Breaking Template Protection: Reconstruction of Face Images from Protected Facial Templates

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

***Abstract*— Face recognition systems tend toward ubiquity and are commonly utilized for security purposes. These systems operate based on facial representations, called templates, extracted by a deep neural network from each face image. However, it has been shown that face recognition templates can be inverted to reconstruct underlying face images, posing new security and privacy threats to face recognition systems. To mitigate such attacks against face recognition systems, several biometric template protection schemes have been proposed in the literature. The ISO/IEC 24745 standard requires each biometric template protection scheme to fulfill several requirements, among which non-invertibility is of the utmost importance. Therefore, each of the proposed template protection schemes in the literature used an ad-hoc approach to investigate the invertibility of the protected templates. In this paper, we consider a scenario where an adversary gains knowledge of a template protection scheme as well as its secrets, and tries to reconstruct a face image using a leaked protected template. We consider different template protection schemes, including BioHashing, MLP-Hashing, and Homomorphic Encryption (HE), and reconstruct face images from protected templates. We also use different state-of-the-art face recognition models in both whitebox and blackbox scenarios. To our knowledge, this is the first work on learning-based reconstruction of face images from protected facial templates.**

I. INTRODUCTION

Face recognition systems are being widely used in different applications which require automatic authentication, such as unlocking cell phones, e-banking, border control, etc. In such systems, a deep neural network is often used to extract some features (also known as *embeddings* or *templates*) from each face image, and then the extracted features are stored in the database of the system which are later used for automatic recognition. However, the extracted facial templates contain privacy-sensitive information about each individual. For example, it has been shown that an adversary can reconstruct the face image of the corresponding subject based on raw facial templates [22], [35], [36].

To prevent such types of attacks against face recognition systems, and in the light of data protection regulations[1] which consider biometric data as sensitive information, several biometric protection (BTP) schemes have been proposed in the literature. The ISO/IEC 24745 standard [12] also establishes four main requirements for each BTP scheme, including renewability, unlinkability, irreversibility, and recognition performance preservation. The ISO/IEC 30136 standard [13] defines different scenarios where the adversary has different levels of knowledge of the biometric system and its secrets to evaluate each of BTP requirements. Among these requirements, the irreversibility of protected templates is of significant importance for each BTP scheme. However, it is always challenging to investigate the invertability of protected templates since BTP schemes have different mechanisms, and therefore for each BTP scheme, a specific inversion method has been used in the literature. In addition, despite general measures in the literature to evaluate linkability of protected templates (such as [42]), there is no general method to investigate invertability of protected templates.

Fig. 1: Sample face images from the FFHQ dataset (first row) and their corresponding reconstructed face images from ArcFace templates protected with Homomorphic Encryption (BFV) in a whitebox attack. The values indicate cosine similarity between templates of the original and reconstructed face images. The decision threshold corresponding to FMR $= 10^{-3}$ is 0.24 for ArcFace on the LFW dataset.

In this paper, we focus on the inversion of face images from protected facial templates. We consider a scenario where the adversary gains knowledge of the template protection scheme as well as its secrets[2] and tries to reconstruct the face image using a leaked protected template. We consider different template protection schemes, including BioHashing,

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

[1]such as the EU General Data Protection Regulation (GDPR) [27]

[2]which is the case in the *full-disclosure* scenario defined in the ISO/IEC 30136 standard [13] for evaluating the invertibility of protected templates.

MLP-Hashing, and Homomorphic Encryption (HE), and reconstruct face images from protected templates. We also use different state-of-the-art face recognition models in both *whitebox* (where the adversary has a complete knowledge of feature extractor) and *blackbox* (where the adversary has a blackbox knowledge of feature extractor) scenarios. Fig. 1 presents sample reconstructed face images from ArcFace templates protected with Homomorphic Encryption (BFV) in a whitebox attack using our method. To our knowledge, this is the first work on the reconstruction of face images from protected facial templates, which is independent of the template protection scheme and can be applied against different protection schemes.

The remainder of this paper is organized as follows. First, we review related work in the literature in Section II. Then, in Section III, we describe the threat model and explain our face reconstruction method. In Section IV, we present our experimental results. Finally, the paper is concluded in Section V.

