This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Face Reconstruction from Face Embeddings using Adapter to a Face Foundation Model

Hatef Otroshi Shahreza[1,2], Anjith George[1], 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,anjith.george,sebastien.marcel}@idiap.ch

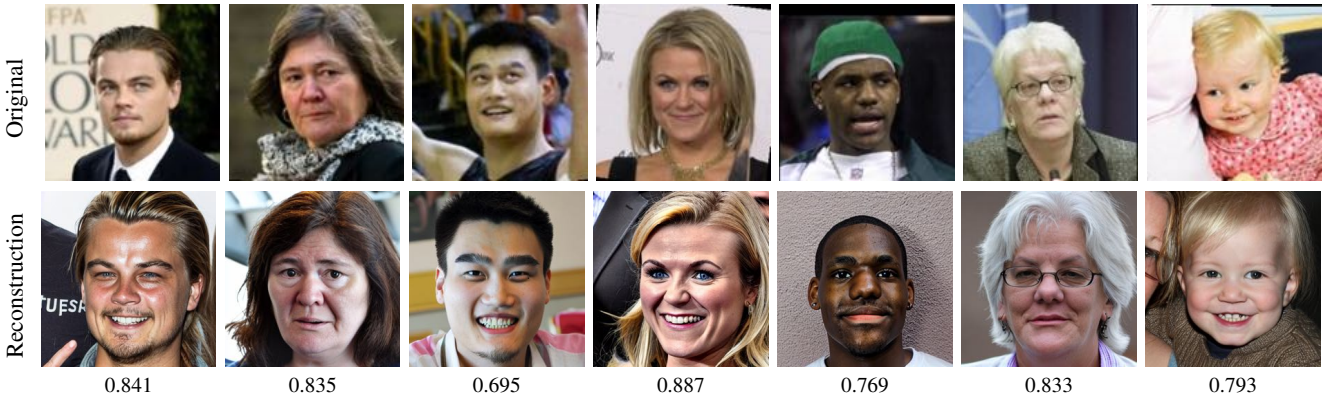

Figure 1. Sample face images from the LFW dataset (first row) and their reconstructed versions by our attack (second row) based on ArcFace embeddings. The values are cosine similarity of embeddings of the original and reconstructed face image.

Abstract

Face recognition systems extract embedding vectors from face images and use these embeddings to verify or identify individuals. Face reconstruction attack (also known as template inversion) refers to reconstructing face images from embeddings and using the reconstructed image to enter a face recognition system. In this paper, we propose to use a face foundation model to reconstruct face images from the embeddings of a blackbox face recognition model. The foundation model is trained with 42M images to generate face images from the facial embeddings of a fixed face recognition model. We propose to use an adapter (called Face Adapter) to translate target embeddings into the embedding space of the foundation model. The generated images are evaluated on different face recognition models and different datasets, demonstrating the effectiveness of our method to translate embeddings of different face recognition models. We also evaluate the transferability of reconstructed face images when attacking different face recognition models. Our experimental results show that our reconstructed face images outperform previous reconstruction attacks against face recognition models. Project Page

1. Introduction

Face recognition systems tend toward ubiquity and are extensively used in various applications that require automatic authentication, ranging from unlocking smartphones to border control. In these systems, a deep neural network is used to extract an embedding vector (also known as a template), which is stored in the database of the system during enrollment. Later during the recognition stage, the extracted embeddings are compared with ones stored in the database of the system. Hence, the extracted embedding vectors play a pivotal role in the automatic recognition of subjects.

With the growing application of face recognition systems, the security of these systems has become a critical topic and attracted considerable attention from the research community [2, 6, 11, 14, 15, 17, 27]. Among different types of attacks against face recognition systems, face reconstruction attacks (also known as template inversion) threaten the security and privacy of subjects. In a face reconstruction attack, the adversary gains access to face embeddings stored in the database of a face recognition system and tries to reconstruct the face image of enrolled users. The reconstructed face images can provide privacy-sensitive (such as age, gender, ethnicity, etc) information about the user, and

in addition, can be used to enter the face recognition system. Existing face reconstruction attacks in the literature can be categorized into two groups: learning-based methods and optimization methods. In learning-based methods, such as [5, 26, 38], a network is trained, on a dataset of pairs of embeddings and images, to reconstruct face images from the embeddings. In optimization-based methods, such as [10, 44], usually a face generator model (such as StyleGAN) is used and through an iterative optimization a latent input is found to generate the reconstructed face images. In each case, the attack requires large computation time for the training (in learning-based methods) or the inference stage (in optimization-based methods).

