Template Inversion Attack using Synthetic Face Images against Real Face Recognition Systems

Hatef Otroshi Shahreza and Sébastien Marcel

Abstract—In this paper, we use synthetic data and propose a new method for template inversion attacks against face recognition systems. We use synthetic data to train a face reconstruction model to generate high-resolution (i.e., 1024×1024) face images from facial templates. To this end, we use a face generator network to generate synthetic face images and extract their facial templates using the face recognition model as our training set. Then, we use the synthesized dataset to learn a mapping from facial templates to the intermediate latent space of the same face generator network. We propose our method for both whitebox and blackbox TI attacks. Our experiments show that the trained model with synthetic data can be used to reconstruct face images from templates extracted from real face images. In our experiments, we compare our method with previous methods in the literature in attacks against different state-of-the-art face recognition models on four different face datasets, including the MOBIO, LFW, AgeDB, and IJB-C datasets, demonstrating the effectiveness of our proposed method on real face recognition datasets. Experimental results show our method outperforms previous methods on high-resolution 2D face reconstruction from facial templates and achieve competitive results with SOTA face reconstruction methods. Furthermore, we conduct practical presentation attacks using the generated face images in digital replay attacks against real face recognition systems, showing the vulnerability of face recognition systems to presentation attacks based on our TI attack (with synthetic train data) on real face datasets.

Index Terms—Face Recognition, Face Reconstructuin, Real Face Image, Synthetic Face Image, Template Inversion

1 Introduction

Automated face recognition (FR) systems are spreading worldwide and are increasingly present in our everyday lives, with applications from unlocking a smartphone to border control checkpoints, etc. Typically, in state-of-the-art (SOTA) FR systems a deep neural network is used to extract some features (also called "*embeddings*" or "*templates*") from face images. These features are stored in the systems' database during the registration stage and are later used for recognition. Thereby, the facial features can represent the face in a compressed space, and thus have important information about the face of each subject.

With the growth of FR systems for authentication and security applications, different types of attacks against FR systems are studied in the literature [1], [2], [3], [4], [5], [6], [7]. Among different potential attacks against FR systems, template inversion (TI) can put both the security and privacy of users in jeopardy. In a TI attack, an adversary gains access to a facial template and aims to invert it to reconstruct the underlying face image. Reconstructing the underlying face image can reveal privacy-sensitive information about subjects. Moreover, the adversary can use the reconstructed face image to enter the system, and thus cause a security threat against the FR system.

On the other side, the recent growth in the development of generative models and synthetic data has created new problems and perspectives in the research landscape of FR systems [8]. In this paper, we focus on TI attacks against FR systems and propose a new method to reconstruct high-resolution (i.e., 1024×1024) face images from facial templates using synthetic data in our training. We use StyleGAN [9] as a face generator network to generate synthetic face images and extract their facial templates. During the generation of synthetic face images for our training dataset, we also keep the intermediate latent codes of StyleGAN when synthesizing each face image. Then, we learn a mapping from facial templates to the intermediate latent space of StyleGAN. The trained model with synthetic data can be used to reconstruct face images from templates

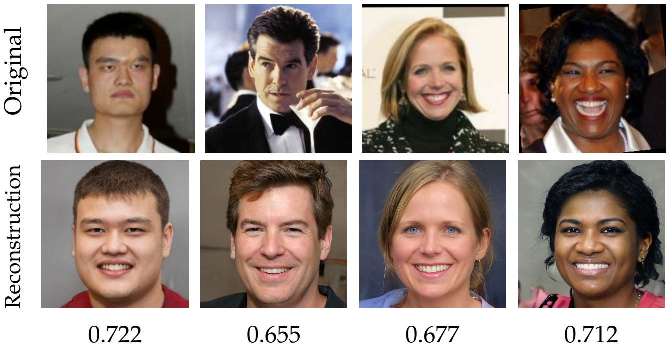


Fig. 1: Sample real face images from the LFW dataset (first row) and their reconstructed images (second row) using from facial templates extracted by ArcFace. The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images. The decision threshold corresponding to FMR $= 10^{-3}$ is 0.24 on the LFW dataset.

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) is also affiliated with École Polytechnique Fédérale de Lausanne (EPFL) and Sébastien Marcel is also affiliated with Université de Lausanne (UNIL).*

extracted from real face images. We propose our method for both *whitebox* (i.e., internal functioning and parameters of feature extractor are known) and *blackbox* (i.e., the internal functioning of feature extractor is unknown) TI attacks. In our experiments, we compare our method with previous methods in the literature in terms of adversary's success attack rate (SAR)[1] in attacks against different state-of-the-art face recognition models on four different face datasets, including the MOBIO [10], LFW [11], AgeDB [12], and IJB-C [13] datasets. In addition, we perform practical presentation attacks using the generated face images in TI attacks against real FR systems. We conduct digital replay attacks and evaluate the vulnerability of FR systems to presentation attacks based on our TI attack (with synthetic train data) on real face datasets. Fig. 1 shows sample face images from the LFW dataset and their corresponding reconstructed face images in *whitebox* attack against ArcFace [14] templates.

To elucidate the contributions of our paper, we list them hereunder:

- We propose a method based on *synthetic data* to reconstruct high-resolution face images from facial templates in TI attacks against FR systems. To train the face reconstruction model is trained using *synthetic data* only and the reconstructed face images can be used in TI attacks against FR systems with real face images.
- Our method can be used in in *whitebox* and *blackbox* TI attacks against FR systems. Experimental results show our method outperforms previous methods on high-resolution 2D face reconstruction from facial templates and achieves competitive results with SOTA face reconstruction methods.
- We provide extensive experimental results on four different FR datasets, demonstrating the effectiveness of our face reconstruction method on real face images. We also conduct presentation attacks using reconstructed face images, which shows the vulnerability of face recognition systems to our TI attacks.

The remainder of the paper is organized as follows. We first review the related work in Section 2. Then, we describe our threat model and present our proposed method in Section 3. We report our experimental results in Section 4 and discuss our results. Finally, we conclude the paper in Section 5.

2 Related Work

Previous template inversion methods in the literature can be categorized into optimization-based and learning-based methods. While optimization-based methods are slow in the inference stage, they do not require training stage, and thus do not need training data. In contrast, learning-based methods are faster in the inference stage, but require training stage and training data. While most learning-based methods in the literature used real training data, synthetic data can be an alternative option, which eases the attack by eliminating the necessity for real training data for the adversary. From the adversary's knowledge of the FR model, TI attacks can also be categorized into *whitebox* and *blackbox* attacks.

In addition to the method basis and training data, previous works in the literature can be categorized based on the output resolution, i.e., low-resolution (e.g., 112×112) or high-resolution (e.g., 1024×1024) reconstructed face images. However, most works in the literature generate low-resolution face images [15], [16], [17], [18], [19], [20], [21]. In [15], an optimization-based method for *whitebox* TI attacks was proposed, where starting from a random noise or a guiding image an iterative gradient-ascend approach is used to generate an image that has a similar facial template. The authors also used multiple regularization terms to generate smooth images. They also used the same loss function to train a convolutional neural network (CNN) to reconstruct face images. For evaluation of their method, they reported only sample reconstructed images and discussed the visual quality of the reconstruction. Similarly, in [16], a learning-based method for low-resolution *whitebox* TI attacks was proposed, where a CNN network was used to reconstruct face images. To train this CNN, several loss terms were used to optimize the pixel-level reconstruction quality, and one loss term used the feature extractor of the *whitebox* FR model to preserve identity in the reconstructed face images. For the security evaluation, the reconstructed face images were injected into the FR system and the adversary's success attack rate was reported.