II. RELATED WORK

In this section, we review the related work in the literature on biometric template protection and reconstruction of face images from facial templates.

A. Biometric Template Protection

In the last two decades, several biometric template protection (BTP) schemes have been proposed in the literature [29], [30]. The main objective of each BTP scheme is to generate new (protected) templates from raw templates that contain less leakage of information from raw (unprotected) biometric data. To satisfy the non-invertiblity of protected templates, each BTP scheme is using some secrets, which are referred to as *key*, along with raw templates in their algorithms. In cancelable biometrics protection methods (such as BioHashing [14], MLP-Hashing [32], IoM-Hashing [15], etc.), a transformation function is used to generate protected templates. The transformer function is dependent on a *key* and raw templates. The generated protected templates are then used instead of raw (unprotected) templates, and the recognition is made by comparing the protected templates [25], [29], [40]. In biometric cryptosystems (such as fuzzy vault [16], fuzzy commitment [17], etc.), a key is either bound with a biometric template or generated from a biometric template. Then, recognition is performed based on the correct retrieval or generation of the key [26], [43]. Some works also used Homomorphic Encryption (HE) to generate protected templates in the ciphertext. Then, the comparison is carried out in the ciphertext, and then the comparison score [4], [9], [41] or decision [2], [3] is decrypted into the plaintext. Since the irreversibility of protected templates is an important requirement of each BTP scheme, each of the proposed methods in the literature has used an ad-hoc approach to investigate the inversion of protected templates.

B. Face Reconstruction from Facial Templates

Several works in the literature explored the reconstruction of face images from facial templates, particularly from raw (unprotected) templates. In general, methods for reconstructing face images from raw facial templates can be categorized into optimisation-based [8], [44] and learning-based [22], [31], [33]–[37], [39]. In optimization-based methods, an iterative algorithm is used to generate a face image that has a similar facial template as the target template. In contrast, in the learning-based methods, a neural network is trained to reconstruct face images from facial templates. The face reconstruction methods can also be categorized based on the required knowledge of the feature extractor into whitebox and blackbox. While some of the face reconstruction methods used pretrained face generator networks [8], [35], [36], [39], [44], some other works trained a convolutional neural network to reconstruct face images from facial templates [22], [31], [33]. In [38], it has been shown that an adversary can reconstruct face images even if a portion of face template is leaked.

In contrast to most work on the reconstruction of face images from raw templates, [19] used the network in [22] to reconstruct binarised facial features. However, no template protection mechanism was considered, and the authors only considered a simple binarisation transformation being applied to raw templates. To our knowledge, no learning-based approach has been used in the literature to reconstruct face images from protected biometric templates. In this paper, we consider the situation where the adversary has access to secrets of the BTP scheme and directly reconstructs face images from protected templates.

III. PROPOSED METHOD

In this section, we present our method to reconstruct face images from protected facial templates. We describe our threat model in Section III-A, where the adversary gains access to a protected template and aims to reconstruct the underlying face image. We also describe our network to reconstruct face images from leaked protected templates in Section III-B.

A. Threat Model

We consider the protected face recognition system with the situation where an adversary gains access to a protected facial template and aims to reconstruct a face image from the leaked protected template and use the reconstructed face image to impersonate. We consider the following properties for the adversary:

- *Adversary's goal:* The adversary aims to impersonate a user enrolled in the FR system.
- *Adversary's knowledge:* The adversary has the following information:
  1) A leaked protected face template $\mathbf{t}_{\text{btp}}$ of a user enrolled in the database of the face recognition system.
  2) Complete knowledge of template protection scheme P and its secrets $\mathbf{k}_{\text{btp}}$.
  3) The whitebox knowledge (including parameters and internal functioning) or blackbox knowledge of the feature extraction model $F(.)$ of the face

recognition system. In the case of the blackbox scenario, the adversary is assumed to have white-box knowledge of an alternative feature extraction model $F_{\text{adv}}(.)$.