Recently, generative models have also absorbed much attention, and there have been significant advancements in the generation capabilities of these models. In a similar vein, different researchers tried to develop large foundation models, which are trained on massive data and can be later used for different applications. Recently, the first face foundation model, called Arc2Face [33], was proposed in which the CLIP [34] and Stable Diffusion [35] models were fine-tuned on 42 million face images from WebFace260M dataset [49]. The resulting model is able to generate different face images from embeddings of a fixed face recognition model while preserving identity information. Therefore, the Arc2Face model can be used as a foundation model in many problems which require an identity-conditioned generator. For example, the authors showed that the proposed method achieves state-of-the-art performance in training face recognition using synthetic data. However, an important question is that *"given the considerable capabilities of foundation models, can face foundation models threaten face recognition systems?"*

In this paper, we focus on face reconstruction attacks and propose a new method to reconstruct face images from different face recognition models using a foundation model. We propose an adapter module (called Face Adapter) that can map the face embeddings to the input space of the foundation model. Then, the foundation model can be used to generate face images from the given embeddings. The adapter module enables the use of a foundation model for embeddings of different face recognition models and prevents the need for training a foundation model for a new face recognition model. Therefore, an adversary can easily leverage the capabilities in the foundation model to perform a successful reconstruction attack in a blackbox scenario. We perform extensive experiments using several face recognition models on different face recognition datasets. Furthermore, we explore the transferability of reconstructed face images to enter a different face recognition system. The experimental results demonstrate the effectiveness of our reconstructed face images in attacking face recognition systems and superiority compared to previous face reconstruction attacks. Fig. 1 illustrates sample reconstructed face images by our attack.

In summary, the contributions of this paper are as follows:

- We propose a simple yet effective adapter module to map the face embeddings to the input of the face foundation model. The adapter module enables the use of a foundation model for embeddings of different face recognition models and prevents the need for training a foundation model for a new face recognition model.
- We propose a new face reconstruction attack against face recognition systems using a foundation model. While the foundation model can generate face images from embeddings of a fixed model, we use this model to reconstruct face images from embeddings of any blackbox model. To our knowledge, this is the first attack against face recognition systems based on foundation models.
- We demonstrate the effectiveness of our reconstructed face images with extensive experiments for different face recognition models on different datasets. We also evaluate the transferability of reconstructed face images for different face recognition models.

In the remainder of the paper, we first review related work in the literature in Sec. 2. Next, we propose our face reconstruction method based on a foundation model in Sec. 3. In Sec. 4, we present our experiments and compare our method with previous attacks in the literature. In Sec. 5, we further discuss different aspects of our attack. Finally, the paper is concluded in Sec. 6.

2. Related Work

Previous methods for reconstructing face images from embeddings rely on training a neural network (learning-based) or optimization techniques (optimization-based). In learning-based methods, a set of face images is used to build a training dataset by extracting their embeddings. Then a network is trained to generate reconstructed face images from embeddings. For example, in [5], a multi-layer perceptron (MLP) and a convolutional neural network (CNN) were trained to generate facial landmark coordinates and face texture, respectively, from the embeddings. Then, a differentiable warping was used to combine estimated landmarks and generated textures and reconstruct a low-resolution face image from embedding. Their method is proposed for the whitebox face reconstruction attack, where the adversary has full knowledge of the face recognition model (i.e., network structure and its parameters). They further trained MLP and CNN by minimizing the distance between embeddings of reconstructed and original face images. However, for a blackbox attack, they trained MLP and CNN individually, and used warping in the inference only. In [26], two new deconvolutional networks, called NbBlock-A and NbBlock-B, are proposed and trained with

pixel loss and perceptual loss to generate low-resolution face images from embeddings in blackbox scenarios. In [12], a bijection-learning-based method is used to train a generative adversarial network (GAN). In whitebox attacks, they trained the GAN model by minimizing the distance between embeddings of original and reconstructed images. For blackbox attacks, they used knowledge distillation to train a student network from the face recognition model, and then used the student network in their method. In [1] a GAN-based face reconstruction network was proposed to generate low-resolution face images in the *blackbox* scenario. However, in contrast to other GAN-based methods, sample reconstructed face images in their paper do not look realistic and suffer from many artifacts. In [37], a CNN-based face reconstruction network was trained with a proxy face recognition network for blackbox face reconstruction attacks against face recognition systems. Therefore, the reconstructed face images were not realistic and had low resolutions.

Several works in the literature used pretrained face generative models and trained a mapping from face embeddings to different layers of generative models. In [36], authors use an adversarial training method to train a mapping from face embeddings to the intermediate layers of StyleGAN3 [21]. While adversarial training leads to an effective mapping, a stable training in an adversarial framework is a difficult task. In [9], the authors proposed to generate some face images using StyleGAN2 [20] and extract the embeddings using the face recognition model. Then, they trained an MLP network to map facial embeddings to the input latent codes of StyleGAN2 [20]. In [39], authors used generated images by StyleGAN3 [21] and trained a mapping from face embeddings of synthetic faces to the intermediate layers of StyleGAN3. In [38] a semi-supervised learning approach was used to find a mapping to the intermediate latent space of EG3D [4]. Compared to other methods based on StyleGAN, the method in [38] can generate face images from different points of view thanks to the generative model which is based on neural radiance fields (NeRF). Therefore, the authors proposed camera parameter optimization to find the best pose which increases the attack rate.

In contrast to most work in the literature, some papers used optimization techniques to find an input to a pretrained StyleGAN model that can generate the reconstructed face images. In [44], a grid-search optimization using the simulated annealing [43] approach is used on the input of StyleGAN2 [20] that generates the reconstructed face image in blackbox attacks. A main drawback of optimization-based methods is their high computation for generating face images. This can even cause difficulty in the evaluation of such methods. For example, in [44], authors reported their evaluation on only 20 face images, because of the extensive computation requirement of the method in the inference stage. In [10], optimization on the input of StyleGAN2 [20] was also used, and the authors used the standard genetic algorithm [42] to solve the optimization.

