G_z goes into the opposite direction. It samples an image x from the data distribution, with size 64x64x3, and generates a vector $\hat{z}$ of size 512 by applying a sequence of convolutions. Finally, the discriminator D takes the results from both generators and apply convolutions in order to obtain the vectors $D(x)$ and $D(z)$. Both vectors are concatenated along the channel axis to generate the prediction of the joint discriminator.

Following the suggestions of Radford *et al.*, MorGAN is trained using the Adam optimizer [27] with value $\beta_1 = 0.5$, instead of Stochastic Gradient Descent (SGD) such as proposed in [19]. The networks are trained with a learning rate of 10^{-5} and a batch size of 65 (samples drawn independently at each update step) during 123 epochs. The networks weights are initialized using an Isotropic Gaussian with $\mu = 0$ and $\sigma = 0.01$ and the biases are initialized to 0.

Algorithm 1 MorGAN algorithmic training procedure

```
$\theta_{\mathrm{G}}, \theta_{\mathrm{D}} \leftarrow$ random initialize network parameters
repeat
    $x^{(1)}, \ldots, x^{(M)} \sim q(x)$    ▷ Sample M images from the data distribution
    $z^{(1)}, \ldots, z^{(M)} \sim p(z)$    ▷ Sample M vectors from a Normal distribution
    $\hat{z}^{(i)} \leftarrow G_z(x^{(i)}), i = 1, \ldots, M$    ▷ Compute $G_z$ generation
    $\tilde{x}^{(j)} \leftarrow G_x(z^{(j)}), j = 1, \ldots, M$    ▷ Compute $G_x$ generation
    $\tilde{x}^{(i)}_{recon} \leftarrow G_x(G_z(x^{(i)})), i = 1, \ldots, M$    ▷ Compute reconstructed $x$
    $\rho_q^{(i)} \leftarrow D(x^{(i)}, \hat{z}^{(i)}), i = 1, \ldots, M$    ▷ Compute discriminator predictions
    $\rho_p^{(j)} \leftarrow D(\tilde{x}^{(j)}, z^{(j)}), j = 1, \ldots, M$
    $\mathcal{L}_{GAN-D} \leftarrow -\frac{1}{M}\sum_{i=1}^{M} \log(\rho_q^{(i)}) - \frac{1}{M}\sum_{j=1}^{M} \log(1 - \rho_p^{(j)}))$
                                                            ▷ Compute discriminator loss
    $\mathcal{L}_{syn} \leftarrow \mathcal{L}_{GAN-G} + \alpha\mathcal{L}_{pixel}$    ▷ Compute generator loss
    $\mathcal{L}_{GAN-G} \leftarrow -\frac{1}{M}\sum_{i=1}^{M} \log(1 - \rho_q^{(i)}) - \frac{1}{M}\sum_{j=1}^{M} \log(\rho_p^{(j)}))$
    $\mathcal{L}_{pixel} = \frac{1}{W \times H}\sum_{k=1}^{W}\sum_{l=1}^{H} \left| x^{(k,l)} - \tilde{x}^{(k,l)}_{recon} \right|$
    $\theta_{\mathrm{D}} \leftarrow \theta_{\mathrm{D}}\text{-} \nabla_{\theta D} \mathcal{L}_{GAN-D}$    ▷ Gradient update on discriminator network
    $\theta_{\mathrm{G}} \leftarrow \theta_{\mathrm{G}}\text{-} \nabla_{\theta G} \mathcal{L}_{syn}$    ▷ Gradient update on generator network
until convergence
```

4. Database

To evaluate the face recognition vulnerability and attack detectability of attacks created by the presented GAN approach, a database of attacks and bona fide images is created. The attacks are created using the MorGAN approach and the conventional LMA approach (used in previous works) to benchmark our findings.

