FACE RECONSTRUCTION FROM PARTIALLY LEAKED FACIAL EMBEDDINGS

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

ABSTRACT

Face recognition systems are widely used in different applications. In such systems, some features (called templates) are extracted from each face image and stored in the system's database. In this paper, we propose an attack against face recognition systems where the adversary gains access to a portion of facial templates and aims to reconstruct the underlying face image. To this end, we train a face reconstruction network to invert partially leaked templates. In our experiments, we evaluate the vulnerability of state-of-the-art face recognition systems on different datasets, including MOBIO, LFW, and AgeDB. Our experiments demonstrate the vulnerability of face recognition systems to template inversion based on a portion of leaked templates. For example, with only 20% of facial templates, our experiments show that an adversary can achieve a success attack rate of 87% on a system based on ArcFace on the LFW dataset configured at the false match rate of 0.1%. To our knowledge, this paper is the first work on the inversion of *partially* leaked facial templates, and paves the way for future studies of attacks against face recognition systems based on partially leaked templates.

Index Terms— Embedding, Face Recognition, Face Reconstruction, Partial Leakage, Template Inversion

1. INTRODUCTION

Face recognition (FR) systems tend towards ubiquity and have been widely used in many real-world applications, including unlocking smartphones, border control, etc. In such systems, a feature extractor model is applied to the input face image, and some features (also known as "*template*" or "*embedding*") are extracted from the face image. During enrollment, these features are stored in the database of the FR system, which are later used for comparison in the recognition stage. Therefore, the extracted facial templates and especially the ones stored in the system, convey the information required to recognize enrolled users. Among different types of attacks against FR systems that have been studied in the literature [1, 2, 3, 4], template inversion (TI) attacks endanger the security and privacy of the users. In a TI attack, the adversary gains access to the templates stored in the database of the system and tries to reconstruct the underlying face images. The reconstructed face image may reveal sensitive information about the enrolled user (i.e., privacy threat). Moreover, the adversary can use the reconstructed face image to impersonate the target user by injecting the reconstructed face image into the FR system (i.e., security threat).

Several papers in the literature proposed different methods to reconstruct face images from facial templates in template inversion attacks. While there are some optimization-based methods [5, 6], most works trained a face reconstruction network to invert facial templates. In the latter case (i.e., learning-based methods), different neural networks are trained to invert facial templates and reconstruct face images. Several works used convolutional neural networks (CNNs) [7, 8, 9, 10] and some other papers used generative models, such as generative adversarial networks (GANs) [11, 12, 13] or generative neural radiance fields (GNeRFs) [14, 15]. In all TI attacks against FR systems investigated in the literature, it is assumed that the adversary gains access to the *complete* facial templates and aims to reconstruct the underlying face image. However, it is also possible in some real-world scenarios that the adversary cannot find a complete template, but rather can reach a part of the template. This can happen due to the limited access of the adversary to the target template or the design of the FR system in which partial leakage is possible. As an example of the latter case, we can consider a FR system with a distributed database, where different parts of each face template are stored on different servers. In such a case, it is possible that an adversary can breach into one server and find *a part* of templates instead of *complete* templates. In this paper, we focus on the inversion of *partially* leaked facial templates and investigate the amount of information required by the adversary for a successful TI attack. To our knowledge, the inversion of *partially* leaked facial templates has not been investigated in previous works.

Our proposed approach for the inversion of partially leaked face templates stems from our previous face reconstruction network proposed in [7]. We train our previous face reconstruction network with the available part of facial templates and use the trained model to invert facial templates in the database

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

of the system. In our experiments, we use the MOBIO, LFW, and AgeDB datasets and evaluate the vulnerability of state-of-the-art (SOTA) FR systems to our TI attack. We consider different leakage percentages for the elements of each target template and investigate the required amount of information for a successful TI attack.

The remainder of this paper is organized as follows. In Section 2, we describe the threat model and explain our face reconstruction method. In Section 3, we present our experimental results. Finally, the paper is concluded in Section 4.

2. METHODOLOGY

In this section, we present our proposed method to invert *partially* leaked facial templates. First, we describe our threat model in Section 2.1, where the adversary gains access to a part of a facial template and aims to reconstruct the underlying face image to impersonate. Next, we describe our face reconstruction network in Section 2.2.

