APSIPA Transactions on Signal and Information Processing, 2025, 14, e200

Original Paper

3D Morphable Master Face: Towards Controllable Wolf Attacks Against 2D and 3D Face Recognition Systems

Siyun Liang[1*], Huy H. Nguyen[2], Satoshi Ikehata[2], Junichi Yamagishi[2] and Isao Echizen[2,3]

[1] *Technical University of Munich, Germany*
[2] *National Institute of Informatics, Japan*
[3] *The University of Tokyo, Japan*

ABSTRACT

Biometric authentication systems are facing increasing threats from artificial intelligence-generated content. Previous research has revealed the vulnerability of 2D face authentication systems to master face attacks, which use GAN-based models to create facial samples capable of matching multiple registered user templates in the database. However, the effectiveness of such attacks in 3D scenarios has not been thoroughly investigated.

In this paper, we present a systematic approach to generate master faces that can compromise both 2D and 3D face recognition systems. It uses a latent variable evolution algorithm with a 3D face morphable model. Notably, our approach achieves, for the first time, controllable and morphable master face attacks on face authentication systems. We explore the effect of facial reenactment and face morphing on enhancing the efficacy of master face attacks and reducing the time required for master face generation. Comprehensive simulations of simultaneous master face attacks based on white-box, gray-box, and black-box scenarios demonstrated that our approach achieves superior attack success rates and has

*Corresponding author: siyun.liang@tum.de

Received 29 November 2024; revised 06 February 2025; accepted 19 February 2025
ISSN 2048-7703; DOI 10.1561/116.20240089

advanced flexibility compared with existing methods, highlighting the importance of defending against master face attacks.

Keywords: 3D master face, face recognition, 3D morphable face model

1 Introduction

Recent developments in artificial intelligence-generated content techniques have brought renewed attention to cybersecurity, particularly concerning biometric authentication systems. Large-scale real-world attacks on remote identity verification have been widely reported, employing adversarial techniques such as deepfake generation [17, 9], facial presentation attacks [16], and video injection attacks [6], which significantly compromise authentication systems [18, 15]. These attacks primarily target face recognition (FR) systems in verification mode, where an attacker attempts to impersonate a legitimate user by presenting altered or synthetic facial data. As a result, they typically require prior knowledge of the victims facial information.

In contrast, the "**wolf attack**" [59] enables attackers to generate generic "**master samples**" that closely resemble multiple enrolled biometric traits within the gallery of the authentication systems. Several studies have successfully created master face samples [42] using GAN-based 2D image generation models [34] without the need for specific victim information. These methods use 2D face recognition systems to assess the similarity between GANs-generated faces and real faces in a database. To improve this similarity, the latent variables input to GANs are iteratively refined, ultimately producing master face samples that effectively cover a wide range of identities in the gallery, thereby revealing the vulnerability of face-based authentication systems to master face attacks. However, previous studies on master face attacks have predominantly focused on 2D scenarios. Such attacks fail with the widespread incorporation of more robust 3D FR systems in contemporary authentication systems.

Friedlander *et al.* [20] introduced the first method for 3D master face generation, which reconstructed the 3D facial geometries from the 2D face images generated by the GAN-based model. They further evaluate the similarity between synthetic faces and real faces using both 2D and 3D FR systems, leveraging this feedback to optimize the latent code. While this approach successfully produces 3D master face samples that perform well in white-box attacks, it is seldom applicable in real-world attacks, which are often gray-box or black-box scenarios.

To make 3D master face attacks applicable in real-world scenarios, the following challenges need to be addressed:

Controllability. While traditional 2D FR systems generally rely on static frontal face images, modern face authentication systems have integrated liveness detection [43] to counter 2D presentation attacks. These systems typically require users to change their facial expressions and/or poses, thereby filtering out static attack samples. However, current 2D and 3D master face generation methods lack flexible controllability. This limitation arises from their reliance on GAN-based models. The entangled nature of latent variable spaces in such models prevents them from generating master faces that enable effective control of facial (semantic) features while maintaining output quality [1]. In addition to the inconvenience associated with facial reenactments, their morphing capabilities are also limited. Interpolation between latent codes can result in unwanted artifacts in the output images.

This highlights the need for a 3D facial template model that offers both robust facial geometry priors and a parametric face space, enabling controllable manipulation. In this work, we use a 3D morphable face model (3DMM) [2] to disentangle shape, appearance, expression, and pose parameters, enabling the production of highly controllable 3D facial samples directly in the 3D space. Compared to GAN-based generators, this attribute disentanglement better preserves critical 3D information, improves the controllability of facial attributes, and facilitates 3D face morphing. Additionally, since the texture space of 3DMM is learned from 3D facial scans, the texture and geometry of the generated master face samples are consistent. In contrast, GAN generates a single 2D image with limited facial information, and the textures and geometries of the reconstructed 3D faces are often misaligned. In contrast, 3DMM-based master faces are more suitable for physical presentation attacks.