- *Adversary's capability:* The adversary can inject the reconstructed face image into the feature extractor of the target face recognition system and bypass the camera. For simplicity and to verify how similar is the reconstructed face image to the original image, we assume that injection is made to a similar system without protection.
- *Adversary's strategy:* Under the above assumptions, the adversary can use the leaked protected template and underlying reconstruct face image $\hat{\mathbf{I}}$ using a face reconstruction method. Then, the adversary can use the reconstructed face image $\hat{\mathbf{I}}$ to inject a query to impersonate.

B. Face Reconstruction

Our network for reconstructing face images from the leaked *protected* templates stems from the network proposed in [33] for reconstructing unprotected facial templates. We consider a dataset of face images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$, with N images. We extract facial templates $\mathbf{t} = F(\mathbf{I})$ from each face image $\mathbf{I}$, and then generate the protected version $\mathbf{t}_{\text{btp}} = P(\mathbf{t}, \mathbf{k}_{\text{btp}})$ with the template protection scheme P using the leaked secret $\mathbf{k}_{\text{btp}}$. The secret $\mathbf{k}_{\text{btp}}$ can be user-specific or identical for all subjects (i.e., application-based key). In either case, it is assumed to be known by the adversary. In case of encryption-based template protection (such as HE), the protected templates $\mathbf{t}_{\text{btp}}$ are in the ciphertext, however since the adversary is assumed to have access to secret key, the adversary can decrypt the protected template into the plaintext[3] as $\mathbf{t}_{\text{btp,adv}}$ and use it for the inversion attack. In other cases where the protection is not based on encyption (such as cancelable biometric schemes), the adversary can directly use the protected template for the inversion attack, i.e., $\mathbf{t}_{\text{btp,adv}} = \mathbf{t}_{\text{btp}}$. Then, the adversary can build the training dataset $\mathcal{D} = \{(\mathbf{t}_{\text{btp,adv},i}, \mathbf{I}_i)\}_{i=1}^{N}$ with N pairs of protected templates $\mathbf{t}_{\text{btp},i}$ and their corresponding face images $\mathbf{I}_i$. We use the network structure in [33], composed of enhanced deconvolution using cascaded convolution and skip connections (shortly, DSCasConv) blocks, and use the protected template (instead of the unprotected template) as the input. We optimize our model with a multi-term loss function, including:

- *Mean Absolute Error (MAE):* To reduce the pixel level reconstruction error, we minimize the ℓ_1 of reconstruction error:
$$\mathcal{L}_{\text{MAE}}(\hat{\mathbf{I}}, \mathbf{I}) = ||\hat{\mathbf{I}} - \mathbf{I}||_1, \tag{1}$$
where $\hat{\mathbf{I}}$ and $\mathbf{I}$ are the reconstructed and original face images, respectively.
- *Dissimilarity Structural Index Metric (DSSIM):* To enhance the quality of the reconstructed image in terms of the Similarity Structural Index Metric (SSIM) [48], we further minimize the DSSIM loss term [28] as follows:
$$\mathcal{L}_{\text{DSSIM}}(\hat{\mathbf{I}}, \mathbf{I}) = \frac{1 - \text{SSIM}(\hat{\mathbf{I}}, \mathbf{I})}{2} \tag{2}$$
- *ID loss:* To preserve the identity information in the reconstructed face image, we use a feature extractor F_{adv} and minimize the distance between the features extracted from the original face $\mathbf{I}$ and reconstructed face $\hat{\mathbf{I}}$ images. We minimize the ℓ_1-norm distance and cosine distance of the extracted templates as follows:
$$\begin{aligned}\mathcal{L}_{\text{ID}}(\hat{\mathbf{I}}, \mathbf{I}) &= \mathcal{L}_{\text{ID},\ell_1}(\hat{\mathbf{I}}, \mathbf{I}) + \mathcal{L}_{\text{ID,cos}}(\hat{\mathbf{I}}, \mathbf{I}) \\ &= ||F_{\text{adv}}(\hat{\mathbf{I}}) - F_{\text{adv}}(\mathbf{Y})||_1 + \frac{-F_{\text{adv}}(\hat{\mathbf{I}}).F_{\text{adv}}(\mathbf{I})}{||F_{\text{adv}}(\hat{\mathbf{I}})||_2.||F_{\text{adv}}(\mathbf{I})||_2}\end{aligned} \tag{3}$$
Similar to [34]–[36], in the whitebox scenario we consider $F_{\text{adv}} = F$, but in the blackbox scenario we assume that the adversary has access to an alternative model F_{adv} and uses it in the loss function.