In [17], a learning-based method is proposed to reconstruct face images. The authors trained a multi-layer perceptron (MLP) and a CNN to estimate landmark coordinates and facial features, respectively. Then, a differentiable warping function was used to combine estimated landmarks and textures to reconstruct the face images. In the *whitebox* scenario, they trained their model end-to-end with a loss function, including a term to minimize the distance between templates of the original and reconstructed face images. However, in the *blackbox* scenario, they trained their MLP and CNN separately and combined the results with the warping function. For the security analysis, they only reported the histogram of similarity of original and reconstructed face images.

In [18], a *blackbox* scenario was considered and a learning-based method was proposed. The authors proposed two networks, called NBNetA and NBNetB, and trained each with two different loss functions, mean absolute error and perceptual loss. Considering the variations in the network structures and loss function, they proposed four different models, called NBNetA-M, NBNetA-P, NBNetB-M, and NBNetB-P. They defined two types of attacks and compared the templates of the reconstructed face image with the same and different images of the same subject and found the success attack rate.

In [23], a learning-based method is used for reconstructing face images from facial templates. The authors used bijection learning and trained a generative adversarial network (GAN) to generate face images. For *whitebox* attacks, they used the FR model to minimize the distance between the templates of the original and reconstructed face images. In the *blackbox* attack, they proposed to use knowledge distillation to mimic the FR model and used the learned model in the training of their GAN model (similar to their *whitebox* attack). For the security analysis, they reported the matching accuracy between an original and a reconstructed

[1] also referred to as attack success rate (ASR).

TABLE 1: Comparison with related work in the literature.

Reference	Method Basis	Training Data	Resolution	Whitebox/Blackbox	Available code
Zhmoginov and Sandler [15]	1) optimization 2) learning	N/A real	low	whitebox	✗
Otroshi Shahreza *et al.* [16]	learning	real	low	whitebox	✓
Cole *et al.* [17]	learning	real	low	both	✗
Mai *et al.* [18]	learning	real	low	blackbox	✓
Doung *et al.* [19]	learning	real	low	both	✗
Akasaka *et al.* [20]	learning + opt.	real	low	blackbox	✗
Ahmad *et al.* [21]	learning	real	low	blackbox	✗
Vendrow and Vendrow [22]	optimization	N/A	high	blackbox	✓
Dong *et al.* [23]	learning	synthetic	high	blackbox	✓
Dong *et al.* [24]	optimization	N/A	high	blackbox	✓
Otroshi Shahreza and Marcel [25]	learning	real+synthetic	high	both	✓
[Ours]	learning	synthetic	high	both	✓

face from another image in each positive pair.

In [20], a *blackbox* method with three steps was proposed. In the first step, the authors trained a general face generator model based on GAN to generate face images from noise vectors. Then, in their second step, they trained a mapping network using a MLP to map the templates of the target FR model (blackbox) to the templates of a known FR model. In the last step, they applied optimization on the input (i.e., noise) of the GAN model to maximize the score of the GAN discriminator (to generate real images) and also maximize the similarity between mapped templates and the templates of the reconstructed face images extracted with the known FR model. For their security evaluation, they evaluated the adversary's success attack rate, but they did not specify the system's operation configuration (e.g., false match rate of the FR system). Similarly, in [21], a GAN-based method is used to reconstruct face images from facial templates in the *blackbox* scenario. They focused on the size of the training set and investigated the amount of images that the adversary needs for training. They assumed that the adversary has access to multiple FR models, and has templates extracted by different models. However, this assumption may not be feasible in the real-world scenario since it is difficult to have access to templates of the same subject extracted by different models. In addition, in their method, the adversary still requires some real face images to use in the training set and they did not investigate the application of synthetic training images.

In contrast to low-resolution methods, there are a few works in the literature [22], [23], [24] to generate high-resolution face images from facial templates. These models use StyleGAN [9], [26], [27] to generate high-resolution face images. StyleGAN is a GAN-based face generator network that can generate high-resolution and realistic face images. It consists of two sub-networks, a mapping network, and a synthesis network. The mapping network takes a random noise $\boldsymbol{z} \in \mathcal{Z}$ in the input and generates an intermediate latent code $\boldsymbol{w} \in \mathcal{W}$, which is then fed to the synthesise network to generate a high-resolution face image. The intermediate latent space is shown to provide more control for editing the synthesized face image [9], [26], [27], [28], [29]. In [23], a *blackbox* learning-based method was used to reconstruct face images using StyleGAN [27]. The authors generated random face images using StyleGAN and extracted facial templates using the FR model. Then, they trained a MLP to map facial templates to the input (noise $\boldsymbol{z}$) of StyleGAN. For the security analysis, they considered two types of attacks similar to [18] and compared the templates of reconstructed and original face images. In addition, they used a commercial off-the-shelf (COTS) presentation attack detection (PAD) system to evaluate the reconstructed face images. However, they did not perform a *practical* presentation attack, where the reconstructed face images needed to be recaptured by a camera.

In contrast to [23] that used learning-based method, in [22], [24] optimization-based methods are proposed to reconstruct high-resolution face images in *blackbox* TI attacks. In [22], authors proposed a hill-climbing approach by a greedy random optimization enhanced with simulated annealing [30] to find input (noise $\boldsymbol{z}$) of StyleGAN for the given facial template. Similarly, In [24] the authors used an optimization-based approach to find input (noise $\boldsymbol{z}$) of StyleGAN, but solved the optimization using the genetic algorithm [31]. For the security analysis, authors in [24] reported similar evaluation as [23], but authors in [22] reconstructed only 20 face images and compared the similarity between templates of the original and reconstructed face images.

In [25], a learning-based method (called GaFaR) was proposed to reconstruct high-resolution and 3D face from facial templates in *whitebox* and *blackox* TI attacks against FR systems. To this end, a semi-supervised and adversarial learning approach using real and synthetic face images was used to train a mapping from facial templates to the intermediate latent space of a generative neural radiance fields (GNeRF) model, from which the adversary can generate reconstructed face images with any arbitrary pose using the renderer part of the GNeRF model. While the frontal reconstructed face image can be used to attack the system, in [25] the adversary can also perform optimization on the camera parameter to find the best pose that enhances the success attack rate. For the security analysis, the reconstructed face images were used to attack the target FR system, and the success attack rate was reported.

Table 1 compares our proposed method with related work in the literature. Compared to most methods in the literature that reconstructed low-resolution face images, our method generates high-resolution and realistic face images.