Table 1. Comparison with related works.

Ref.	Resolution	Method	Available code
[5]	low	learning-based (MLP+CNN)	✗
[26]	low	learning-based (CNN)	✓
[12]	low	learning-based (GAN)	✗
[1]	low	learning-based (GAN)	✗
[37]	low	learning-based (CNN)	✓
[36]	high	learning-based (StyleGAN)	✓
[9]	high	learning-based (StyleGAN)	✓
[39]	high	learning-based (StyleGAN)	✓
[38]	high	learning-based (EG3D)	✓
[44]	high	optimization-based (StyleGAN)	✓
[10]	high	optimization-based (StyleGAN)	✓
[Ours]	**high**	**learning-based (Arc2Face)**	✓

Tab. 1 compares our proposed method with the previous works in the literature. All previous methods require large computation, either in the training (learning-based methods) or inference stage (optimization-based methods). However, our method benefits from the capabilities of a foundation model and requires less computation resources for the adversary. We should note that our method also requires training of the adapter module. However, compared to previous learning-based methods which require end-to-end learning, training the adapter module in our method is stable and computationally efficient. To our knowledge, our proposed method is the first attack against face recognition systems using face foundation models.

3. Face Reconstruction Attack using Foundation Models

3.1. Threat Model

We consider a face reconstruction attack against a face recognition system based on the following threat model:

- *Adversary's goal*: The adversary aims to reconstruct a face image from a leaked face embedding $\boldsymbol{e}_{\text{leaked}}$, and use the reconstructed face image to enter a target face recognition system F_{target} system.
- *Adversary's knowledge:* The adversary is assumed to have the following information:
  - The adversary knows a face embedding $\boldsymbol{e}_{\text{leaked}}$, which is leaked from database of a victim face recognition system F_{victim}.
  - The adversary has a blackbox access (such as SDK, API, etc.) of the feature extractor model F_{victim} of the face recognition system from which the embedding $\boldsymbol{e}_{\text{leaked}}$ is leaked.
- *Adversary's capability:* The adversary can use the reconstructed face image to enter the target face recognition system. For simplicity, we assume that the adversary can

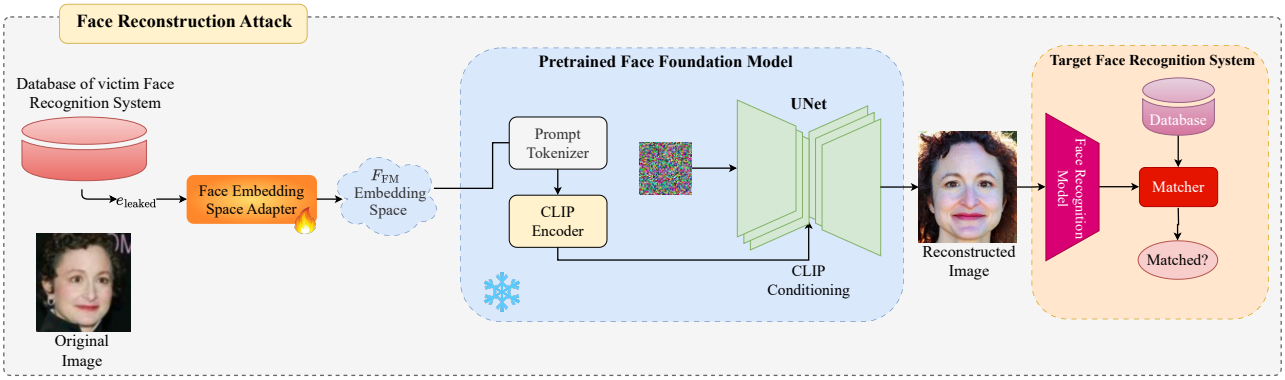

Figure 2. Block diagram for our face reconstruction attack

inject the reconstructed face image as a query to the target face recognition system.

- *Adversary's strategy:* The adversary's strategy is to use a foundation model to reconstruct the underlying face image from the leaked embedding $\boldsymbol{e}_{\text{leaked}}$. Then, the adversary can use the reconstructed face images to enter the target face recognition system F_{target} (Fig. 2).

We should note that in our threat model, the victim face recognition system F_{victim} from which the embedding $\boldsymbol{e}_{\text{leaked}}$ is leaked can be the same or different than the target face recognition system F_{target} system. The latter case (i.e., $F_{\text{victim}} \neq F_{\text{target}}$) is a more difficult scenario, where we assume that the same subject is enrolled in another face recognition system with a different feature extractor, and we would like to see if the reconstructed face images from embeddings of F_{victim} are *transferable* and can be recognized by a different face recognition model (F_{target}).

3.2. Face Reconstruction

As discussed earlier, Arc2Face [33] is capable of generating identity-consistent images given an embedding of a specific face recognition model F_{FM}. Their pipeline includes a modified CLIP [34] model to accept face embeddings from the face recognition model (F_{FM}) and a Stable Diffusion [35] UNet decoder. This pipeline is end-to-end trained on an upscaled version of WebFace42M [49], resulting in the generation of high-quality, identity-consistent images. It is evident that this model can be readily used for inverting the embeddings of F_{FM}, as it is specifically trained for those embeddings. The face recognition model F_{FM} used to train Arc2Face has a ResNet100 backbone and is trained with WebFace42M [49].

