Vulnerability of State-of-the-Art Face Recognition Models to Template Inversion Attack

Hatef Otroshi Shahreza, Vedrana Krivokuća Hahn, and Sébastien Marcel

***Abstract*—Face recognition systems use the templates (extracted from users' face images) stored in the system's database for recognition. In a template inversion attack, the adversary gains access to the stored templates and tries to enter the system using images reconstructed from those templates. In this paper, we propose a framework to evaluate the vulnerability of face recognition systems to template inversion attacks. We build our framework upon a real-world scenario and measure the vulnerability of the system in terms of the adversary's success attack rate in entering the system using the reconstructed face images. We propose a face reconstruction network based on a new block called "enhanced deconvolution using cascaded convolution and skip connections" (shortly, *DSCasConv*), and train it with a multi-term loss function. We use our framework to evaluate the vulnerability of state-of-the-art face recognition models, with different network structures and loss functions (in total 31 models), on the MOBIO, LFW, and AgeDB face datasets. Our experiments show that the reconstructed face images can be used to enter the system, which threatens the system's security. Additionally, the reconstructed face images may reveal important information about each user's identity, such as race, gender, and age, and hence jeopardize the users' privacy.**

***Index Terms*—biometrics, face recognition, face reconstruction, embedding, template inversion, vulnerability evaluation.**

I. Introduction

FACE recognition (FR) systems are being widely used in different applications and are among the most popular biometric recognition systems. In particular, in the recent years, FR is increasingly used as a secure authentication tool in a broad range of applications such as smart phone unlocking[1], border control[2], etc. In addition to the security purposes, FR is used for entertainment[3] and also in social media[4].

Despite the recent increase in the number of FR applications, there is a growing concern regarding the privacy of users in such systems. For this reason, Facebook for example, announced in November 2021 that the company is going to shut down its FR system and "delete more than a billion people's individual facial recognition templates"[5]. Furthermore, data protection regulations, such as the European Union General Data Protection Regulation (EU-GDPR)[6] have been put in place to legally protect our digital identities, which highlights the fact that biometric data is sensitive information and must be protected.

Fig. 1: Sample face images from the FFHQ dataset and their corresponding reconstructed images, using our template inversion method in a face recognition system based on the iResNet100-ArcFace model. The values below each image show the cosine similarity between the corresponding face templates. The decision threshold corresponding to FMR $= 10^{-3}$ is 0.24 on the LFW dataset.

Generally, in state-of-the-art (SOTA) FR systems, a deep convolutional neural network (CNN) is applied to the input face images to extract features (called face "templates" or "embeddings"). These features are extracted from each user's face and are stored as reference templates in the system's database during the enrollment stage. Then, during the recognition stage, similar features are extracted from the user's face, and the resulting probe template is compared with the reference templates. Since the face templates convey important information about the facial characteristics, and therefore identities, of the corresponding users, attacks on biometric recognition systems may jeopardize both the users' privacy and the system's security.

To address the potential threats to FR systems, different types of attack have been considered in the literature, and the vulnerability of FR systems to such attacks has been studied [1]–[4]. Amongst potential attacks to FR systems, template inversion[7] (TI) significantly endangers the users' privacy. In a TI attack, the adversary has access to the templates stored in the system's database, and they try to invert these templates to reconstruct the underlying face images. Then, the adversary may use the reconstructed face image to impersonate a user by injecting this image into the system and bypassing the camera [5]. The adversary can also use the reconstructed

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

Authors are with the Biometrics Security and Privacy Group of Idiap Research Institute, Martigny, Switzerland. Hatef Otroshi Shahreza (hatef.otroshi@epfl.ch) and Sébastien Marcel are also affiliated with École Polytechnique Fédérale de Lausanne (EPFL) and Université de Lausanne (UNIL), respectively.

[1]https://apple.co/3mLGCYV
[2]https://cnet.co/3sG8qSd
[3]https://bit.ly/3FKPBks
[4]https://bit.ly/3Jr8PxH
[5]https://bit.ly/3pEdjcv
[6]https://bit.ly/3FGy1y9

[7]Also known as template reconstruction.

face image as a presentation attack to the FR system. So, a successful TI attack can lead not only to breaking the security by accessing the system, but also to a reconstructed image of each user's face, which may reveal the identity of the user or at least some important information such as race, gender, age, etc. For instance, Fig. 1 shows sample face images from the FFHQ [6] database and their corresponding reconstructed images after our TI attack on a FR system using the iResNet100-ArcFace [7] model. As illustrated in this figure, the reconstructed faces can be recognized by the system since the similarity of features extracted from the original and reconstructed faces may be higher than the system's decision threshold. Moreover, we observe that the reconstructed faces reveal important information about each user's identity (e.g., gender, age, race). Hence, a TI attack is a crucial security and privacy threat to biometric recognition systems, which demands specific attention.

In this paper, we focus on TI attack on FR systems and propose a framework to evaluate the vulnerability of FR systems to this type of attack. To this end, we propose a new neural network to reconstruct face images from face templates. Our neural network is based on a new block, called "enhanced deconvolution using cascaded convolution and skip connections" (shortly, *DSCasConv*). DSCasConv generates outputs with the same size as the deconvolution layer, while enhancing the deconvolution output through several cascaded convolutional layers with skip connections. Each convoultional layer in DSCasConv enhances the output of previous layer, and in total, the residual cascaded convolutional layers recover and improve local dependencies of deconvolution output. We train our face reconstruction network with a multi-term loss function. In particular, to improve the TI attack, we use a loss term that minimizes the mean absolute error and cosine distance between the templates extracted from the original and reconstructed face images. For vulnerability evaluation, we consider the real-world scenario where the adversary gains access to the system's database and wants to enter the system by inverting the enrolled face templates and impersonating the enrolled users by employing the reconstructed face images. We consider the case where the adversary can bypass the camera and inject the reconstructed face image into the system. Our experiments show that the face images reconstructed by our TI network can be recognized by FR systems, and they also provide crucial identity information about the users enrolled in the system database. In our experiments, we evaluate the vulnerability of SOTA FR methods to our TI attack. We consider different SOTA FR methods with various network structures (different "backbones") as well as with different loss functions (different "heads"). In summary, the contributions of our paper are as follows:

- We propose a framework to evaluate the vulnerability of face recognition systems to template inversion attack. We define an evaluation protocol based on the real-world scenario, where the adversary gains access to the system's database and tries to invert the stored templates, to reconstruct the underlying face images, and impersonate the enrolled users using the reconstructed face image. We investigate the vulnerability of SOTA FR methods (including different network structures and different loss functions) to our TI attack in the proposed vulnerability evaluation framework.
- We propose a new network to reconstruct face images from face templates using a new block, called DSCasConv. Each DSCasConv block includes a deconvolutional layer followed by cascaded convolutional layers and skip connections. The residual cascaded convolutional layers recover and improve local dependencies of deconvolution output.
- We use a multi-term loss function, including a loss term which improves the TI attack and minimizes the mean absolute error and cosine distance between the templates extracted from the original and reconstructed face images.

In the following sections, we first review the related works in section II. Next, in section III, we describe our proposed framework to analyse the vulnerability of FR systems to our TI attack, and in section IV we discuss our experimental results. Finally, the paper is concluded in section V.

II. Related Works

There are two scenarios when attempting to invert a face template, based on what information is available about the FR model: *whitebox* scenario and *blackbox* scenario. In the whitebox scenario, the FR model is fully available (i.e., model parameters and internal functioning are known). However, in the blackbox scenario, there is no information about the parameters and internal functioning of the FR model, but we can use the FR model to generate templates from the given face image (e.g., using the Software Development Kit (SDK) of the FR system). Accordingly, in the blackbox scenario, the adversary does not have access to exact values of gradient of the input to the FR model, and therefore it is difficult to use the FR model either in a continuous gradient-based optimization[8] process or in a loss function for training a TI neural network.

In [8], authors proposed an iterative gradient ascent-based algorithm to reconstruct face images. They started from random noise or a guiding image, and optimized the template extracted from that. In addition, they used total-variation and Laplacian pyramid gradient normalization [31] to generate a smooth image. Furthermore, they minimized the ℓ_2 distance between the intermediate layers of the reconstructed image and the guiding image to enforce face orientations. As another approach, they also proposed a neural network including deconvolutional layers to reconstruct face images from facial templates, and they trained their deconvolutional neural network with the same loss function. Therefore, they did not use the original face image even in the training of their reconstruction network. The main drawback of their method is that the reconstructed face images look similar to the guiding image and do not reveal any privacy-sensitive information about the underlying user. We should also note that they only discussed the visual reconstruction quality and did not provide

[8]Some works also use the FR model for other types of optimization (i.e., non-gradient-based) in blackbox attacks. For instance, in [30] a particle swarm optimization was used.

TABLE I: Comparison with related works

Reference	method basis	Target FR model	whitebox /blackbox	Security Evaluation	Available source code
[8]	1) optimization 2) learning	FaceNet [9]	whitebox	None	✗
[10]	learning	FaceNet [9] VGG-Face [11]	whitebox + blackbox	histogram of scores (between original and reconstructed face images) on LFW [12] (only for FaceNet)	✗
[13]	learning	FaceNet [9]	blackbox	attack1 (against the original images) and attack2 (against a different face image of the same user) on LFW [12], Color FERET v2.0 [14], and FRGC [15]	✓
[16]	learning	ArcFace [7] FaceNet [9] SphereFace+ [17]	whitebox (+ blackbox by distillation of FR model)	matching accuracy (by substituting an image in each positive pair with its reconstructed version) on LFW [12], AgeDB [18], and CFP-FP [19]	✗
[20]	learning	ArcFace [7]	blackbox	attack1 (against the original images) and attack2 (against a different face image of the same user) on LFW [12] + 1 COTS PAD	✓
[21]	optimization	FaceNet [9]	blackbox	cosine and ℓ_2 distance between templates of the original and reconstructed face images on 20 random images of FFHQ [6]	✓
[22]	optimization	ArcFace [7] (3 different backbones)	blackbox	attack1 (against the original images) and attack2 (against a different face image of the same user) on LFW [12] and Color FERET v2.0 [14]+ 3 COTS PAD	✓
[23]	learning	ArcFace [7] ElasticFace [24] (3 different backbones)	whitebox + blackbox	SAR in entering FR systems by injecting the reconstructed face image on MOBIO [25], LFW [12]	✓
[26]	learning	ArcFace [7] ElasticFace [24] (4 different backbones)	whitebox + blackbox	SAR in entering FR system using the reconstructed face images from 3D reconstruction by 1) injection 2) presentation attack on MOBIO [25], LFW [12]	✓
Ours	learning	ArcFace [7] ElasticFace [24] AdaFace [27] EdgeFace [28] PocketNet [29] 17 SOTA backbones 9 SOTA heads	whitebox	SAR (at different FMRs) in entering FR system by injecting inverted versions of the templates which are stored in system's database on MOBIO [25], LFW [12], and AgeDB [18]	✓

any TI vulnerability evaluation of applying their methods on a FR system.