We use a weighted summation of these loss terms as our total loss:
$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \gamma_1 \mathcal{L}_{\text{DSSIM}} + \gamma_2 \mathcal{L}_{\text{ID}} \tag{4}$$
We experimentally found that the choice of $\gamma_1 = 0.75$, and $\gamma_2 = 0.025$ achieves the best performance.

IV. EXPERIMENTS

In this section, we present our experimental results and discuss our findings. In Section IV-A, we describe our experimental setup. Then, we present our experimental result in the reconstruction of protected templates in Section IV-B. Finally, we discuss our findings in Section IV-C.

A. Experimental Setup

Biometric template protection schemes: We consider different biometric template protection schemes, including two cancelable biometric schemes as well as a template protection method based on Homomorphic Encryption (HE). For cancelable biometric, we consider BioHashing [14] (which is a simple and popular scheme) and MLP-Hash [32] (which is a recently proposed scheme). We consider the protected systems with these schemes to be operating with a user-specific key setting, and thus, the adversary knows the key for the leaked facial template. For the HE-based method, different algorithms have been used for biometric template protection. For instance, HE based on the CKKS scheme supports floating-point encryption, and thus decryption of the protected template using the leaked template will lead to the original unprotected template. In contrast, some other schemes, such as BFV, support integers and, therefore, require quantization before encryption. That means the decryption of the protected templates leads to quantized templates

[3]In such cases, the protected templates in the ciphertext $\mathbf{t}_{\text{btp,ciphertext}} = \text{Enc}(M(\mathbf{t}), \mathbf{k}_{\text{btp}})$ are generated by encrypting the mapped template $M(\mathbf{t})$, where $M(.)$ is a transformation function which is specific to the encryption algorithm (e.g., quantization). Therefore, decrypting the protected version into the plaintext willl lead to $\mathbf{t}_{\text{btp,plaintext}} = \text{Dec}(\mathbf{t}_{\text{btp,ciphertext}}, \mathbf{k}_{\text{btp}}) = M(\mathbf{t})$, which is the mapped version of unprotected template, and thus $\mathbf{t}_{\text{btp,adv}} = M(\mathbf{t})$. We should note that $\text{Enc}(.,.)$ and $\text{Dec}(.,.)$ denote encryption and decryption, respectively.

TABLE I: Recognition performance of face recognition models used in our experiments in terms of true match rate (TMR) at the thresholds correspond to false match rates (FMRs) of 10^{-2} and 10^{-3} evaluated on the MOBIO, LFW, and AgeDB datasets. The values are in percentage.

FR model	MOBIO		LFW		AgeDB	
	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
ElasticFace	100.00	100.00	96.87	94.70	98.20	97.57
AttentionNet	99.71	97.73	84.27	72.77	97.93	96.90
HRNet	98.98	98.23	89.30	78.43	97.67	96.23
RepVGG	98.75	95.80	77.20	58.07	95.93	93.93
Swin	99.75	98.98	91.70	87.83	98.03	97.10

in plaintext. In our experiments, we consider the protection based on HE schemes that require quantized templates. We should note that in the HE-based protection, the secret key (i.e., private key) is often the same for all subjects.

Face recognition models: We use state-of-the-art face reconstruction models including ArcFace [6] and ElasticFace [5] as well as four different face recognition models with state-of-the-art backbones from FaceX-Zoo [46], including AttentionNet [45], HRNet [47], RepVGG [7], and Swin [21]. We use the pretrained models of each of these network trained on the MS-Celeb-1M dataset [10]. Table I compares the recognition performance of these models on the MOBIO, LFW, and AgeDB datasets.

