Template Inversion Attack against Face Recognition Systems using 3D Face Reconstruction

Hatef Otroshi Shahreza[1,2] and Sébastien Marcel[1,3]
[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3]Université de Lausanne (UNIL), Lausanne, Switzerland
{hatef.otroshi,sebastien.marcel}@idiap.ch

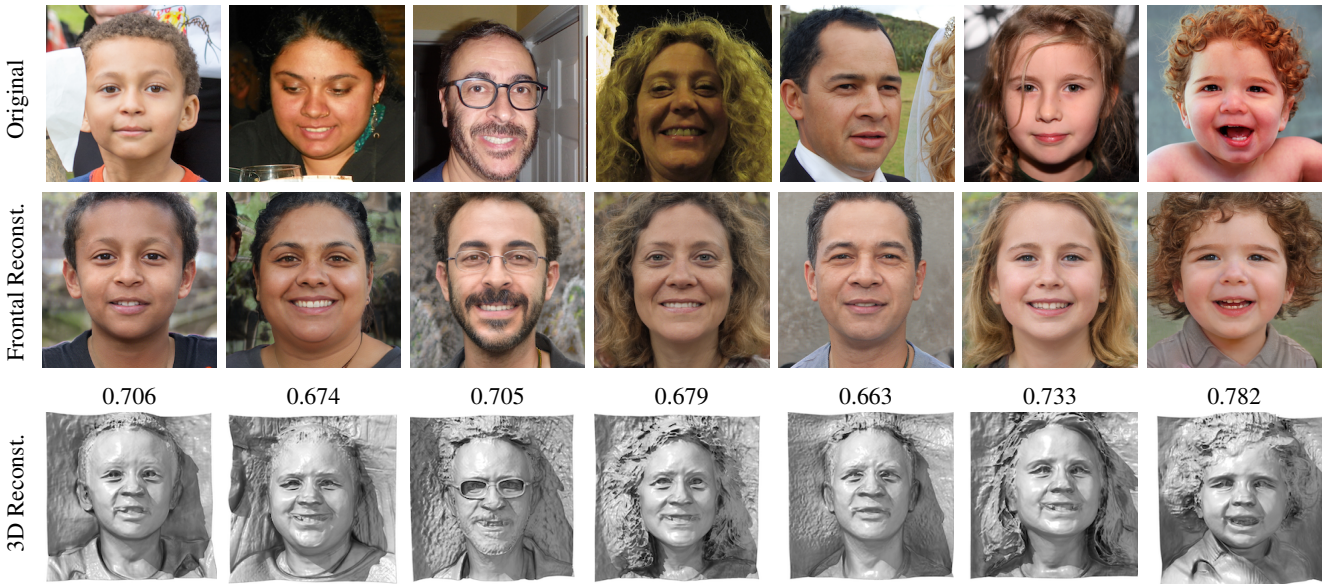

Figure 1: Sample face images from the FFHQ dataset (first row) as well as their corresponding 3D (third row) and frontal 2D reconstruction (second row) from facial templates in the whitebox *template inversion attack* against ArcFace. Values show the cosine similarity between the templates of the original and frontal reconstructed face images.

Abstract

Face recognition systems are increasingly being used in different applications. In such systems, some features (also known as embeddings or templates) are extracted from each face image. Then, the extracted templates are stored in the system's database during the enrollment stage and are later used for recognition. In this paper, we focus on template inversion attacks against face recognition systems and introduce a novel method (dubbed GaFaR) to reconstruct 3D face from facial templates. To this end, we use a geometry-aware generator network based on generative neural radiance fields (GNeRF), and learn a mapping from facial templates to the intermediate latent space of the generator network. We train our network with a semi-supervised learning approach using real and synthetic images simultaneously. For the real training data, we use a Generative Adversarial Network (GAN) based framework to learn the distribution of the latent space. For the synthetic training data, where we have the true latent code, we directly train in the latent space of the generator network. In addition, during the inference stage, we also propose optimization on the camera parameters to generate face images to improve the success attack rate (up to 17.14% in our experiments). We evaluate the performance of our method in the whitebox and blackbox attacks against state-of-the-art face recognition models on the LFW and MOBIO datasets. To our knowledge, this paper is the first work on 3D face reconstruction from facial templates. The project page is available at: https://www.idiap.ch/paper/gafar

1. Introduction

Automatic face recognition (FR) has become a well-known biometric authentication tool which has been widely used in different applications, including smartphone locks, border controls, etc. In such systems, generally, some features (also referred to as facial embeddings or facial templates) are extracted from users and stored in the system's database during the enrollment stage. In the recognition stage, either verification or identification, similar features are extracted from the user and are compared with the features in the system's database. Therefore, facial features play the main role in automatic face recognition systems and convey important information about users.

Template inversion (TI) attack against FR systems refers to an attempt by an adversary to reconstruct face images from the templates stored in the system's database. Then, the adversary may find sensitive information about the enrolled users and also can use the reconstructed face image to impersonate and enter the system. Therefore, compared to most attacks on FR systems which threaten the security of the system [19, 3, 22, 18, 35, 34, 10], TI attack jeopardizes both security and privacy of users, and thus requires further study. In this paper, we introduce a novel method (called geometry-aware face reconstruction, shortly *GaFaR*) for the TI attack in FR systems to reconstruct a 3D face from facial features (extracted from 2D FR models). The 3D face reconstruction provides further information than 2D face reconstruction, and in particular can be used to generate face image from any pose to improve the attack to the FR system. To our knowledge, this is the first paper on 3D face reconstruction from facial templates.

Recently, neural radiance fields (NeRF) [38] absorbed considerable attention in computer vision society due to remarkable results in the novel-view synthesis problem. Built upon NeRF, generative NeRF (GNeRF) methods such as [37, 44, 56, 6, 8, 41, 42, 20, 55, 7, 12, 48, 43] combine conditional NeRF with generative models for geometry-aware image generation tasks. In these methods, a generative model, such as a generative adversarial network (GAN), is used to embed the shape and appearance of an object in a latent space. Then, the latent code of the GAN, along with the camera parameters, are fed to a NeRF for the rendering process. Among GNeRF methods, there are several works, such as [8, 41, 42, 20, 55, 7, 12, 48], for geometry-aware 3D face generation, which can generate face images from different views.