In [10], authors used a multi-layer perceptron (MLP) to find facial landmark geometry (optimized with mean squared error) and a convolutional neural network, including transposed convolutional layers, to generate face texture (optimized with mean absolute error) from the given templates. Then, they used a differentiable warping to combine estimated landmarks and textures to reconstruct the face images. In a whitebox scenario, they trained all the networks simultaneously using mean squared error (for landmark estimation), mean absolute error (for texture generation), and cosine distance between the input templates and the templates extracted from the reconstructed face (for the reconstructed face). However, in the blackbox scenario, they trained their MLP and CNN separately, and used the warping function only in the inference stage. From the security aspect, they only reported and compared the histogram of scores between templates extracted from the reconstructed images and original faces.

In [13], authors considered a blackbox scenario and trained convolutional neural networks to reconstruct face images. Inspired by DenseNet [32] and MemNet [33], they proposed two new blocks, named neighborly deconvolution blocks A/B (shortly, NbBlock-A and NbBlock-B), each of which includes deconvolutional and convolutional layers but differ in the concatenations. Based on these two blocks, they proposed two reconstruction networks, called NbNet-A and NbNet-B, by stacking corresponding NbBlocks. Moreover, they considered two different loss functions, pixel loss (mean absolute error of the reconstructed and original images) and perceptual loss (square of ℓ_2 norm of features at a middle layer of VGG-19 [34] when given the reconstructed and original images), and trained different networks with only one of these loss functions (resulting in four face reconstruction models in total). For the vulnerability analysis, they evaluated their reconstruction models against two types of attacks. In the first type of attack, they compared the reconstructed images against the original images, and in the second type of attack, they compared the reconstructed images against a different face image of the same user.

In [16], authors proposed a generative adversarial network (GAN) framework to reconstruct face images from face templates using bijection learning. They used the generator structure of PO-GAN [35] (with convolutional blocks) along with a feature conditional branch (with fully connected layers) and trained their generator network with a multi-term loss function, including an adversarial term (for optimizing the generator in a GAN-based framework), a bijection term (to learn distances in the bijection space), a distillation term (weighted summation of distances between layers of FR for the reconstructed and original face images), and a reconstruction term (mean absolute error of the reconstructed and original face images). In a whitebox scenario, they used the FR model to calculate the distillation loss term; however, in a blackbox scenario, they trained a new neural network (called student network) to mimic the FR network, and they used the new network instead of the FR model to calculate the distillation loss term in training their generator network. While they

reported the performance for their blackbox scenario to be close to that in the whitebox scenario, they have not provided more details (and no published source code), such as the network structure and training data, for learning the student networks. For the vulnerability analysis, they reported the matching accuracy by substituting an image in each positive pair with its reconstructed version and keeping the other image in both whitebox and blackbox scenarios.

In [21], authors considered a blackbox scenario and proposed a greedy random optimization using simulated annealing [36] on the latent space of StyleGAN [6] that can generate a reconstructed face image. In their proposed method, they repeatedly generated new guesses (for latent vectors of StyleGAN) by adding random noises to the previous guess, and compared the templates of the generated image by StyleGAN with the original templates. If the new guess could generate face image with a template closer to the original template, they updated their best guess with the new guess. For the vulnerability analysis, they randomly selected only 20 images from the the FFHQ [6] dataset and calculated cosine and ℓ_2 distance between the templates of the original and reconstructed face images. Similarly, in [22], an optimization on the latent space of StyleGAN was proposed, and authors used the standard Genetic Algorithm (GA) [37] to solve this optimization. For the security evaluation, they considered two types of attacks (similar to [13]) and evaluated the reconstructed face images. In addition, they used three commercial-off-the-shelf (COTS) presentation attack detection (PAD) systems to evaluate the reconstructed face images. Along the same lines of using StyleGAN for face reconstruction, Dong *et al.* [20] trained a MLP regression using the mean squared error (MSE) loss function to map the templates to the StyleGAN latent vector. For the security evaluation, they evaluated the reconstructed face images in two types of attacks (similar to [13]) and using a COTS PAD. Similarly, in [23] a learning-based approach was used to map facial templates to the intermediate latent space of StyleGAN using adversarial training. The method can be applied for both whitebox and blackbox scenarios. In the whitebox scenario, the same feature extractor was used in the loss function, but in the blackbox scenario a different FR model (which adversary has access) was instead used in the loss function. For the security evaluation, the reconstructed face images were injected to the FR system to evaluate the vulnerability of the FR system. The authors also explored important area in the reconstructed face image and investigated transferability of reconstructed face images.

In [26], a geometry-aware face reconstruction method (called GaFaR) was proposed to reconstruct 3D face from facial templates. They used a semi-supervised learning approach to learn a mapping from facial templates to the intermediate latent space of a face generator network based on Generative Neural Radiance Fields (GNeRF). In the supervised learning part, synthetic training images were used, where the intermediate latent codes are available, and in the unsupervised learning part, a GAN-based learning was deployed to learn the intermediate latent space. Then, the mapped intermediate latent codes were used along with camera parameters to generate reconstructed face image through generator and renderer part of GNeRF model. Therefore, after finding the mapped intermediate latent code, the adversary can generate frontal reconstruction of face image. In addition, the adversary can generate any arbitrary pose using the GNeRF model, and thus can find the pose image that maximize the success attack rate using greed search (GS) or continuous optimization (CO) on the camera parameters of the GNeRF model. Similar to [23], the method can be applied for both whitebox and blackbox scenarios (with a proxy FR model in the loss function for blackbox). For security evaluation, the reconstructed face images were used to inject as query to the target FR system and the transferability across different FR systems were also evaluated. In addition, the reconstructed face images were used to perform practical presentation attack using digital reply and printed photograph with different settings.

Table I compares our work with the previous works in the literature on the inversion of templates extracted by deep FR models. In summary, our contributions include proposing a novel neural network structure (based on the proposed DSCas-Conv block) and using a multi-term loss function to train our network. In addition, we consider a real-world attack to FR systems and propose an open-source evaluation framework to evaluate vulnerability of FR systems. We evaluate the vulnerability of 31 different SOTA FR models (including different backbones and different heads), in terms of an adversary's Success Attack Rate[9] (SAR) in our framework. The SAR refers to the attacker's success rate when they attempt to impersonate an enrolled user of the FR system, using the image reconstructed by inverting the user's template. The SAR is computed at different False Match Rate (FMR) configurations, which refer to the FR system's decision thresholds. It is also noteworthy that the source code of all the experiments in this paper is publicly available to help other researchers reproduce our results as well as to allow them to use our framework to evaluate the vulnerability of their own FR systems to a TI attack[10].

III. Proposed Framework

In this section, we introduce our proposed framework (as depicted in Fig. 2), to evaluate the vulnerability of FR systems to a TI attack. First, we describe the threat model that we consider in this study, in section III-A. Then, we propose our face reconstruction network in section III-B. Finally, we describe our vulnerability evaluation protocol in section III-C.

A. Threat model

To evaluate the vulnerability of a given FR system, we need to first define the threat model that characterises the adversary on which we wish to base our vulnerability analysis [2], [13], [38]. Considering a real-world attack scenario[11], we define the following properties for the adversary:

[9]Also referred to as Attack Success Rate (ASR).

[10]Source code: https://gitlab.idiap.ch/bob/bob.paper.tifs2024_face_ti

[11]We should note that our threat model is aligned with the *full-disclosure* scenario defined in the ISO/IEC 30136 standard [38] for evaluating invertibility of biometric templates.

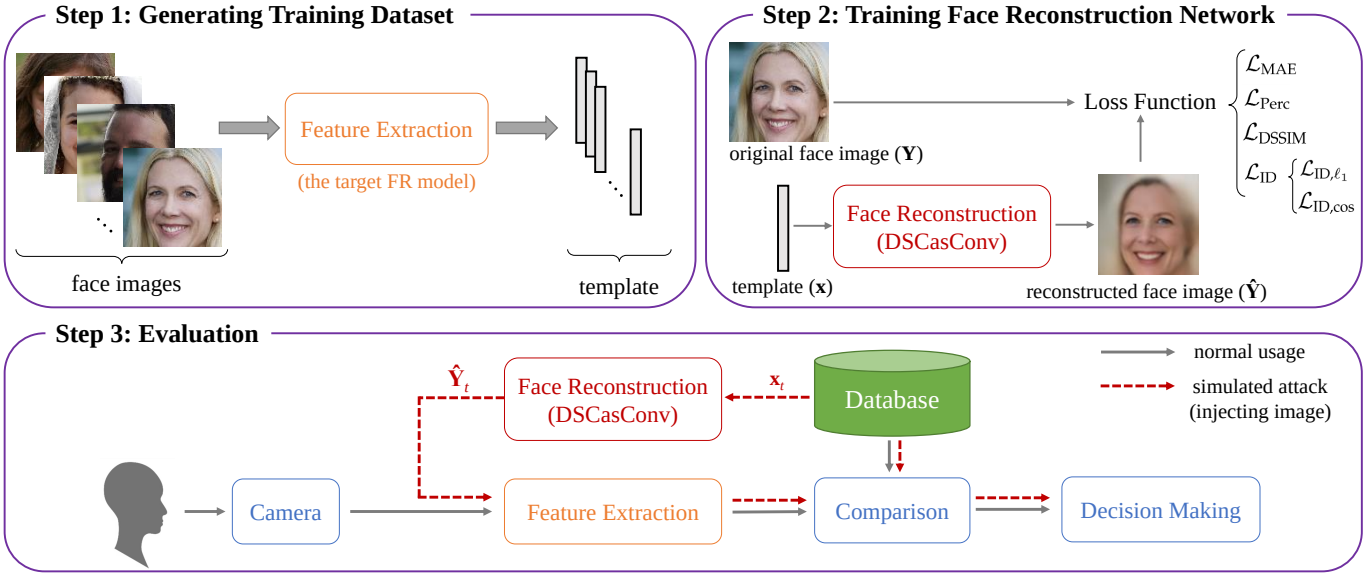


Fig. 2: Block diagram of the proposed framework

- *Adversary's goal:* The attacker aims to impersonate a user enrolled in the target FR system.
- *Adversary's knowledge:* The attacker is assumed to have only the following information:
  1) The target face templates $\mathbf{x}_t$ of a user enrolled in the system's database.
  2) The whitebox knowledge (including structure and internal values, e.g., CNN parameters) of the feature extraction model $F(.)$, which can be used to generate a face template $\mathbf{x} = F(\mathbf{I})$ from a face image $\mathbf{I}$.