Basic database, filtering, and pre-processing: To enable the selection of similar faces from a wide range of identities, we selected a large database, the CelebA [32]. CelebA is composed of 202,599 face images of 10,177 identities and 40 attribute binary vectors. The images in this dataset cover large pose variations and background clutter. The size of the images is 178 x 218 pixels. The images more relevant to the attack scenario must be the ones that would at least cover the frontal image condition in the International Civil Aviation Organization (ICAO) travel document face image standard [21]. Based on that, we filter out all non-frontal images by initially detecting the central coordinate of the eyes and the upper coordinate of the nose. The two distances between each of the two eyes and the nose landmarks are calculated, and if the difference between these distances was more than 0.05 of each other, the image is neglected. This filtering lowers the database size to 103,480 images. Further filtering was performed based on the provided attributes, images labeled as blurry, with glasses, or with hat, was also filtered out. Resulting in a final database of 89,341 images of 8,673 identities. Face area in these images were cropped after an alignment step based on the detected outer eyes and nose landmarks.

Database structure and pairs selection: Building the database started by manually choosing 500 images of 500 identities, split evenly between males and females. These 500 images were chosen to have neutral impression, good illumination quality, and no occlusion. The 500 initial images/identities are noted here as key bona fide reference images, where R-K-ID$_i$ is the i_{th} image of this group. For each of these images, the two most similar identities in the rest of the filtered database was chosen to be paired with it. The identities appearing in one pair are not considered in any other pair. The similarity was measured by the Euclidean distance between the OpenFace representations [2] of the face images as explained in Section 5. These pairs are used to create attack images that can be verified as both identities. We chose similar identities to pair because in a real attack scenario, the attacker would chose a similar face image for morphing. The selected (by similarity)images are noted as secondary reference images (R-S-ID$_i$) and total to 1000 images of 1000 identities. This resulted in 1500 reference bona fide images that build 1000 morphing attacks pairs. For each of the 1500 reference identities, a second image was chosen to be a probe image (P-ID$_i$). Each of the 1000 morphing attacks pairs is used to create an attack.

Creating the morphing attacks: Morphing of the *Landmark based attacks (LMA)* is performed by detecting 68 landmarks on the face as proposed in [25]. The mean face points for each image are calculated and each image is then warped to sit on these coordinates after performing perform the Delaunay Triangulation [29]. Only the facial area is morphed and stitched into one of the original morphed images [33]. *MorGAN attacks* are created for the same image pairs using the approach presented in Section 3. Section 5 present additional information on experimental details of the MorGAN implementation.

Resulting database: The 1000 pairs between the 500 key references and the 1000 secondary references (each key with two secondary images) were used to create the morphing attacks, once using the conventional LMA approach, and once using our MoGAN approach. The database resulted in a total of 1500 bona fide references(500 key and 1000 secondary), 1500 bona fide probes, 1000 LMA morphing attacks, and 1000 MorGAN morphing attacks. The database is split into a disjoint (identity and image) and equal train and test sets. These sets are based on a random split of the initial 500 key reference images and therefore, each includes 750 bona fide references, 750 bona fide probes, and 500 attack images from each of both attack types. Because of computational and structural limitations, the MorGAN attack images are of 64x64 pixels size (below the ICAO recommendations). To enable a fair comparison, the bona fide images and the LMA attacks are downsized to the same resolution. To avoid biasing the attack detection to image codec format, all images were stored with the same codec. Examples of the resulting image attacks and the original images creating these attacks are presented in Figure 2, where the visual identity preserving of both identities are clear in the MorGAN attacks.

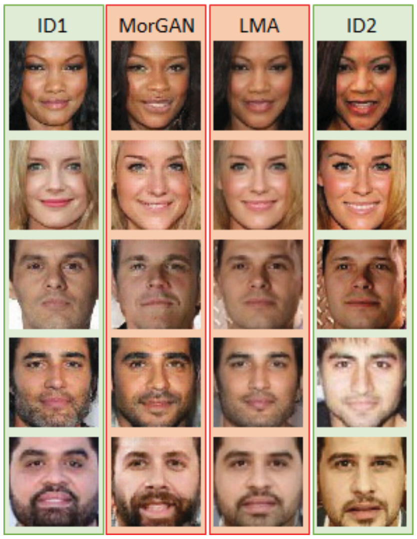

Figure 2: Examples of the morphing attacks created by our MorGAN approach and the LMA approach. Original reference images are on the right and left.

