Supplement for "Face Reconstruction from Face Embeddings using Adapters 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

Abstract

In this supplement, we report further ablation studies to investigate the effectiveness of the adapter module and its structure. As a new experiment, we feed the leaked embeddings directly to the foundation model and evaluate the generated face images, which demonstrates the effectiveness of our adapter module. We also consider adapter module with different numbers of layers and evaluate the performance of generated images. Furthermore, we report the success attack rate for smaller values of False Match Rate (FMR), such as 10^{-4} and 10^{-5} on the MOBIO dataset, which includes images captured with mobile devices. Project Page: https://www.idiap.ch/paper/face_adapter

1. Ablation Study on the Effectiveness of the Proposed Adapter Module

To investigate the effectiveness of our proposed adapter module, we perform an ablation study, where we feed the embeddings from different feature extractors directly to the foundation model (without the adapter module). In this case, the input is not from the same space as the F_{FM}, and therefore we can expect that the foundation model will generate images that do not have similar identities. Tab. 1 of this supplement compares the success attack rate (SAR) of the reconstructed face images with and without the adapter module. As the results in this table show, without using the adapter module the SAR values are almost zero on all benchmark datasets. Fig. 1 of this supplement also illustrates sample face images using the foundation model with and without the adapter module. This experiment demonstrates the effectiveness of our adapter module, which enables us to use the foundation model with any arbitrary face recognition model. For the particular case of ArcFace as target/victim model, we should emphasize that the model is different with the face recognition of foundation model F_{FM}. As described in section 4.1 of the paper, the face recognition model F_{FM} used in the foundation model is trained with WebFace42M, while the model used as victim or target face recognition model is trained with MS-Celeb1M dataset. 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. Hence, as the results in Tab. 1 show, if we do not use the adapter module, the foundation model cannot successfully reconstruct face images.

Table 1. Evaluation of *blackbox* attacks against different face recognition models in terms of success attack rate (SAR) **with and without adapter module** at the false match rate (FMR) of 10^{-3} on the LFW dataset. In this experiment, the target face recognition model is the same as in the system from which the embeddings are leaked (i.e., $F_{\text{victim}} = F_{\text{target}}$).

	Victim Face Recognition					
	ArcFace	**ElasticFace**	**AttentionNet**	**HRNet**	**RepVGG**	**Swin**
with adapter	95.71	93.18	78.71	80.66	67.11	88.44
without adapter	0.07	0.01	0.02	0.01	0.01	0.05