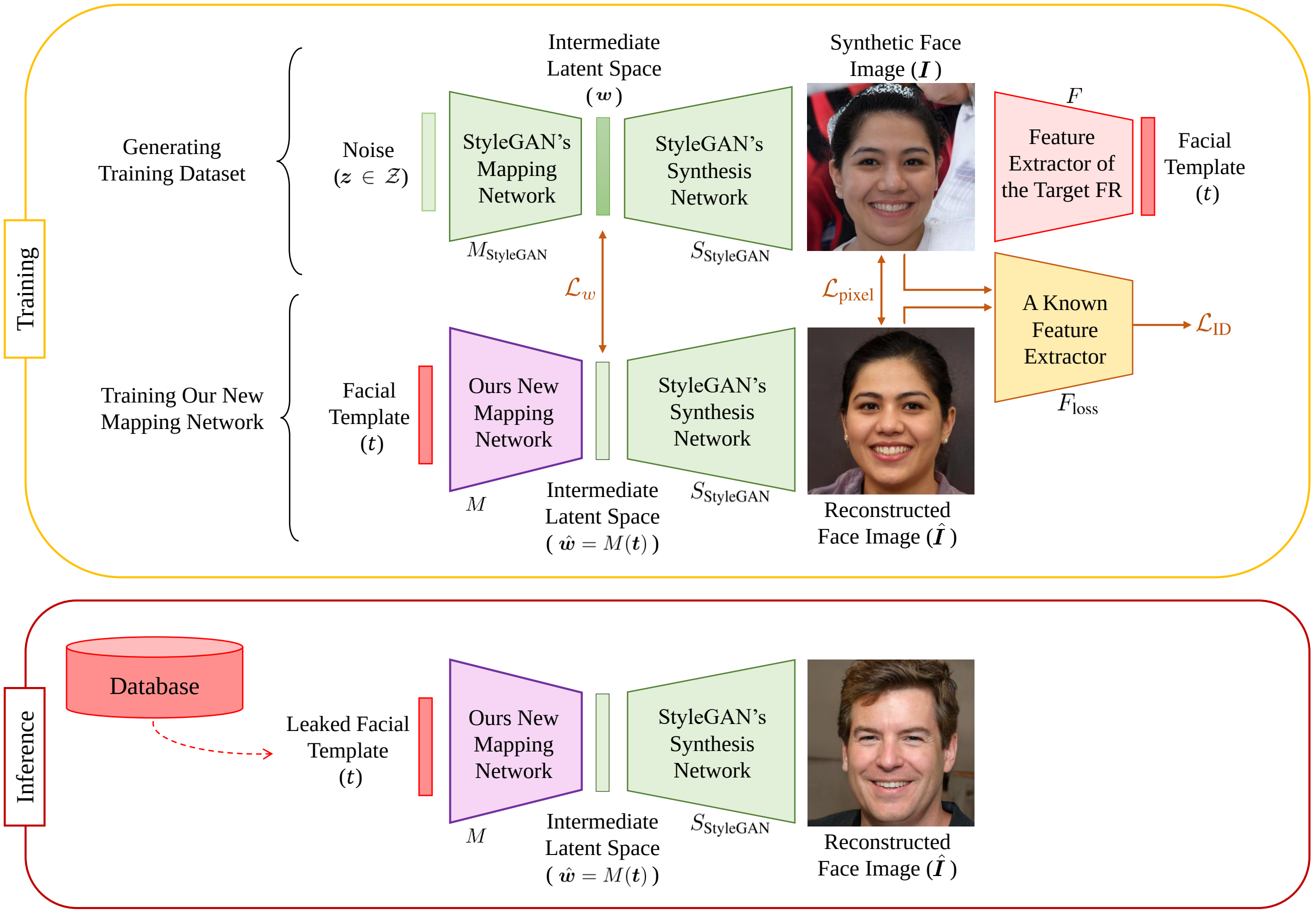

Training
Generating Training Dataset
Noise ($z \in \mathcal{Z}$)
StyleGAN's Mapping Network
M_{StyleGAN}
Intermediate Latent Space (w)
StyleGAN's Synthesis Network
S_{StyleGAN}
Synthetic Face Image (I)
F
Feature Extractor of the Target FR
Facial Template (t)
$\mathcal{L}_w$
$\mathcal{L}_{\text{pixel}}$
A Known Feature Extractor
F_{loss}
$\mathcal{L}_{\text{ID}}$
Training Our New Mapping Network
Facial Template (t)
Ours New Mapping Network
M
Intermediate Latent Space ($\hat{w} = M(t)$)
StyleGAN's Synthesis Network
S_{StyleGAN}
Reconstructed Face Image ($\hat{I}$)
Inference
Database
Leaked Facial Template (t)
Ours New Mapping Network
M
Intermediate Latent Space ($\hat{w} = M(t)$)
StyleGAN's Synthesis Network
S_{StyleGAN}
Reconstructed Face Image ($\hat{I}$)

Fig. 2: Block diagram of the proposed method in the training and inference stages.

The low-resolution reconstruction has several limitations for practical attacks and cannot be used for presentation attacks based on the reconstructed low-resolution face images. However, a few methods [22], [23], [24] were proposed for high-resolution face reconstruction from facial templates, where facial templates were mapped to noise vectors in the input space $\mathcal{Z}$ of StyleGAN. In contrast, we train a mapping form facial templates to the *intermediate* latent space $\mathcal{W}$ of StyleGAN, which is shown to have more control over the generated face image. In [25], facial templates were also mapped to the intermediate latent space of a GNeRF model, however training such mapping was proposed using a semi-supervised and adversarial learning approach, and included real face images in the training. We should highlight that in contrast to most works in the literature, which use real face images for training the face reconstruction networks, we use *synthetic* face images as training data. Using synthetic data has two merits: first, the adversary does not need to find a dataset of real face images to use for training. Second, since we generate the training dataset, we can have the corresponding intermediate latent code for each face image, and thus use the correct intermediate latent codes directly in our training. This helps train our mapping to the *intermediate* latent space of StyleGAN. In contrast to most methods that work only for *whitebox* or *blackbox* scenarios, our method can be applied for both *whitebox* and *blackbox* TI attacks. Our experiments also show that our method outperforms previous methods on high-resolution face reconstruction in the literature in terms of the adversary's success attack rate in TI attacks against FR systems.

3 Methodology

We assume the threat model described in Sec. 3.1 and use the proposed method in Sec. 3.2 to reconstruct face images from facial templates.

3.1 Threat Model

We consider the scenario where the adversary gains access to the templates from the database of the FR system and aims to invert the template to impersonate. We assume the threat model with the following properties:

- *Adversary's Goal:* The adversary's goal is to reconstruct face images from face templates stored in the database of the FR systems and use the reconstructed face image to impersonate into the FR system.
- *Adversary's Knowledge:* We assume that the adversary has the following knowledge:
  1) The adversary has access to templates from the database of the FR system.
  2) The adversary also has a whitebox or blackbox knowledge of the feature extractor of the FR system.

In the case of the blackbox knowledge, we assume that the adversary has the whitebox knowledge of another FR model.

3) The adversary also has access to a general face generator network.

- *Adversary's Capability:* We consider two scenarios:
  1) The adversary can use the reconstructed face image from the TI attack to inject it as a query into the feature extractor of the FR system.
  2) The adversary can use the reconstructed face image to perform a presentation attack to enter the FR system.
- *Adversary's Strategy:* The adversary trains a face reconstruction network to invert facial templates. Then, the adversary uses the trained network to reconstruct face images from the leaked facial templates. The adversary can use the reconstructed face images to inject a query into the system or conduct a presentation attack.

3.2 Proposed Face Reconstruction Method

To reconstruct face images from facial templates, we consider the situation where the adversary has access to a pretrained face generator model such as StyleGAN [9]. As described in Section 2, the StyleGAN model is composed of two sub-networks, a mapping network and a synthesis network. Let us denote the mapping network with M_{StyleGAN} and the synthesis network with S_{StyleGAN}. The mapping network gets a random noise vector $\boldsymbol{z} \in \mathcal{Z}$ as input and generates an *intermediate* latent code $\boldsymbol{w} = M_{\text{StyleGAN}}(\boldsymbol{z}) \in \mathcal{W}$, which is then fed to the synthesis network to generate the face image $\boldsymbol{I} = S_{\text{StyleGAN}}(\boldsymbol{w})$.

To generate a training dataset for learning a face reconstruction network, we use the StyleGAN model to generate synthetic face images and extract facial templates from the synthesized face images. To this end, we sample K noise $\{\boldsymbol{z}_i | \boldsymbol{z} \in \mathcal{Z} \sim \mathcal{N}(0, \mathbb{I}), i = 1, \dots, K\}$ from Gaussian distribution $\mathcal{N}(0, \mathbb{I})$ for the input of StyleGAN. Next, we generate corresponding intermediate latent codes $\boldsymbol{w}_i = M_{\text{StyleGAN}}(\boldsymbol{z}_i)$ and synthetic images $\boldsymbol{I}_i = S_{\text{StyleGAN}}(\boldsymbol{w}_i)$, and then use the feature extractor of the target FR system[2] to extract facial templates $\boldsymbol{t}_i = F(\boldsymbol{I}_i)$ from our synthetic face images. Finally, we can have our training dataset $\mathcal{D} = \{(\boldsymbol{I}_i, \boldsymbol{t}_i, \boldsymbol{w}_i)|, i = 1, \dots, K\}$ which has triples of synthetic face images $\boldsymbol{I}_i$ as well as their corresponding facial templates $\boldsymbol{t}_i$ and the StyleGAN intermediate latent codes $\boldsymbol{w}_i$.