In our proposed TI method, we use a geometry-aware face generator network based on GNeRF, and learn a new mapping from facial templates to the intermediate latent space of the generator network. We train our network using real and synthesized images simultaneously with a semi-supervised learning approach. For real training data where we do not have the corresponding latent code (i.e., unsupervised), we use a GAN-based framework to learn the distribution of the latent space. For the synthetic training data where we have true values of latent codes (i.e., supervised), we directly learn the latent space. Because we have 3D reconstructed face, we can use any arbitrary pose to inject the the FR system for the attack. Therefore, during the inference stage, we also use optimization on the camera parameters to improve the attack against the FR system. We evaluate the performance of our proposed method in the *whitebox* (i.e., where the adversary knows the internal functioning and parameters of the FR model) and *blackbox* (i.e., where the adversary does not have information about the internal functioning of the FR model) attacks against state-of-the-art (SOTA) FR systems. For evaluation, we consider FR systems with the LFW [23] and MOBIO [36] datasets and evaluate our attack if the reconstructed face can be used to enter the system. Fig. 1 illustrates sample face images from FFHQ [26] dataset and their 3D reconstruction from ArcFace [11] templates using our attack.

To elaborate on the contributions of this paper, we list them hereunder:

- We propose a novel template inversion method to reconstruct 3D faces from the facial templates of a face recognition system. To our knowledge, this paper is the first work on 3D face reconstruction from facial templates.

- We use a geometry-aware generator network based on GNeRF, and learn a mapping from facial templates to the intermediate latent space of the generator network. We train our mapping network using a GAN-based framework and a semi-supervised learning approach using real and synthetic face images.

- We use optimization on the camera parameters in the geometry-aware generator network during the inference stage. The optimization of camera parameters can find a pose that improves the success attack rate.

The remainder of the paper is organized as follows. First, we review the related works in the literature in Sec. 2. Then, we present our proposed method in Sec. 3. Next, we describe our experiments in Sec. 4 and discuss limitations and ethical considerations in Sec. 5. Finally, the paper is concluded in Sec. 6.

2. Related Works

Generally, the methods in the literature for reconstructing face from facial templates generate 2D face images. In [57], authors proposed two methods based on optimization and learning to reconstruct 2D low-resolution face images from facial templates in the whitebox scenario. In the

Table 1: Comparison with related works.

Ref.	2D/3D	Resolution	Whitebox/ Blackbox	Method Basis	Available code
[57]	2D	low	whitebox	1) optimization 2) learning	✗
[9]	2D	low	both*	learning	✗
[33]	2D	low	blackbox	learning	✓
[16]	2D	low	both†	learning	✗
[49]	2D	low	both†	learning	✗
[1]	2D	low	blackbox	learning + opt.	✗
[14]	2D	high	blackbox	learning	✓
[51]	2D	high	blackbox	optimization	✓
[15]	2D	high	blackbox	optimization	✗
[Ours]	**3D**	**high**	**both‡**	**learning**	✓

*The method is based on the *whitebox* attack, and is also applied in the *blackbox* scenario by removing a loss term that required the FR model.

†The method is based on the *whitebox* attack, and is extended to the *blackbox* with knowledge distillation of the FR model.

‡The method is based on the *whitebox* attack, and is extended to *blackbox* using a different FR model.

optimization-based method, they used a gradient ascent-based algorithm to reconstruct the face image, and optimized the generated image to minimize a multi-term loss function, including the ℓ_2 distance between target templates and templates of the reconstructed face image. In addition, they applied total-variation and Laplacian pyramid gradient normalization [5] on the reconstructed image to generate a smooth image. In their learning-based method, they used a deconvolutional neural network to reconstruct face images and trained it with the same loss function that they used in their optimization-based method.

In [9], a multi-layer perceptron (MLP) is trained to find facial landmark coordinates (optimized using the mean squared error) from the given template. They also trained a convolutional neural network (CNN) to generate face texture (optimized using mean absolute error) from the target template. Then, they used a differentiable warping to combine estimated landmarks (from MLP) and textures (from CNN) and reconstruct low-resolution face images. In the *whitebox* scenario, they further optimized MLP and CNN by minimizing the distance between templates of reconstructed and original face images. However, in the *blackbox* scenario, they trained MLP and CNN separately, and used warping in the inference only.

In [33], two new deconvolutional networks are proposed, called NbBlock-A and NbBlock-B, and are trained with pixel loss (ℓ_1 norm of reconstruction error) and perceptual loss (distance of middle layers of VGG-19[46] when given the reconstructed and original face images) to reconstruct low-resolution face images in the blackbox scenario. In [16] and [49], a same bijection-learning-based method is used to train GAN models with PO-GAN [25] and TransGAN [24] structure, respectively. The method is based on *whitebox* attack and for training the GAN model, authors also minimized the distance between target templates and templates extracted from the reconstructed face images using the FR model. In the *blackbox* attack, they proposed to use the distillation of knowledge to train a student network that mimics the target FR model and used the student network in their method. However, they did not report any detail (and no published source code) about the training of the student network (e.g., network structure, etc.).

In [1], a 3-step method is proposed to reconstruct low-resolution face images in the *blacbox* scenario. First, they trained a GAN for general face generation. In the second step, they trained a MLP to map target templates to embeddings of a known FR model. In the final step, they found a latent code in the input of their generator (of their GAN) which generates a face image that maximizes two terms; the discriminator score (for being a real face image) and cosine similarity between the mapped embedding and the embeddings extracted by the known FR model.