Dataset: We use the Flickr-Faces-HQ (FFHQ) [18] dataset for training our face reconstruction model. The FFHQ dataset contains 70,000 face images that we randomly split into 90% training and 10% validation. We also evaluate our models on the MOBIO [23], LFW [11], and AgeDB [24] datasets. We build protected face recognition systems using the mentioned face recognition model and BTP schemes on each of our evaluation datasets. Then, we use our corresponding reconstruction model trained on FFHQ to invert enrolled protected templates and reconstruct face images. We inject the reconstructed face image as a query to the system to evaluate the performance of face reconstruction in terms of an adversary's Success Attack Rate (SAR) in entering an unprotected face recognition system with the same feature extractor when the system is configured at False Match Rate (FMR) of 10^{-3}.

Implementation: We use the PyTorch package and Bob toolbox [1] in our implementations. We train our face reconstruction networks using the Adam [20] optimizer with the initial learning rate of 10^{-3}, and we decrease the learning rate every 10 epochs, by a factor of 0.5. The source code of our experiments is publicly available[4].

B. Face Reconstruction from Protected Templates

We consider face recognition systems protected with BioHashing, MLP-Hash, and HE and assume that the adversary knows the template protection scheme and its secrets. We train our face reconstruction network and use the protected templates stored in the database of the face recognition system to reconstruct the face images. We use ArcFace as F_{adv} and evaluate the performance of our method in attack against protected templates of different face recognition models. Table II and Table III report the adversary's success attack rate in entering a face recognition with the same feature extractor on false match rates (FMRs) of 10^{-2} and 10^{-3} on the MOBIO, LFW, and AgeDB datasets. We should note that since we use ArcFace as F_{adv}, the attacks against ArcFace are whitebox attacks but against other face recognition models are blackbox attacks. As the results in these tables shows, the reconstructed face images by inverting protected templates using our method achieve significant performance in attacks against protected templates. Fig. 1 shows sample reconstructed face images using our method in the reconstruction of ArcFace templates protected with Homomorphic Encryption (HE) in a whitebox attack. Fig. 2 also shows sample reconstructed face images using our method in the reconstruction of ElasticFace templates protected with different template protection schemes in blackbox attacks using ArcFace as F_{adv}. As the sample reconstructed face images show, inversion of protected templates can reveal important information about underlying subjects.

To further explore the effect of F_{adv}, we consider HE-protected templates and use ElasticFace as F_{adv}. In addition, we consider the whitebox scenario, where the adversary has access to the feature extractor of the face recognition model and uses it as F_{adv}. As the results in Table IV show, using the same feature extractor (i.e., whitebox attack) or different feature extractor (i.e., ArcFace or ElasticFace in blackbox attacks), the reconstructed face images achieve very similar performances. Even in some cases, such as attacks against AttentionNet, we can see that the blackbox attack using ArcFace as F_{adv} achieves better performance than the whitebox attack. This observation can be interpreted considering the superior performance of ArcFace compared to other face recognition models used in our experiments, as reported in Table I. Therefore, we can expect that ArcFace enhances the reconstruction when it used as F_{adv}.

C. Discussion

Our experiments in Section IV-B show that if the template protection scheme and its secrets are known, then an adversary can reconstruct face images from protected facial templates. We considered different feature extractors protected with different template protection schemes and evaluated

[4]https://gitlab.idiap.ch/bob/bob.paper.fg2024_breaking_btp.

TABLE II: Performance of reconstructed face images from protected templates in attacking a face recognition system with same feature extractor evaluated on the MOBIO, LFW, and AgeDB datasets for the false match rate of $\mathbf{10^{-2}}$. The ArcFace model is used as F_{adv}, and thus the attacks against ArcFace are **whitebox** but against other face recognition models are in **blackbox** (denoted as cell color in gray). The values are in percentage.