5. Experimental setup

MorGAN implementation details: MorGAN consists of three neural networks trained simultaneously: generator *G* (encoder G_z and decoder G_x) and discriminator *D*. The networks are built on multiple convolution, deconvolution and ReLU layers; and use batch normalization for easing the min-max game. The generator G_x is trained to produce 64x64x3 images given a 512-dimensional random vector. The generator G_z is trained in the opposite direction. The training is performed on the filtered, alligned, and cropped, CelebA database introduced in Section 4. MorGAN framework is run on a NVIDIA GeForce GTX 1050 Ti GPU. Taking a batch size of 65, while epoch takes 1 hour and 20 minutes to be completed. The networks are trained during 123 epochs,summing up to a training time is of around 7 days. On the other hand, once MorGAN is trained, the generation of a morph sample takes around $2 \sim 3$ seconds.

Vulnerability of face recognition: To investigate the strength of the MorGAN face morphing attacks, we demonstrate the vulnerability of face recognition algorithms to these attacks, along with the baseline LMA attacks. This vulnerability of two pre-trained face recognition systems are tested, the OpenFace as described in [2] and the VGG-face as described in [36]. OpenFace [2] is a general-purpose face recognition library based on the deep neural network (DNN) FaceNet architecture suggested in [46]. Given an aligned and cropped face image, this pre-trained network produces a highly discriminant representation (feature vector) of 128 elements. On the other hand, VGG-Face [36] is based on the VGG-Very-Deep-16 CNN. The feature representation is extracted from the last max pooling layer which gives an output of 7x7x512. Face representations, whether from VGG or OpenFace, are compared by calculating the Euclidean distance between two representation vectors.

This vulnerability is discussed by showing the comparison score distributions of imposter, genuine, and morphed face attack comparison to each of the probe original identities contained in the morph. To measure the attacks ability to simultaneously match both original identities, we plot the comparison scores between the morphing attacks and their two original identities images in the probe set. These comparisons scores are shown with respect to the threshold at the equal error rate (EER) operational point to get a relative measure of the attack success. The comparisons are made of images from reference and probe sets. The genuine and imposter comparisons are results of the 1500 bona fide references cross-compared (NxN) with the 1500 probes. The morphing attacks score distribution is based on comparing the 1000 morphed images, each with their corresponding two identities in the probe set. Detailed information on the data are presented in Section 4. The face recognition vulnerability experiments will be noted by the attack type (LMA or MorGAN) and the face recognition approach (OpenFace or VGG), this results in four experiments (OpenFace-LMA, OpenFace-MorGAN, VGG-LMA, VGG-MorGAN).

Detectability of MorGAN attacks: Experiments are performed to measure the performance of morphing attack detection of both types of attacks, LMA and MorGAN, assuming the attacks are known (detector is trained on the same type of attacks). Moreover, to simulate a realistic scenario where systems are only trained on known attacks (e.g. LMA attacks as the current state-of-the-art situation) and faced by unknown attacks, we perform evaluation of the systems detecting unknown attacks. To enable a wider range of conceptual evaluation and more diverse coverage to the state-of-the-art attack detection, we considered image feature extraction methods of two different natures. One is the hand crafted classical image descriptors, the Local Binary Pattern Histogram (LBPH) [1]. The second is based on transferable deep-CNN features. Both types of features were previously utilized for the detection of face morphing attacks of similar nature to LMA [38][40]. The LBPH features are extracted from the cropped face image. A histogram is calculated for each block of an 8x8 grid of blocks in the face image. These histograms are concatenated to produce the final feature vector describing the image. Each LBP is extracted within a radius of one pixel and eight neighbor pixels. The transferable deep-CNN features are

