Inversion of Deep Facial Templates using Synthetic Data

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

Figure 1. Sample face images from the LFW dataset and their reconstructed images using our template inversion method 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. It is noteworthy that while our proposed face reconstruction network is trained on synthetic data the reconstruction is generalizable on facial templates extracted from real face images. The decision threshold corresponding to FMR $= 10^{-3}$ is 0.24 on the LFW dataset, and thus all these reconstructed face images pass this threshold.

Abstract

In this paper, we use synthetic data and propose a new method to reconstruct high-resolution face images from facial templates in a template inversion attack against face recognition systems. We use a pre-trained face generator network to generate synthetic face images, and then learn a mapping from the facial templates to the intermediate latent space of the face generator network. We train our mapping network with a multi-term loss function. During the inference stage, we use our mapping network to map facial templates to the intermediate latent code and then generate high-quality face images using the face generator network. We propose our method for whitebox and blackbox template inversion attacks against face recognition systems. We use our model (trained on synthetic data) to evaluate the vulnerability of state-of-the-art face recognition systems on real face datasets, including Labeled Faces in the Wild (LFW) and MOBIO datasets. Experimental results show the vulnerability of the state-of-the-art face recognition system to our template inversion attack. Our experiments also show that our template inversion method outperforms previous methods in the literature. The source code of our experiments is publicly available to facilitate reproducibility of our work.

1. Introduction

Applications of face recognition (FR) for automatic authentication purposes are spreading and range from personal (e.g., smartphone lock, e-banking, etc.) to large-scale (e.g., border control, national identity system, etc.) utilizations. Typically, in automatic face recognition systems, some features (also called *templates* or *embeddings*) are extracted from face images and are stored in the system's database in the enrollment stage, which are later used for comparison. Among different types of attacks are which are studied in the literature against FR systems [14, 5, 15, 13, 24, 23], a template inversion (TI) attack threatens both security and privacy of users. It is because, in a TI attack, an adversary can reconstruct the face images of enrolled users. The re-

constructed face images can reveal privacy-sensitive information of the underlying user and also can be used by the attacker to impersonate and enter the FR system.

In this paper, we focus on a TI attack against FR systems and propose a new method to reconstruct high-resolution face images from facial templates using a pre-trained face generator network. We use StyleGAN [18] as a face generation network and generate synthetic face images. Then, we build our training set by extracting face templates from the synthesized face images. We also keep the intermediate latent codes in the face generator network when synthesizing each face image in our training set. We learn a mapping from facial templates to the intermediate latent space of the StyleGAN model using a multi-term loss function. In the inference stage, we use the trained mapping to generate an intermediate latent code for StyleGAN and use the remaining part of the StyleGAN network to generate the reconstructed face image. We propose our method for whitebox (where the adversary knows the parameters and internal functioning of the feature extractor of the FR system) and blackbox (where the adversary does not have knowledge about the internal functioning of the feature extractor of the FR system) template inversion attacks against face recognition systems. We evaluate the vulnerability of state-of-the-art (SOTA) FR systems to our TI attack on two datasets of real face images, including Labeled Faces in the Wild (LFW) [16] and MOBIO [25] datasets. While our model is trained on the synthetic data, the experimental results show that on real data our model outperforms previous methods in the literature. Our experiments also show the vulnerability of SOTA FR systems to our TI attack. Fig. 1 illustrates sample face images from the LFW [16] dataset and their corresponding reconstructed face images.

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

- We propose a new method to reconstruct high-resolution face images from facial templates. We use a pre-trained face generator network to synthesize face images and extract facial templates from the synthesized images as our training set. We also keep the intermediate latent codes in the face generator network when synthesizing each face image in our training set.
- We use our synthesized training set and learn a mapping from facial templates to the intermediate latent space of the face generator network.
- While we train our network on the synthetic face images, in the inference stage we use templates extracted from the real face images and generate intermediate latent code using our trained mapping network. The generated intermediate latent code is used by the remaining part of the face generator network to reconstruct the face image.