After generating our dataset $\mathcal{D}$, we can use this dataset to train a new mapping network $M(.)$ to project the facial template t to the intermediate latent code $\hat{\boldsymbol{w}} = M(\boldsymbol{t})$ in $\mathcal{W}$ space of StyleGAN. Then, we use the the intermediate latent code $\hat{\boldsymbol{w}} = M(\boldsymbol{t})$ as input to the synthesis network of StyleGAN S_{StyleGAN} to generate the reconstructed face image $\hat{\boldsymbol{I}} = S_{\text{StyleGAN}}(\hat{\boldsymbol{w}})$. We train our new mapping network $M(.)$ with parameters θ_M using the following multi-term loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_w + \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}, \tag{1}$$

where $\mathcal{L}_w$, $\mathcal{L}_{\text{pixel}}$ and $\mathcal{L}_{\text{ID}}$ are the intermediate latent space loss, pixel loss, and ID loss, respectively, and are defined as follows:

$$\mathcal{L}_w = \|\boldsymbol{w} - M(\boldsymbol{t})\|_2^2, \tag{2}$$

$$\mathcal{L}_{\text{pixel}} = \left\|\boldsymbol{I} - S_{\text{StyleGAN}}(M(\boldsymbol{t}))\right\|_2^2, \tag{3}$$

$$\mathcal{L}_{\text{ID}} = \left\|F_{\text{loss}}(\boldsymbol{I}) - F_{\text{loss}}(\hat{\boldsymbol{I}})\right\|_2^2. \tag{4}$$

Algorithm 1 Training process in our proposed method.

```
Require: : $n_{\text{epoch}}$: number of epochs, $n_{\text{iteration}}$: number of
    iterations in each epoch, $\alpha$: learning rate.
procedure TRAINING
    Initialize weights $\theta_M$ of our new mapping network
    for epoch $= 1, ..., n_{\text{epoch}}$ do
        for itr $= 1, ..., n_{\text{iteration}}$ do
            Sample a batch of random noise vectors:
                $\boldsymbol{z} \in \mathcal{Z} \sim \mathcal{N}(0, \mathbb{I})$
            Generate training data:
                $\boldsymbol{w} = M_{\text{StyleGAN}}(\boldsymbol{z})$
                $\boldsymbol{I} = S_{\text{StyleGAN}}(\boldsymbol{w})$
                $\boldsymbol{t} = F(\boldsymbol{I})$
            Reconstruct image from template $\boldsymbol{t}$:
                $\hat{\boldsymbol{w}} = M(\boldsymbol{t})$
                $\hat{\boldsymbol{I}} = S_{\text{StyleGAN}}(\hat{\boldsymbol{w}})$
            Calculate loss $\mathcal{L}_{\text{total}}$ and optimize $\theta_M$:
                $g_{\theta_M} \leftarrow \nabla_{\theta_M} \mathcal{L}_{\text{total}}$
                $\theta_M \leftarrow \theta_M - \alpha \cdot \text{Adam}(\theta_M, g_{\theta_M})$
        end for
    end for
end procedure
```

The intermediate latent space loss ($\mathcal{L}_w$) is used to minimize the error in the estimated intermediate latent code $\hat{\boldsymbol{w}} = M(\boldsymbol{t})$ in the intermediate latent space $\mathcal{W}$ of StyleGAN. Since we use synthetic face images, we have the correct values of intermediate latent codes w to calculate $\mathcal{L}_w$. The pixel loss ($\mathcal{L}_{\text{pixel}}$) is also applied to minimize the pixel-level reconstruction error for the reconstructed face image $\hat{\boldsymbol{I}} = S_{\text{StyleGAN}}(M(\boldsymbol{t}))$ compared to the original image $\boldsymbol{I}$. Finally, the ID loss is used to optimize the similarity between the facial templates extracted from the reconstructed and original face images using a FR feature extractor $F_{\text{loss}}(.)$. In the whitebox TI attack, the adversary can use the same feature extractor as the one in the target FR system (i.e., F) as $F_{\text{loss}}(.)$; however, in the blackbox scenario, the adversary needs to use a different feature extractor that has access to[3]. Therefore, in blackbox TI attacks, $F_{\text{loss}}(.)$ is different from the target FR system[4]. Algorithm 1 summarizes our training process and Fig. 2 illustrates the block diagram of our proposed face reconstruction method. In our experiments, we generate 25,000 synthetic face images for our training

[2] As mentioned in our threat model in Section 3.1, we only need the *blackbox* knowledge of target FR model, and it is not necessary to have the *whitebox* knowledge.

[3] Note that the adversary needs to have whitebox knowledge of feature extractor used in F_{loss} to be able to calculate gradients in optimizing loss function for training face reconstruction model.

[4] Note that the alternate model is only used in the *blackbox* scenario and is only applied for $F_{\text{loss}}(.)$ in the loss function $\mathcal{L}_{\text{ID}}$ (not to extract the initial templates t). In both whitebox and blackbox scenarios, feature extractor of the target FR system (i.e., F) is always used to extract the initial templates t in generating training dataset $\mathcal{D}$.

TABLE 2: Recognition performance of FR models in terms of true match rate (TMR) at false match rates (FMRs) of 10^{-2} and 10^{-3} on the MOBIO, LFW, AgeDB, and IJB-C datasets. The values are in percentage.

FR model	MOBIO		LFW		AgeDB		IJB-C	
	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
ArcFace	100.00	99.98	97.60	96.40	98.33	98.07	95.29	90.90
ElasticFace	100.00	100.00	96.87	94.70	98.20	97.57	93.73	84.70
AttentionNet	99.71	97.73	84.27	72.77	97.93	96.90	92.65	82.43
HRNet	98.98	98.23	89.30	78.43	97.67	96.23	89.68	78.25
RepVGG	98.75	95.80	77.20	58.07	95.93	93.93	87.67	77.42
Swin	99.75	98.98	91.70	87.83	98.03	97.10	93.40	89.35

TABLE 3: Comparison with previous TI methods in attacks against SOTA FR models in terms of success attack rate (in percentage) at systems' $\mathbf{FMR = 10^{-2}}$ on the MOBIO, LFW, AgeDB, and IJB-C datasets . For attacks using our method, we use ArcFace and ElasticFace as F_{loss} in our loss function. The best two values in attack against each system is embolden.

method	MOBIO						LFW					
	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin
NBNetA-M [18]	2.86	10.0	4.76	4.76	6.19	6.67	14.30	37.13	10.37	20.19	10.64	13.18
NBNetA-P [18]	23.81	60.95	15.24	14.29	44.76	30.48	35.61	60.05	6.80	16.83	26.44	25.92
NBNetB-M [18]	20.95	30.0	21.43	25.24	21.43	27.62	26.91	52.99	17.62	31.74	18.18	27.00
NBNetB-P [18]	49.05	70.95	66.67	64.76	51.43	71.43	61.66	81.74	43.42	56.30	38.12	61.02
Dong *et al.* [23]	24.29	34.76	38.57	16.19	24.76	18.10	28.21	34.56	19.17	24.87	14.76	26.62
Vendrow and Vendrow [22]	69.52	74.29	55.71	43.81	39.52	70.00	77.00	79.37	46.52	49.52	22.4	66.07
Dong *et al.* [24]	87.62	90.95	80.48	71.90	44.29	82.38	87.26	89.00	55.40	59.46	28.60	69.07
GaFaR [25]	95.71	91.90	89.05	87.62	87.14	**96.19**	89.27	89.78	56.57	67.64	46.89	78.91
GaFaR + GS [25]	**97.62**	**93.33**	**90.00**	**90.00**	**90.95**	**96.19**	**90.77**	91.28	**62.03**	**72.28**	**51.27**	**81.39**
[Ours] (F_{loss}= Els.Face)	88.57	**92.38**	87.14	83.33	82.38	**93.33**	84.70	**92.28**	60.75	70.78	49.78	75.09
[Ours] (F_{loss}= ArcFace)	**96.67**	**93.33**	**90.48**	**91.43**	**86.67**	**93.33**	**92.32**	**92.71**	**67.49**	**77.23**	**56.30**	**78.60**