To attack *any* blackbox face recognition system, one would need to redo the entire process of adapting a large model on a large-scale dataset, which is impractical. In this work, we propose a simple approach to adapt this model for reconstructing images from embeddings of any face recognition system (F_{victim}).

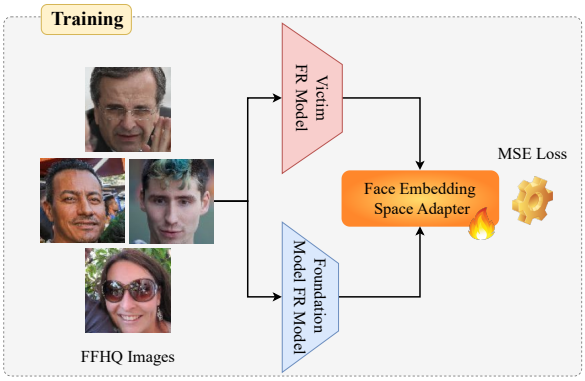

Figure 3. Training process of the adapter module. The parameters of the models are obtained by minimizing MSE loss between embeddings extracted by the (blackbox) victim face recognition model and the face recognition model of the foundation model.

The main idea is to develop an adapter module that can transform the embeddings from the space of the leaked embedding into the embedding space of F_{FM}, i.e.,

$$\boldsymbol{e}_{\text{leaked}} \mapsto \boldsymbol{e}_{\text{FM}} \tag{1}$$

Essentially, we propose to learn a mapping $\mathcal{M}$ which would map the leaked embeddings to the embedding space of F_{FM}.

$$\boldsymbol{e}_{\text{FM}} = \mathcal{M}(\boldsymbol{e}_{\text{leaked}}) \tag{2}$$

We implement this mapping function, $\mathcal{M}$, as a simple linear layer. The parameters of this adapter layer can be learned using a set of pairs of embeddings extracted from F_{victim} and F_{FM}. For training the adapter network, we use the Adam [23] optimizer with a learning rate of 10^{-3} for 20 epochs, employing Mean Square Error (MSE) as the loss function. It is important to note that we just need blackbox access to the victim model in this process (Fig. 3).

Given the adapter, inverting a leaked embedding from a face recognition system involves first projecting it into the embedding space of F_{FM} and then reconstructing it with the Arc2Face foundation model. Fig. 2 depicts our face reconstruction attack.

4. Experiments

4.1. Experimental Setup

Face Recognition models In our experiments, we consider different state-of-the-art face recognition models including ArcFace [7], ElasticFace [3] with ResNet100 backbones as well as four different face recognition models with state-of-the-art backbones from FaceX-Zoo [47], including AttentionNet [45], HRNet [46], RepVGG [8] and Swin [25]. Tab. 2 reports the recognition performance of these models. All of these models are trained on the MS-Celeb1M dataset [16]. Note that the face recognition model F_{FM} used in the Arc2Face model is also trained with ArcFace loss function, but on the WebFace42M [49] dataset. In our experiments, we denote the face recognition model used in the foundation model (i.e., trained with WebFace42M) with F_{FM}, and refer to the one trained with MS-Celeb1M dataset [16] as ArcFace for a victim or target face recognition model. Although both the models are trained with the ArcFace loss function, the latent spaces of these models are not aligned due to the different initialization and training datasets.

Datasets To train the adapter module, we use the Flickr-Faces-HQ (FFHQ) dataset [19], which consists of 70,000 high-quality images collected from the internet (without identity labels). All 70,000 images from the FFHQ dataset are cropped and aligned using MTCNN [48]. We use 60,000 images for training and reserve the remaining 10,000 as the test set.

To evaluate the vulnerability of face recognition models, we use three benchmarking datasets, including MOBIO [28], Labeled Faces in the Wild (LFW) [18] and AgeDB [29] datasets. The MOBIO dataset includes face images of 150 people captured using mobile devices in 12 sessions for each person. The LFW dataset consists of 13,233 face images of 5,749 people collected from the internet, where 1,680 people have two or more images. The AgeDB [29] dataset includes 16,488 images (in different ages for each subject) of 568 famous people.

Evaluation Protocol For evaluation, we build a face recognition system using one of the mentioned feature extractors as F_{victim} and train an adapter module for the corresponding feature extractor. We use the benchmarking datasets for the verification task (i.e., 1:1 matching) and consider the enrolled embedding as leaked embedding which is used by the adversary for reconstruction attack. After reconstruction, we inject the reconstructed face image as a query to the feature extractor of the target system and evaluate the adversary's success attack rate (SAR) to enter the target system. We should note that the target face recognition system may have the same or different (i.e., transferability evaluation) feature extractor model that the system from which the embeddings are leaked.

Table 2. Recognition performance of face recognition models in terms of true match rate (TMR) at false match rate (FMR) of 10^{-3} evaluated on the MOBIO, LFW, and AgeDB datasets.

Face Recognition	MOBIO	LFW	AgeDB
ArcFace	99.98	96.40	98.07
ElasticFace	100.00	94.70	97.57
AttentionNet	97.73	72.77	96.90
HRNet	98.23	78.43	96.23
RepVGG	95.80	58.07	93.93
Swin	98.98	87.83	97.10

Implementation and Source Code Our experiments, including training of adapter and comparison with previous methods, are conducted on a system equipped with a single NVIDIA RTX 3090 GPU. For the face foundation model, we use a pretrained model of Arc2Face [33] in our experiments and generate 512×512 resolution face images. The source code of our experiments is publicly available[1].