Dataset	BTP	Face Recognition					
		ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
MOBIO	BioHashing	100.0	100.0	98.57	99.05	95.71	100.0
	MLP-Hash	96.67	91.9	84.29	82.86	75.24	94.76
	HE	100.0	100.0	100.0	99.05	98.1	99.52
LFW	BioHashing	95.74	96.34	79.73	85.73	69.31	90.18
	MLP-Hash	88.2	91.0	58.09	66.04	47.62	77.61
	HE	97.18	96.57	82.87	87.52	73.07	91.65
AgeDB	BioHashing	83.23	88.13	73.51	71.63	67.44	89.79
	MLP-Hash	62.89	64.73	32.94	32.74	28.42	58.8
	HE	92.75	90.88	77.99	78.61	72.93	91.8

TABLE III: Performance of reconstructed face images from protected templates in attacking a face recognition system with same feature extractor evaluated on the MOBIO, LFW, and AgeDB datasets for the false match rate of $\mathbf{10^{-3}}$. The ArcFace model is used as F_{adv}, and thus the attack against ArcFace is **whitebox** and against other face recognition models are in **blackbox** (denoted as cell color in gray). The values are in percentage.

Dataset	BTP	Face Recognition					
		ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
MOBIO	BioHashing	99.52	100.0	91.43	95.24	89.05	100.0
	MLP-Hash	78.57	80.95	56.67	59.05	47.62	85.71
	HE	100.0	99.52	97.62	97.14	96.19	99.52
LFW	BioHashing	92.69	92.79	66.4	68.29	56.3	85.21
	MLP-Hash	77.51	77.1	30.44	30.83	24.49	63.62
	HE	95.92	93.91	72.54	73.51	60.58	87.52
AgeDB	BioHashing	62.47	76.81	49.34	44.57	50.92	73.54
	MLP-Hash	36.37	45.34	13.05	11.8	15.03	31.68
	HE	82.35	82.1	56.34	53.43	57.4	79.63

TABLE IV: Performance of reconstructed face images from HE-protected templates in attacking a face recognition system with same feature extractor using different models as F_{adv}, evaluated on the MOBIO, LFW, and AgeDB datasets for the false match rate of 10^{-3}. In case F_{adv} is the same as target face recognition model, the adversary is assumed to have knowledge of the target model and thus the attack is *whitebox*; otherwise, the attack is *blackbox* (denoted as cell color in gray). The values are in percentage.

Dataset	$\mathbf{F}_{\text{adv}}$	Face Recognition					
		ArcFace	ElasticFace	AttentionNet	HRNet	RepVGG	Swin
MOBIO	**ArcFace**	100.0	99.52	97.62	97.14	96.19	99.52
	ElasticFace	100.0	99.52	97.62	97.14	96.19	99.52
	same (i.e., whitebox)	100.0	99.52	95.24	97.14	96.19	99.52
LFW	**ArcFace**	95.92	93.91	72.54	73.51	60.58	87.52
	ElasticFace	95.92	93.91	71.25	73.51	60.58	87.02
	same (i.e., whitebox)	95.92	93.91	71.38	73.51	60.58	87.61
AgeDB	**ArcFace**	82.35	82.1	56.34	53.43	57.4	79.63
	ElasticFace	82.35	82.1	54.05	53.42	57.4	79.0
	same (i.e., whitebox)	82.35	82.1	55.21	53.42	57.4	79.41

them on the MOBIO, LFW, and AgeDB datasets. However, the reconstructed face images from different feature extractors have different performances when comparing for the same protection scheme and the same dataset. For most cases, the model with higher recognition performance in Table I is more vulnerable to reconstruction attack. Similarly, if the performance of a model is better on a dataset in Table I, the attack rates are higher on that dataset in Table II and Table III. Comparing different template protection schemes, we can see that in most cases, protected templates with HE are the most invertible, but protected templates with MLP-Hash are more robust to inversion and lead to the lowest invertibility.