	AgeDB						IJB-C					
	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin	ArcFace	Els.Face	Att.Net	HRNet	RepVGG	Swin
NBNetA-M [18]	2.56	8.44	1.85	2.45	2.85	1.89	3.29	6.91	1.84	0.79	1.38	4.98
NBNetA-P [18]	9.31	20.07	2.43	1.54	10.14	4.72	11.19	16.11	1.19	0.44	5.68	10.75
NBNetB-M [18]	5.40	14.56	3.83	3.68	4.72	3.70	7.76	13.83	4.47	2.22	2.77	12.89
NBNetB-P [18]	23.89	44.47	17.19	14.83	18.62	21.48	28.22	42.51	19.26	11.77	15.4	37.81
Dong *et al.* [23]	9.13	12.11	7.58	6.02	6.82	7.62	7.80	6.54	5.77	2.64	3.5	10.79
Vendrow and Vendrow [22]	44.75	52.17	35.48	24.65	27.39	40.43	38.33	37.91	22.68	15.45	16.61	38.52
Dong *et al.* [24]	58.80	66.10	36.82	32.45	14.98	37.81	57.29	59.22	33.00	25.94	10.86	50.01
GaFaR [25]	63.30	63.45	33.23	31.56	31.71	49.17	69.18	61.16	45.34	37.79	37.92	69.17
GaFaR + GS [25]	**67.86**	68.82	40.26	38.53	38.78	**55.20**	**73.72**	66.46	**51.31**	**44.14**	**44.02**	**73.76**
[Ours] (F_{loss}= Els.Face)	57.94	**77.28**	**43.28**	**41.95**	**44.07**	51.78	63.33	**71.05**	49.22	42.25	43.67	66.56
[Ours] (F_{loss}= ArcFace)	**73.78**	**76.98**	**49.92**	**50.57**	**48.36**	**59.95**	**78.75**	**71.84**	**55.03**	**49.60**	**51.01**	**71.26**

set and use Adam optimizer [32] with the learning rate of 10^{-4}. In the inference stage, we use our trained mapping network to project the facial template to the intermediate space of StyleGAN, and then use the synthesis network to generate the reconstructed face image.

4 Experiments

In this section, we present our experiments and discuss our results. First, in Section 4.1 we describe our experimental setup. Then, in Section 4.2 we compare the performance of our method with previous methods in the literature in TI attacks against SOTA FR models. In Section 4.3, we report our vulnerability evaluation of FR systems to practical presentation attacks using the reconstructed face images from TI attacks. We discuss further our experimental results in Section 4.4.

4.1 Experimental Setup

4.1.1 Face Recognition Models

We use SOTA FR systems and evaluate their vulnerability to our TI attack on real face images. In our experiments, we consider ArcFace [14], ElasticFace [33], and also four FR models with different SOTA backbones from FaceX-Zoo [34], including AttentionNet [35], HRNet [36], RepVGG [37], and Swin [38]. Table 2 presents the recognition performances of these models.

4.1.2 Evaluation Datasets

We evaluate the performance of our face reconstruction network on real face datasets, including MOBIO [10], Labeled Faced in the Wild (LFW) [11], AgeDB [12], and IARPA Janus Benchmark-C (IJB-C) [13] datasets. The MOBIO dataset includes face images of 150 subjects captured using mobile devices in 12 sessions. The LFW database consists of 13,233 face images of 5,749 people, where 1,680 people have two or more images. The AgeDB dataset consists of 16,488 images of 568 subjects (famous people) with the average age range of 50.3 years. The IJB-C dataset includes 31,334 images of 3,531 subjects.

4.1.3 Evaluation Protocol

To evaluate the vulnerability of the FR system to TI attacks, we consider the situation in which the adversary gains access to the database of the FR system and reconstructs the underlying face images to enter the system. As described in

TABLE 4: Comparison with previous TI methods in attacks against SOTA FR models in terms of success attack rate (in percentage) at systems' $\mathbf{FMR = 10^{-3}}$ on the MOBIO, LFW, AgeDB, and IJB-C datasets . For attacks using our method, we use ArcFace and ElasticFace as F_{loss} in our loss function. The best two values in attack against each system is embolden.

method	**MOBIO**						**LFW**					
	ArcFace	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**	**ArcFace**	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**
NBNetA-M [18]	0	2.38	0	0	0	0	4.32	10.90	1.24	1.60	1.13	3.82
NBNetA-P [18]	4.76	16.19	0.48	0	14.29	7.14	16.83	26.98	0.66	1.44	5.72	9.70
NBNetB-M [18]	1.90	3.80	3.33	7.14	3.33	8.57	10.98	21.44	3.22	4.47	3.21	11.23
NBNetB-P [18]	15.24	43.81	31.90	26.67	23.81	44.29	40.26	58.16	16.29	18.42	15.24	40.76
Dong *et al.* [23]	3.33	8.10	10.48	6.67	9.05	3.33	13.21	12.61	3.90	4.07	3.22	12.38
Vendrow and Vendrow [22]	29.05	43.81	27.14	26.67	20.95	45.24	57.70	53.03	21.12	18.85	9.62	46.84
Dong *et al.* [24]	61.43	76.67	42.86	49.05	20.00	65.71	74.48	73.67	32.07	31.73	10.89	53.59
GaFaR [25]	82.86	84.76	72.38	76.67	**72.86**	**89.05**	79.84	74.54	33.59	37.80	25.40	**67.11**
GaFaR + GS [25]	**85.23**	86.62	**80.00**	**83.80**	**73.33**	**93.33**	**82.52**	78.67	**38.42**	**43.27**	**29.84**	**70.82**
[Ours] (F_{loss}= Els.Face)	80.00	**87.62**	78.10	78.10	68.57	79.05	71.31	**80.41**	36.92	43.13	29.33	61.63
[Ours] (F_{loss}= ArcFace)	**84.76**	**86.67**	**81.90**	**85.24**	70.95	84.76	**85.01**	**81.70**	**43.58**	**50.04**	**35.75**	66.57

	AgeDB						**IJB-C**					
	ArcFace	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**	**ArcFace**	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**
NBNetA-M [18]	0.81	2.55	0.22	0.38	0.44	0.27	0.03	0.12	0.03	0.01	0.04	0.67
NBNetA-P [18]	3.99	8.93	0.34	0.14	3.71	1.02	0.40	0.30	0	0	0.31	1.58
NBNetB-M [18]	1.88	6.28	0.50	0.77	1.06	0.68	0.36	0.32	0.13	0.06	0.13	2.40
NBNetB-P [18]	13.18	28.94	5.08	5.61	7.93	8.75	3.15	4.1	1.27	1.12	2.32	15.11
Dong *et al.* [23]	3.94	4.88	1.58	1.97	2.23	2.48	0.40	0.13	0.21	0.09	0.26	2.45
Vendrow and Vendrow [22]	29.65	34.89	15.06	12.02	14.49	21.10	7.49	5.41	2.86	2.17	4.13	20.73
Dong *et al.* [24]	43.22	48.98	17.22	17.89	7.07	21.40	19.31	16.90	4.82	4.36	2.43	30.91
GaFaR [25]	48.94	47.37	14.59	17.09	18.02	30.05	29.77	19.28	17.10	12.99	17.42	50.75
GaFaR + GS [25]	**53.10**	53.10	18.76	22.40	**24.01**	**35.20**	**34.92**	24.69	**23.12**	**18.08**	**22.80**	**58.02**
[Ours] (F_{loss}= Els.Face)	42.35	**61.84**	**19.42**	**24.12**	**28.22**	30.86	25.93	**33.42**	16.08	15.46	19.58	46.04
[Ours] (F_{loss}= ArcFace)	**60.03**	**62.43**	**25.81**	**30.84**	19.15	**37.43**	**45.42**	**32.73**	**21.27**	**21.33**	**28.40**	**52.58**

Section 3.1, we consider two scenarios in our threat model. In the first scenario, we consider the situation where the adversary can inject the reconstructed face images from the TI attack into the feature extractor of the target FR system. In the second scenario, we consider the situation in which the adversary performs a presentation attack using the reconstructed face images from the TI attack. In each case, we evaluate the vulnerability of the FR system in terms of the adversary's success attack rate (SAR) in entering the system using the reconstructed face images from the TI attack.