4.2. Analyze

Blackbox Attack We train adapter modules for different face recognition models and use the adapter module along with the foundation model to reconstruct face images from embeddings of the MOBIO, LFW, and AgeDB datasets. Tab. 3 compares the performance of reconstructed face images with previous methods in the literature in blackbox face reconstruction attacks against different face recognition models. As the results in this table show our method outperforms previous methods in the literature for different face recognition models and on different datasets. Comparing the results with the recognition performance in Tab. 2, we can observe that a face recognition model with better accuracy is more vulnerable to face reconstruction attacks. Fig. 4 presents sample reconstructed face images in blackbox attacks against different face recognition models using our method. The reconstructed images are of high quality and accurately match the identities of the original images.

Transferability Evaluation To evaluate the transferability of reconstructed face images, we use the generated images to enter a different face recognition system (i.e.,

[1]Project page: https://www.idiap.ch/paper/face_adapter

Table 3. Evaluation of *blackbox* attacks against face recognition models at FMR=10^{-3} on the MOBIO, LFW, and AgeDB datasets in terms of success attack rate (SAR). In this experiment, the target face recognition model is the same model from which the embeddings are leaked (i.e., $F_{\text{victim}} = F_{\text{target}}$).

Dataset	Method	Victim Face Recognition					
		ArcFace	El.Face	Att.Net	HRNet	RepVGG	Swin
MOBIO	NBNetA-M [26]	2.86	10.0	4.76	4.76	6.19	6.67
	NBNetA-P [26]	23.81	60.95	15.24	14.29	44.76	30.48
	NBNetB-M [26]	20.95	30.0	21.43	25.24	21.43	27.62
	NBNetB-P [26]	49.05	70.95	66.67	64.76	51.43	71.43
	BIOSIG 2021 [9]	24.29	34.76	38.57	16.19	24.76	18.10
	NeurIPSW 2021 [44]	69.52	74.29	55.71	43.81	39.52	70.00
	Com. & Sec. 2023 [10]	87.62	90.95	80.48	71.90	44.29	82.38
	IJCB 2023 [32]	80.00	86.67	81.90	85.24	70.95	84.76
	NeurIPS 2023 [36]	81.90	89.05	81.43	86.19	58.10	90.48
	GaFaR [38]	47.62	84.76	72.38	76.67	72.86	89.05
	GaFaR + GS [38]	64.76	86.62	80.00	83.80	73.33	93.33
	ICIP 2023 [37]	95.71	98.10	92.38	97.62	85.71	99.52
	Face Adapter [Ours]	**100.0**	**99.05**	**99.52**	**99.52**	**98.57**	**99.52**
LFW	NBNetA-M [26]	14.30	37.13	10.37	20.19	10.64	13.18
	NBNetA-P [26]	35.61	60.05	6.80	16.83	26.44	25.92
	NBNetB-M [26]	26.91	52.99	17.62	31.74	18.18	27.00
	NBNetB-P [26]	61.66	81.74	43.42	56.30	38.12	61.02
	BIOSIG 2021 [9]	28.21	34.56	19.17	24.87	14.76	26.62
	NeurIPSW 2021 [44]	77.00	79.37	46.52	49.52	22.4	66.07
	Com. & Sec. 2023 [10]	87.26	89.00	55.40	59.46	28.60	69.07
	IJCB 2023 [32]	71.31	81.70	43.58	50.04	35.75	66.57
	NeurIPS 2023 [36]	77.13	83.43	45.02	48.22	24.34	61.95
	GaFaR [38]	51.78	74.54	33.59	37.80	25.40	67.11
	GaFaR + GS [38]	61.56	78.67	38.42	43.27	29.84	70.82
	ICIP 2023 [37]	90.67	92.27	62.20	68.99	44.56	82.21
	Face Adapter [Ours]	**95.71**	**93.18**	**78.71**	**80.66**	**67.11**	**88.44**
AgeDB	NBNetA-M [26]	0.81	2.55	0.22	0.38	0.44	0.27
	NBNetA-P [26]	3.99	8.92	0.34	0.14	3.71	1.02
	NBNetB-M [26]	1.88	6.27	0.50	0.77	1.06	0.68
	NBNetB-P [26]	13.18	28.94	5.08	5.61	7.92	8.75
	BIOSIG 2021 [9]	3.93	4.88	1.58	1.97	2.22	2.48
	NeurIPSW 2021 [44]	29.64	34.89	15.06	12.02	14.49	21.10
	Com. & Sec. 2023 [10]	58.80	66.10	36.82	32.45	14.98	37.81
	IJCB 2023 [32]	57.94	76.98	49.92	50.57	48.36	59.95
	NeurIPS 2023 [36]	48.57	64.89	26.26	29.39	19.15	30.63
	GaFaR [38]	21.67	47.37	14.59	17.09	18.02	30.05
	GaFaR + GS [38]	28.94	53.10	18.76	22.40	24.01	35.20
	ICIP 2023 [37]	58.07	79.20	46.21	46.75	52.29	62.75
	Face Adapter [Ours]	**86.42**	**85.64**	**84.31**	**78.43**	**79.43**	**85.69**

Table 4. Transferability evaluation (i.e., $F_{\text{victim}} \neq F_{\text{target}}$) of reconstructed face images from *blackbox* attacks against ArcFace embeddings (F_{victim}) for entering a *different target* face recognition system (F_{target}) at FMR=10^{-3} on the MOBIO, LFW, and AgeDB datasets in terms of success attack rate (SAR).