  However, the attacker is assumed not to have any other information, neither from the target system nor from the target template. In particular, the attacker is assumed not to have the following information:
  - Any additional information or prior knowledge about the identity of the target template, including age, gender, etc.
  - Any information about the training set of the feature extraction model. Therefore, the attacker is assumed not to be able to use the same (or similar) dataset to learn TI.
  - Any knowledge about the comparison and decision making submodules of the target system, including the similarity score function and the system's decision threshold.
- *Adversary's capability:* The attacker is assumed to have the following capabilities:
  1) The attacker can inject the reconstructed face images directly into the feature extractor of the target system and bypass the sensor (i.e., camera).
  2) For each target template, the attacker is allowed only one attempt to enter the system.
- *Adversary's strategy:* Under the above assumptions, the attacker can reconstruct face image $\hat{\mathbf{Y}}_t = G_W(\mathbf{x}_t)$ from the target template $\mathbf{x}_t$ using a reconstruction model $G_W(.)$. Then, the attacker can use the reconstructed face image $\hat{\mathbf{Y}}_t$ as a query to enter the target FR system. The weights W of the reconstruction model $G_W(.)$ can be learned using a dataset of face images and their corresponding face templates extracted by the feature extractor model $F(.)$.

B. Face Reconstruction Network

In this section, we introduce our neural network $G_W(.)$ to invert a face template $\mathbf{x}$ and reconstruct a face image $\hat{\mathbf{Y}} = G_W(\mathbf{x})$. To train our network, we first need to generate training data, including pairs of face template $\mathbf{x}$ and face image $\mathbf{Y}$, which is described in section III-B1. We train our network with a multi-term loss function as described in section III-B2. Our network structure, which includes multiple DSCasConv blocks, is described in section III-B3.

1) Generating Training Data: To generate our training dataset, let us assume that we have a dataset of face images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$, where $\mathbf{I}_i$ and N indicate the ith image and the total number of images, respectively. Also, let us assume that we have the coordinates of the facial landmarks (e.g., eyes), $\mathbf{L}_i$, for image $\mathbf{I}_i$ in the dataset $\mathcal{I}$. In addition, let $A(\mathbf{I}_i, \mathbf{L}_i)$ denote the function used prior to feature extraction, which accepts $\mathbf{I}_i$ and $\mathbf{L}_i$ as inputs and returns an aligned and cropped face image. We can generate our training dataset of aligned images and associated templates, $\mathcal{D}$, by extracting facial features from all face images in the face dataset $\mathcal{I}$ after alignment:

$$\mathcal{D} = \{([F \circ A](\mathbf{I}_i, \mathbf{L}_i), A(\mathbf{I}_i, \mathbf{L}_i))\}_{i=1}^{N}, \tag{1}$$

where $F(.)$ indicates the feature extraction model.

However, our experiments show that an augmented dataset can improve the generalization and performance of our TI network. So, we augment the dataset $\mathcal{I}$ and generate a new

dataset $\mathcal{I}_a = \{\mathbf{I}_{a,j}\}_{j=1}^{M}$ using a random transformation function $T(.)$, where M is the number of images in the augmented dataset $\mathcal{I}_a$, and $\mathbf{I}_{a,j}$ is an image augmented by $T(.)$, i.e., $\mathbf{I}_{a,j} = T(\mathbf{I}_k)$, $0 \leq k < N$. To increase the robustness of the inversion network, we also add random noise to the coordinates of landmarks before feature extraction. However, in our augmented training dataset, $\mathcal{D}_a$, we pair up each extracted feature with its corresponding aligned face using the original values of landmark coordinates (i.e., without noise). Hence, we generate the augmented training dataset $\mathcal{D}_a$ as follows:

$$\mathcal{D}_a = \{([F \circ A](\mathbf{I}_{a,j}, \mathbf{L}_j + \mathbf{N}_j), A(\mathbf{I}_{a,j}, \mathbf{L}_j))\}_{j=1}^{M}, \quad (2)$$

where $\mathbf{N}_j$ is random noise with a uniform distribution in $(-\delta, \delta)$. We consider $\delta = 4$ in our experiments. It is worth mentioning that using the original facial landmark coordinates in feature extraction helps our inversion network to generate all images with the same alignment and thus eliminates the additional work of finding the landmark coordinates and reconstructing face images in different locations. Adding noise to the coordinates of landmarks before feature extraction increases the variation in the template space, thereby enhancing the robustness of our inversion network to the alignment.

It is worth mentioning that for the random transformation function $T(.)$ in this paper, we use a random combination of the following transformations: random PCA color augmentation [39], randomly adjusting contrast, randomly adjusting brightness, Gaussian blurring (random standard deviation), and JPEG compression (random compression rate).

For simplicity, in the rest of the paper, the augmented training dataset in Eq. 2 is denoted as $\mathcal{D}_a = \{(\mathbf{x}_j, \mathbf{Y}_j)\}_{j=1}^{M}$, where $\mathbf{x}_j = [F \circ A](\mathbf{I}_{a,j}, \mathbf{L}_j + \mathbf{N}_j)$ and $\mathbf{Y}_j = A(\mathbf{I}_{a,j}, \mathbf{L}_j)$.

2) Loss Function: To train the reconstruction network, $G_W(.)$, we optimize its weights W using loss function $\mathcal{L}(.,.)$ on the augmented training dataset $\mathcal{D}_a$ such that:

$$W^* = \underset{W}{\operatorname{argmin}} \underset{(\mathbf{x},\mathbf{Y})\in\mathcal{D}_a}{\mathbb{E}} \mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y}), \quad (3)$$

where $(\mathbf{x}, \mathbf{Y})$ denotes a pair of face template $\mathbf{x}$ and face image $\mathbf{Y}$ in our augmented training dataset $\mathcal{D}_a$, and $\hat{\mathbf{Y}} = G_W(\mathbf{x})$ is the reconstructed face image. To this end, we define a multi-term loss function including:

- *Mean Absolute Error (MAE):* To help the network to generate a face image that is similar to the original image, we use the Mean Absolute Error (MAE) loss term, which includes the ℓ_1-norm of the reconstruction error:

$$\mathcal{L}_{\text{MAE}}(\hat{\mathbf{Y}}, \mathbf{Y}) = ||\hat{\mathbf{Y}} - \mathbf{Y}||_1 \quad (4)$$

- *Dissimilarity Structural Index Metric (DSSIM):* In addition to MAE of the reconstructed face, we maximize the objective quality of the reconstructed image. To this end, we use the Similarity Structural Index Metric (SSIM) [40] of the reconstructed image and optimize the DSSIM loss term as follows:

$$\mathcal{L}_{\text{DSSIM}}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1 - \text{SSIM}(\hat{\mathbf{Y}}, \mathbf{Y})}{2} \quad (5)$$

- *Perceptual Loss:* In addition to DSSIM, we use a perceptual loss by minimizing the ℓ_1-norm of the difference between the features extracted from $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ by a convolutional neural network trained on ImageNet [41]. This helps the model to generate images with a similar representation of the original image (i.e., face). In this paper, we use a pre-trained VGG-16 [34] model and consider its middle feature maps to calculate the perceptual loss. Let us denote the feature mapping of VGG-16 as $P(.)$. Then the perceptual loss can be expressed as:

$$\mathcal{L}_{\text{Perc}}(\hat{\mathbf{Y}}, \mathbf{Y}) = ||P(\hat{\mathbf{Y}}) - P(\mathbf{Y})||_1 \quad (6)$$

- *ID loss:* In addition to the aforementioned loss terms, we would like the templates extracted from the reconstructed face image to be close to the templates of the original face image, to increase the chances of a successful TI attack. So, we minimize the distance between the templates extracted from the reconstructed face $\hat{\mathbf{Y}}$ and original face $\mathbf{Y}$. To achieve this, we minimize the ℓ_1-norm of the difference between the extracted features and also maximize their cosine similarity. Thereby, we define ID loss with two terms as follows:

$$\begin{aligned}\mathcal{L}_{\text{ID}}(\hat{\mathbf{Y}}, \mathbf{Y}) &= \mathcal{L}_{\text{ID},\ell_1}(\hat{\mathbf{Y}}, \mathbf{Y}) + \mathcal{L}_{\text{ID,cos}}(\hat{\mathbf{Y}}, \mathbf{Y}) \\ &= \underbrace{||F(\hat{\mathbf{Y}}) - F(\mathbf{Y})||_1}_{\text{minimizing } \ell_1\text{-norm}} + \underbrace{\frac{-F(\hat{\mathbf{Y}}).F(\mathbf{Y})}{||F(\hat{\mathbf{Y}})||_2.||F(\mathbf{Y})||_2}}_{\text{maximizing cosine similarity}}\end{aligned} \quad (7)$$

We use a linear combination of the above loss terms as the total loss:

$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \alpha_1 \mathcal{L}_{\text{DSSIM}} + \alpha_2 \mathcal{L}_{\text{Perc}} + \alpha_3 \mathcal{L}_{\text{ID}}, \quad (8)$$

where α_1, α_2, and α_3 are weights of each loss term. To train the proposed face reconstruction network with this multi-term loss function, we use the Adam [42] optimizer with the initial learning rate of 10^{-3} and decrease the learning rate by a factor of 0.5 every 10 epochs.