4.1.4 Implementation Details

We use the Bob[5] toolbox [39], [40] to build the pipelines for the FR systems in our experiments and also evaluate the TI attacks against FR systems. We also use the PyTorch package and trained our models on a system equipped with an NVIDIA GeForce RTX$^{\text{TM}}$ 3090. We use the pretrained model of StyleGAN3[6] to generate 1024×1024 high-resolution face images. The source codes of our experiments are publicly available to facilitate the reproducibility of our results[7].

4.2 Comparison with Previous TI Methods

We compare the performance of our face reconstruction method with state-of-the-art TI methods in the literature, including NBNetA-M [18], NBNetA-P [18], NBNetB-M [18], NBNetB-P [18], Dong *et al.* [23], Vendrow and Vendrow [22], Dong *et al.* [24], GaFaR [25], and GaFaR+GS [25]. Among these methods and according to Section 2, Dong *et al.* [23], Vendrow and Vendrow [22], and Dong *et al.* [24] used StyleGAN to reconstruct high-resolution face images. We consider the scenario where the adversary can inject the reconstructed face image as a query to the feature extractor of the target FR system. Table 3 compare the performance of different methods in terms of the adversary's success attack rate (SAR) in TI attacks against SOTA FR systems at the system FMR of 10^{-2} on the MOBIO, LFW, AgeDB, and IJB-C datasets. Table 4 reports similar results for the system threshold corresponding to FMR of 10^{-3} on the MOBIO, LFW, AgeDB, and IJB-C datasets. As the results in these tables show, while our method is trained on synthetic data, it achieves high SAR in TI attacks against FR systems. Furthermore, compared to other methods, our experimental results show that our method achieves superior performance than SOTA TI methods on high-resolution face reconstruction. In particular, compared to Dong *et al.* [23], Vendrow and Vendrow [22], and Dong *et al.* [24] which also used StyleGAN for face reconstruction our method achieves higher SAR values. Compared to GaFaR+GS [25], our method achieve competitive performance. However, we should note that GaFaR+GS [25] use a geometry-aware network to reconstruct 3D face images and then find the best pose for each image to enhance SAR.

In our experiments in Table 3 and Table 4, we use two FR models, ArcFace and ElasticFace, as F_{loss} in our proposed method. As the results in these table shows, face reconstruction with ArcFace achieves higher SAR values in our method. Besides, the recognition performances in Table 2 also show that ArcFace has a better recognition accuracy. Therefore, a model with a better recognition accuracy can more help training in our proposed method and lead to better reconstruction performance.

Fig. 3 illustrates sample face images from the LFW

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

[6] Available at https://github.com/NVlabs/stylegan3

[7] https://gitlab.idiap.ch/bob/bob.paper.tbiom2024_face_ti

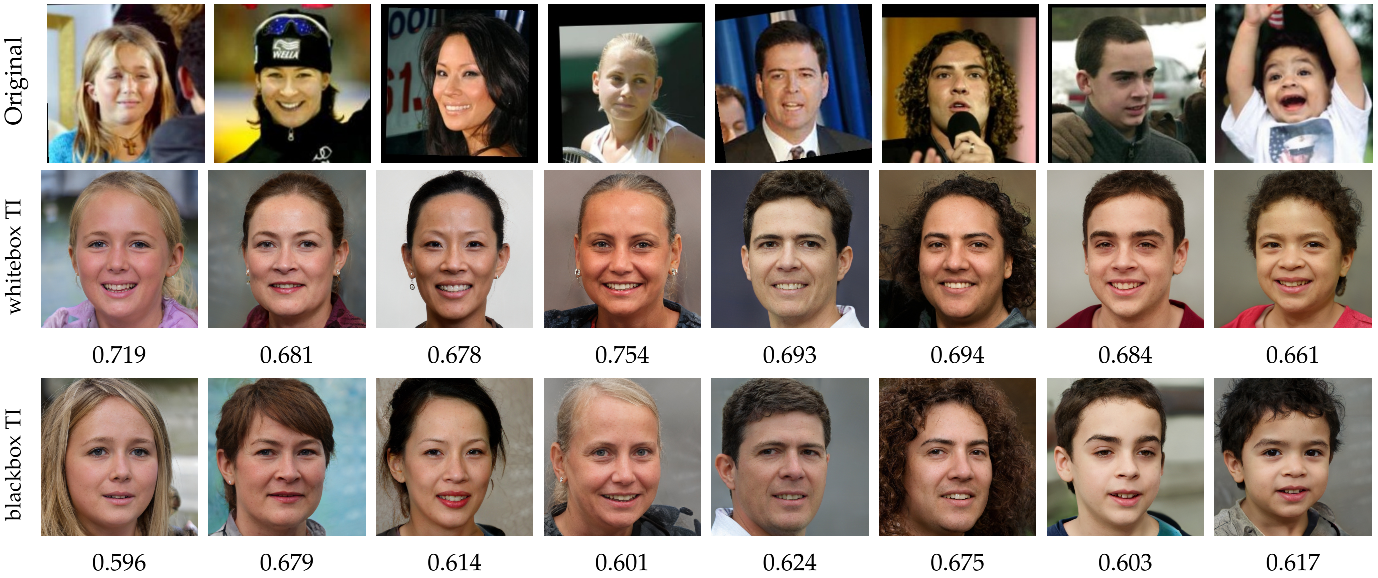
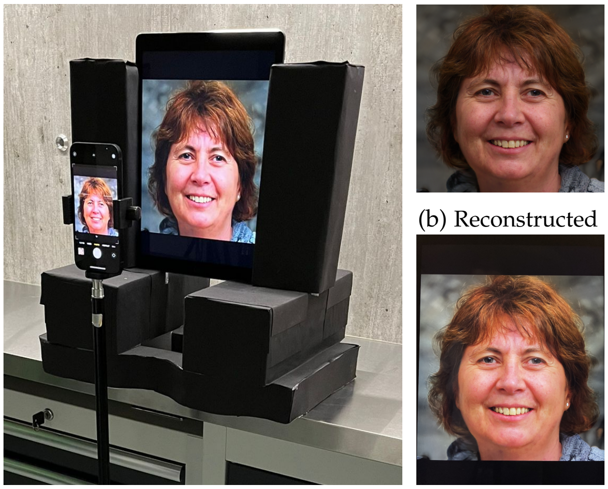


Fig. 3: Sample real face images from the LFW dataset (first row) and their reconstructed images from ArcFace templates in whitebox (second row) and blackbox (third row). The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images. The decision threshold corresponding to FMR $= 10^{-3}$ is 0.24 on the LFW dataset, and thus all these reconstructed images pass this threshold.



(b) Reconstructed

(a) Our Setup

(c) Captured

Fig. 4: (a) Presentation attack evaluation setup, (b) The reconstructed face image from TI attack that is used for presentation attack, (c) The captured image by the smartphone camera (iPhone 12) in our presentation attack.

dataset and their corresponding reconstructed face image from ArcFace in the whitebox (using ArcFace as F_{loss}) and blackbox (using ElasticFace as F_{loss}) TI attacks. The reconstructed face images are realistic and reveal important privacy-sensitive information about underlying users (such as gender, ethnicity, etc.). In addition, the reconstructed face images have similar facial templates to the templates of the original face images and can be recognized as the same subject by the FR system with the FMR of 10^{-3} on the LFW dataset. We should note that the reconstructed face images are also *high-resolution* (i.e., 1024×1024) and can be used for presentation attack, which is discussed in Section 4.3.