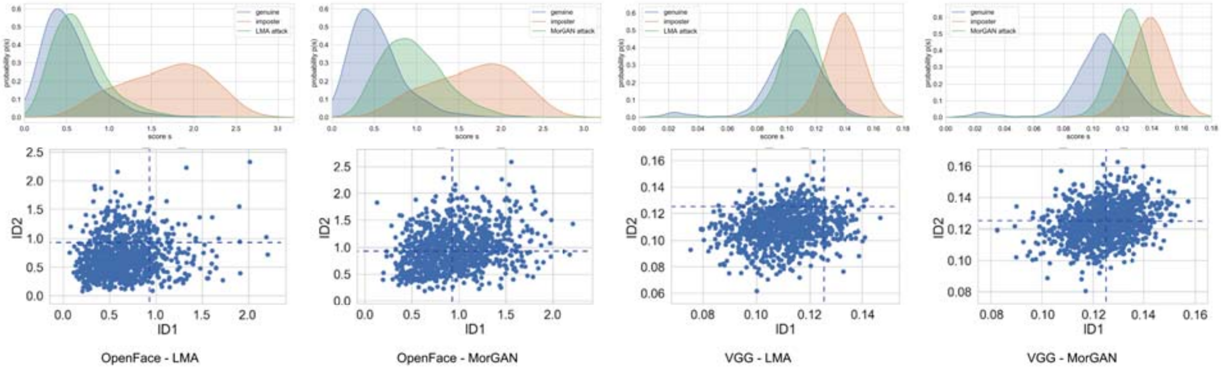

Figure 3: Vulnerability of two face recognition approaches (OpenFace & VGG) to attacks of our LMA and MorGAN databases. Top: the comparison score (dissimilarity) distributions of genuine (blue), imposter (red), and attack (green) comparisons. Bottom: morphing attacks comparison scores in comparison to the dotted line representing the threshold at EER.

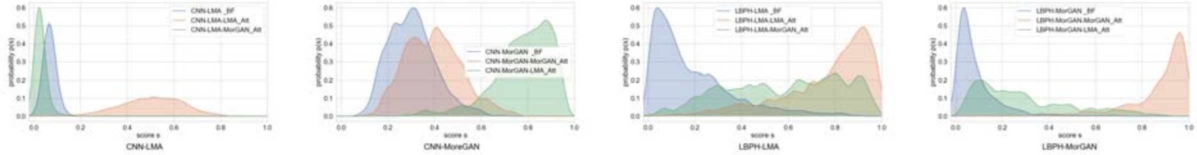

Figure 4: Detectability of our two morphing attacks databases using CNN-based and LBPH features, for the two cases where the detector is trained on LMA or MorGAN attacks. Bona fide in blue, known attacks in red, and unknown attacks in green.

extracted using the well performing, and relatively small OpenFace NN4.SMALL2 model [2]. The extracted feature vector from an image, whether from CNN or LBPH, is classified by a support vector machine (SVM) classifier, to be originated from a morphed or a bona fide image. The SVM utilized a Linear kernel. The SVM hyperparameters are found using Bayesian optimization. The SVM classifier produces a decision score that represent the confidence degree of the input image being a morphed one rather than a bona fide one. The training and evaluation were performed on the training and testing sets subsequently, as defined in Section 4. The attack detection experiments are noted by the feature type used (LBPH or CNN), by the type of attack used for training the detector (LMA or MorGAN), and by the images used for evaluating the detector (LMA, MorGAN, bona fide (BF)), e.g. a detection experiment noted as LBPH-LMA-MorGAN is for a detector based on LBPH feature and trained on LMA trying to detect MorGAN attacks. When the same type of attack is used for training and testing, we will refer to the attack as ”known attack”. Otherwise, we will refer to it as ”unknown attack”.