3) Network Structure: To reconstruct face images from their corresponding templates, we can use deconvolution layers to build our face reconstruction network (e.g., [8], [13]). However, since deconvolution acts as upsampling, the deconvolution output suffers from insufficient detail [13]. In particular, similar to upsampling, in deconvolution, local dependencies are weakened, which leads to a blurry output. To address these shortcomings, we propose a new block, called "enhanced deconvolution using cascaded convolutions and skip connections" (shortly, *DSCasConv*), which generates outputs with the same size as the deconvolution layer. In the proposed block, we apply multiple cascaded convolutional layers on the deconvolution output. Considering the significant effect of residual learning [43], we also use skip connections to further enhance the output by forcing the convolutional layers to learn residuals. In addition, skip connections can enhance the gradient flow and prevent gradient vanishing problem [43] in our deep DSCasConv block. Hence, we use a skip connection for each of the convolution layers as well as a skip connection over all cascaded convolution layers in our DSCasConv block. Using cascaded convolutional layers with skip connections after deconvolution can recover and improve local dependencies, and therefore result in sharper

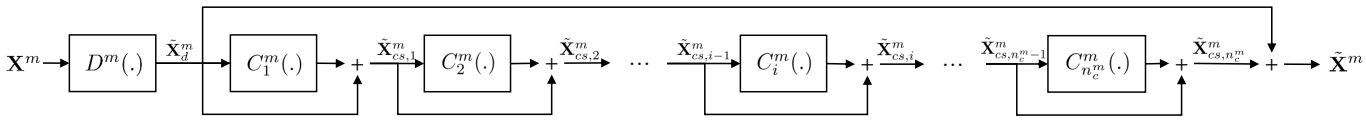


Fig. 3: Block diagram of the mth DSCasConv block

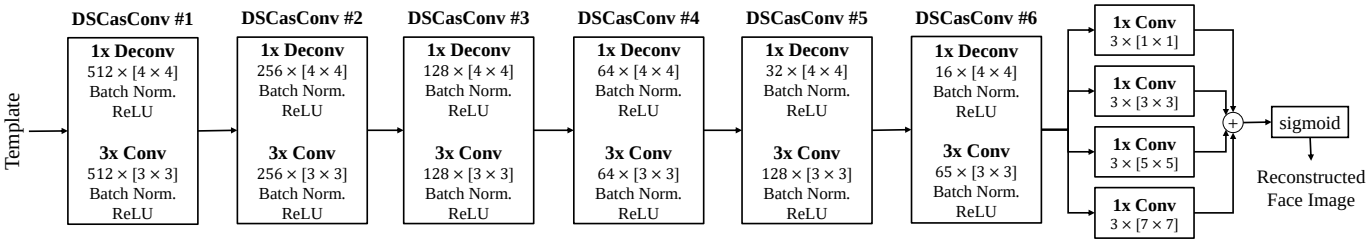


Fig. 4: Structure of the proposed face reconstruction network

and more detailed outputs. Indeed, each residual convolutional layer enhances the result of previous layers, and in total, the residual cascaded convolutional layers improves the deconvolution output.

To formulate the proposed block, let $\mathbf{X}^m$ and $\tilde{\mathbf{X}}^m$ denote the input and output of the mth DSCasConv block, respectively. Assume that the mth DSCasConv block consists of the deconvolution operator, $D^m(.)$, and a set of convolution operations, $\mathcal{C}_m = \{C_i^m(.)|i = 1, .., n_c^m\}$, where n_c^m is the number of convolution operations at the mth DSCasConv block. Let us define $\tilde{\mathbf{X}}_d^m = D^m(\mathbf{X}^m)$ as the output of the deconvolution operator and $\tilde{\mathbf{X}}_{cs,i}^m$ as the summation of the ith convolution operation and its corresponding skip connection as follows:

$$\tilde{\mathbf{X}}_{cs,i}^m = \begin{cases} \tilde{\mathbf{X}}_{cs,i-1}^m + C_i^m(\tilde{\mathbf{X}}_{cs,i-1}^m) & \text{if } i > 0 \\ \tilde{\mathbf{X}}_d^m & \text{if } i = 0 \end{cases}. \tag{9}$$

Then, we define the output of the mth DSCasConv block as below:

$$\tilde{\mathbf{X}}^m = \tilde{\mathbf{X}}_d^m + \tilde{\mathbf{X}}_{cs,n_c^m}^m \tag{10}$$

Fig.3 illustrates the block diagram of the mth DSCasConv block.

We build our network with 6 DSCasConv blocks (each includes 1 deconvolution and 3 convolution operations) with 512, 256, 128, 64, 32, 16 filters, respectively. For deconvolution and convolution layers in our DSCasConv blocks, we use kernels of sizes 4 and 3, respectively. In addition, we use Batch Normalization [44] and a rectified linear unit (ReLU) after each deconvolution and convolution operation in our DSCasConv blocks. Finally, we pass the output of the last DSCasConv block to 4 parallel convolutional layers with different kernel sizes (including sizes of 1, 3, 5, and 7), which are added and passed through a sigmoid function to generate the final reconstructed face. Fig. 4 depicts the general structure of our face reconstruction neural network.

C. TI Vulnerability Evaluation Protocol

To evaluate the vulnerability of a FR model to a TI attack, we consider a real-world scenario based on the assumptions described in section III-A. To this end, we consider a FR system with several enrolled users. Based on our threat model, we assume that the attacker can access the system's database and aims to invert the enrolled templates to reconstruct the underlying face images. The images are then injected into the feature extractor to impersonate the enrolled users, and therefore, enter the system. We should note that there might be several templates for each user stored in the system's database, but according to our threat model, the attacker does not have any knowledge of this.

To train the face reconstruction network, based on our threat model we assume that the attacker does not have any information about the training set of the feature extractor. Therefore, the attacker uses a different training dataset for the inversion model. So, we have three different datasets in our evaluation: 1) dataset used for training the feature extractor (by the system designer), 2) dataset enrolled in the FR system, and 3) dataset used for training the TI model (by the attacker).

We assume that the attacker trains the inversion model (as in section III-B), then uses the inversion model to invert the enrolled templates, and injects the reconstructed face images into the system. Again according to our threat model, the attacker is allowed only one attempt to enter the system for each inverted target template. Hence, in our evaluation, for each template stored in the system's database, we invert the template and reconstruct the face image. Then, we extract the template from the reconstructed face image and find the system's comparison score between this template and the corresponding reference templates. If the score is greater than the system's threshold, the attack is considered successful, meaning that the attacker can enter the system. Hence, similar to the Receiver Operating Characteristic (ROC) plot, we use the comparison scores of inverted templates to plot the Success Attack Rate (SAR) versus the system's False Match Rate (FMR) by changing the system's decision threshold in the impostors' score range. This plot can be used to compare the vulnerability of FR models at different FMRs. Fig. 2 illustrates the general block diagram of the proposed TI evaluation

framework.

IV. Experiments

In this section, we describe the experiments used to evaluate our framework and analyze the vulnerability of SOTA FR models using this framework. First, in section IV-A, we describe the experimental setup and the FR models used in our experiments. Next, as a primary experiment, we evaluate the vulnerability of the iResNet100-ArcFace [7] model, which is a well-known SOTA FR model, in section IV-B. Next, we compare our proposed face reconstruction method with previous methods in the literature in section IV-C against the iResNet100-ArcFace [7] model. Then, in section IV-D, we provide an ablation study on the effect of our network structure (section IV-D2) and our loss function (section IV-D1) on the performance of the primary experiment. After evaluating our proposed framework, in section IV-E we evaluate the vulnerability of different SOTA FR models, with different backbones (section IV-E1) and different heads (section IV-E2), using our framework. Finally, we discuss the experimental findings in section IV-F.

A. Experimental Setup

In our experiments, we evaluate the vulnerability of SOTA FR models to our TI attack. For the primary experiment, we use the iResNet100-ArcFace[12] [7] model to study the performance of our face reconstruction network and compare the proposed face reconstruction network with previous methods in the literature. Furthermore, we evaluate the vulnerability of different FR model backbones, including MobileFaceNet [45], ResNet [43], SE-ResNet [46], HRNet [47], EfficientNet [48], GhostNet [49], AttentionNet [50], TF-NAS [51], ResNeS t [52], ReXNet [53], RepVGG [54], LightCNN [55], and Swin [56]. Moreover, we evaluate the vulnerability of FR models which are trained with different loss functions (different heads), including AM-Softmax [57], ArcFace [7], AdaCos [58], AdaM-Softmax [59], CircleLoss [60], CurricularFace [61], MV-Softmax [62], NPCFace [63], and MagFace [64]. We should note that when we compare the vulnerability of different backbones, we use the same head (i.e., MV-Softmax [62]) for all the models, and when we compare the vulnerability of different heads, we use the same backbone (i.e., MobileFaceNet [45]) for all the models. In addition to iResNet100-ArcFace and also SOTA backbones and heads, we also evaluate the vulnerability of four other pretrained SOTA FR models in the literature, including ElasticFace [24], AdaFace [27], EdgeFace [28], and PocketNet [29]. All the aforementioned FR models, except EdgeFace, are trained on the MS-Celeb-1M dataset [65], and EdgeFace is trained on the WebFace260M dataset [66].

To train our face reconstruction network for each FR model, we use the FFHQ dataset [6], which consists of 70,000 face images (with no identity labels) and includes variations in terms of age, ethnicity, accessories, and image background. For a fair comparison, we train each of the face reconstruction networks with 90 epochs. We use a random 90% portion of the FFHQ dataset to generate the training set as explained in section III-B1, by generating 10 augmented images for each original image. The remaining 10% portion is used for validation. After training our face reconstruction network, we use the MOBIO [25], Labeled Faced in the Wild (LFW) [12], and AgeDB [18] datasets to build the FR systems and evaluate their vulnerability in our framework. The MOBIO dataset is a bimodal dataset including audio and face data acquired using mobile devices from 150 people. We use the *development* subset of the *mobio-all* protocol[13] in our experiments. The LFW database includes 13,233 images of 5,749 people, where 1,680 people have two or more images. We use the *View 2* protocol[14] to evaluate the models. The AgeDB [18] dataset contains 16,488 images of 568 famous people. The minimum and maximum age in this dataset are 1 and 101, respectively, and the average age range for each subject is 50.3 years. We use the *30-year* protocol (i.e., the age difference of each pair's faces is equal to 30) in our experiments.

TABLE II: Recognition performance (TMR) and vulnerability to template inversion attack (SAR) of the iResNet100-ArcFace model, at FMR $= 10^{-2}$ and FMR $= 10^{-3}$ on the MOBIO, LFW, and AgeDB datasets.

Dataset	$\mathbf{FMR = 10^{-2}}$		$\mathbf{FMR = 10^{-3}}$	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
MOBIO	100.00	100.00	100.00	100.00
LFW	97.70	97.44	96.63	96.48
AgeDB	96.97	94.03	93.77	84.60

In our experiments, we use the Bob[15] toolbox [67] to both build the FR systems and develop the TI evaluation framework. We use the PyTorch package to train the face reconstruction networks. For the implementation of iResNet100-ArcFace, we use its official pretrained (InsightFace)[16] model ported into PyTorch[17]. For the other SOTA FR models with different backbones and different heads, we use the FaceX-Zoo[18] [68] toolbox. For other FR models, we use the pretrained models from their corresponding repository. In our experiments, we consider $\alpha_1 = 0.75$, $\alpha_2 = 0.02$, and $\alpha_3 = 0.025$ as the weights of our loss function in Eq. 8. We also provide an ablation study on the effect of these weights in section IV-D1. The source code of the experiments is publicly available to help reproduce the results[19].

B. Primary Experiment

As a primary experiment, we evaluate the vulnerability of the iResNet100-ArcFace [7] model using our framework, on the MOBIO, LFW, and AgeDB datasets. To this end, as

[12]iResNet100 backbone trained with ArcFace loss.