2.1. Threat Model

We consider the situation where the adversary gains access to a portion of a target facial template and aims to invert the *partially* leaked template to impersonate into the FR system. Let $\mathbf{t}_c = F(\mathbf{I}) \in \mathbb{R}^d$ denote the *complete* facial template with d dimensions extracted from the face image $\mathbf{I}$ using the feature extractor $F(.)$. Also, let us assume that the adversary has access to a portion $\tilde{\mathbf{t}}_{\mathcal{M}}$ of the *complete* facial temple $\mathbf{t}_c$ with indices $M \subseteq \{1, 2, ..., d\}$. We define 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 is assumed to have the following information:
  1. A portion $\tilde{\mathbf{t}}_{\mathcal{M}}$ of the target face template $\mathbf{t}$ of a user enrolled in the system's database.
  2. The set $\mathcal{M}$ of indices of the known elements in the *partially* leaked template $\tilde{\mathbf{t}}_{\mathcal{M}}$.
  3. The whitebox knowledge (including parameters and internal functioning) of the feature extraction model $F(.)$ of the FR system.
- *Adversary's capability:* The adversary can inject the reconstructed face image $\hat{\mathbf{I}}$ from the inversion of the *partially* leaked template $\tilde{\mathbf{t}}_{\mathcal{M}}$ directly into the feature extractor of the target system and bypass the camera.
- *Adversary's strategy:* Under the above assumptions, the adversary can invert the *partially* leaked template $\tilde{\mathbf{t}}_{\mathcal{M}}$ and reconstruct face image $\hat{\mathbf{I}}$ using a face reconstruction method. Then, the adversary can inject the reconstructed face image $\hat{\mathbf{I}}$ as a query to enter the target FR system.

2.2. Face Reconstruction

Our method to reconstruct face images from the *partially* leaked templates stems from our previous face reconstruction network using a *complete* leaked template proposed in [7]. To train our network, we consider a dataset of face images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$, where $\mathbf{I}_i$ and N indicate the ith image and the total number of images, respectively. We use the data augmentation (randomly adjusting contrast and brightness, Gaussian blurring, and JPEG compression) and generate our training dataset $\mathcal{D} = \{(\tilde{\mathbf{t}}_{\mathcal{M},i}, \mathbf{I}_i)\}_{i=1}^{K}$ with K pairs of *partial* templates $\tilde{\mathbf{t}}_{\mathcal{M},i}$ and face images $\mathbf{I}_i$. To generate partial template $\tilde{\mathbf{t}}_{\mathcal{M},i}$, we first extract complete template $\mathbf{t}_{c,i} = [F \circ A](\mathbf{I}_{a,i})$ from augmented face image $\mathbf{I}_{a,i}$, where $A(.)$ indicates the face alignment function. Then, we keep only elements of known indices $\mathcal{M}$ from the complete template $\mathbf{t}_{c,i}$ as the *partial* template $\tilde{\mathbf{t}}_{\mathcal{M},i}$.

We use a similar network structure as proposed in [7] with a skip connection over convolutional blocks, but the input is a partially leaked template (instead of *complete* template). We optimize our model with a multi-term loss function, including:

- *Mean Absolute Error (MAE):* We use Mean Absolute Error (MAE) loss term on the reconstructed face images, to minimize the pixel level reconstruction error:

$$\mathcal{L}_{\text{MAE}}(\hat{\mathbf{I}}, \mathbf{I}) = ||\hat{\mathbf{I}} - \mathbf{I}||_1, \tag{1}$$

  where $\mathbf{I}$ and $\hat{\mathbf{I}}$ are the original and reconstructed face images, respectively.

- *Dissimilarity Structural Index Metric (DSSIM):* In addition to MAE loss, we enhance the objective quality of the reconstructed image in terms of the Similarity Structural Index Metric (SSIM) [16] by optimizing the DSSIM loss term [17] as follows:

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

- *Perceptual Loss:* To further enhance the reconstruction quality, we minimize the ℓ_1-norm distance between the features extracted from $\mathbf{I}$ and $\hat{\mathbf{I}}$ by a CNN trained on ImageNet. To this end, we extract the middle feature maps of a pre-trained VGG-16 [18] model. Let us denote the feature mapping of VGG-16 as $P(.)$, then the perceptual loss is as follows:

$$\mathcal{L}_{\text{Perc}}(\hat{\mathbf{I}}, \mathbf{I}) = ||P(\hat{\mathbf{I}}) - P(\mathbf{I})||_1 \tag{3}$$