We should highlight that using synthetic face images as training data in our proposed method has two merits: first, the adversary does not need to find a dataset of real face images to use for training. Second, we can have corresponding latent code for each face image and use it directly in our training.

The remainder of the paper is organized as follows. We first review the related works in the literature in Sec. 2. Then, we describe our proposed method in Sec. 3, and present our experimental results in Sec. 4. Finally, the paper is concluded in Sec 5.

2. Related Works

TI methods in the literature can be categorized based on the resolution of the reconstructed face images (i.e., low-resolution or high-resolution) and also the adversary's knowledge of the feature extractor of the FR system (i.e., whitebox or blackbox)

Most of the TI methods proposed in the literature generate low-resolution face images [33, 26, 7, 22, 12, 10, 29, 11, 2, 1]. In [33], authors proposed a whitebox method to generate low-resolution face images. They used a gradient-ascend optimization with regularization terms (to smooth the reconstructed face images) to generate images that have similar facial templates using a guiding image or random initialization. They also trained a deconvolution neural network with the same loss function. Similarly, in [26] a low-resolution whitebox method based on deconvolution neural network was proposed. The authors used a multi-term loss function, where several loss terms improved the image-level reconstruction and one term enhanced the facial templates of the reconstructed face image using the whitebox model of FR model. In [7], authors trained a multi-layer perceptron (MLP) and a convolutional neural network (CNN) to estimate landmarks and face textures from facial templates. Then they applied a differentiable warping to combine estimated facial landmarks and textures and reconstruct low-resolution facial images in both whitebox and blackbox TI attacks. In [22], authors proposed two CNN-based networks, called NBNet-A and NBNet-B to generate low-resolution face images in the blackbox TI attack. They trained their models with two different loss functions (pixel loss and perceptual loss using a pre-trained VGG-19 [27]) and proposed four different face reconstruction networks (two network structures and two different loss functions).

In [12, 2, 1], generative adversarial networks (GANs) are used to reconstruct low-resolution face images from facial templates. In [12], Pro-GAN [17] is trained to generate face images from facial templates in a bijection-learning framework. While their method is proposed based on whitebox scenario, they use knowledge distillation to train a student network and use the trained student network in their method. However, no details about their knowledge distil-

Table 1. Template Inversion methods in the literature.

Reference	Method Basis	Resolution	Whitebox/Blackbox	Available code
Zhmoginov and Sandler [33]	1) optimization 2) learning	low	whitebox	✗
Otroshi Shahreza *et al.* [26]	learning	low	whitebox	✓
Cole *et al.* [7]	learning	low	both	✗
Mai *et al.* [22]	learning	low	blackbox	✓
Doung *et al.* [12]	learning	low	both	✗
Akasaka *et al.* [2]	learning + opt.	low	blackbox	✗
Ahmad *et al.* [1]	learning	low	blackbox	✗
Vendrow and Vendrow [29]	optimization	high	blackbox	✓
Dong *et al.* [10]	learning	high	blackbox	✓
Dong *et al.* [11]	optimization	high	blackbox	✓
[Ours]	learning	high	both	✓

lation (e.g., the structure of the student network) are available. In [2], authors trained a generic GAN-based face generation model. Then, they trained a network to transfer the target facial templates to templates of a known FR model. Finally, they generated the reconstructed face image in the blackbox scenario by optimizing the generated face image in their GAN to have the same facial templates extracted by the known FR model. In [1], a GAN-based method is proposed to reconstruct facial images in the blackbox scenario. They investigated the size of training set of face images that the adversary needs in the training. However, all the training face images in their experiments are real face data and they did not consider that the adversary can use synthetic face images.

In contrast to low-resolution face reconstruction, there are few methods in the literature that generate high-resolution face images [29, 10, 11]. In [29], authors used a grid-search on the input (noise) of StyleGAN [19] to find the input (noise) vector that can generate a face image with a similar facial template in the blackbox TI attack. Similarly, in [11], authors applied optimization on the input of StyleGAN [19] but solved the optimization with the Genetic algorithm [28]. Instead of optimization, in [10] authors used a learning-based approach and trained a network to find the input (noise) vector of StyleGAN [19] from facial templates. However, the main drawback of all these methods is that they find a vector in the input of StyleGAN that is a random Gaussian noise in the main structure of StyleGAN and has less control over the generated output compared to intermediate layers of StyleGAN.