[13]The implementation of the *mobio-all* protocol for the MOBIO dataset is available at https://gitlab.idiap.ch/bob/bob.db.mobio

[14]The implementation of the *View 2* protocol for the LFW dataset is available at https://gitlab.idiap.ch/bob/bob.db.lfw

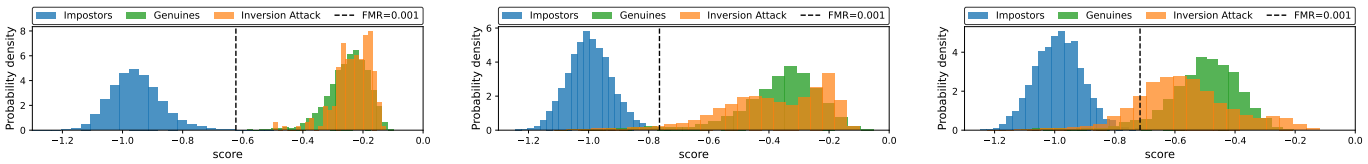
[15]Available at https://www.idiap.ch/software/bob

[16]Available at: https://github.com/deepinsight/insightface

[17]Available at https://gitlab.idiap.ch/bob/bob.bio.face

[18]Available at https://github.com/JDAI-CV/FaceX-Zoo

[19]Source code: https://gitlab.idiap.ch/bob/bob.paper.tifs2023_face_ti



(a) MOBIO (b) LFW (c) AgeDB

Fig. 5: Histogram of scores (negative cosine distance) for genuine templates and impostor templates, as well as scores for the templates extracted from reconstructed face images when injecting the reconstructed face images into the system using (a) MOBIO, (b) LFW, and (c) AgeDB datasets.

discussed in section III-C, we train our face reconstruction network using the FFHQ dataset. Fig. 1 depicts the reconstructed face images of sample faces in our validation set from the FFHQ dataset. Next, we use our face reconstruction network to invert the facial templates stored in the FR system and inject the reconstructed face images into the system. Fig. 5 illustrates the histogram of scores for genuine and impostor templates, as well as scores for the templates extracted from reconstructed face images when injecting the reconstructed face images into the FR system. As this figure shows, the scores between the templates extracted from the reconstructed face images and the reference templates enrolled in the system's database are close to the genuine scores and, therefore, are likely to break the system. Table II reports the recognition performance and vulnerability of the iResNet100-ArcFace model to a TI attack in terms of True Match Rate (TMR) and SAR, respectively, at FMR $= 10^{-2}$ and FMR $= 10^{-3}$, on the MOBIO, LFW, and AgeDB datasets. As this table shows, while the iResNet100-ArcFace model achieves high recognition performance on the MOBIO, LFW, and AgeDB datasets, it is seriously vulnerable to our whitebox TI attack.

C. Comparison with Previous Methods

We compare the performance of our face reconstruction network with the methods proposed[20] in [8], [13], [20]–[23], [26] in attacks against iResNet100-ArcFace model. As mentioned in Table I, [13], [20], [21] are based on the blackbox scenario[21] and [8] is based on the whitebox scenario. Methods in [23], [26] can be used for both whitebox and blackbox scenarios, and we use their whtebox implementation in our experiments. For each method, we train a separate face reconstruction networks using FFHQ dataset, and use the trained network to invert facial templates stored in the FR system. We use the reconstructed face images to inject into the system and evaluate vulnerability of iResNet100-ArcFace on the MOBIO, LFW, and AgeDB datasets using our framework. Table III compares the performance of these different methods in terms of SAR in attacks against the iResNet100-ArcFace model on the MOBIO, LFW, and AgeDB datasets. As this table shows,

TABLE III: Comparison of different face reconstruction methods against the iResNet100-ArcFace model in terms of SAR at FMR=10^{-2} and FMR=10^{-3} on the MOBIO, LFW, and AgeDB datasets.

method	MOBIO		LFW		AgeDB	
	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
Zhmoginov and Sandler [8]	100.00	85.71	93.01	85.87	82.40	54.18
NBNetA-M [13]	2.86	0.48	16.06	5.35	3.75	0.42
NBNetA-P [13]	15.24	1.43	29.61	12.16	8.26	1.14
NBNetB-M [13]	19.52	0.48	26.10	10.79	6.06	0.49
NBNetB-P [13]	51.90	21.9	60.33	39.49	21.56	5.18
Dong *et al.* [20]	24.85	3.33	28.21	13.21	9.56	1.80
Vendrow and Vendrow [21]	69.52	29.05	77.00	57.70	40.94	16.56
Dong *et al.* [22]	85.71	58.57	87.25	75.31	58.79	43.22
Otroshi Shahreza and Marcel [23]	100.00	92.38	93.64	86.82	75.87	62.08
GaFaR [26]	95.71	82.86	89.27	79.84	63.30	48.94
GaFaR+GS [26]	97.62	85.23	90.77	82.52	67.86	53.10
GaFaR+CO [26]	97.62	89.05	91.87	84.25	71.95	58.00
[Ours]	**100.00**	**100.00**	**97.44**	**96.48**	**94.03**	**84.60**

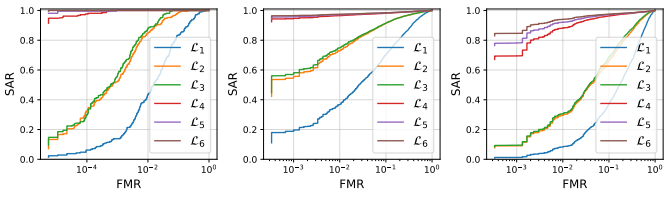
our proposed method outperforms previous methods in [8], [13], [20]–[23], [26]. In particular, our method achieves better performance compared to low-resolution face reconstruction methods (i.e., [8], [13]). This is achieved as the result of our network structure and loss function which are further studied in section IV-D. Comparing other methods which generate high-resolution face images (i.e., [20]–[23]) or 3D face (i.e., [26]), the reconstructed face images by our method still achieve superior performance, which elaborates on a trade-off between resolution of reconstructed face images and performance in terms of SAR in our method and these methods in the literature. Among different face reconstruction methods in the literature, [23], [26] achieve the best performance after our proposed method in attack against FR systems and generate high-resolution and 3D face, respectively.

D. Ablation Study

In this section, we describe our ablation study on the effect of network structure (section IV-D2) and loss function (section IV-D1) on the face reconstruction performance. In our ablation studies, we consider a FR system based on the iResNet100-

[20]The source codes of other methods in Table I are not publicly available and we could not reproduce their results.

[21]We should note that having access to whitebox model is a realistic assumption in many practical cases, and the question remains how the adversary can perform a successful attack in whitebox scenario. In such a case, the adversary may use whitbox or blackbox methods to reconstruct face images. In the case of blackbox methods, not using the knowledge of the available whitebox model is the limitation of blackbox methods in whitebox attacks, but the adversary can still use blackbox methods for the attack.



(a) MOBIO (b) LFW (c) AgeDB

Fig. 6: Effect of loss function (as in Eq.11) on the performance of our face reconstruction network in a template inversion attack against a face recognition system based on the iResNet100-ArcFace model evaluated using (a) MOBIO, (b) LFW, and (c) AgeDB datasets.

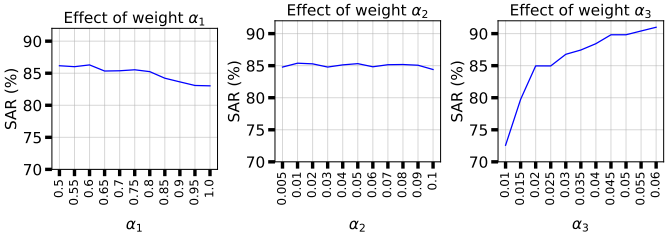


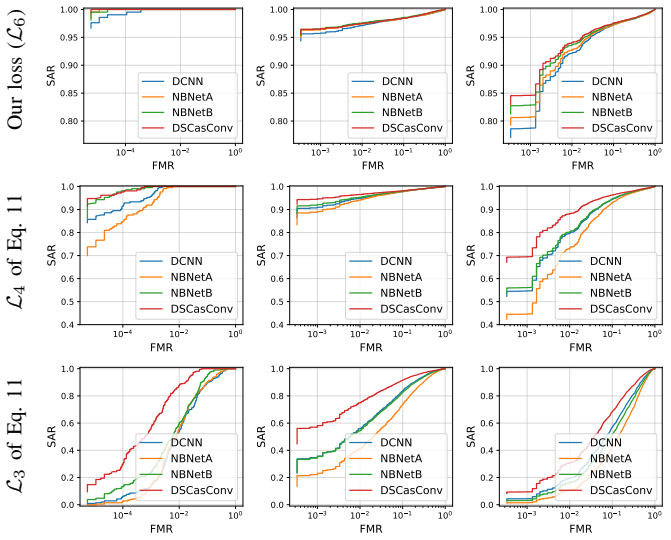
Fig. 7: Effect of weights (α_1, α_2, and α_3) in our loss function (as in Eq. 8) on the performance of our face reconstruction network in a template inversion attack against a face recognition system based on the iResNet100-ArcFace model evaluated on the AgeDB dataset.

ArcFace model, and we evaluate the SAR over different values of the system's FMR using the MOBIO, LFW, and AgeDB datasets.