The performance of the morphed face detection is presented as a trade-off between two error rates, the Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER) as defined by the ISO/IEC 30107-3 [22] and advised by recent works [44]. Here, the APCER is the proportion of morphed face presentations incorrectly classified as bona fide presentations. The BPCER is the proportion of bona fide presentations incorrectly classified as morphed face attacks. The detection decision thresholds producing fixed APCER rates on known attacks are calculated. These threshold represent possible decision thresholds chosen for system deployment and they depend on the detection performance of known attacks. BPCER values achieved at fixed APCER rates are reported to enable direct comparison between different solutions at a wide range of operation points, these BPCER rates only depend on the bona fide images, and thus are the same for known and unknown attacks. APCER rates of unknown attacks are reported on the thresholds previously assigned for the fixed known attacks APCER rates, to measure the detectability of unknown attacks. Lower values of BPCER and APCER indicates higher detection performance.

6. Results

Vulnerability of face recognition: Figure 3 presents information about the vulnerability of the VGG-Face and OpenFace face recognition solutions to both LMA and MorGAN attacks. In the top of the figure 3 one can see the comparison scores distributions produced by the genuine, imposter, and attack comparisons. As expected, the distributions of genuine and imposter comparisons are quite separated in both face recognition solutions, which indicates correct verification decision. When it comes to attacks, a successful attack imitates a genuine scenario, which means that it should produce comparison scores distribution similar to the genuine one. Knowing that, and under the different experiment scenarios, it is noticeable that the LMA attacks produces scores that fall almost completely in the genuine distribution range, which indicates strong attacks. On the other hand, MorGAN attacks produce scores that are relatively less similar to the genuine one, however, still relatively separate from the imposter. This means that the MorGAN attacks are weaker than the LMA ones, however, still make successful attacks on both face recognition solutions by having considerable portion of the scores laying in the genuine distribution range. Knowing that both attacks can be successful, although with different degrees of success, we have to make sure that this works to attack both identities involved in the attack with the same degree. Without this property, that attack is not relevant in the real scenario. To demonstrate this property, we plot the comparison score between the attacks and the first involved identity vs. the one with the second identity in the bottom of Figure 3. The dotted lines in these plots represent the threshold value that achieves equal error rate, where false acceptance rate equals false rejection rate. This helps putting the plotted scores in perspective knowing that lower scores are stronger attacks. Ideally, if the morphing is performing as expected, most of the comparison scores will occur in the diagonal line, meaning that they produce similar similarity to both of the fused identities. This is the case in both LMA and MorGAN. From the attack strength (success) perspective, the same conclusions can be made as from the one made from the plots in the top of Figure 3.

APCER	**0.1%**	**1%**	**5%**	**10%**	**20%**	**30%**
CNN-LMA-LMA	0%	0%	0%	0%	0%	0%
CNN-MorGAN-MorGAN	98.7%	91.3%	82.3%	68%	54.1%	39.5%
LBPH-LMA-LMA	52.3%	36.1%	14.1%	8%	3.2%	1.6%
LBPH-MorGAN-MorGAN	2.5%	1.2%	0.4%	0%	0%	0%

Table 1: BPCER rates achieved at fixed APCER rates of known attacks.

APCER	**0.1%**	**1%**	**5%**	**10%**	**20%**	**30%**
CNN-LMA-MorGAN	100%	100%	100%	100%	100%	100%
CNN-MorGAN-LMA	0%	0%	0%	0%	0.2%	0.2%
LBPH-LMA-MorGAN	2.2%	7%	23.4%	36.4%	51.4%	62.8%
LBPH-MorGAN-LMA	68.6%	74.8%	88.2%	100%	100%	100%

Table 2: APCER values of unknown attacks achieved at the detection decision thresholds that produced certain fixed APCER of known attacks.