Dataset	Method	Target Face Recognition				
		El.Face	Att.Net	HRNet	RepVGG	Swin
MOBIO	NBNetA-M [26]	0.0	0.0	0.0	0.0	0.0
	NBNetA-P [26]	1.43	2.86	1.43	1.90	0.95
	NBNetB-M [26]	4.29	3.81	4.29	1.90	4.29
	NBNetB-P [26]	10.95	6.67	6.19	2.38	13.33
	BIOSIG 2021 [9]	3.81	2.86	3.33	2.38	4.29
	NeurIPSW 2021 [44]	17.14	5.71	10.48	3.81	13.81
	Com. & Sec. 2023 [10]	37.61	26.67	30.48	19.52	40.00
	IJCB 2023 [32]	74.76	59.52	59.52	55.24	68.57
	NeurIPS 2023 [36]	73.81	54.29	57.14	44.29	67.14
	GaFaR [38]	52.38	29.05	31.90	28.10	43.33
	GaFaR + GS [38]	55.24	36.19	41.90	32.86	50.00
	ICIP 2023 [37]	90.48	73.33	75.24	61.43	72.38
	Face Adapter [Ours]	**99.52**	**97.62**	**97.62**	**84.29**	**98.57**
LFW	NBNetA-M [26]	3.32	0.51	0.41	0.23	1.60
	NBNetA-P [26]	8.56	1.22	1.44	0.76	4.79
	NBNetB-M [26]	13.09	1.89	2.03	0.86	8.44
	NBNetB-P [26]	34.41	7.21	7.77	4.06	23.82
	BIOSIG 2021 [9]	6.25	2.12	1.70	1.39	6.22
	NeurIPSW 2021 [44]	29.06	9.79	9.51	4.48	20.75
	Com. & Sec. 2023 [10]	50.47	24.16	24.72	13.99	41.27
	IJCB 2023 [32]	62.33	25.59	25.91	14.72	44.91
	NeurIPS 2023 [36]	68.05	28.87	28.45	16.90	28.28
	GaFaR [38]	43.59	14.58	14.77	8.42	29.69
	GaFaR + GS [38]	47.57	17.13	17.35	10.11	33.16
	ICIP 2023 [37]	51.93	87.28	56.62	31.23	72.73
	Face Adapter [Ours]	**92.40**	**73.40**	**76.24**	**52.54**	**86.70**
AgeDB	NBNetA-M [26]	0.73	0.04	0.12	0.20	0.10
	NBNetA-P [26]	2.84	0.57	0.41	0.87	0.72
	NBNetB-M [26]	2.21	0.53	0.67	0.70	0.52
	NBNetB-P [26]	13.15	3.92	3.72	4.13	5.32
	BIOSIG 2021 [9]	2.30	1.07	1.08	1.24	1.33
	NeurIPSW 2021 [44]	17.74	8.19	7.84	8.14	9.80
	Com. & Sec. 2023 [10]	30.52	17.26	16.27	18.79	18.87
	IJCB 2023 [32]	39.39	16.11	15.29	18.09	21.83
	NeurIPS 2023 [36]	41.30	16.10	14.77	17.33	22.04
	GaFaR [38]	20.13	6.24	5.94	6.60	8.08
	GaFaR + GS [38]	23.85	8.17	8.07	9.51	10.57
	ICIP 2023 [37]	58.44	37.78	32.28	42.15	41.56
	Face Adapter [Ours]	**80.92**	**74.48**	**67.77**	**75.16**	**78.33**

$F_{\text{victim}} \neq F_{\text{target}}$). Tab. 4 reports the transferability using reconstructed face images from ArcFace embeddings (i.e., F_{victim}) when attacking a *different* target face recognition system (F_{target}), and compares it with different methods in the literature. As the results in this table show, the reconstructed face images are transferable and can be also used to enter a different face recognition system. The results in this table also show that our method outperforms previous methods in transferability evaluation and demonstrates the effectiveness of our reconstruction attack. Comparing the results in Tab. 4 with our evaluation of blackbox attacks in Tab. 3, we observe that for all methods, the transferability evaluation leads to a lower success rate. We would like to stress that the transferability of reconstructed images simulates a very difficult attack scenario and further sheds light on the vulnerability of face recognition systems.