1) Ablation Study on the Loss Function: To evaluate the effect of each loss term, we train different face reconstruction networks with different loss functions. Considering our multi-term loss function in Eq. 8, let us denote linear combinations of different loss terms as follows:

$$\begin{aligned}
\mathcal{L}_1 &= \mathcal{L}_{\text{MAE}},\\
\mathcal{L}_2 &= \mathcal{L}_{\text{MAE}} + \alpha_1\mathcal{L}_{\text{DSSIM}},\\
\mathcal{L}_3 &= \mathcal{L}_{\text{MAE}} + \alpha_1\mathcal{L}_{\text{DSSIM}} + \alpha_2\mathcal{L}_{\text{Perc}},\\
\mathcal{L}_4 &= \mathcal{L}_{\text{MAE}} + \alpha_1\mathcal{L}_{\text{DSSIM}} + \alpha_2\mathcal{L}_{\text{Perc}} + \alpha_3\mathcal{L}_{\text{ID},\ell_1},\\
\mathcal{L}_5 &= \mathcal{L}_{\text{MAE}} + \alpha_1\mathcal{L}_{\text{DSSIM}} + \alpha_2\mathcal{L}_{\text{Perc}} + \alpha_3\mathcal{L}_{\text{ID,cos}},\\
\mathcal{L}_6 &= \mathcal{L}_{\text{MAE}} + \alpha_1\mathcal{L}_{\text{DSSIM}} + \alpha_2\mathcal{L}_{\text{Perc}} + \alpha_3\mathcal{L}_{\text{ID}},
\end{aligned} \tag{11}$$

where $\mathcal{L}_{\text{ID}} = \mathcal{L}_{\text{ID},\ell_1} + \mathcal{L}_{\text{ID,cos}}$ as in Eq. 7. Fig. 6 compares the performance of face reconstruction networks trained with these different loss functions. As this figure shows, each of the terms enhances the performance of the face reconstruction network. In particular, using either of the terms in ID loss improves the performance, but using both results in the best performance. However, we should note that ID loss terms require full knowledge of the FR model (i.e., whitebox scenario), which is the assumption we make in the evaluations presented in this paper. To further investigate the effect of weights α_1, α_2, and α_3 in our loss function, we perform an ablation study where we change the value of each of these weights and keep the other ones unchanged. Fig. 7 shows the effect of these weight in the performance of our method on the AgeDB dataset. The



(a) MOBIO (b) LFW (c) AgeDB

Fig. 8: Effect of network structure on the face reconstruction performance, when trained with our loss function (first row) as well as $\mathcal{L}_4$ (second row) and $\mathcal{L}_3$ (third row) of Eq. 11, in a template inversion attack against a face recognition system based on the iResNet100-ArcFace model, evaluated using (a) MOBIO, (b) LFW, and (c) AgeDB datasets.

(a) original (b) Deconv (c) NBNet-A (d) NBNet-B (e) DSCasConv

Fig. 9: Sample face images from the FFHQ dataset and their reconstructed images from iResNet100-ArcFace templates using face reconstruction networks based on different blocks: (a) original image, (b) deconvolution, (c) NBNet-A, (d) NBNet-B, and (e) DSCasConv.

results in this figure show that compared to other weights, α_3 is very sensitive and has a significant effect on the performance of our model. In other words, ID loss has the most contribution to the performance of our method, which is aligned with our ablation study in Fig. 6.

2) Ablation Study on the Network Structure: For evaluating the efficacy of the proposed network, we train several face reconstruction networks with similar structures but built

TABLE IV: Comparison of face recognition models with different SOTA backbones and the same head in terms of the number of parameters, MACs, recognition performance (TMR), and vulnerability to template inversion (SAR) at FMR $= 10^{-3}$ on the MOBIO, LFW, and AgeDB datasets.

Model	Params	MACs	MOBIO		LFW		AgeDB	
			↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
MobileFaceNet	1.19M	227.57M	89.86	98.57	53.47	67.49	65.70	69.48
Resnet50	43.57M	6.31G	98.00	100.00	72.87	72.77	89.67	75.97
Resnet152	71.14M	12.33G	98.00	100.00	77.07	73.96	90.63	72.94
HRNet	70.63M	4.35G	98.34	99.52	79.77	79.88	88.50	68.70
EfficientNet-B0	33.44M	77.83M	94.13	99.05	61.53	70.44	69.37	66.49
TF-NAS-A	39.59M	534.41M	92.59	99.52	62.87	68.80	78.27	69.37
LightCNN-29	11.60M	2.84G	87.76	93.81	52.70	66.15	56.17	57.40
GhostNet	26.76M	194.49M	87.51	100.00	55.00	65.05	49.90	48.75
Attention-56	98.96M	6.34G	98.75	99.52	67.33	69.05	87.30	67.04
Attention-92	134.56M	10.62G	98.12	99.05	73.00	74.32	92.13	72.46
ResNeSt50	76.79M	5.55G	99.02	97.62	89.37	86.97	91.53	71.51
ReXNet	15.20M	429.64M	92.15	97.62	67.57	73.30	70.03	62.70
RepVGG-A0	39.94M	1.55G	89.77	99.05	45.43	57.87	72.37	66.01
RepVGG-B0	46.65M	3.44G	93.58	95.71	44.80	52.85	83.60	75.85
RepVGG-B1	106.75M	13.21G	96.87	98.10	62.70	62.66	85.17	65.91
Swin-T	46.74M	4.37G	96.78	100.00	79.97	83.58	90.30	85.11
Swin-S	68.01M	8.53G	99.02	100.00	88.07	89.81	91.23	83.94

with different blocks[22], including typical deconvolution block, NBNet-A block [13], NBNet-B block [13], and DSCasConv block. We train these networks with $\mathcal{L}_3$, $\mathcal{L}_4$, and $\mathcal{L}_6$ (our loss function) of Eq. 11. Fig. 8 compares the performance of these networks in terms of SAR over different values of FMR, evaluated on the MOBIO, LFW, and AgeDB datasets. As this figure shows, due to the dominant effect of our loss function, these blocks achieve competitive performance when trained with this loss function ($\mathcal{L}_6$ of Eq. 11). However, when trained with $\mathcal{L}_4$, our proposed network achieves the best performance on the LFW dataset and competitive performance with NBNet-B on the MOBIO dataset. Finally, when using loss $\mathcal{L}_3$, our proposed network clearly outperforms other network structures on both the MOBIO, LFW, and AgeDB datasets.

Fig. 9 illustrates sample reconstructed face images using face reconstruction networks with deconvolution, NBNet-A, NBNet-B, and our proposed DSCasConv blocks trained with the same loss function (i.e., Eq. 8). As the results in this figure show, the reconstructed face images using the network with DSCasConv blocks have better visual quality and fewer visual artifacts.

E. TI Vulnerability Analysis of SOTA FR Models

In this section, we evaluate the vulnerability of SOTA FR models to a TI attack using our proposed framework, on the MOBIO, LFW, and AgeDB datasets. The vulnerability of FR models with different SOTA backbones and the same head is evaluated in section IV-E1. We then evaluate the

[22]We should note that the typical deconvolution block is used in [8]. Also, [20]–[23] used StyleGAN which generates high-resolution images and is not comparable to our reconstructed face images. Similarly, [26] used a GNeRF model to generate 3D face which is neither directly comparable to our method.

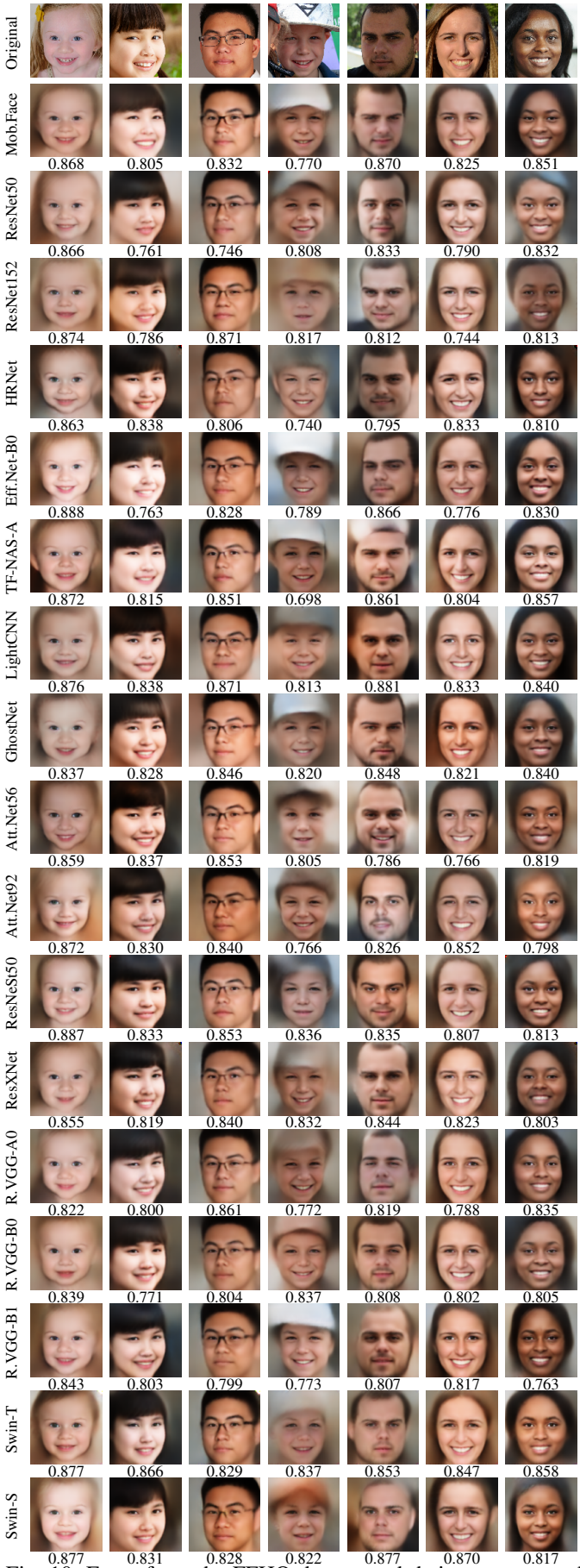


Fig. 10: Faces from the FFHQ dataset and their reconstructed versions using our TI method for different backbones. The values show the cosine similarity between the templates.

TABLE V: Comparison of face recognition models with different SOTA heads and the same backbone (MobileFaceNet) in terms of recognition performance (TMR) and vulnerability to template inversion (SAR) at FMR $= 10^{-3}$ on the MOBIO, LFW, and AgeDB datasets.

Model	MOBIO		LFW		AgeDB	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
AM-Softmax	92.29	98.57	63.37	72.37	56.90	63.92
AdaM-Softmax	89.57	98.57	66.03	73.85	50.87	58.33
AdaCos	84.83	97.62	61.33	70.69	48.20	56.33
ArcFace	88.73	98.57	60.43	71.17	66.90	71.60
MV-Softmax	89.86	98.57	53.47	67.49	65.70	69.48
CurricularFace	87.51	97.62	48.87	66.28	49.20	59.38
CircleLoss	87.98	99.52	45.43	63.87	63.70	71.72
NPCFace	87.07	98.1	63.93	72.98	60.87	68.53
MagFace	88.00	97.62	59.23	70.74	63.20	70.21

vulnerability of FR models with different SOTA heads and the same backbone in section IV-E2.

1) Different Backbones: Table IV compares FR models with different SOTA backbones and the same head (i.e., MV-Softmax) in terms of the number of parameters and the number of multiply/accumulate operations (MACs)[23], as well as recognition performance (i.e., TMR) and vulnerability to TI (i.e., SAR) at FMR $= 10^{-3}$ on the MOBIO, LFW, and AgeDB datasets. Fig. 10 also compares the reconstructed face images of these models from the validation subset of the FFHQ dataset. The values below each image in this figure report the cosine similarity between the templates extracted by the corresponding FR model from the original image and the reconstructed image.