In contrast to most works in the literature which generate low-resolution face images, recently there have been few works to generate high-resolution 2D face images. In [51], a grid-search optimization using the simulated annealing [50] approach is used on the latent vector (i.e., input noise) of StyleGAN2 [28] to find latent codes that can generate face images which have templates similar to the target templates. However, their proposed method is computationally expensive, and they reported their evaluation on only 20 face images. In [15], a similar optimization to [51] on the latent vector of StyleGAN2 [28] is considered, but it is solved using the standard genetic algorithm [47]. In contrast to [51, 15] which are based on optimization, in [14], a learning-based method is proposed in the *blackbox* scenario. The authors proposed to generate some face images using StyleGAN2 [28] and extract the templates using the FR model. Then, they trained a MLP to map facial templates to the input latent codes of StyleGAN2 [28].

Tab. 1 compares our proposed method with the previous works in the literature. To our knowledge, our proposed method is the first work on 3D face reconstruction from facial templates (extracted from 2D face recognition models). Furthermore, unlike most works in the literature, our method generates high-resolution face images. Last but not least, our method can be used in both *whitebox* and *blackbox* attacks against FR systems.

3. Proposed Method

In this paper, we consider a threat model as described in Sec. 3.1 and use the proposed 3D face reconstruction method (dubbed *GaFaR*) described in Sec. 3.2. After training and in the inference stage, we use an optimization on the camera parameters as described in Sec. 3.3 to generate a 2D face image with a pose that can lead to a higher success attack rate (SAR) and inject the face image into the system. Fig. 2 depicts the block-diagram of the proposed method.

3.1. Threat Model

We consider an attack against FR systems, where the adversary gains access to the system's database and aims to generate a 3D reconstruction of enrolled users. To this end, the adversary trains a network to reconstruct a 3D face given a target facial template. Then, the adversary can use the 3D reconstruction to impersonate users into the system (i.e., using a 3D face mask, etc.). For simplicity, we consider that the adversary can inject a 2D face image (from the 3D reconstructed face) into the system. Hence, the adversary uses an appropriate pose to generate a 2D face image for the attack. According to the adversary's knowledge of the FR model, we consider both *whitebox* and *blackbox* attacks against FR systems. In the *whitebox* scenario, we assume that the adversary has complete knowledge of the FR model from which the templates are leaked. However, in the *blackbox* scenario, we assume that the adversary does not know the internal functioning of the FR model and can only use the FR model to generate facial template for any face image. In addition, we assume that in the *blackbox* scenario, the adversary has access to another FR model and knows its internal parameters and functioning.

3.2. 3D Face Reconstruction

To reconstruct 3D faces from facial templates, we consider a geometry-aware face generator network based on GNeRF, such as EG3D [7], and learn a mapping $M_{\text{GaFaR}} : \mathcal{T} \rightarrow \mathcal{W}$ from the facial templates $\boldsymbol{t} \in \mathcal{T}$ to the intermediate latent space $\mathcal{W}$ of the GNeRF model. Then, we use the remainder of GNeRF model $G(.)$ to synthesize an image $\hat{I} = G(\hat{\boldsymbol{w}}, \boldsymbol{c})$ from an arbitrary view using the mapped intermediate latent vector $\hat{\boldsymbol{w}}$ and camera parameters $\boldsymbol{c}$. We consider a network with two fully-connected layers with the Leaky ReLU activation functions for our mapping network M_{GaFaR}.

We train the mapping network M_{GaFaR} with a *semi-supervised* approach using synthetic and real data. For the real data, we consider a set of real face images $\{\boldsymbol{I}_{\text{real},i}\}_{i=0}^{N}$ and extract the facial template $\boldsymbol{t}_{\text{real},i} = F(\boldsymbol{I}_{\text{real},i})$ from each face image $\boldsymbol{I}_{\text{real},i}$ using the FR model $F(.)$. As we do not have the true value of intermediate latent space $\mathcal{W}$ of the GNeRF model to generate the real face image, we consider training our mapping network using this data as *unsupervised* learning. On the other hand, for the synthetic data, we use the pre-trained geometry-aware face generator network to generate a set of random face images $\{\boldsymbol{I}_{\text{syn},i}\}_{i=0}^{M}$. Therefore, as opposed to real data, we have the true value of intermediate latent space $\boldsymbol{w} \in \mathcal{W}$ to generate the same synthetic face image. Hence, we consider training our mapping network using the synthetic data as *supervised* learning. We train our mapping network simultaneously using real and synthetic training data as follows:

Unsupervised learning using real data To train our mapping network $M_{\text{GaFaR}}(.)$ with the real data, we use a GAN-based framework based on Wasserstein GAN (WGAN) [2] algorithm to learn the distribution of intermediate latent space $\mathcal{W}$ of the GNeRF model. In this framework, our mapping network M_{GaFaR} generates a latent code $\hat{\boldsymbol{w}} = M_{\text{GaFaR}}([\boldsymbol{n}, \boldsymbol{t}])$ using the facial template $\boldsymbol{t}$ and a random vector $\boldsymbol{n} \in \mathcal{N}$. We can also generate real latent codes $\boldsymbol{w} = M(\boldsymbol{z}) \in \mathcal{W}$ for training our GAN using the GNeRF mapping function M and a random vector $\boldsymbol{z} \in \mathcal{Z}$. Then, we can use a critic network $C(.)$ to score the latent codes generated by our mapping $M_{\text{GaFaR}}(.)$ and GNeRF mapping $M(.)$. Therefore, we can train the critic network $C(.)$ and our mapping M_{GaFaR} using the following loss functions:

$$\mathcal{L}_{C}^{\text{WGAN}} = \mathbb{E}_{\boldsymbol{w} \sim M(\boldsymbol{z})}[C(\boldsymbol{w})] - \mathbb{E}_{\hat{\boldsymbol{w}} \sim M_{\text{GaFaR}}([\boldsymbol{n},\boldsymbol{t}])}[C(\hat{\boldsymbol{w}})] \tag{1}$$

$$\mathcal{L}_{M_{\text{GaFaR}}}^{\text{WGAN}} = \mathbb{E}_{\hat{\boldsymbol{w}} \sim M_{\text{GaFaR}}([\boldsymbol{n},\boldsymbol{t}])}[C(\hat{\boldsymbol{w}})] \tag{2}$$