Cross-modality. Master faces are rooted in the imbalanced distribution of features within the FR system [41]. Deep learning-based FR systems often suffer from non-uniform distributions in the feature space. Consequently, if a face falls within a dense cluster in the feature space, its likelihood of being falsely matched to other samples within that cluster increases. Training a master face can be regarded as approaching the densest cluster within the feature space of the FR system. However, acquiring a 3D master face that can compromise 2D and 3D FR systems simultaneously is extremely difficult because these dense clusters may not align between two systems, making it challenging to pinpoint cross-modal clusters of these vulnerable faces.

To this end, we propose a latent variable evolution (LVE) algorithm to iteratively optimize the disentangled shape and appearance of latent vectors of our 3DMM generator, using an objective function to calculate the joint false matching rate (FMR) of the generated faces based on their similarity with the facial data of the training set for optimization.

Generalizability. Real-world attacks are generally grey-box or black-box attacks, which means that the FR system targeted by the attacker may be

different from the one used to train the master face, and the distribution of the target face gallery may deviate from the face dataset used for training. As a result, the generated master faces may be difficult to generalize, leading to the failure of the attack.

While using multiple master faces in dictionary attacks manner [20] could alleviate this generalizability problem, current methods are so time-consuming that it takes up to an entire day to generate even a single master face, not to mention generating a set of multiple master faces. In our work, we substituted tedious multiple master faces generation by morphing a few master faces. The morph of two master faces produced with our framework retains the capability of multiple false matching due to smoothly bridging the matching space between the source master faces. Hence, only a small number of master faces generated from the training set are needed to obtain a large number of new master face samples via morphing, which greatly reduces the time required to train a large master face set and enlarges the potential coverage for real-world attack purposes.

In summary, our work introduces a novel framework for generating 3D master faces that can effectively compromise both 2D and 3D face recognition systems. By leveraging a 3DMM, we achieve a high level of controllability over facial attributes, which is critical for bypassing advanced face authentication systems. Additionally, we propose an LVE algorithm to enhance the cross-modality performance of generated master faces, ensuring their effectiveness across different FR systems. Finally, our approach addresses the generalizability challenge by efficiently generating a diverse set of master faces through morphing techniques, significantly reducing computational time while expanding the potential for real-world application.

2 Related Work

2.1 Face Recognition Systems

The last decade has seen the rapid development of deep learning methods for 2D face recognition. An important milestone was the introduction of the DeepFace model [57], which achieved an impressive accuracy rate of 97.35% on the LFW benchmark [32], approaching human-level performance. Subsequently, the application of convolutional neural networks (CNNs) to FR systems flourished. Schroff *et al.* presented the FaceNet model [50], which was trained with a triplet loss function on a GoogLeNet architecture. Liu *et al.* instead proposed a novel angular softmax loss [39]. Further, Wang *et al.* [60] and Deng *et al.* [11] addressed the optimization challenges of this loss with additive cosine and angular margin. More recent research has explored adaptive loss functions [36], including the adaptive margin for image quality.

In contrast, 3D FR systems, known for their superior performance in challenging cases compared with their 2D counterparts, have received less attention in deep learning-based research. This is partly due to the scarcity and privacy sensitivity of 3D facial training data. The first CNN-based 3D FR model [35] involved fine-tuning the pre-trained 2D VGGFace model [4] with facial depth maps. Gilani and Mian [23] combined public-available 3D face datasets to create a comprehensive one for training a CNN-based model called FR3DNet from scratch. To address the challenges posed by the lack of high-quality training data, Mu *et al.* [40] proposed Led3D, an open-source lightweight CNN model that uses low-quality depth images captured using a Kinect sensor for training, achieving state-of-the-art performance.

Recognizing that real-world 3D face acquisition often involves live capture using commercial range cameras rather than static, high-resolution 3D scanners in a lab setting, we consider Led3D to be a suitable 3D FR system for simulating authentic, real-world scenarios. Additionally, inspired by Kim *et al.* [35], we fine-tuned a commonly used 2D FR system called ArcFace [11], which was initially trained on an IResNet [13] backbone, by using a high-resolution FaceScape [62] dataset. The incorporation of these two FR systems enables us to simulate a broader range of situations in real-life authentication scenarios.

2.2 Face Generation

Among the various generative models for creating 2D facial images, the generative adversarial network (GAN) [24] framework is noteworthy. GAN can be conceptualized as a two-player minimax game between the generator and the discriminator. The generator is a differentiable function that transforms an initial latent vector into a data sample, striving to generate data that closely resembles real training data. In contrast, the discriminator is trained to differentiate between samples generated by the generator and real training data. An important development was that of StyleGAN [33], which includes a mapping network that separates content and style information, leading to improved control over the appearance of generated images.