Detectability of MorGAN attacks: Table 1 presents the detectability of both, LMA and MorGAN attacks, given that they are known, i.e. the detector is trained on the same type of attacks. CNN-based features performed very good in detecting LMA attacks. However, it performed very poorly in detecting MorGAN attacks. On the contrary, LBPH features performed good in detecting MorGAN attacks and less so for LMA attack. The failure of CNN-based features in detecting MorGAN attacks might be due to the fact that deep learning structure that is trained to extract the features acts in a similar manner to the deep learning structure optimized to generate the MorGAN images. To measure the generalization ability of these detectors on unknown attacks, Table 2 presents the APCER values produced by unknown attacks, given that the system was configured (decision threshold) on fixed APCER values of the known attacks. Here, it is very noticeable that the MorGAN attacks were very hard to detect by systems trained on the LMA attacks. This was the case for both CNN-based and LBPH features, although CNN-based features performs worse again with MorGAN attacks. The same scenario can be seen when an LBPH-based solution is trained on MorGAN and faced by LMA attacks. On the contrary, detecting LMA attacks by a MorGAN-trained CNN-based detector performed quite well. Figure 4 presents a deeper look into the discussed results by visualizing the detection score distribution for the different detection experiment settings. One can notice that in the case of LMA-trained solution based on CNN features, MorGAN attacks appeared completely in the bona fide range, which points out a great vulnerability in such morphing detection systems. Generally, CNN-based features performs badly in detecting MorGAN attacks, whether known or unknown. Morphing detection solution in practiced have to be designed in a way that anticipates new methods of creating the morphing attacks. Such novel attacks are yet to be studied with attack detection approaches that consider pairing with a live probe [7].

7. Conclusion

Previous works dealing with face morphing detection were limited to attacks created by manipulation based on localized facial landmarks. On the contrary, this work successfully presented a novel face morphing attack approach, based on automatic image generation using a specially designed GAN. We presented a large novel database containing our proposed attacks and landmark-based attacks. Face recognition vulnerability of two recognition approach to the novel attacks was also analyzed, along with a reference vulnerability analysis based on landmark-based attacks. We also studied the detectability of the novel attacks, pointing out the high vulnerability of the evaluated morphing detection approaches to such attacks, especially when they are previously unknown.

Acknowledgment

This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within CRISP.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In T. Pajdla and J. Matas, editors, *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I*, volume 3021 of *Lecture Notes in Computer Science*, pages 469–481. Springer, 2004.

[2] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[3] G. Antipov, M. Baccouche, and J. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*, pages 2089–2093. IEEE, 2017.

[4] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning*, pages 226–234, 2014.

[5] R. Bolle and S. Pankanti. *Biometrics, Personal Identification in Networked Society: Personal Identification in Networked Society*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[6] C. Chu, A. Zhmoginov, and M. Sandler. Cyclegan: a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017.

[7] N. Damer, V. Boller, Y. Wainakh, F. Boutros, P. Terhörst, A. Braun, and A. Kuijper. Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts. In *Pattern Recognition - 40th German Conference, GCPR 2018, Stuttgart, Germany, October 10-12, 2018, Proceedings*, Lecture Notes in Computer Science. Springer, 2018.

[8] N. Damer, V. Boller, Y. Wainakh, S. von den Berken, P. Terhörst, A. Braun, and A. Kuijper. CrazyFaces: Unassisted circumvention of watchlist face identification. In *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Los Angeles, California, USA, October 22-25, 2018*. IEEE, 2018.

[9] N. Damer and K. Dimitrov. Practical view on face presentation attack detection. In R. C. Wilson, E. R. Hancock, and W. A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[10] T. I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa. Disguise detection and face recognition in visible and thermal spectrums. In J. Fiérrez, A. Kumar, M. Vatsa, R. N. J. Veldhuis, and J. Ortega-Garcia, editors, *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, pages 1–8. IEEE, 2013.

[11] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[12] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[13] M. Ferrara, R. Cappelli, and D. Maltoni. On the feasibility of creating double-identity fingerprints. *IEEE Trans. Information Forensics and Security*, 12(4):892–900, 2017.

[14] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics, Clearwater, IJCB 2014, FL, USA, September 29 - October 2, 2014*, pages 1–7. IEEE, 2014.

[15] M. Ferrara, A. Franco, and D. Maltoni. On the effects of image alterations on face recognition accuracy. In T. Bourlai, editor, *Face Recognition Across the Imaging Spectrum*, pages 195–222. Springer, 2016.

[16] M. Ferrara, A. Franco, and D. Maltoni. Face demorphing. *IEEE Trans. Information Forensics and Security*, 13(4):1008–1017, 2018.