Table 1 compares our proposed method with previous methods in the literature. Compared to most methods in the literature that generate low-resolution face image, our method generates high-resolution and realistic face images. In contrast to previous methods that generate high-resolution face images using StyleGAN by finding an input (noise) vector to reconstruct the face image [29, 10, 11], we train a mapping form facial templates to the *intermediate* latent space of StyleGAN, which is shown to have more control on the generated face image. We also propose our method for both whitebox and blackbox TI attacks. Furthermore, our experiments show that our method outperforms previous methods in the literature in terms of the adversary's success attack rate.

3. Proposed Method

We consider the threat model as described in Sec. 3.1 and use the proposed face reconstruction method in Sec. 3.2 to invert facial templates.

3.1. Threat Model

We assume a TI attack against FR systems with the following threat model:

- *Adversary's Goal:* The adversary aims to invert face templates stored in the database of the FR systems and impersonate.
- *Adversary's Knowledge:* The adversary has access to the database of the face recognition system (complete or partial) and also has whitebox or blackbox knowledge of the feature extractor of the FR system.
- *Adversary's Capability:* The adversary can use the reconstructed face image to inject to the feature extractor of the system as a query.
- *Adversary's Strategy:* The adversary plans to train a face reconstruction network and invert the facial templates. The adversary then uses the reconstructed face image to impersonate by injecting the reconstructed face image into the FR system.

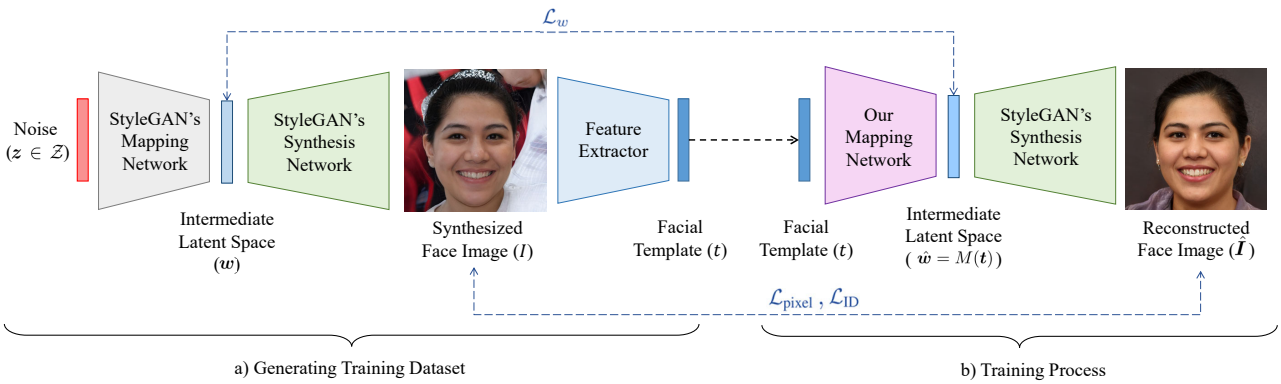

Figure 2. Block diagram of the proposed method

3.2. Template Inversion method

We assume that the adversary has access to a pre-trained face generator network such as StyleGAN [18]. StyleGAN is composed of two networks, a mapping network, and a synthesize network. The mapping network M_{StyleGAN} takes a random noise $\boldsymbol{z} \in \mathcal{Z}$ in its input and generates an intermediate latent code $\boldsymbol{w} = M_{\text{StyleGAN}}(\boldsymbol{z}) \in \mathcal{W}$. Then, the intermediate latent code $\boldsymbol{w}$ is fed to the synthesize network S_{StyleGAN} to generate the face image $\boldsymbol{I} = S_{\text{StyleGAN}}(\boldsymbol{w})$.