TABLE 5: Vulnerability of the FR system with ArcFace model against presentation attack using reconstructed face images from whitebox and blackbox TI attacks in terms of SAR/IAPMR at system FMRs of 10^{-2} and 10^{-3} on the MOBIO dataset. The values are in percentage.

Attack Type	Eval. Scenario	Camera	SAR/IAPMR	
			FMR=10^{-2}	FMR=10^{-3}
whitebox	injection	N/A	96.67	84.76
	Replay Attack	iPhone12	98.10	85.24
		Galaxy S9	96.19	86.19
		Redmi A9	96.67	86.67
blackbox	injection	N/A	88.57	80.76
	Replay Attack	iPhone12	89.04	78.09
		Galaxy S9	88.09	77.62
		Redmi A9	88.57	80.00

4.3 Presentation Attack using Reconstructed Face Images

As another experiment, we consider the situation where the adversary can reconstruct face images from facial templates and use the reconstructed face images to perform a presentation attack to impersonate into the FR system. To this end, we use the reconstructed face images to display with a tablet (Apple iPad Pro) and take it in front of a camera as the sensor of the FR system. We use cameras of three different smartphones, including Apple iPhone 12, Samsung Galaxy S9, and Xiaomi Redmi 9A, in our experiment. Fig 4 illustrates our presentation attack evaluation setup.

We consider *whitebox* (using ArcFace as F_{loss}) and *blackbox* (using ElasticFace as F_{loss}) TI attacks against ArcFace templates on the MOBIO dataset and reconstruct facial templates using our proposed method. Fig. 5 shows sample face images from the MOBIO dataset and their corresponding

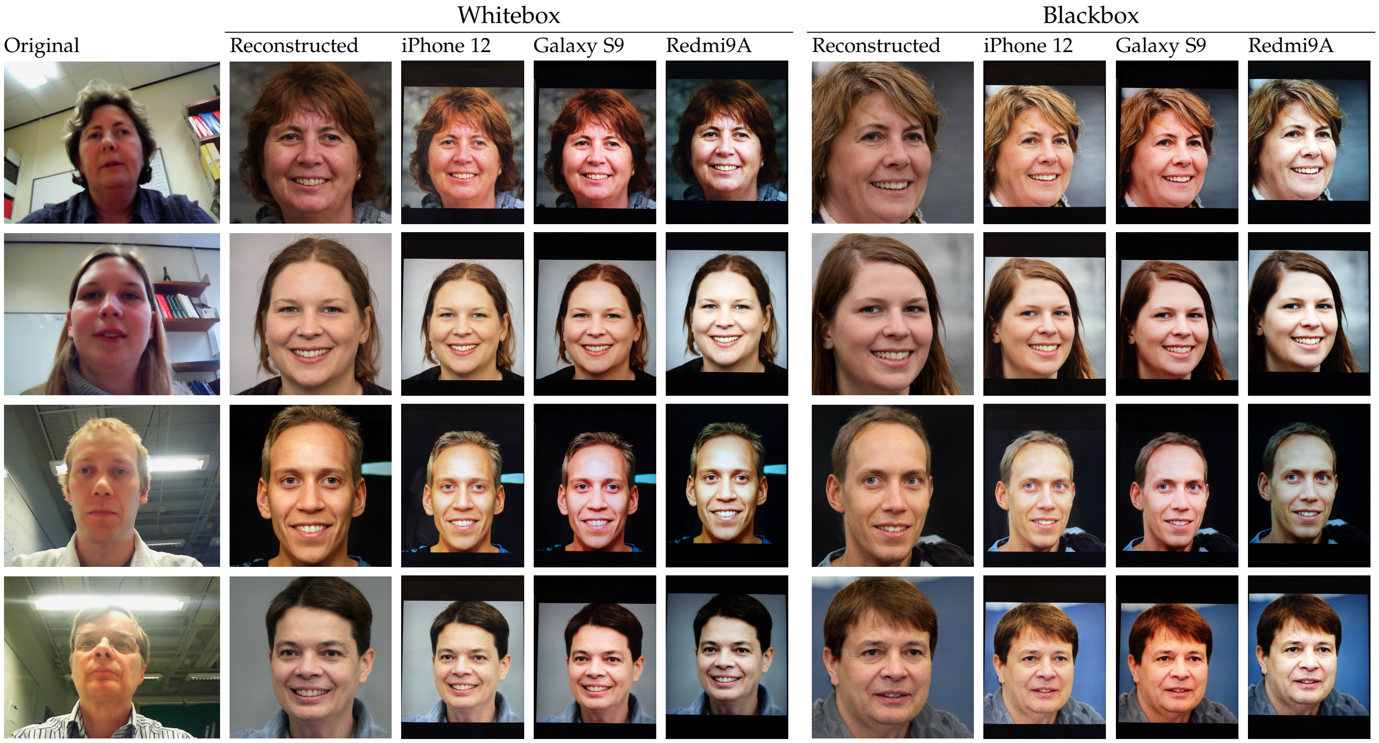


Fig. 5: Sample face images from the MOBIO dataset and their corresponding reconstructed face images in *whitebox* and *blackbox* attacks as well as the captured images from our digital replay attack using cameras of different smartphones.

reconstructed face images in *whitebox* and *blackbox* attacks as well as the captured images from our digital replay attack using different smartphones. We evaluate the performance of replay attacks using our reconstructed face images in terms of adversary's SAR[8]. Table 5 reports the vulnerability of a FR system with ArcFace model to replay attacks using reconstructed face images in *whitebox* and *blackbox* TI attacks with our method for FMRs of 10^{-2} and 10^{-3} on the MOBIO dataset. As the results in this table show, our reconstructed face images achieve high SAR values when captured with different smartphones. Also, the results in this table show that our replay attacks achieve comparable performance with TI attacks using the injection of reconstructed face images. This experiment demonstrates the vulnerability of real FR systems to the reconstructed face images using our method.

4.4 Discussion

While our experiments in Section 4.2 show that our proposed method achieves state-of-the-art performance, there are still some failure cases where the reconstructed face images do not match the original face image. Fig. 6 illustrates some sample reconstructed face images in the whitebox TI attack against ArcFace on the LFW dataset that do not match the original face images, and therefore the attack is not successful. As shown in this figure, some of the failure cases correspond to people with dark skin color or the eldery. As a matter of fact, the StyleGAN model has been trained on the FFHQ dataset, which is not a balanced face dataset and has a bias on some demography groups. In addition, the face recognition model used in this attack (ArcFace) is also shown to have bias [42].

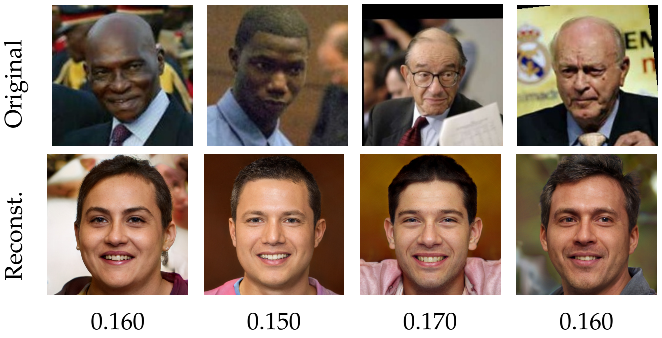


Fig. 6: Sample failure cases from the LFW dataset (first row) and their corresponding (second row) reconstructed face images using our method in the *whitebox* TI attack against ArcFace. The values below each image show the cosine similarity between templates of original and reconstructed face images. The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images.