- *ID loss:* To preserve the identity in the reconstructed face image, we minimize the distance between the complete templates extracted from the reconstructed face $\hat{\mathbf{I}}$ and original face $\mathbf{I}$ images. To this end, we minimize the ℓ_1-norm distance and maximize the cosine similarity of the extracted templates as follows:

$$\mathcal{L}_{\text{ID}}(\hat{\mathbf{I}}, \mathbf{I}) = \mathcal{L}_{\text{ID},\ell_1}(\hat{\mathbf{I}}, \mathbf{I}) + \mathcal{L}_{\text{ID,cos}}(\hat{\mathbf{I}}, \mathbf{I})$$
$$= \underbrace{||F(\hat{\mathbf{I}}) - F(\mathbf{Y})||_1}_{\text{minimizing } \ell_1\text{-norm}} + \underbrace{\frac{-F(\hat{\mathbf{I}}).F(\mathbf{I})}{||F(\hat{\mathbf{I}})||_2.||F(\mathbf{I})||_2}}_{\text{maximizing cosine similarity}} \quad (4)$$

It is noteworthy that in this loss, we extract *complete* template from each face image using face recognition model F.

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

$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \gamma_1 \mathcal{L}_{\text{DSSIM}} + \gamma_2 \mathcal{L}_{\text{Perceptual}} + \gamma_3 \mathcal{L}_{\text{ID}} \quad (5)$$

We experimentally found that the choice of $\gamma_1 = 0.75$, $\gamma_2 = 0.02$, and $\gamma_3 = 0.025$ performs the best. We train our network using the Adam [19] optimizer with the initial learning rate of 10^{-3}, and we decrease the learning rate by a factor of 0.5 every 10 epochs.

3. EXPERIMENTS

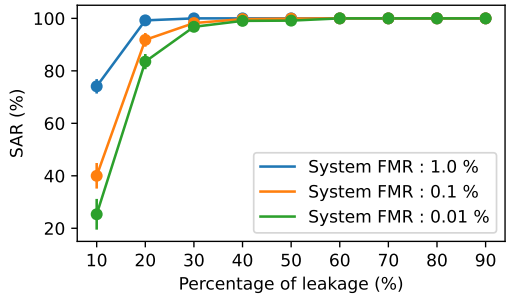
3.1. Experimental Setup

We use two different SOTA FR models in our experiments, including ArcFace [20] and ElasticFace [21]. We consider different percentages of the elements of each facial template being leaked and evaluate the vulnerability of FR systems models based on these FR models on different datasets against our TI attacks using *partially* leaked templates. For each percentage of *partially* leaked templates, we randomly select some elements in the facial templates and remove them from the *complete* template to achieve the *partial* template. We use five different random seeds for implementing each percentage of *partially* leaked templates in our experiments.

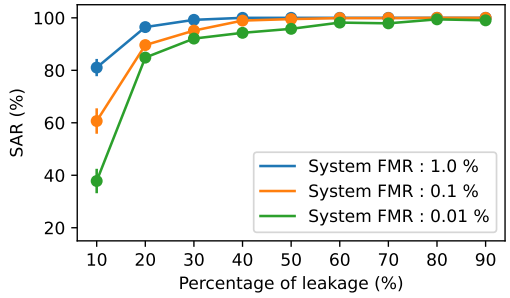
We use the Flickr-Faces-HQ (FFHQ) dataset [22], to train our face reconstruction network. The FFHQ dataset consists of 70,000 face images (with no identity labels) with different variations in terms of age, ethnicity, accessories, and image background. We randomly split the FFHQ dataset to train (90%) and validation (10%), and train each model for 85 epochs. After training our face reconstruction networks, we evaluate the trained models in TI attacks against FR systems on the MOBIO [23], Labeled Faced in the Wild (LFW) [24], and AgeDB [25] datasets. The MOBIO dataset includes audio data and face images captured using mobile devices from 150 people. In our experiments, we use the *development* subset of the *mobio-all* protocol. The LFW database includes 13,233 face images of 5,749 subjects (1,680 people have two or more face images). We use the *View 2* protocol in our experiments. The AgeDB [25] dataset includes 16,488 images of 568 famous people. There is a wide age variation in the AgeDB dataset, with the minimum and maximum ages of 1 and 101, respectively, and an average age range of 50.3 years. In our experiments, we use the *30-year* protocol for the AgeDB dataset. Table 1 reports the recognition performance of ArcFace and ElasticFace on our evaluation datasets.

Table 1: Recognition performance of face recognition models in terms of true match rate (TMR) at the thresholds correspond to false match rates (FMRs) of 1% and 0.1% evaluated on the MOBIO, LFW, and AgeDB datasets.