In addition to the WGAN training, we use the generated face image $\hat{\boldsymbol{I}} = G(M_{\text{GaFaR}}([\boldsymbol{n}, \boldsymbol{t}]), \boldsymbol{c})$ to optimize our mapping network using the following loss function:

$$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}, \tag{3}$$

where $\mathcal{L}^{\text{Pixel}}$ and $\mathcal{L}^{\text{ID}}$ are pixel loss and ID loss, respectively, and are defined as:

$$\mathcal{L}^{\text{Pixel}} = \mathbb{E}_{\hat{\boldsymbol{w}} \sim M_{\text{GaFaR}}([\boldsymbol{n},\boldsymbol{t}])}[\|\boldsymbol{I} - G(\hat{\boldsymbol{w}}, \boldsymbol{c})\|_2^2] \tag{4}$$

$$\mathcal{L}^{\text{ID}} = \mathbb{E}_{\hat{\boldsymbol{w}} \sim M_{\text{GaFaR}}([\boldsymbol{n},\boldsymbol{t}])}[\|F_{\text{loss}}(\boldsymbol{I}) - F_{\text{loss}}(G(\hat{\boldsymbol{w}}, \boldsymbol{c}))\|_2^2] \tag{5}$$

The pixel loss minimizes the pixel-level reconstruction error and ID loss helps the network to generate face images with a similar facial templates extracted by F_{loss}. To calculate the ID loss in Eq. (5), we use the same FR model as the one by which leaked templates are calculated (i.e., F) in the *whitebox* scenario. However, in the *blackbox* scenario, we use another FR model that we assume the adversary has access to. The critic network is a fully-connected network with three layers and Leaky ReLU activation function.

Supervised learning using synthetic data To train our mapping network $M_{\text{GaFaR}}(.)$ with synthetic data, we can directly learn the GNeRF intermediate latent code $\boldsymbol{w} = M(\boldsymbol{z})$. In addition to directly learning $\boldsymbol{w}$, we use the generated face image to train our mapping network by minimizing the following multi-term loss function:

$$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^{w} + \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}, \tag{6}$$

where $\mathcal{L}^{\text{Pixel}}$ and $\mathcal{L}^{\text{ID}}$ are pixel loss and ID loss, respectively. Furthermore, $\mathcal{L}^{w}$ is w-loss to directly learn the latent space