Training Adapter Module and the Effect of Number of Images Compared to other methods which require long training (learning-based methods) or inference (optimization-based methods) time, our method relies on training a light adapter module to use a foundation model for face reconstruction attacks. Our approach for training the adapter module involves estimating a linear layer with

Figure 4. Sample face images from the dataset (first row) and their corresponding reconstructed face images using various methods. The values show the cosine similarity between embeddings of original and reconstructed face images.

dimensions[2] of 512×512. Once we obtain the embeddings from the F_{FM} model and the victim model from a number of images, the computation required to estimate these parameters is minimal. To assess the model's performance with varying amounts of training data, we conduct an ablation study by changing the number of training images. For this evaluation, we use images from the FFHQ dataset and evaluate the performance of adapters trained with different quantities of training samples in face reconstruction attacks. Tab. 5 compares the performance of adapter modules trained with different numbers of images when used in blackbox attacks against different face recognition models. As the results in this table show with more number of images the performance of the adapter module increases. However, 10,000 images are almost enough to train the adapter module with close to maximum performance. Compared with previous methods in Tab. 3 with only 600 training samples, the adapter module can outperform previous methods in the literature. This table also reports the required time to train the adapter module for different numbers of images on a system equipped with an NVIDIA GeForce RTX™ 3090. In fact, in contrast to the optimization methods, the training of the adapter module is required to be conducted only once for each model, and then the adversary can use the trained adapter to reconstruct unlimited leaked embeddings in a fraction of a second. Compared to previous learning-based methods, the training time for our adapter module is almost negligible to the training time of all previous learning-based methods. We should note that we report further ablation study in the supplement.

[2]Note that face recognition models used in our experiment as well as the one for foundation model F_{FM} have embeddings with 512 dimensions. However, a face recognition model with a different embedding dimension can still be used and requires training an adapter module with corresponding dimensions.

Table 5. Ablation study on the performance of adapter module when trained with different numbers of training images and used in *blackbox* attacks against different face recognition models at FMR=10^{-3} on the LFW dataset in terms of success attack rate (SAR). In this experiment, the target model is the same model from which the embeddings are leaked. The training time of the adapter module is also reported in seconds.

Number of Images	Training Time (sec)	Face Recognition					
		ArcFace	El.Face	Att.Net	HRNet	RepVGG	Swin
500	0.32	89.30	80.2	57.40	48.04	43.70	73.79
600	0.38	90.78	82.88	64.01	56.88	50.07	78.59
800	0.49	92.66	86.28	70.64	64.92	55.36	82.35
1000	0.63	93.60	88.27	73.12	69.59	57.80	84.25
2000	1.23	94.82	91.41	77.02	77.52	63.03	87.25
5000	3.05	95.40	92.18	78.28	80.41	65.45	88.32
10000	6.05	95.69	92.95	78.53	80.55	66.71	87.97
15000	9.10	95.69	92.49	78.32	80.64	66.86	88.45

5. Discussion

Failure Cases Fig. 4 presents samples that have been successfully reconstructed and identified with a high match score as the original identities. Conversely, Fig. 5 illustrates some failure cases of our approach in blackbox attack against ArcFace embeddings. Although the model attempts to preserve the identity, it does not constrain the age, leading to inconsistencies. The model also struggles with extreme expressions and poses, and in some instances, it fails to preserve the ethnicity of the original samples. While the

Figure 5. Sample failure cases: Images from the LFW dataset (first row) and their corresponding reconstructed face images in black-box attack against ArcFace. The values show the cosine similarity between embeddings of original and reconstructed face images.

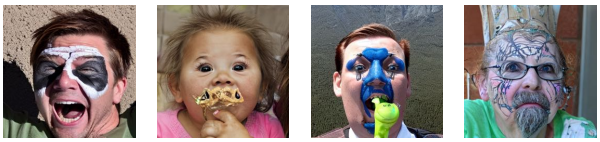

Figure 6. Samples of reconstructed face images with artifacts.

reconstructions are effective in terms of attack success rates, these limitations suggest that the face embedding may not capture sufficient details for certain attributes.

Another group of failure cases corresponds to images that have artifacts in the generation using the Arc2Face model. Fig. 6 illustrates such cases, where the generated face images do not look realistic. The generation of these samples can be caused by an error in translating embeddings by our adapter module. In such a case, the generated embedding by our adapter module may not fit the distribution of embedding space of the foundation model, and therefore the generator may generate out of distribution image.

Societal Impacts In this paper, we proposed an efficient adapter module to map the embeddings of any face recognition model to the embeddings space of Arc2Face as a foundation model. The reconstructed face images by our method have high quality and outperform the state-of-the-art in the literature, demonstrating the effectiveness of our adapter module. Our adapter module can be used in future foundation models which are based on embeddings of a preset face recognition model. Our experiments also show that the reconstructed face images by our method using a foundation can achieve high success attack rates when attacking a face recognition system. The reconstructed face images are also transferable and can be used to enter any other system in which the subjects are enrolled. Our results also shed light on the critical vulnerability of face recognition systems to face reconstruction attacks. While this work is conducted with the motivation of showing the vulnerability of face recognition systems, we do not condone using our method with the intent of attacking a real face recognition system. In a similar vein, data protection regulations, e.g., the European Union General Data Protection Regulation (EU-GDPR) [13], classify biometric data as sensitive information, and impose legal obligations to protect them. To address and mitigate such attacks against biometric systems, several template protection methods are proposed in the literature [22, 24, 30, 31, 40, 41]. We should also note that the project on which the work has been conducted has undergone and passed an Internal Ethical Review Board (IRB).