FR model	MOBIO		LFW		AgeDB	
	FMR=1%	FMR=0.1%	FMR=1%	FMR=0.1%	FMR=1%	FMR=0.1%
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



(a) ArcFace



(b) ElasticFace

Fig. 1: Vulnerability of face recognition systems to inversion of partially leaked templates on the MOBIO dataset for systems configured at different false match rate (FMR) values: (a) ArcFace and (b) ElasticFace.

We use the Bob[1] toolboxto build the FR systems and evaluate TI attacks in the evaluation framework introduced in [7]. We use the PyTorch package to train our face reconstruction networks. We use pretrained ArcFace and ElasticFace models on MS-Celeb-1M [26] with IResNet100 backbones. The dimension of facial templates for these FR models is 512, and the resolution of reconstructed face images is 112×112. The source code from our experiments is publicly available to help reproduce our results[2].

3.2. Analysis

To evaluate the vulnerability of FR systems to TI attacks using partially leaked templates, we consider different leakage percentages and evaluate the adversary's success attack rate (SAR). Table 2 reports the SAR of TI attacks from partial

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

[2] Available at https://gitlab.idiap.ch/bob/bob.paper.icassp2024_face_ti_partial

Table 2: Vulnerability of face recognition systems to TI attack from different percentage of template leakage on the MOBIO, LFW, and AgeDB datasets (configured at FMR=0.1%). Cells are color coded according the SAR value between low SAR (indicated with light pink) and high SAR (indicated with dark pink).

Dataset	Face Recognition	SAR at different percentage of template leakage								
		10%	20%	30%	40%	50%	60%	70%	80%	90%
MOBIO	ArcFace	40.0±4.87	91.81±2.62	98.21±0.97	99.71±0.38	99.9±0.19	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
	ElasticFace	60.67±4.84	89.62±1.02	95.14±0.63	98.95±0.82	99.52±0.52	99.9±0.19	99.9±0.19	100.0±0.0	100.0±0.0
LFW	ArcFace	51.58±2.76	87.32±0.67	84.79±37.35	95.45±0.17	96.13±0.11	96.36±0.05	96.42±0.09	96.58±0.08	96.57±0.03
	ElasticFace	55.57±0.53	79.59±11.59	91.26±0.19	91.85±0.5	93.84±0.0	93.99±0.13	94.22±0.01	94.35±0.13	94.42±0.14
AgeDB	ArcFace	12.26±2.08	47.76±1.17	65.29±1.36	76.99±0.73	80.89±0.38	82.81±0.53	83.55±0.36	84.45±0.37	84.77±0.38
	ElasticFace	21.74±0.28	54.23±0.64	69.37±0.48	77.38±0.65	80.9±0.51	82.73±0.17	83.37±0.37	83.74±0.22	84.54±0.14

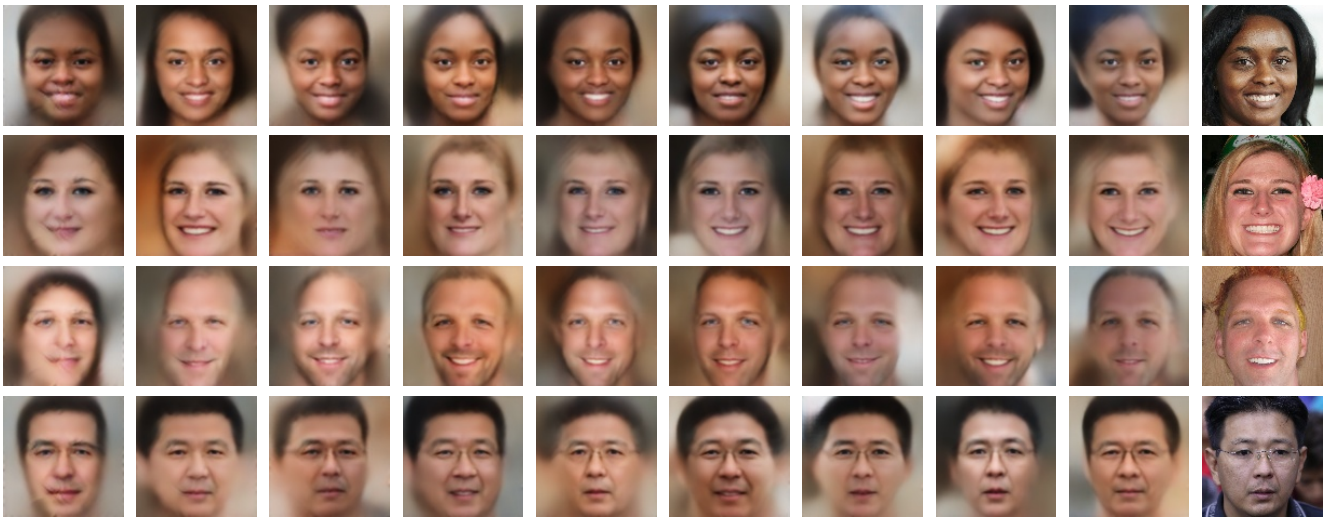


Fig. 2: Sample face images from FFHQ dataset and their corresponding reconstructed face images from different percentages of leaked facial templates extracted by ArcFace.