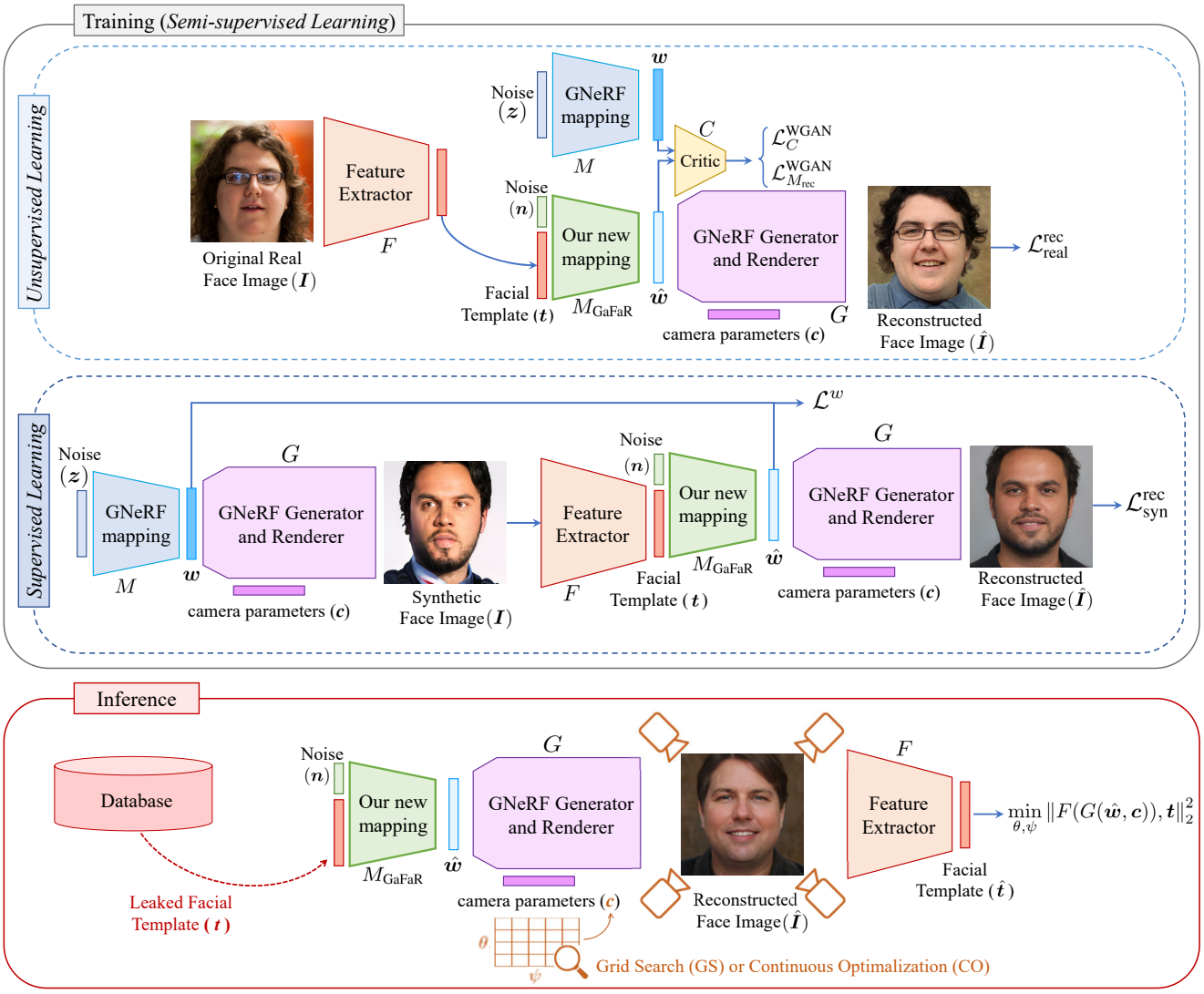

Figure 2: Block-diagram of the proposed method: In the training process, a *semi-supervised* approach is used to learn our mapping M_{GaFaR} (depicted with a green block) from the facial templates to the intermediate latent space of the GNeRF model using real data (*unsupervised*, where we don't have the corresponding latent code) and synthetic data (*supervised*, where we have the corresponding latent code $\boldsymbol{w}$), simultaneously. During the inference stage, the leaked template $\boldsymbol{t}$ is given to our mapping network to find corresponding latent code $\hat{\boldsymbol{w}} = M_{\text{GaFaR}}([\boldsymbol{n}, \boldsymbol{t}])$. Then, $\hat{\boldsymbol{w}}$ along with camera parameters $\boldsymbol{c}$ are given to the generator and renderer of GNeRF G to generate a reconstructed face image $\hat{\boldsymbol{I}} = G(\hat{\boldsymbol{w}}, \boldsymbol{c})$. Finally, an optimization (either grid search or continuous optimization) is applied on two of the camera parameters, θ and ψ, from $\boldsymbol{c}$, to find the best pose that minimizes the distance between leaked template $\boldsymbol{t}$ and the template of reconstructed face image.

by minimizing mean squared error between $\boldsymbol{w}$ and $\hat{\boldsymbol{w}} = M_{\text{GaFaR}}([\boldsymbol{n}, \boldsymbol{t}])$ as follows:

$$\mathcal{L}^w = \mathbb{E}_{\boldsymbol{w} \sim M(\boldsymbol{z})}[\|\boldsymbol{w} - M_{\text{GaFaR}}([\boldsymbol{n}, \boldsymbol{t}])\|_2^2] \tag{7}$$

3.3. Camera Parameters Optimization

During the inference stage, after generating a 3D reconstruction of face using our method described in Sec. 3.2, we optimize the camera parameters to find a pose that increases the SAR. Among different camera parameters $\boldsymbol{c}$, the parameters which correspond to the camera rotation can change the pose of the generated face image. Note that by varying the camera rotations, we aim to change the pitch and yaw rotations of the 3D reconstructed face and do not modify the roll rotation. In fact, the roll rotation of the face will be eliminated through face alignment as a preprocess prior to feature extraction in the FR system. To optimize camera parameters, we consider two different approaches as follows:

Grid Search (GS) For the grid search, we consider predefined steps to change the camera pitch $\theta \in \Theta$ and yaw $\psi \in \Psi$ rotations of camera parameters $\boldsymbol{c}$. We generate the 2D face image for all values of camera rotation steps (θ_{step} and ψ_{step}) and find the facial template for each generated image. Then, we select the face image $\hat{\boldsymbol{I}} = G(M_{\text{GaFaR}}([\boldsymbol{n}, \boldsymbol{t}]), \boldsymbol{c})$ that has a template $\hat{\boldsymbol{t}} = F(\hat{\boldsymbol{I}})$ with the minimum mean squared error with the target template $\boldsymbol{t}$. Note that the grid search can be applied in both *whitebox* and *blackbox* scenarios using the FR model F.

Continuous Optimization (CO) In our continuous optimization approach, we start from the frontal image and use Adam [30] optimizer to solve the following minimization using $\hat{\boldsymbol{w}} = M_{\text{GaFaR}}([\boldsymbol{n}, \boldsymbol{t}])$:

$$\min_{\theta,\psi} \|F(G(\hat{\boldsymbol{w}}, \boldsymbol{c})), \boldsymbol{t}\|_2^2, \tag{8}$$

In this optimization, we find the parameters θ and ψ for camera rotation that lead to a face image with templates close to the target template $\boldsymbol{t}$. In contrast to the grid search approach, the continuous optimization can be applied only in the *whitebox* scenario.

4. Experiments

4.1. Experimental Setup

Face Recognition models In our experiments, we consider different SOTA FR models including ArcFace [11], ElasticFace [4] as well as four different FR models with SOTA backbones from FaceX-Zoo [53], including AttentionNet [52], HRNet [54], RepVGG [13] and Swin [32]. The recognition performance of these models are reported in the supplementary material.

Datasets All the mentioned FR models are trained on the MS-Celeb1M dataset [21]. To train the face reconstruction network, we assume that adversary does not have any knowledge of the dataset used in training FR model, and uses a different dataset. To this end, we consider the Flickr-Faces-HQ (FFHQ) dataset [27], which consists of 70,000 high-quality images crawled from internet (with no identity label), for training our face reconstruction network. We randomly split FFHQ into train (90%) and test (10%) set.

To evaluate the vulnerability of FR models, we use two other datasets, including Labeled Faces in the Wild (LFW) [23] and MOBIO [36] datasets. The LFW dataset consists of 13,233 face images of 5,749 people collected from internet, where 1,680 people have two or more images. The MOBIO dataset includes face images of 150 people captured using mobile devices in 12 sessions for each person[1].

[1]The results on AgeDB [39] dataset is also reported in the supplementary material.

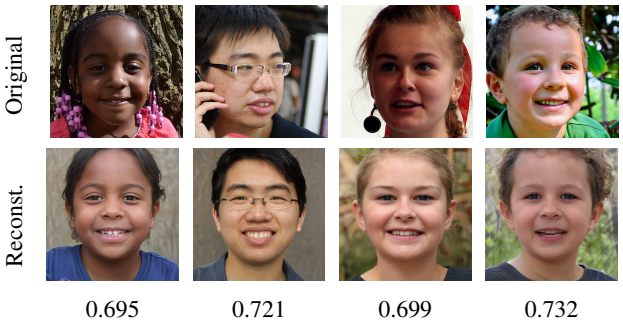

Figure 3: Sample face images from the FFHQ dataset (first row) and their corresponding frontal (second row) reconstructed face images using our method in the *blackbox* attack against ElasticFace using ArcFace for F_{loss}. The values show the cosine similarity between templates of original and frontal reconstructed face images.