One of the limitations of our work is that the adversary needs to train a face reconstruction network for each set of secrets. In protected systems that have the same secrets for all subjects (such as in HE), the adversary needs only to train a single face reconstruction network. However, to

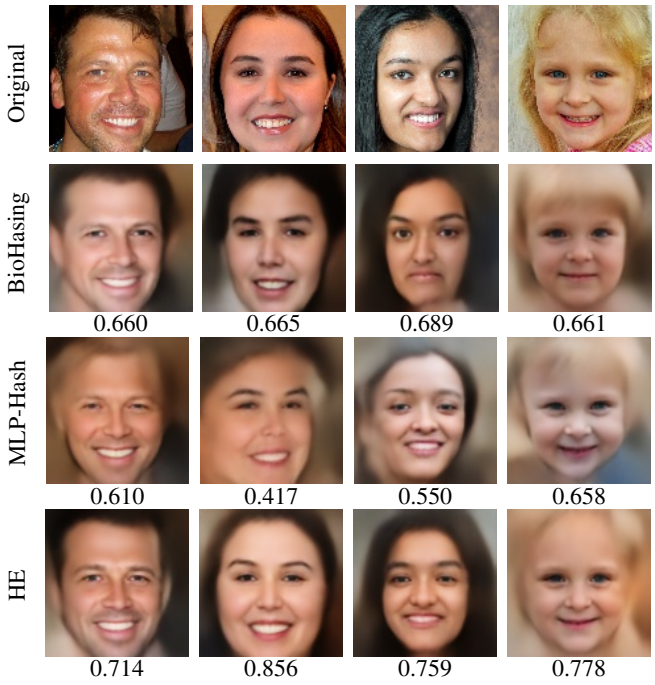

Fig. 2: Sample face images from the FFHQ dataset (first row) and their corresponding reconstructed face images from ElasticFace templates protected with BioHashing (second row), MLP-Hash (third row), and Homomorphic Encryption (fourth row) in blackbox attacks. The values indicate cosine similarity between templates of the original and reconstructed face images. The decision threshold corresponding to FMR $= 10^{-3}$ is 0.29 for ELasticFace on the LFW dataset.

attack a protected system with a user-specific key setting, the adversary needs to train different face reconstruction models for each leaked protected template. This limitation, however, can be resolved if the adversary applies optimization-based approaches described in Section II-B, which do not require gradients in their method.

All in all, our experiments in this paper demonstrate that protected templates are still vulnerable to template inversion attacks. In particular, if an adversary gains access to the template protection scheme and its secrets, they can reconstruct face images directly from the protected template. The recent social concerns on privacy in biometric systems and in the view of data protection regulations demand more future work toward secure and protected biometric systems. We would like to highlight that while in this paper we show the vulnerability of protected face recognition systems, we do not condone using our work with the intent of attack to real face recognition systems. In fact, the main motivation for this work is to demonstrate such a vulnerability in the protected face recognition systems and to encourage the scientific community to develop the next generation of safe and protected face recognition systems. As the results in this paper show, the disclosure of protection keys poses critical vulnerability in the protected face recognition systems and necessary measures should be taken for the secrecy of protection keys. We should also note that the project on which the work has been conducted has passed an Internal Ethical Review Board (IRB).

V. CONCLUSION

In this paper, we proposed the first learning-based method for the inversion of protected templates, which can be used for different protection mechanisms. We considered a protected face recognition system in a situation where an adversary gains knowledge of the template protection scheme and its secrets and tries to reconstruct the face image using a leaked protected template. To this end, we trained a neural network to generate face images from protected facial templates. In our experiments, we considered different template protection schemes, including BioHashing, MLP-Hashing, and Homomorphic Encryption (HE), and reconstructed face images from protected templates. We also used different state-of-the-art face recognition models and inverted their protected templates in both whitebox and blackbox scenarios. The experimental results show that our method can be used to reconstruct face images from templates protected by different template protection schemes and shed light on the vulnerability of protected face recognition systems to template inversion attacks. Considering the importance of template protection in the light of data protection regulations, our proposed method and experimental results pave the way for evaluating the vulnerability of protected face recognition systems and shed light on the necessity of more research toward robust and protected biometric systems.

REFERENCES

[1] 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.

[2] A. Bassit, F. Hahn, J. Peeters, T. Kevenaar, R. Veldhuis, and A. Peter. Fast and accurate likelihood ratio-based biometric verification secure against malicious adversaries. *IEEE transactions on information forensics and security*, 16:5045–5060, 2021.

[3] A. Bassit, F. Hahn, R. Veldhuis, and A. Peter. Improved multiplication-free biometric recognition under encryption. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023.

[4] V. N. Boddeti. Secure face matching using fully homomorphic encryption. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018.