2) Different Heads: Table V compares FR models with different SOTA heads and the same backbone (i.e., MobileFaceNet[24]) in terms of recognition performance (i.e., TMR) and vulnerability to TI (i.e., SAR) at FMR $= 10^{-3}$ on the MOBIO, LFW, and AgeDB datasets. Fig. 11 also compares the reconstructed face images of these models from the validation subset of the FFHQ dataset. Similarly to Fig. 10, the values below each image in Fig. 11 report the cosine similarity between templates extracted from the original image and the reconstructed image by the corresponding FR model.

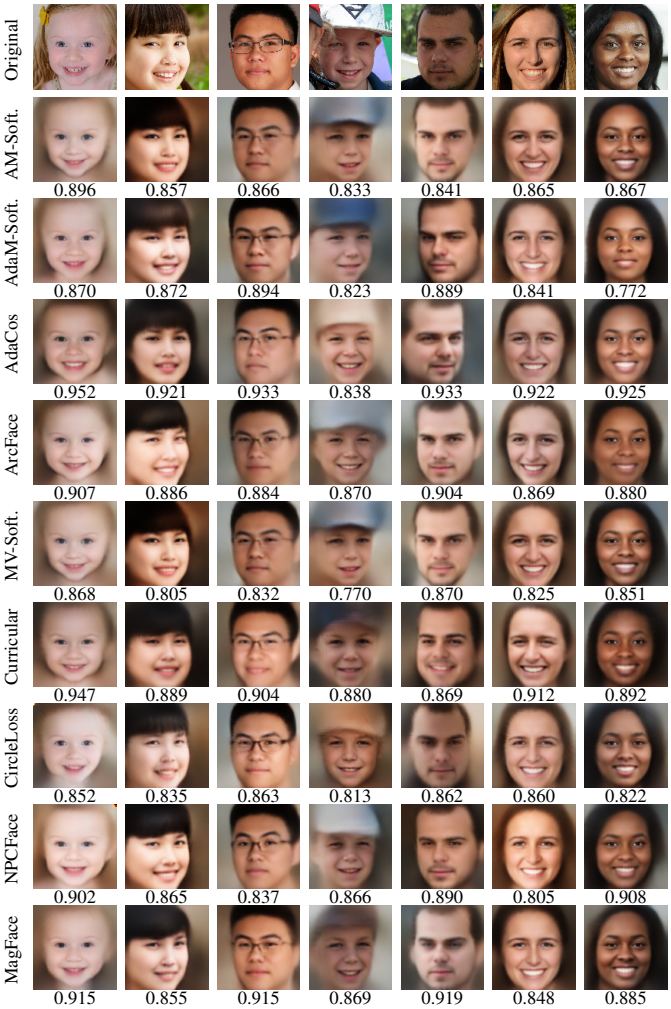


Fig. 11: Faces from the FFHQ dataset and their reconstructed versions using our TI method for different heads. The values below each image show the cosine similarity between the corresponding templates.

F. Discussion

Our experiments in sections IV-B to IV-E show the privacy and security threat of a TI attack to FR systems. In particular, Fig. 1, Fig. 10, and Fig. 11 suggest that the reconstructed face images reveal important information about the users, including race, gender, age, etc. In addition, as shown by the relatively high SAR values in Fig. 6, Fig. 8, Table IV, and Table V, the reconstructed face images can be used to enter the system by impersonating the corresponding enrolled users, which threatens the security of the FR system. In many cases in Table II, Table IV, and Table V, the values of SAR are even higher than the values for system recognition performance in terms of TMR. This is due to the fact that in our evaluation framework, the templates of reconstructed face images are compared to the templates of original face images (i.e., reference templates stored in the system's database). However, in the evaluation of recognition performance (i.e., TMR) other samples of the enrolled users are used to enter the system. Therefore, a good reconstructed face image may have a higher chance than another sample of the same subject to enter the system. Our experiments in section IV-C show that the proposed face reconstruction achieves higher SAR values than previous methods in the literature [8], [13], [20], [21] in TI attacks against FR systems.

In our threat model in section III-A, we consider the case where the attacker is assumed not to have any other information about the target FR system except the feature extractor. In particular, we assume that the attacker does not have any knowledge about the comparison and decision making submodules of the target system. However, in our experiments in sections IV-B-IV-E, we used negative cosine distance as

[23]The values for the number of parameters and the number of MACs are from [68].

[24]which has the least number of parameters in Table IV.

TABLE VI: Comparison of SAR against FR systems with iResNet100-ArcFace model and using different similarity score functions at FMR=10^{-3} on the MOBIO, LFW, and AgeDB datasets. In each case, the value of distance for probe and reference comparison is multiplied by -1 to get a similarity score.

Function	MOBIO		LFW		AgeDB	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
Cosine distance	100.00	100.00	96.63	96.48	93.77	84.60
Euclidean distance	99.98	100.00	87.27	81.61	86.1	77.74
Manhattan (L1) distance	99.98	100.00	86.13	80.02	85.93	77.34
Correlation distance	100.00	100.00	96.63	96.46	93.70	84.51
Canberra distance	100.00	100.00	95.63	95.22	91.80	77.42
Bray-Curtis distance	100.00	100.00	96.07	96.16	93.50	83.06

TABLE VII: Complexity comparison of different network structures.

Network	Params	Exe Time (ms)	
		CPU	GPU
DCNN	6.98M	4.87	0.10
NBNet-A	4.13M	3.59	0.24
NBNet-B	5.23M	4.43	0.33
DSCasConv	16.44M	13.69	0.57

the similarity score between reference and probe templates. To evaluate the effect of this assumption in the performance of our proposed method, as another experiment, we consider different functions[25] for the comparison and decision making submodules of the target system. Table VI compares the recognition performance of the iResNet100-ArcFace model and its vulnerability to our attack on the MOBIO, LFW, and AgeDB datasets. As this table shows, regardless of scoring function, the values for SAR are considerably high for each case and comparable to the value of the system's recognition performance.

Our ablation study in section IV-D shows that our loss function and proposed network structure are very effective at reconstructing the underlying face images from their enrolled face templates. In addition to the experiment in section IV-D1, which shows the effectiveness of the proposed loss function, our experiments in section IV-D2 also confirm that using $\mathcal{L}_{\text{ID,cos}}$ plus $\mathcal{L}_{\text{ID},\ell_1}$ (as in Eq. 7) improves the reconstruction such that all the studied network structures achieve competitive performance. However, when using weaker loss functions such as $\mathcal{L}_3$ and $\mathcal{L}_4$ of Eq. 11, our network structure was generally found to outperform other network structures. Sample reconstructed face images in Fig. 9 show that DSCasConv can result in better perceptual reconstruction quality and fewer visual artifacts. We also compare the complexity and execution times of the different network structures studied in section IV-D2. Table VII compares the network complexity in terms of the number of parameters and the average inference execution time (milliseconds) in the reconstruction of 112×112 face images from 512 dimensional templates, using a system equipped with an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and an NVIDIA GeForce GTX 1080 Ti GPU. As this table shows, with a similar number of blocks, our network has the highest number of parameters and the slowest execution time, because compared to other structures, DSCasConv block has additional convolutional operations (i.e., multiple cascaded convolutional layers with skip connections). Therefore, there is a trade-off between the complexity and the reconstruction performance in our proposed network. However, we should note that the number of parameters in our face reconstruction network is still much smaller than that of almost all SOTA FR models in Table IV, and therefore the TI network is inverting FR models using a lower network capacity (in terms of the number of parameters).

[25]Implementations of all these scoring functions are available in the SciPy package: https://scipy.org

TABLE VIII: Comparison of performance of our face reconstruction network when trained on different datasets (FFHQ and CASIA-WebFace) in template inversion attack (SAR) against FR with the iResNet100-ArcFace model, evaluated on the MOBIO and LFW datasets.

	$\mathbf{FMR=10^{-2}}$		$\mathbf{FMR=10^{-3}}$	
	FFHQ	CASIA-WebFace	FFHQ	CASIA-WebFace
MOBIO	100.00	100.00	100.00	100.00
LFW	97.44	97.84	96.48	97.03

TABLE IX: Comparison of reconstruction quality of the proposed network trained with ($\mathcal{D}_a$) and without ($\mathcal{D}$) data augmentation, on the validation set of FFHQ.

Training Data	↑SSIM	↓FID	$\downarrow\mathcal{L}_{\text{Perc}}$
w data aug. ($\mathcal{D}_a$)	0.37	114.66	2.69
wo data aug. ($\mathcal{D}$)	0.35	149.93	2.77

To investigate the effect of training data, as another experiment, we train our face reconstruction network using the CASIA-WebFace dataset [69]. Table VIII compares the performance of models trained with FFHQ and CASIA-WebFace. As the results in this table show, the performance of our method in terms of SAR remains comparable when trained with different datasets. There is, however, a slightly better performance when the model is trained with CASIA-WebFace. This may be due to the fact that CASIA-WebFace contains 494,414 face images while FFHQ contains 70,000 images. Therefore, when the model is trained with CASIA-WebFace, there are more variations in the images, and thus the model can be more generalizable in the test stage. In addition, while FFHQ has high-quality images, the quality of images in CASIA-WebFace is more similar to the test datasets, which can also contribute to improvement in the performance of the face reconstruction model on the test set when the model is trained on CASIA-WebFace.

It is noteworthy that in all our experiments, we used the data augmentation method described in section III-B1. Fig. 12 compares sample face images reconstructed from iResNet100-ArcFace templates using our network trained over training set $\mathcal{D}$ as in Eq. 1 (*without* data augmentation) and also over the augmented training set $\mathcal{D}_a$ as in Eq. 2. As depicted in Fig.

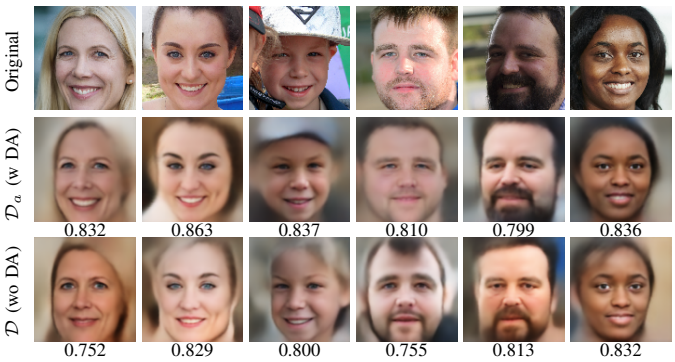


Fig. 12: Sample face images from the FFHQ dataset (first row) and their corresponding reconstructed image using our face reconstruction network trained *with* (second row) and *without* (third row) data augmentation in template inversion of a face recognition system based on the iResNet100-ArcFace model. The values below each image show the cosine similarity between the corresponding templates.