Evaluation Protocol For evaluation with each of the LFW and MOBIO datasets, we build a separate FR system, and use the reference templates (i.e., enrolled in the system's database) as input to our face reconstruction method. Then, we inject the reconstructed face image as a query to the system and evaluate the adversary's success attack rate (SAR) to enter the FR system. We should note that for each image, we use only one reconstructed face image as a query to the FR system and evaluate the SAR.

Implementation Details and Source Code In our experiments, we use the grid search (in both *whitebox* and *blackbox* attacks) and continuous optimization (in *whitebox* attacks only) to optimize camera parameters, as described in Sec. 3.3, in the inference stage. For grid search, we consider $\psi \in [-45°, +45°]$ and $\theta \in [-30°, +30°]$ for a 11×11 grid with step sizes of $\psi_{\text{step}} = 9°$ and $\theta_{\text{step}} = 6°$. In the continuous optimization approach, we use 121 iterations of Adam optimizer [30] with the learning rate of 10^{-2}. An ablation study on these hyperparamters and the corresponding execution times are reported in the supplementary material. All our models (for different FR models and different scenarios) are trained in 15 epochs (each takes around 2 days) on a system equipped with a single NVIDIA RTX 3090 GPU. For GNeRF, we used a pretrained model of EG3D [7] in our experiments and generate 512×512 resolution face images. The input noise $\boldsymbol{z}$ to GNeRF model has 512 dimensions and the noise $\boldsymbol{n}$ in the input of our mapping network M_{GaFaR} has 16 dimensions. For unsupervised learning using real face images (FFHQ), we assume that camera parameters for real face images as frontal view. The source code of our experiments are publicly available[2].

[2]https://www.idiap.ch/paper/gafar

Table 2: Evaluation of *blackbox* attack against SOTA FR models at systems' FMR=10^{-3} on the LFW and MOBIO datasets in terms of success attack rate (SAR). For attacks using our proposed method, we use ArcFace and also ElasticFace as F_{loss} to calculate the ID loss in Eq. (5). The values are in percentage.

	LFW						MOBIO					
	ArcFace	**ElsFace**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**	**ArcFace**	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**
NBNetA-M [33]	4.32	10.90	1.24	1.60	1.13	3.82	0	2.38	0	0	0	0
NBNetA-P [33]	16.83	26.98	0.66	1.44	5.72	9.70	4.76	16.19	0.48	0	14.29	7.14
NBNetB-M [33]	10.98	21.44	3.22	4.47	3.21	11.23	1.90	3.80	3.33	7.14	3.33	8.57
NBNetB-P [33]	40.26	58.16	16.29	18.42	15.24	40.76	15.24	43.81	31.90	26.67	23.81	44.29
Dong *et al.* [14]	13.21	12.61	3.90	4.07	3.22	12.38	3.33	8.10	10.48	6.67	9.05	3.33
Vendrow and Vendrow [51]	57.70	53.03	21.12	18.85	9.62	46.84	29.05	43.81	27.14	26.67	20.95	45.24
[F_{loss}= Els.Face] GaFaR	51.78	-	18.07	11.68	11.63	47.15	47.62	-	54.29	43.33	45.71	70.48
[F_{loss}= Els.Face] GaFaR + GS	**61.56**	-	23.56	17.21	15.82	54.08	**64.76**	-	63.81	55.23	58.57	82.38
[F_{loss}= ArcFace] GaFaR	-	74.54	33.59	37.80	25.40	67.11	-	84.76	72.38	76.67	72.86	89.05
[F_{loss}= ArcFace] GaFaR + GS	-	**78.67**	**38.42**	**43.27**	**29.84**	**70.82**	-	**86.62**	**80.00**	**83.80**	**73.33**	**93.33**

Table 3: Evaluation of *whitebox* attack against SOTA FR models at systems' FMR $= 10^{-3}$ on the LFW and MOBIO datasets in terms of success attack rate (SAR). The values are in percentage.

	LFW						MOBIO					
	ArcFace	**ElsFace**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**	**ArcFace**	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**
GaFaR	79.84	62.32	27.00	31.87	17.33	74.08	82.86	81.43	64.29	71.43	53.81	94.76
GaFaR + GS	82.52	68.71	32.42	37.48	19.83	76.06	85.23	82.38	70.47	81.42	59.52	94.29
GaFaR + CO	**84.25**	**71.69**	**34.53**	**39.53**	**20.44**	**77.36**	**89.05**	**84.29**	**75.71**	**81.43**	**62.86**	**95.71**

4.2. Analyze

Blackbox Scenario To evaluate the proposed method in the *blackbox* attack against SOTA FR models, we consider ArcFace and also ElasticFace as F_{loss} to calculate the ID loss in Eq. (5). Tab. 2 compares the performance of the proposed method with *blackbox* face reconstruction methods in the literature in terms of SAR for systems configured at false match rate (FMR) of 10^{-3}. Similar results for FMR $= 10^{-2}$ are reported in the supplementary material. As this table shows, the frontal reconstruction by our method (i.e, GaFaR) achieves superior performance than previous methods in the literature. Furthermore, camera parameter optimization (i.e., GaFaR+GS) improves the performance of our attack up to 17.14% compared to our frontal face reconstruction (i.e, GaFaR). Comparing the use of ArcFace and ElasticFace as F_{loss}, the performance of attacks with the ArcFace model is better. This is due to the fact that ArcFace has a higher recognition accuracy than ElasticFace. Fig. 3 shows sample face images reconstructed from ElasticFace templates in the *blackbox* attack (using ArcFace for F_{loss}).