The adversary can use the pre-trained StyleGAN to first generate a training dataset of face images and their corresponding facial templates. To this end, the adversary can use the StyleGAN network to generate several facial images and then use the feature extractor to extract facial templates. Let us assume that the adversary could generate a dataset $\{\boldsymbol{I}_i | i = 1, .., N\}$ where N is the number of generated face images. Then, the adversary can build $\mathcal{D} = \{(\boldsymbol{t}_i, \boldsymbol{I}_i) | i = 1, .., N\}$ where $\boldsymbol{t}_i = F(\boldsymbol{I}_i)$ is the face template extracted using feature extractor $M(.)$ from face image $\boldsymbol{I}_i$.

After generating the training dataset $\mathcal{D}$ of the synthetic face images and their corresponding facial templates, the adversary can use this dataset to train a new mapping network $M(.)$ that generates the intermediate latent code $\hat{\boldsymbol{w}} = M(\boldsymbol{t})$ for face template t and then use the synthesize network S_{StyleGAN} to generate face image $\hat{\boldsymbol{I}} = S_{\text{StyleGAN}}(\hat{\boldsymbol{w}})$. We propose training the weights of the mapping network $F(.)$ 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 latent code loss, pixel loss, and ID loss, respectively, and are defined as follows:

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

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

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

The latent code loss ($\mathcal{L}_w$) minimizes the error of the generated intermediate latent code $\hat{\boldsymbol{w}} = M(\boldsymbol{t})$ in the intermediate latent space $\mathcal{W}$ of StyleGAN. The pixel loss ($\mathcal{L}_{\text{pixel}}$) minimizes the pixel-level reconstruction error for the generated face image $\hat{\boldsymbol{I}} = S_{\text{StyleGAN}}(M(\boldsymbol{t}))$. The ID loss minimizes the distance between the facial templates extracted from the original and the reconstructed face images using feature extractor $F_{\text{loss}}(.)$. In the whitebox scenario, the adversary can use the feature extractor of the target FR system (i.e., $F(.)$), however in the blackbox scenario, the adversary can use another feature extractor that has access to as $F_{\text{loss}}(.)$. Fig. 2 depicts the block diagram of the proposed method.

4. Experiments

4.1. Experimental Setup

We consider SOTA FR systems as target systems and evaluate their vulnerability to our TI attack. We use ArcFace [8], ElasticFace [6], and also different FR models with SOTA backbones from FaceX-Zoo [31], including AttentionNet [30], HRNet [32], RepVGG [9], and Swin [21]. Table 2 reports the recognition performances of these models.

To evaluate the vulnerability of these FR models, we use the MOBIO [25] and Labeled Faces in the Wild (LFW) [16] datasets. The MOBIO dataset includes face images of 150 subjects captured using mobile devices in 12 sessions (6-11 samples in each session). The LFW dataset contains 13,233 face images of 5,749 subjects collected from the internet, in which 1,680 subjects have two or more images. For each of the MOBIO or LFW datasets, we build a FR system and then invert the enrolled facial templates to reconstruct face

Table 2. Recognition performance of face recognition models in terms of true match rate (TMR) at FR system false match rates (FMRs) of 10^{-2} and 10^{-3} evaluated on the MOBIO and LFW datasets. The TMR values are in percentage.

model	MOBIO		LFW	
	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
ArcFace	100.00	99.98	97.60	96.40
ElasticFace	100.00	100.00	96.87	94.70
AttentionNet	99.71	97.73	84.27	72.77
HRNet	98.98	98.23	89.30	78.43
RepVGG	98.75	95.80	77.20	58.07
Swin	99.75	98.98	91.70	87.83

images. Next, according to our threat model described in Sec. 3.1, we inject the reconstructed face images into the feature extractor of the FR system as a query and evaluate the adversary's success attack rate (SAR) at different false match rates (FMRs) of the FR system.