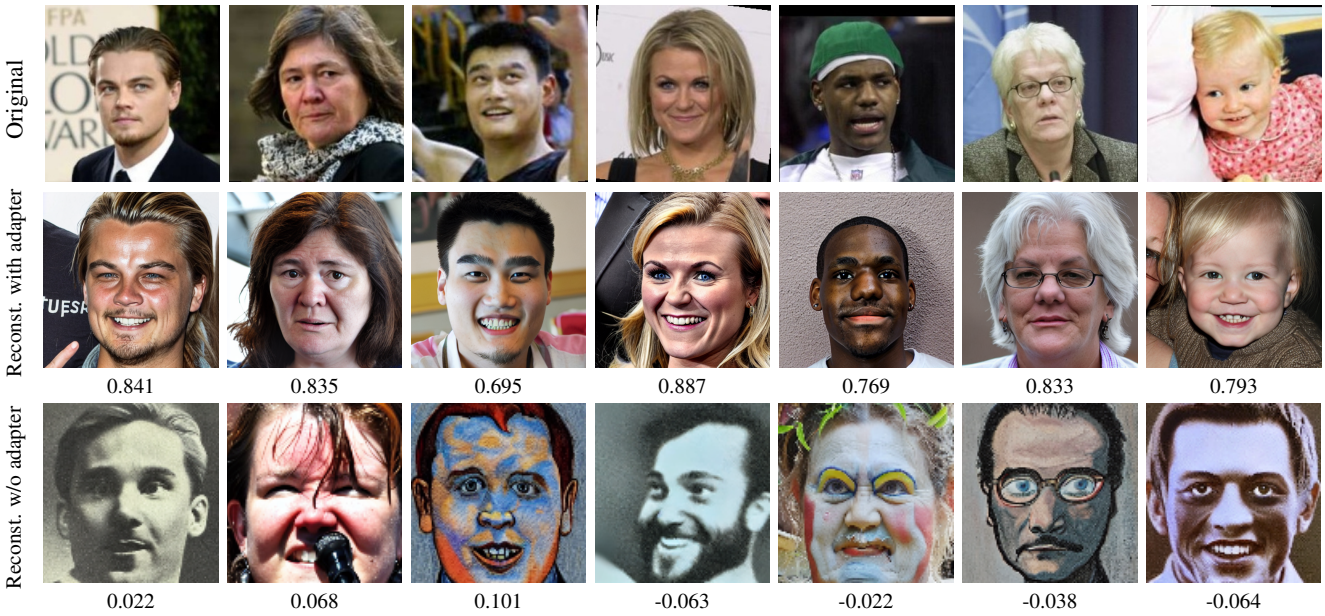

Figure 1. Sample face images from the LFW dataset (first row) and their reconstructed versions using adapter (second row) and without adapter (third row) from ArcFace embeddings. The values are cosine similarity of embeddings of the original and reconstructed face image.

2. Ablation Study on the Structure of Adapter Module

In our experiments in the paper, we considered the adapter module containing a single-layer network. As another experiment, we perform an ablation study on the structure of the adapter module with different numbers of layers. We consider the adapter module with two-layer and three-layer network structures (with the same number of neurons in the middle layers). Tab. 2 of this supplement reports the performance of the reconstructed face images for adapters with different numbers of layers on the LFW dataset. As the results in this table show only a single linear can achieve very good performance and increasing the depth of the network does not improve the performance. We can explain this with the fact that we assume the face embeddings of each face recognition on a hypersphere, where for each face recognition model different identities cover the surface of the corresponding hypersphere and have angular distances between each two identities (calculated with cosine similarity of embeddings for the matching). Therefore, we almost need to perform a simple linear transformation to map the hypersphere of embeddings for different face recognition models. However, increasing the dimension of the adapter module increases the complexity of the adapter and causes it to overfit.

Table 2. Evaluation of *blackbox* attacks against different face recognition models in terms of success attack rate (SAR) using **different structures for adapter module** at the false match rate (FMR) of 10^{-3} on the LFW dataset. In this experiment, the target face recognition model is the same as in the system from which the embeddings are leaked (i.e., $F_{\text{victim}} = F_{\text{target}}$).

Adapter	**Victim Face Recognition**					
	ArcFace	**ElasticFace**	**AttentionNet**	**HRNet**	**RepVGG**	**Swin**
1-layer	95.71	93.18	78.71	80.66	67.11	88.44
2-layer	95.70	92.53	78.03	79.16	64.66	87.84
3-layer	94.07	88.07	70.41	61.94	55.59	81.61

3. Reconstruction Performance for Low Values of FMR

In our experiments, we reported the success attack rate (SAR) for face recognition systems configured at False Match Rate (FMR) of 10^{-3}. Tab. 3 of this supplement reports the vulnerability of face recognition models for lower values of FMR on the MOBIO dataset (captured with mobile devices). As the results in this table show, for lower FMR values, the reconstructed

Table 3. Evaluation of *blackbox* attacks against different face recognition models at **different false match rate (FMR) values** on the MOBIO dataset in terms of success attack rate (SAR). In this experiment, the target face recognition model is the same as in the system from which the embeddings are leaked (i.e., $F_{\text{victim}} = F_{\text{target}}$).

FMR	**Victim Face Recognition**					
	ArcFace	**ElasticFace**	**AttentionNet**	**HRNet**	**RepVGG**	**Swin**
10^{-3}	100.0	99.05	99.52	99.52	98.57	99.52
10^{-4}	100.0	98.57	97.62	99.05	97.14	99.52
10^{-5}	100.0	97.14	96.19	98.57	95.24	99.05

face images can still achieve considerable SAR values, which indicates serious vulnerability in face recognition systems.