Whitebox Scenario Tab. 3 reports the performance of the proposed method in the *whitebox* attack[3] in terms of SAR for FR systems configured at false match rate (FMR) of 10^{-3}. Similar results for FMR $= 10^{-2}$ are reported in the supplementary material. According to this table, all these FR models are highly vulnerable to our attack. Furthermore, the camera parameter optimization improves the performance of GaFaR. Comparing grid search with contin-

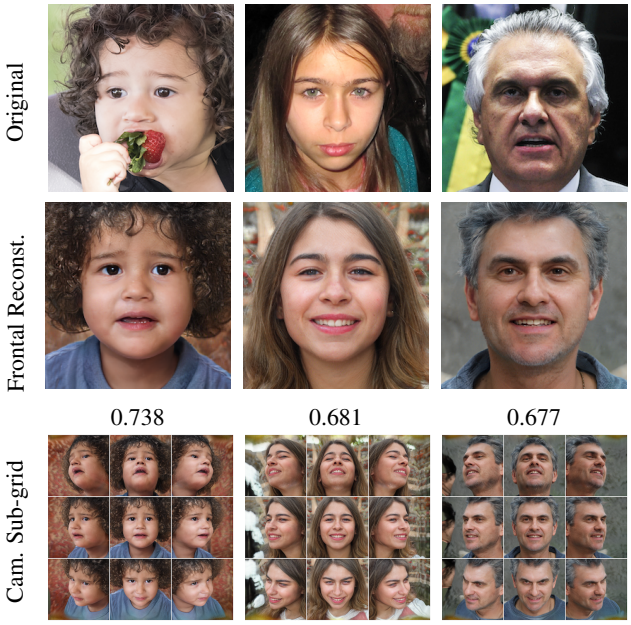

Figure 4: Sample face images from the FFHQ dataset (first row) and their corresponding frontal (second row) and camera parameters sub-grid (third row) reconstructed face images using our method in the *whitebox* attack against ArcFace. The values show the cosine similarity between templates of original and frontal reconstructed face images.

[3]The *whitebox* attack methods reported in Tab. 1 do not have available source code, and we could not reproduce their results.

Table 4: Ablation study on the proposed *semi-supervised* learning approach and evaluation of loss terms in the *white-box* attack against ArcFace model in terms of success attack rate (SAR) on the LFW and MOBIO datasets. The SAR values are in percentage and for an attack without any camera parameter optimization (GS/CO).

approach	Loss Functions	LFW		MOBIO	
		FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
supervised	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^{w} + \mathcal{L}^{\text{Pixel}} + \mathcal{L}^{\text{ID}}$	83.80	69.467	90.96	82.38
	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^{w} + \mathcal{L}^{\text{Pixel}}$	31.75	13.92	43.81	8.57
	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^{w} + \mathcal{L}^{\text{ID}}$	0.86	0.30	0	0
	$\mathcal{L}_{\text{syn}}^{\text{rec}} = \mathcal{L}^{w}$	33.69	15.43	32.38	9.52
unsupervised	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}_{\text{Pixel}} + \mathcal{L}_{\text{ID}}$ [without WGAN]	0.44	0.15	0	0
	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}_{\text{Pixel}} + \mathcal{L}_{\text{ID}}$ [with WGAN]	67.72	45.76	70.48	31.90
	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}_{\text{ID}}$ [with WGAN]	54.51	30.83	52.86	19.52
	$\mathcal{L}_{\text{real}}^{\text{rec}} = \mathcal{L}_{\text{Pixel}}$ [with WGAN]	2.21	0.40	0	0
semi-supervised	Eqs. (1) to (3) and (6)	**89.27**	**79.84**	**95.71**	**82.86**

uous optimization, results show that with the same number of iterations continuous optimization achieves better performance. In addition, comparing these results with recognition performance of FR models (available in the supplementary material), we conclude that models with higher recognition accuracy are more vulnerable to our attack. Comparing results in Tab. 2 and Tab. 3, we observe that when we use ArcFace as F_{loss} in our loss function, *blackbox* attacks achieve better results than *whitebox* attacks for most cases, which can be explained considering the superior recognition performance of ArcFace than other FR models. Fig. 4 shows sample face images reconstructed from ArcFace templates in the *whitebox* attack. This figure also presents a grid of reconstructed face images with different poses[4].

Ablation Study We evaluate the effect of our *semi-supervised* learning approach in our proposed method compared to fully *supervised* learning and fully *unsupervised* learning approach. In each case, we evaluate the effect of each loss function too. Furthermore, for the fully *unsupervised* learning approach, we evaluate the effect of GAN learning which we used in our method. Tab. 4 reports our ablation study in *whitebox* attack against ArcFace model on the LFW and MOBIO datasets in terms of SAR at system's FMR of 10^{-2} and 10^{-3}. As these results show, the proposed *semi-supervised* approach achieves a better performance than fully *supervised* learning and fully *unsupervised* learning approaches. Furthermore, each of our loss terms has an important effect on the performance. In particular, using WGAN for real data, where we don't have the true value of intermediate latent codes, helps the network to learn the distribution of GNeRF intermediate latent space $\mathcal{W}$. Otherwise, the generated latent code by the mapping network will be out of distribution, and therefore generator part of GNeRF will not generate face-like images.

[4]Samples of reconstructed face images with the camera parameters grid used in our grid search optimization are presented in the supplementary material.

Figure 5: Sample failure cases from the LFW dataset (first row) and their corresponding frontal (second row) reconstructed face images using our method in the *whitebox* attack against ArcFace. The values below each image show the cosine similarity between templates of original and reconstructed face images.

5. Discussion

Limitations In spite of the considerable success attack rate of our method in our experiments, the reconstructed face images from some facial templates fail to attack the system. Fig. 5 illustrates sample failed cases in the *whitebox* attack against ArcFace. As these sample images show, the failed reconstructions reflect a bias in the generated faces for dark skin and also old people. Such a bias in final results can be caused by inherent biases of each of the datasets used to train the FR model, GNeRF model, and our face reconstruction model. In addition to bias, comparing attacks against different FR models in Tabs. 2 and 3 shows that FR models with worse recognition performance have lower SAR and more failed cases.

Ethical considerations This work is proposing a new method for a TI attack against FR models. The 3D reconstructed face can be used to generate 3D face masks or 2D printed photographs for presentation attacks against FR models. In addition to the security threats, the reconstructed face images can reveal important privacy-sensitive information of enrolled users, such as age, gender, ethnicity, etc. We do not condone using our work with the intent of attack to *real* FR systems. As a matter of fact, the main motivation for this work is to show such a vulnerability in the FR systems, and to encourage the scientific community to develop and propose the next generation of safe and protected FR systems. Along the same lines, data protection regulations, such as the European Union General Data Protection Regulation (EU-GDPR) [17], consider biometric data as sensi-

tive information, and put legal obligations to protect them. To this end and to mitigate such threats, several biometric template protection schemes are also proposed in the literature [40, 29, 31, 45]. We should also note that the project on which the work has been conducted has passed an Internal Ethical Review Board (IRB).