We use the Bob[1] toolbox [4, 3] to build the pipelines for the FR systems and evaluate the TI attacks against FR systems. We also use the PyTorch package and trained our template inversion models using Adam optimizer [20] with the learning rate of 10^{-4} on a system equipped with an NVIDIA GeForce RTX^{TM} 3090. In our experiments, we use the pre-trained model of StyleGAN3[2] to generate 1024×1024 high-resolution face images. We generated 25,000 synthetic face images for our training set in our experiments. The source code of our experiments is publicly available to help researchers reproduce our results[3].

4.2. Comparison with previous methods

We compare the performance of our proposed method with previous works in the literature with available source code, including NBNet-A-M [22], NBNet-A-P [22], NBNet-B-M [22], NBNet-B-P [22], Vendrow and Vendrow [29], Dong *et al.* [10], and Dong *et al.* [11]. Table 3 and Table 4 compare the performance of our method with these methods in terms of adversary's success attack rate (SAR) against different SOTA FR systems at FMRs of 10^{-2} and 10^{-3}, respectively, on the MOBIO and LFW datasets. For our method, we use ArcFace and ElasticFace as F_{loss} in our loss function (Eq. 4) to reconstruct face images from facial templates extracted from different FR systems and train a separate model for each FR model. As the results in these tables show, our method outperforms previous methods in the literature. In particular, compared to [29, 10, 11] which used StyleGAN to generate high-resolution and realistic face images our method achieves superior performance. Comparing the results in these two

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

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

[3]Source code: https://gitlab.idiap.ch/bob/bob.paper.ijcb2023_face_ti

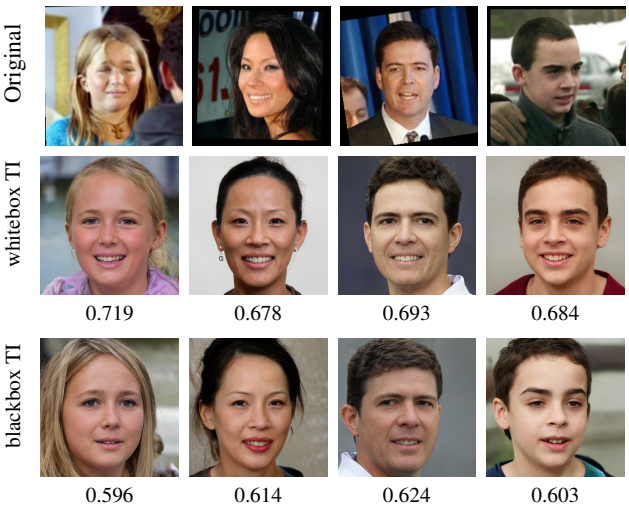

Figure 3. Sample real face images from the LFW dataset (first row) and their reconstructed images from ArcFace templates in whitebox (second row) and blackbox (third row). The values below each image show the cosine similarity between the corresponding templates of original and reconstructed face images.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

tables with the recognition performances of FR systems reported in Table 2, we observe that a FR system with a higher recognition accuracy is more vulnerable to our attack. Comparing the results of ArcFace and ElasticFace in the loss function of our method, the results show that ArcFace leads to better SAR values, which may be due to the fact that ArcFace has a better recognition performance than ElasticFace as shown in Table 2. Fig. 3 illustrates sample face images from LFW dataset and their corresponding reconstructed face images in whitebox and blackbox TI attacks using ArcFace templates.

4.3. Ablation Study

To evaluate the effect of each loss term in our proposed method, we perform an ablation study, where we train different mapping networks with different loss functions and evaluate the adversary's SAR in a TI attack against a FR system. To this end, we consider a whitebox TI attack against a FR system based on ArcFace on the MOBIO and LFW datasets. Table 5 reports the effect of each loss term in our proposed method. As the results in this table show, each term in our loss function improves the reconstructed face images in TI attacks against FR systems. In particular, we observe that using the latent code loss (i.e., $\mathcal{L}_w$) helps the training compared to using all other terms except the latent code loss term. This also highlights the use of synthetic data in our proposed method where we have the correct latent code for each single image in our synthetic training dataset. When using the latent code loss, our ID loss