templates of ArcFace and ElasticFace on the MOBIO, LFW, and AgeDB datasets for FR systems configured at false match rate (FMR) of 0.1%. As the results in this table show, if 30% of facial templates are leaked, the adversary can achieve a considerably high SAR value. For the leakage of less than 30%, the SAR is dropping while still achieving considerable SAR at 20% of leakage. This experiment demonstrates that a portion of facial templates can still be useful to represent and reconstruct face images.

Fig.1 illustrates the SAR for different leakage percentages of ArcFace and ElasticFace templates on the MOBIO dataset for FMR values of 1%, 0.1%, and 0.01%. This plot also confirms that the SAR is significant if the adversary can access at least 30% of templates. The partially leaked template can then be used by the adversary to reconstruct the underlying face image, which can be recognized as the same user by the FR model. Fig.2 depicts sample face images from the validation set of the FFHQ dataset and the corresponding reconstructed face images from partial templates with different leakage percentages. As the results in this figure show, the sample reconstructed face images can reveal privacy-sensitive information about the underlying user, such as age, gender, etc. As expected, the reconstruction is weakened if the adversary gains less portion of the template.

4. CONCLUSION

In this paper, we focus on the inversion of *partially* leaked facial templates and investigate the amount of information required by the adversary for a successful TI attack. We extended our previous TI approach for training a face reconstruction network with the leaked portion of facial templates and then used the trained model to invert facial templates in the database of the FR system. In our experiments, we considered ArcFace and ElasticFace models and used the MOBIO, LFW, and AgeDB datasets to evaluate the vulnerability of FR systems to TI attacks based on partial leakage of facial templates. We investigated different leakage percentages that the adversary requires to achieve successful reconstruction. We show that with 30% of facial templates, the adversary can achieve significant SAR, and the reconstructed face images from partial templates reveal privacy-sensitive information. Our results motivate future studies of potential more sophisticated attacks against FR systems based on a portion of facial templates.

5. REFERENCES

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

[2] Abdenour Hadid, Nicholas Evans, Sebastien Marcel, and Julian Fierrez, "Biometrics systems under spoofing attack: an evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

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

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

[5] Edward Vendrow and Joshua Vendrow, "Realistic face reconstruction from deep embeddings," in *Proceedings of NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[6] Xingbo Dong, Zhihui Miao, Lan Ma, Jiajun Shen, Zhe Jin, Zhenhua Guo, and Andrew Beng Jin Teoh, "Reconstruct face from features based on genetic algorithm using gan generator as a distribution constraint," *Computers & Security*, vol. 125, pp. 103026, 2023.

[7] Hatef Otroshi Shahreza, Vedrana Krivokuća Hahn, and Sébastien 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.

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

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

[10] Hatef Otroshi Shahreza and Sébastien Marcel, "Blackbox face reconstruction from deep facial embeddings using a different face recognition model," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 2435–2439.

[11] Sohaib Ahmad, Kaleel Mahmood, and Benjamin 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.

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

[13] Hatef Otroshi Shahreza and Sébastien Marcel, "Face reconstruction from facial templates by learning latent space of a generator network," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[14] Hatef Otroshi Shahreza and Sébastien 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.

[15] Hatef Otroshi Shahreza and Sébastien Marcel, "Template inversion attack against face recognition systems using 3d face reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19662–19672.

[16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[17] Sahar Sadrizadeh, Hatef Otroshi-Shahreza, and Farokh Marvasti, "Impulsive noise removal via a blind cnn enhanced by an iterative post-processing," *Signal Processing*, vol. 192, pp. 108378, 2022.

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

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

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

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

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

[23] Chris McCool, Roy Wallace, Mitchell McLaren, Laurent El Shafey, and Sébastien Marcel, "Session variability modelling for face authentication," *IET Biometrics*, vol. 2, no. 3, pp. 117–129, Sept. 2013.

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

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

[26] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.