Despite such failure cases, SOTA FR models are still significantly vulnerable to our TI attacks as shown in Section 4.2 and Section 4.3. To investigate the effect of each loss term in our proposed method, we implement an ablation study and train our mapping network with different

[8] The ISO/IEC 30107-3 standard [41] suggests to refer to the adversary's success attack rate in evaluations of presentation attacks as Impostor Attack Presentation Match Rate (IAPMR). However, for consistency with our previous experimental results, we use SAR to report the success attack rate in our presentation attack (replay attack) evaluation.

TABLE 6: Ablation study on the effect of loss terms in our proposed method in whitebox attack against ArcFace in terms of SAR for a FR system with FMRs of 10^{-2} and 10^{-3} on the MOBIO and LFW datasets.

Loss function	**MOBIO**		**LFW**	
	FMR=10^{-2}	**FMR=10^{-3}**	**FMR=10^{-2}**	**FMR=10^{-3}**
$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}$	0	0	0.32	0.02
$\mathcal{L}_{\text{total}} = \mathcal{L}_{w}$	43.81	13.80	47.69	27.54
$\mathcal{L}_{\text{total}} = \mathcal{L}_{w} + \mathcal{L}_{\text{pixel}}$	40.00	13.81	45.61	25.98
$\mathcal{L}_{\text{total}} = \mathcal{L}_{w} + \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}$	97.62	89.05	92.89	85.84

loss functions. We consider the *whitebox* TI attack against a FR system with ArcFace and evaluate the adversary's SAR on the MOBIO and LFW datasets. Table 6 reports our ablation study on the effect of each loss term in our proposed method. As the results in this table show, each of our loss terms contributes to the reconstruction of our TI attacks against FR systems. Using the latent space loss is particularly important as it has a significant effect on the training compared to using all other terms except the latent code loss term. When using the latent space loss, our ID loss also considerably enhances the reconstruction compared to other cases in which we do not use the ID loss. However, the pixel-level loss slightly degrades the reconstruction in terms of SAR, but it reduces the pixel-level errors (e.g., hair color, etc.) in the reconstructed face images.

In our method, we use *synthetic* data to train our face reconstruction network and use the trained network in TI attacks against FR systems with *real* face images. Our experiments in Section 4.2 show that our proposed method outperforms SOTA TI methods in the literature on high-resolution face reconstruction. The results also indicate the vulnerability of SOTA FR models to our TI attacks. Our experiments in Section 4.3 also demonstrate that our reconstructed face images can be used to perform presentation attacks by the adversary, and can achieve high SAR values. We should note that this work is conducted with the motivation of showing the vulnerability of FR systems, and we do not condone misuse of our work for the intention of attacking real FR systems. As a matter of fact, to mitigate TI attacks against FR systems and in the light of data protection regulations such as European Union General Data Protection Regulation (EU-GDPR) [43], several biometric template protection schemes are proposed in the literature [44], [45], [46], [47], [48]. In particular, the ISO/IEC 24745 standard [49] considers irreversibility of protected templates as one of the main properties of biometric template protection methods. According to this property, it should be infeasible for an adversary to invert protected templates and reconstruct the corresponding unprotected biometric templates.

5 Conclusion

In this paper, we used *synthetic* data and proposed a new method to reconstruct high-resolution (i.e., 1024×1024) face images from facial templates in TI attacks against FR systems. We used a face generator network to generate synthetic face images and extracted their facial templates to build our training dataset. Then, we used our generated training dataset to learn a mapping from facial templates to the intermediate latent space of the face generator network using a multi-term loss function. We proposed our method for both whitebox and blackbox TI attacks against FR systems and evaluated our model (trained with synthetic data) in TI attacks against FR systems with real face images. We provided extensive experiments on four different face datasets, including the MOBIO, LFW, AgeDB, and IJB-C datasets, demonstrating the superiority of our proposed method compared to SOTA TI methods on high-resolution face reconstruction. Moreover, we used the reconstructed face images from our TI attacks to perform digital replay attacks against real FR systems, showing the vulnerability of FR systems to presentation attacks based on the reconstructed face images with our model (trained only with synthetic train data) on real face datasets.

Acknowledgments

The authors would like to thank Karine Vaucher (Idiap Research Institute, Switzerland) for her help in conducting data collection in the presentation attack experiments.

References

[1] A. K. Jain, D. Deb, and J. J. Engelsma, "Biometrics: Trust, but verify," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 303–323, 2021.

[2] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, "Mipgan—generating strong and high quality morphing attacks using identity prior driven gan," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 365–383, 2021.

[3] 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.

[4] 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.

[5] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: an evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

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

[7] H. Otroshi Shahreza and S. Marcel, "Inversion of deep facial templates using synthetic data," in *Proceedings of the IEEE International Joint Conference on Biometric (IJCB)*, 2023.

[8] F. Boutros, V. Struc, J. Fierrez, and N. Damer, "Synthetic data for face recognition: Current state and future prospects," *Image and Vision Computing*, p. 104688, 2023.

[9] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[10] 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.

[11] 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.

[12] 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.

[13] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 158–165.

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] M. Akasaka, S. Maeda, Y. Sato, M. Nishigaki, and T. Ohki, "Model-free template reconstruction attack with feature converter," in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2022, pp. 1–5.

[21] S. Ahmad, K. Mahmood, and B. Fuller, "Inverting biometric models with fewer samples: Incorporating the output of multiple models," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022, pp. 1–11.

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

[23] 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.

[24] 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.

[25] H. O. Shahreza and S. Marcel, "Template inversion attack against face recognition systems using 3d face reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 19 662–19 672.

[26] 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 (CVPR)*, 2019, pp. 4401–4410.

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[28] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano, "Hyperstyle: Stylegan inversion with hypernetworks for real image editing," in *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, 2022, pp. 18 511–18 521.

[29] X. Hu, Q. Huang, Z. Shi, S. Li, C. Gao, L. Sun, and Q. Li, "Style transformer for image inversion and editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 337–11 346.

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

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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 (CVPR)*, 2021, pp. 13 733–13 742.

[38] 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," *arXiv preprint arXiv:2103.14030*, 2021.

[39] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *Proceedings of the 20th ACM Conference on Multimedia Systems (ACMMM)*, Oct. 2012.

[40] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, Aug. 2017.

[41] *ISO/IEC 30107-3:2017(E) Information technology – Biometric presentation attack detection – Part 3: Testing and reporting*, International Organization for Standardization International Standard, Jun. 2017.

[42] T. de Freitas Pereira and S. Marcel, "Fairness in biometrics: a figure of merit to assess biometric verification systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2021.

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

[44] K. Nandakumar and A. K. Jain, "Biometric template protection: Bridging the performance gap between theory and practice," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 88–100, 2015.

[45] P. Kaur, N. Kumar, and M. Singh, "Biometric cryptosystems: a comprehensive survey," *Multimedia Tools and Applications*, pp. 1–56, 2022.

[46] M. Sandhya and M. V. Prasad, "Biometric template protection: A systematic literature review of approaches and modalities," in *Biometric Security and Privacy*. Springer, 2017, pp. 323–370.

[47] H. O. Shahreza, V. K. Hahn, and S. Marcel, "Mlp-hash: Protecting face templates via hashing of randomized multi-layer perceptron," in *Proceedings of the 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 605–609.

[48] H. O. Shahreza, C. Rathgeb, D. Osorio-Roig, V. K. Hahn, S. Marcel, and C. Busch, "Hybrid protection of biometric templates by combining homomorphic encryption and cancelable biometrics," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022.

[49] ISO/IEC JTC1 SC27 Security Techniques, *ISO/IEC 24745:2022. Information Technology - Security Techniques - Biometric Information Protection*, International Organization for Standardization, 2022.

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), Switzerland, 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.

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.