12, the reconstruction network trained *with* data augmentation generates face images with better visual reconstruction quality. Table IX also compares the quality of the reconstructed images in terms of $\mathcal{L}_{\text{Perc}}$ as in Eq.6, SSIM [40], and Fréchet Inception Distance (FID) [70] for the validation data of the FFHQ dataset. As this table shows, the network trained *with* data augmentation generates images of better quality.

Last but not least, the experiments in section IV-E1 and section IV-E2 show the vulnerability of SOTA FR models to our TI attack. Comparing Table IV and Table V, we can see that changing the FR model backbone seems to have more effect than changing the head, on the recognition performance (in terms of TMR) of the FR model and also its TI vulnerability (in terms of SAR). For example, in the case of the LFW dataset, changing the head of the FR model can change the TMR and SAR in the range of $45.43\% - 66.03\%$ and $63.87\% - 73.85\%$, respectively, while changing the backbone of the FR model can change the TMR and SAR in the range of $44.80\% - 89.37\%$ and $52.85\% - 89.81\%$, respectively. Therefore, not only are the ranges of change in recognition performance (TMR) and TI vulnerability (SAR) larger when the backbone is changed, but the maximum value of each range is also greater. Table IV and Table V further suggest that models with higher recognition performance are more likely to be more vulnerable to this type of attack. As another experiment, we evaluate the vulnerability of four other pretrained SOTA FR models in the literature, including ElasticFace [24] and AdaFace [27] as two SOTA FR models as well as EdgeFace [28] and PocketNet [29] as two SOTA lightweight FR models. Table X reports the vulnerability of these pretrained state-of-the-art face recognition models in the literature (including iResNet100-ArcFace) on the MOBIO, LFW, and AgeDB datasets. As the results in this table show, all these models are highly vulnerable to TI attacks. Since these models also have high recognition performance, this table also supports the hypothesis that models with higher recognition performance are more likely to be more vulnerable to this type of attack.

TABLE X: Vulnerability evaluation of state-of-the-art pretrained face recognition models in terms of SAR at FMR=10^{-3} on the MOBIO, LFW, and AgeDB datasets.

method	MOBIO		LFW		AgeDB	
	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)	↑TMR(%)	↓SAR(%)
ArcFace [7]	100.00	100.00	97.44	96.48	94.03	84.60
ElasticFace [24]	100.00	99.52	94.70	94.33	92.30	87.26
AdaFace [27]	100.00	99.52	98.40	95.85	95.10	68.39
EdgeFace-S [28]	99.48	100.00	92.70	91.27	75.23	64.67
PocketNet-S [29]	99.34	100.00	90.27	91.09	76.43	78.09

V. Conclusion

In this paper, we proposed a framework for evaluating the vulnerability of FR systems to a TI attack. In our threat model, we considered a real-world scenario where the adversary gains access to the system's database and tries to invert the stored templates, to reconstruct the underlying face images. Then, the adversary attempts to inject the reconstructed face images into the FR system. We proposed a face reconstruction (TI) network based on a new block, DSCasConv, and trained our network with a multi-term loss function. We measured the vulnerability of FR systems to our TI attack in terms of the Success Attack Rate (SAR). Our ablation study using the iResNet100-ArcFace model shows that our loss function and our proposed network structure are highly effective at reconstructing the underlying face images from the corresponding face templates. In addition to the iResNet100-ArcFace model, we evaluated the vulnerability of SOTA FR models (with different backbones and different heads) to our TI method on the MOBIO, LFW, and AgeDB datasets. The experiments show that FR models with higher recognition performance tend to be more vulnerable to this type of attack. Furthermore, changing the backbone may have more effect than changing the head on the vulnerability of the FR models. Our experiments also confirm that the reconstructed face images may reveal important information about each user, including race, gender, age, etc. Therefore, a TI attack, in addition to being a security threat to the FR system itself, can be also considered as a privacy threat to FR systems' users.

References

[1] J. Galbally, C. McCool, J. Fierrez, S. Marcel, and J. Ortega-Garcia, "On the vulnerability of face verification systems to hill-climbing attacks," *Pattern Recognition*, vol. 43, no. 3, pp. 1027–1038, 2010.

[2] B. Biggio, P. Russu, L. Didaci, F. Roli *et al.*, "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 31–41, 2015.

[3] S. Marcel, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Springer, 2023.

[4] H. O. Shahreza, V. K. Hahn, and S. Marcel, "Face reconstruction from deep facial embeddings using a convolutional neural network," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1211–1215.

[5] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[7] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] A. Zhmoginov and M. Sandler, "Inverting face embeddings with convolutional neural networks," *arXiv preprint arXiv:1606.04189*, 2016.

[9] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[10] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing normalized faces from facial identity features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3703–3712.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2015.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[13] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, "On the reconstruction of face images from deep face templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 5, pp. 1188–1202, 2018.

[14] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[15] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 947–954.

[16] C. N. Duong, T.-D. Truong, K. Luu, K. G. Quach, H. Bui, and K. 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*, 2020, pp. 6132–6141.

[17] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, and L. Song, "Learning towards minimum hyperspherical energy," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 6225–6236.

[18] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–59.

[19] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[20] X. Dong, Z. Jin, Z. Guo, and A. B. J. Teoh, "Towards generating high definition face images from deep templates," in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2021, pp. 1–11.

[21] E. Vendrow and J. Vendrow, "Realistic face reconstruction from deep embeddings," in *Proceedings of NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[22] X. Dong, Z. Miao, L. Ma, J. Shen, Z. Jin, Z. Guo, and A. B. J. Teoh, "Reconstruct face from features based on genetic algorithm using gan generator as a distribution constraint," *Computers & Security*, vol. 125, p. 103026, 2023.

[23] H. Otroshi Shahreza and S. Marcel, "Face reconstruction from facial templates by learning latent space of a generator network," *Advances in Neural Information Processing Systems*, vol. 36, pp. 12 703–12 720, 2024.

[24] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1578–1587.

[25] C. McCool, R. Wallace, M. McLaren, L. El Shafey, and S. Marcel, "Session variability modelling for face authentication," *IET Biometrics*, vol. 2, no. 3, pp. 117–129, Sep. 2013.

[26] H. O. Shahreza and S. Marcel, "Comprehensive vulnerability evaluation of face recognition systems to template inversion attacks via 3d face reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[27] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 750–18 759.

[28] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel, "Edgeface: Efficient face recognition model for edge devices," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[29] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper, "Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation," *IEEE Access*, vol. 10, pp. 46 823–46 833, 2022.

[30] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.

[31] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," in *Readings in computer vision*. Elsevier, 1987, pp. 671–679.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[33] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4539–4547.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[35] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[36] P. J. Van Laarhoven and E. H. Aarts, "Simulated annealing," in *Simulated annealing: Theory and applications*. Springer, 1987, pp. 7–15.

[37] M. Srinivas and L. M. Patnaik, "Genetic algorithms: A survey," *computer*, vol. 27, no. 6, pp. 17–26, 1994.

[38] *ISO/IEC 30136:2018(E) Information technology – Security techniques – Performance testing of biometric template protection schemes*, International Organization for Standardization International Standard, Jun. 2018.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California., USA, May 2015.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML)*, Lille, France, Jul. 2015, pp. 448–456.

[45] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[47] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[48] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[49] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580–1589.

[50] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[51] Y. Hu, X. Wu, and R. He, "Tf-nas: Rethinking three search freedoms of latency-constrained differentiable neural architecture search," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 123–139.

[52] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2736–2746.

[53] D. Han, S. Yun, B. Heo, and Y. Yoo, "Rethinking channel dimensions for efficient model design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 732–741.

[54] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 733–13 742.

[55] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[56] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[57] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[58] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "Adacos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 823–10 832.

[59] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, "Adaptiveface: Adaptive margin and sampling for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 947–11 956.

[60] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6398–6407.

[61] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5901–5910.

[62] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 241–12 248.

[63] D. Zeng, H. Shi, H. Du, J. Wang, Z. Lei, and T. Mei, "Npcface: A negative-positive cooperation supervision for training large-scale face recognition," *arXiv preprint arXiv:2007.10172*, 2020.

[64] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 225–14 234.

[65] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[66] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 492–10 502.

[67] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel, "Continuously reproducing toolchains in pattern recognition and machine learning experiments," in *ICML 2017 Reproducibility in Machine Learning Workshop*, 2017, pp. 1–8. [Online]. Available: https://openreview.net/forum?id=BJDDItGX-

[68] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, "Facex-zoo: A pytorch toolbox for face recognition," in *Proceedings of the 29th ACM international conference on Multimedia*, 2021.

[69] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[70] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

Hatef Otroshi Shahreza received the B.Sc. degree (Hons.) in electrical engineering from the University of Kashan, Iran, in 2016, and the M.Sc. degree in electrical engineering from the Sharif University of Technology, Iran, in 2018. He is currently pursuing the Ph.D. degree with the École Polytechnique Fédérale de Lausanne (EPFL) and is a Research Assistant with the Biometrics Security and Privacy Group, Idiap Research Institute, Switzerland, where he received H2020 Marie Skłodowska-Curie Fellowship (TReSPAsS-ETN) for his doctoral program. During his Ph.D., Hatef also experienced 6 months as a visiting scholar with the Biometrics and Internet Security Research Group at Hochschule Darmstadt, Germany. He is also the winner of the European Association for Biometrics (EAB) Research Award 2023. His research interests include deep learning, computer vision, biometrics, and biometric template protection.

Vedrana Krivokuća Hahn is a Research Associate at Idiap Research Institute (Switzerland). She received her PhD degree from the University of Auckland (New Zealand) in 2015. Vedrana works primarily on investigating biometric template protection methods and evaluation techniques. She has also been involved in consultation on the use and evaluation of face recognition systems in practice, and is currently contributing towards producing a digital, secure, and user-friendly personal data platform for European citizens. In addition to her research work, Vedrana teaches a full biometrics course as part of Idiap's Masters in Artificial Intelligence programme, and she is particularly interested in promoting ethical uses of biometric technologies.

Sébastien Marcel heads the Biometrics Security and Privacy group at Idiap Research Institute (Switzerland) and conducts research on face recognition, speaker recognition, vein recognition, attack detection (presentation attacks, morphing attacks, deepfakes) and template protection. He received his Ph.D. degree in signal processing from Université de Rennes I in France (2000) at CNET, the research center of France Telecom (now Orange Labs). He is Professor at the University of Lausanne (School of Criminal Justice) and a lecturer at the École Polytechnique Fédérale de Lausanne. He is also the Director of the Swiss Center for Biometrics Research and Testing, which conducts certifications of biometric products.