6. Conclusion

The foundation models are trained with a large volume of data and can be used in different tasks. In this paper, we used a face foundation model and proposed a new face reconstruction attack against face recognition systems. While the foundation model can generate face images given embeddings of a preset face recognition model, we trained an efficient adapter module that can map embeddings of a blackbox face recognition model to the embedding space of the foundation model. Then, we can use the foundation model to generate the face images from embeddings of various face recognition models. Our experiments show that our method outperforms previous face reconstruction methods in the literature and demonstrate the effectiveness of our approach. In particular, the adversary requires a limited computation resource to train the adapter module and then conduct the face reconstruction attack. The reconstructed face images by our method are realistic and have high quality. In addition, the reconstructed images are transferable and can be used to enter any other system in which the same subject is enrolled. Our results show the serious vulnerability of face recognition systems to face reconstruction attacks. Moreover, the results achieved by training a light adapter module in conjunction with a pretrained foundation model shed light on the potential of foundation models and demonstrate that the application of a pretrained foundation model can be easily used with different face recognition models. Hence, this paper paves the way for future studies on the application of foundation models for various problems related to face recognition systems.

Acknowledgment

This research is based upon work funded by the Hasler foundation through the Responsible Face Recognition (SAFER) project as well as the H2020 TReSPAsS-ETN Marie Skłodowska-Curie-Curie early training network under agreement 860813. This work was also supported by the Swiss Center for Biometrics Research & Testing at Idiap Research Institute.

References

[1] 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)*, pages 1–11. IEEE, 2022. 3

[2] 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*, 32(5):31–41, 2015. 1

[3] 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)*, pages 1578–1587, 2022. 5

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 3

[5] 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)*, pages 3703–3712, 2017. 2, 3

[6] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 1

[7] 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. 5

[8] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian 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. 5

[9] Xingbo Dong, Zhe Jin, Zhenhua Guo, and Andrew Beng Jin Teoh. Towards generating high definition face images from deep templates. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–11. IEEE, 2021. 3, 6

[10] 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*, 125: 103026, 2023. 2, 3, 6

[11] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7714–7722, 2019. 1

[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)*, pages 6132–6141, 2020. 3

[13] European Council. Regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation), 2016. 8

[14] Javier Galbally, Chris McCool, Julian Fierrez, Sebastien Marcel, and Javier Ortega-Garcia. On the vulnerability of face verification systems to hill-climbing attacks. *Pattern Recognition*, 43(3):1027–1038, 2010. 1

[15] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014. 1

[16] 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*, pages 87–102. Springer, 2016. 5

[17] Abdenour Hadid, Nicholas Evans, Sebastien Marcel, and Julian Fierrez. Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. *IEEE Signal Processing Magazine*, 32(5):20–30, 2015. 1

[18] 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. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 5

[19] 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)*, pages 4401–4410, 2019. 5

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo 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. 3

[21] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[22] Prabhjot Kaur, Nitin Kumar, and Maheep Singh. Biometric cryptosystems: a comprehensive survey. *Multimedia Tools and Applications*, pages 1–56, 2022. 8

[23] 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, 2015. 4

[24] Nitin Kumar et al. Cancelable biometrics: a comprehensive survey. *Artificial Intelligence Review*, 53(5):3403–3446, 2020. 8

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin trans-

former: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 5

[26] 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*, 41(5):1188–1202, 2018. 2, 3, 6

[27] Sébastien Marcel, Julian Fierrez, and Nicholas Evans. *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Springer Nature, 2023. 1

[28] Chris McCool, Roy Wallace, Mitchell McLaren, Laurent El Shafey, and Sébastien Marcel. Session variability modelling for face authentication. *IET Biometrics*, 2(3):117–129, 2013. 5

[29] 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*, pages 51–59, 2017. 5

[30] Karthik Nandakumar and Anil K Jain. Biometric template protection: Bridging the performance gap between theory and practice. *IEEE Signal Processing Magazine*, 32(5):88–100, 2015. 8

[31] Hatef Otroshi Shahreza. *On the Information in Deep Biometric Templates: from Vulnerability of Unprotected Templates to Leakage in Protected Templates*. PhD thesis, EPFL, 2024. 8

[32] Hatef Otroshi Shahreza and Sébastien Marcel. Inversion of deep facial templates using synthetic data. In *Proceedings of the IEEE International Joint Conference on Biometric*, 2023. 6

[33] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model of human faces. In *European Conference on Computer Vision*, 2024. 2, 4, 5

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[36] 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. 3, 6

[37] 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)*, pages 2435–2439. IEEE, 2023. 3, 6

[38] 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. 2, 3, 6

[39] Hatef Otroshi Shahreza and Sébastien Marcel. Template inversion attack using synthetic face images against real face recognition systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 3

[40] Hatef Otroshi Shahreza, Christian Rathgeb, Dailé Osorio-Roig, Vedrana Krivokuća Hahn, Sébastien Marcel, and Christoph Busch. Hybrid protection of biometric templates by combining homomorphic encryption and cancelable biometrics. In *Proceedings of the 2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2022. 8

[41] Hatef Otroshi Shahreza, Vedrana Krivokuća Hahn, and Sébastien 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. 8

[42] Mandavilli Srinivas and Lalit M Patnaik. Genetic algorithms: A survey. *Computer*, 27(6):17–26, 1994. 3

[43] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987. 3

[44] Edward Vendrow and Joshua Vendrow. Realistic face reconstruction from deep embeddings. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021. 2, 3, 6

[45] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 5

[46] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 5

[47] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM international conference on Multimedia*, 2021. 5

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 5

[49] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 2, 4, 5