[5] 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.

[6] 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.

[7] 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*, pages 13733–13742, 2021.

[8] 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.

[9] P. Drozdowski, N. Buchmann, C. Rathgeb, M. Margraf, and C. Busch. On the application of homomorphic encryption to face identification. In *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2019.

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

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

[12] ISO/IEC 24745:2022(E) Information technologyy, cybersecurity and privacy protection – Biometric information protection, Feb. 2022.

[13] ISO/IEC 30136:2018(E) Information technology – Security techniques – Performance testing of biometric template protection schemes, June 2018.

[14] A. T. B. Jin, D. N. C. Ling, and A. Goh. Biohashing: two factor authentication featuring fingerprint data and tokenised random number. *Pattern Recognition*, 37(11):2245–2255, 2004.

[15] Z. Jin, J. Y. Hwang, Y.-L. Lai, S. Kim, and A. B. J. Teoh. Ranking-based locality sensitive hashing-enabled cancelable biometrics: Index-of-max hashing. *IEEE Transactions on Information Forensics and Security*, 13(2):393–407, 2017.

[16] A. Juels and M. Sudan. A fuzzy vault scheme. *Designs, Codes and Cryptography*, 38(2):237–257, 2006.

[17] A. Juels and M. Wattenberg. A fuzzy commitment scheme. In *Proceedings of the 6th ACM Conference on Computer and Communications Security*, pages 28–36, 1999.

[18] 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)*, pages 4401–4410, 2019.

[19] D. Keller, M. Osadchy, and O. Dunkelman. Inverting binarizations of facial templates produced by deep learning (and its implications). *IEEE Transactions on Information Forensics and Security*, 16:4184–4196, 2021.

[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] 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.

[24] 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*, pages 51–59, 2017.

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

[26] C. Rathgeb, J. Merkle, J. Scholz, B. Tams, and V. Nesterowicz. Deep face fuzzy vault: Implementation and performance. *Computers & Security*, 113:102539, 2022.

[27] G. D. P. Regulation. Regulation EU 2016/679 of the european parliament and of the council of 27 april 2016. *Official Journal of the European Union*, 2016.

[28] S. Sadrizadeh, H. Otroshi-Shahreza, and F. Marvasti. Impulsive noise removal via a blind cnn enhanced by an iterative post-processing. *Signal Processing*, 192:108378, 2022.

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

[30] A. Sarkar and B. K. Singh. A review on performance, security and various biometric template protection schemes for biometric authentication systems. *Multimedia Tools and Applications*, pages 1–56, 2020.

[31] 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.

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

[33] H. O. Shahreza, V. K. Hahn, and S. Marcel. Vulnerability of state-of-the-art face recognition models to template inversion attack. *IEEE Transactions on Information Forensics and Security*, 2024.

[34] H. O. Shahreza and S. Marcel. Blackbox face reconstruction from deep facial embeddings using a different face recognition model. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2435–2439. IEEE, 2023.

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

[36] H. O. Shahreza and S. Marcel. Face reconstruction from facial templates by learning latent space of a generator network. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[37] 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*, pages 19662–19672, 2023.

[38] H. O. Shahreza and S. Marcel. Face reconstruction from partially leaked facial embeddings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

[39] H. O. Shahreza and S. Marcel. Template inversion attack using synthetic face images against real face recognition systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[40] H. O. Shahreza, P. Melzi, D. Osorio-Roig, C. Rathgeb, C. Busch, S. Marcel, R. Tolosana, and R. Vera-Rodriguez. Benchmarking of cancelable biometrics for deep templates. *arXiv preprint arXiv:2302.13286*, 2023.

[41] 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 *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022.

[42] H. O. Shahreza, Y. Y. Shkel, and S. Marcel. Measuring linkability of protected biometric templates using maximal leakage. *IEEE Transactions on Information Forensics and Security*, 2023.

[43] U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain. Biometric cryptosystems: issues and challenges. *Proceedings of the IEEE*, 92(6):948–960, 2004.

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

[45] 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.

[46] 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*, pages 3779–3782, 2021.

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

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.