[17] D. Frossard. VGG in TensorFlow, url: http://www.cs.toronto.edu/ frossard/post/vgg16/ (visited on 08/10/2018).

[18] Z. Gan, L. Chen, W. Wang, Y. Pu, Y. Zhang, H. Liu, C. Li, and L. Carin. Triangle generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 5247–5256, 2017.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[20] R. Huang, S. Zhang, T. Li, R. He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017.

[21] International Civil Aviation Organisation (ICAO). ICAO Draft Technical Report: Portrait quality (reference facial images for MRTD). technical report, Version 0.9, 2017.

[22] International Organization for Standardization. ISO/IEC DIS 30107-3:2016: Information Technology Biometric presentation attack detection Part 3: Testing and reporting. Standard, 2017.

[23] O. Kähm and N. Damer. 2d face liveness detection: An overview. In A. Brömme and C. Busch, editors, *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group, Darmstadt, Germany, September 6-7, 2012*, volume 197 of *LNI*, pages 1–12. IEEE/GI, 2012.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[25] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[26] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[29] D.-T. Lee and B. J. Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980.

[30] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[31] X. Liu, B. V. Kumar, Y. Ge, C. Yang, J. You, and P. Jia. Normalized face image generation with perceptron generative adversarial networks. In *Identity, Security, and Behavior Analysis (ISBA), 2018 IEEE 4th International Conference on*, pages 1–8. IEEE, 2018.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[33] S. Mallick. Face morph using opencv c++ / python. https://www.learnopencv.com/face-morph-using-opencv-cpp-python/, 2016.

[34] Markets and Markets. Facial Recognition Market by Component (Software Tools and Services), Technology, Use Case (Emotion Recognition, Attendance Tracking and Monitoring, Access Control, Law Enforcement), End-User, and Region - Global Forecast to 2022. Report, November 2017.

[35] T. Neubert. Face morphing detection: An approach based on image degradation analysis. In C. Kraetzer, Y. Shi, J. Dittmann, and H. J. Kim, editors, *Digital Forensics and Watermarking - 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings*, volume 10431 of *Lecture Notes in Computer Science*, pages 93–106. Springer, 2017.

[36] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 41.1–41.12. BMVA Press, 2015.

[37] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[38] R. Ramachandra, K. B. Raja, and C. Busch. Detecting morphed face images. In *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*, pages 1–7. IEEE, 2016.

[39] R. Ramachandra, K. B. Raja, S. Venkatesh, and C. Busch. Face morphing versus face averaging: Vulnerability and detection. In *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*, pages 555–563. IEEE, 2017.

[40] R. Ramachandra, K. B. Raja, S. Venkatesh, and C. Busch. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Honolulu, HI, USA, July 21-26, 2017*, pages 1822–1830. IEEE Computer Society, 2017.

[41] C. Rathgeb and C. Busch. On the feasibility of creating morphed iris-codes. In *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*, pages 152–157. IEEE, 2017.

[42] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[43] D. J. Robertson, R. S. S. Kramer, and A. M. Burton. Fraudulent id using face morphs: Experiments on human and automatic recognition. *PLOS ONE*, 12(3):1–12, 03 2017.

[44] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. J. Veldhuis, L. J. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel, R. Breithaupt, R. Ramachandra, and C. Busch. Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb, and A. Uhl, editors, *International Conference of the Biometrics Special Interest Group, BIOSIG 2017, Darmstadt, Germany, September 20-22, 2017*, volume P-270 of *LNI*, pages 149–159. GI / IEEE, 2017.

[45] U. Scherhag, R. Ramachandra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *5th International Workshop on Biometrics and Forensics, IWBF 2017, Coventry, United Kingdom, April 4-5, 2017*, pages 1–6. IEEE, 2017.

[46] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

[47] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *CoRR*, abs/1801.00349, 2018.

[48] T. T. D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, et al. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016.

[49] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. *CoRR*, abs/1511.01844, 2015.

[50] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint*, 2017.