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

method	**MOBIO**						**LFW**					
	ArcFace	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**	**ArcFace**	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**
NBNetA-M [22]	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 [22]	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 [22]	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 [22]	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.* [10]	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 [29]	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.* [11]	87.62	90.95	80.48	71.90	44.29	82.38	**87.26**	89.00	55.40	59.46	28.60	69.07
[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**

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

method	**MOBIO**						**LFW**					
	ArcFace	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**	**ArcFace**	**Els.Face**	**Att.Net**	**HRNet**	**RepVGG**	**Swin**
NBNetA-M [22]	0	2.38	0	0	0	0	4.32	10.90	1.24	1.60	1.13	3.82
NBNetA-P [22]	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 [22]	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 [22]	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.* [10]	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 [29]	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.* [11]	61.43	76.67	42.86	49.05	20.00	65.71	**74.48**	73.67	32.07	31.73	10.89	53.59
[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**

Table 5. Ablation study on the effect of each loss term in whitebox attack against ArcFace in terms of SAR for a 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}**
$\mathcal{L}_{\text{total}} = \mathcal{L}_w$	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
$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}$	0	0	0.32	0.02

also significantly improves the reconstruction compared to other cases without ID loss. The pixel-level loss, however, slightly degrades the SAR values but reduces the pixel-level errors (e.g., hair color, etc.) in the reconstructed face images.

5. Conclusion

In this paper, we proposed a new method to reconstruct high-resolution face images from facial templates in TI attacks against FR systems. We used StyleGAN as a pre-trained face generator network to synthesize a set of face images. Then, we built our training set by extracting facial templates from the synthesized face images. We trained a mapping network from facial templates to the intermediate latent space of StyleGAN using a multi-term loss function. We used the trained mapping network to generate an intermediate latent code for each facial template and generate the reconstructed face image using the generated intermediate latent code through the remaining network of StyleGAN. We evaluated our proposed method on the real face images from the MOBIO and LFW datasets. Our experiments show the vulnerability of SOTA face recognition systems to our TI attack. Experimental results also show that our template inversion method outperforms previous methods in the literature.

We should note that in our experiments, we evaluated the vulnerability of FR systems by injecting the reconstructed face image into the feature extractor of FR systems. However, the injection of reconstructed face images may not be feasible in practice in attacks against real face recognition systems. Therefore, it is necessary to evaluate the vulnerability of FR systems in more practical scenarios, such as presentation attacks using reconstructed face images, which can be studied in future work.

Acknowledgments

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

References

[1] 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)*, pages 1–11. IEEE, 2022.

[2] 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)*, pages 1–5. IEEE, 2022.

[3] 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*, pages 1–8, 2017.

[4] 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.

[5] 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*, 32(5):31–41, 2015.

[6] 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)*, pages 1578–1587, 2022.

[7] 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*, pages 3703–3712, 2017.

[8] 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.

[9] 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)*, pages 13733–13742, 2021.

[10] 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)*, pages 1–11. IEEE, 2021.

[11] 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*, 125:103026, 2023.

[12] 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*, pages 6132–6141, 2020.

[13] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[14] 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*, 43(3):1027–1038, 2010.

[15] A. Hadid, N. Evans, S. Marcel, and J. Fierrez. Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. *IEEE Signal Processing Magazine*, 32(5):20–30, 2015.

[16] 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. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[17] 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.

[18] 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*, 34, 2021.

[19] 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)*, pages 8110–8119, 2020.

[20] 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.

[21] 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.

[22] 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*, 41(5):1188–1202, 2018.

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

[24] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans. *Handbook of biometric anti-spoofing: Presentation attack detection*, volume 2. Springer, 2019.

[25] C. McCool, R. Wallace, M. McLaren, L. El Shafey, and S. Marcel. Session variability modelling for face authentication. *IET Biometrics*, 2(3):117–129, Sept. 2013.

[26] 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)*, pages 1211–1215. IEEE, 2022.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] M. Srinivas and L. M. Patnaik. Genetic algorithms: A survey. *computer*, 27(6):17–26, 1994.

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

[30] 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*, pages 3156–3164, 2017.

[31] 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.

[32] 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.

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