6. Conclusion

In this paper, we proposed a new method (dubbed *GaFaR*) to generate 3D face reconstruction from facial templates for a TI attack against FR models. We used a geometry-aware face generation network based on GNeRF and trained a mapping from facial templates to the intermediate latent space of the GNeRF model using a *semi-supervised* learning approach. To train our model, we used synthetic and real face images. For the synthetic training data, we had the latent code of each face image and could train our mapping with *supervised* learning. For the real training data, we used a GAN-based framework to learn the distribution of latent space. In the inference stage, we used optimization on the camera parameters to find the pose which increases the success attack rate. We evaluated our method on the *whitebox* and *blackbox* attacks against SOTA FR models on the LFW and MOBIO datasets. To our knowledge, this paper is the first work on 3D face reconstruction from facial templates (extracted from 2D face recognition models).

Acknowledgment

This research is based upon work supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813).

References

[1] Muku Akasaka, Soshi Maeda, Yuya Sato, Masakatsu Nishigaki, and Tetsushi Ohki. Model-free template reconstruction attack with feature converter. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2022.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 214–223. PMLR, 2017.

[3] Battista Biggio, Paolo Russu, Luca Didaci, Fabio Roli, et al. Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective. *IEEE Signal Processing Magazine*, 32(5):31–41, 2015.

[4] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1587, 2022.

[5] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in Computer Vision*, pages 671–679. Elsevier, 1987.

[6] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3981–3990, 2022.

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022.

[8] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5799–5809, 2021.

[9] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3703–3712, 2017.

[10] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[11] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10673–10683, 2022.

[13] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13733–13742, 2021.

[14] Xingbo Dong, Zhe Jin, Zhenhua Guo, and Andrew Beng Jin Teoh. Towards generating high definition face images from deep templates. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–11. IEEE, 2021.

[15] Xingbo Dong, Zhihui Miao, Lan Ma, Jiajun Shen, Zhe Jin, Zhenhua Guo, and Andrew Beng Jin Teoh. Reconstruct face from features using gan generator as a distribution constraint. *arXiv preprint arXiv:2206.04295*, 2022.

[16] Chi Nhan Duong, Thanh-Dat Truong, Khoa Luu, Kha Gia Quach, Hung Bui, and Kaushik Roy. Vec2face: Unveil human faces from their blackbox features in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6132–6141, 2020.

[17] European Council. Regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation), April 2016.

[18] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[19] Javier Galbally, Chris McCool, Julian Fierrez, Sebastien Marcel, and Javier Ortega-Garcia. On the vulnerability of face verification systems to hill-climbing attacks. *Pattern Recognition*, 43(3):1027–1038, 2010.

[20] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, pages 1–25, 2022.

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102. Springer, 2016.

[22] Abdenour Hadid, Nicholas Evans, Sebastien Marcel, and Julian Fierrez. Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. *IEEE Signal Processing Magazine*, 32(5):20–30, 2015.

[23] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[24] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021.

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020.

[29] Prabhjot Kaur, Nitin Kumar, and Maheep Singh. Biometric cryptosystems: a comprehensive survey. *Multimedia Tools and Applications*, pages 1–56, 2022.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California., USA, May 2015.

[31] Nitin Kumar et al. Cancelable biometrics: a comprehensive survey. *Artificial Intelligence Review*, 53(5):3403–3446, 2020.

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 10012–10022, 2021.

[33] Guangcan Mai, Kai Cao, Pong C Yuen, and Anil K Jain. On the reconstruction of face images from deep face templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1188–1202, 2018.

[34] Sébastien Marcel, Julian Fierrez, and Nicholas Evans. *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Springer Nature, 2023.

[35] Sébastien Marcel, Mark S Nixon, Julian Fierrez, and Nicholas Evans. *Handbook of biometric anti-spoofing: Presentation attack detection*, volume 2. Springer, 2019.

[36] Chris McCool, Roy Wallace, Mitchell McLaren, Laurent El Shafey, and Sébastien Marcel. Session variability modelling for face authentication. *IET Biometrics*, 2(3):117–129, Sept. 2013.

[37] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6351–6361, 2021.

[38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020.

[39] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–59, 2017.

[40] Karthik Nandakumar and Anil K Jain. Biometric template protection: Bridging the performance gap between theory and practice. *IEEE Signal Processing Magazine*, 32(5):88–100, 2015.

[41] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, 2021.

[42] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, 2022.

[43] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1558–1567, 2022.

[44] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

[45] Hatef Otroshi Shahreza, Christian Rathgeb, Dailé Osorio-Roig, Vedrana Krivokuća Hahn, Sébastien Marcel, and Christoph Busch. Hybrid protection of biometric templates by combining homomorphic encryption and cancelable biometrics. In *Proceedings of the 2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022.

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] Mandavilli Srinivas and Lalit M Patnaik. Genetic algorithms: A survey. *Computer*, 27(6):17–26, 1994.

[48] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022.

[49] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Marios Savvides, and Khoa Luu. Vec2face-v2: Unveil human faces from their blackbox features via attention-based network in face recognition, 2022.

[50] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[51] Edward Vendrow and Joshua Vendrow. Realistic face reconstruction from deep embeddings. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[52] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2017.

[53] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[54] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[55] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18430–18439, 2022.

[56] Jichao Zhang, Enver Sangineto, Hao Tang, Aliaksandr Siarohin, Zhun Zhong, Nicu Sebe, and Wei Wang. 3d-aware semantic-guided generative model for human synthesis. *arXiv preprint arXiv:2112.01422*, 2021.

[57] Andrey Zhmoginov and Mark Sandler. Inverting face embeddings with convolutional neural networks. *arXiv preprint arXiv:1606.04189*, 2016.