Our research emphasizes 3D face generation methods, particularly those involving the widely used 3DMM [2]. This model disentangles facial components such as shape, appearance, and expressions, facilitating statistical capture of variations and tasks like facial reenactment. The preprocessing stage establishes point-to-point correspondence within the training database, which enables meaningful combinations of faces and face generation through coefficient sampling [14]. Furthermore, analysis-by-synthesis techniques allow for the estimation of these coefficients directly from 2D images, making it a foundational approach for single-image 3D face reconstruction.

Recent non-linear extensions of 3DMM have been developed using autoencoder-based [70, 46] and GAN-based architectures [55, 7, 51]. These ap-

proaches significantly enhance single-image 3D face reconstruction. For instance, DECA [19] introduces expression-conditioned displacement models learned in a self-supervised manner, enabling both high-fidelity 3D face reconstruction and realistic facial animation from in-the-wild images. More recently, researchers have explored combining 3DMM with advanced neural 3D representations such as Neural Radiance Fields [69, 21] and 3D Gaussian Splattings [61]. These hybrid approaches enhance dynamic head reconstruction from monocular video by leveraging both the parametric control of 3DMM and the view-dependent rendering capabilities of neural representations[22, 52]. These advancements in 3DMM introduce new security risks. The ability to generate highly realistic and dynamically controllable synthetic faces increases the vulnerability of face recognition-based authentication systems, posing new challenges for biometric security.

2.3 Master Attack

The wolf attack, also known as the master attack, was introduced by Une *et al.* [59]. This attack aims to create a generic sample capable of falsely matching multiple enrolled subjects in a biometric authentication system's gallery. Initially applied to fingerprint-based authentication systems [3], this concept was further extended to face-based authentication systems [42]. Recent research [53, 41] analyzed master faces, exploring their properties and assessing their generalizability across different datasets and 2D FR systems.

3 Proposed Method

3.1 Overview

Our training process for 3D master face generation is illustrated in Figure 1. The collected authentic human templates in the training set are denoted as $\mathcal{T}_h$. While numerous publicly available 3D facial datasets primarily consist of human face meshes, many 3D FR systems use depth images as input rather than the entire mesh. To accommodate this, we developed a data preprocessing pipeline labeled P, which is detailed in Section 4.1. This pipeline transforms $\mathcal{T}_h$ into RGB and Depth (RGB-D) image pairs for each facial scan.

Face authentication systems use FR models to encode input images into lower-dimensional feature representations. For 2D FR, the function $f_{\text{2d}} : \mathbb{R}^{W\times H\times 3} \rightarrow \mathbb{R}^d$ maps color images to a d-dimensional space. A similar function is used for 3D FR, utilizing depth images as input instead.

The face matching function $m : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \{0, 1\}$ is used to predict whether the embeddings of the two inputs correspond to the same identity. This matching function is conditioned on a chosen threshold θ specific to the

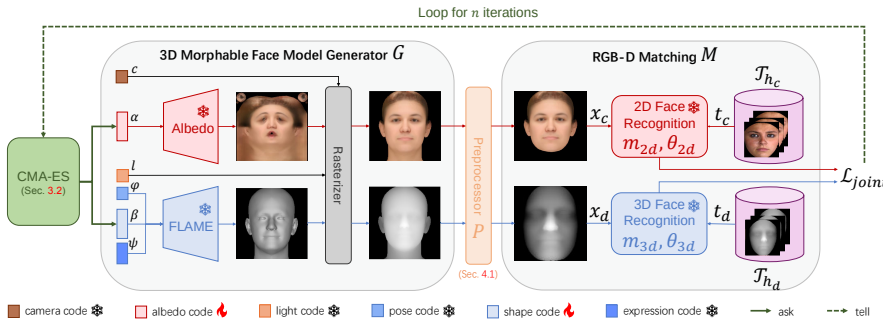


Figure 1: The Latent Variable Evolution process with 3D Morphable Face Model. The CMA-ES optimizer iteratively updates the albedo code α and the shape code β to generate a 3D master face that maximizes the joint false matching rate across both 2D and 3D FR systems. The FLAME model and Albedo model of a 3DMM produce a synthesized face mesh and texture, which are then rasterized and passed through the preprocessor P to create RGB-D images. These images are fed into 2D and 3D FR, where they are compared against a gallery of authentic faces $\mathcal{T}_h$. The ask-and-tell loop continues for n iterations, ultimately yielding a 3D master face that simultaneously compromises both modalities.

selected similarity metric, in our case, the cosine similarity metric between feature embeddings. However, our work necessitated the simultaneous consideration of RGB-D matching, leading to a more complex matching function:

$$M(\boldsymbol{t}_1, \boldsymbol{t}_2, \theta_{2\mathrm{d}}, \theta_{3\mathrm{d}}) = m_{2\mathrm{d}} \cap m_{3\mathrm{d}} \rightarrow \{0, 1\}, \tag{1}$$

where the two matching functions are:

$$m_{2\mathrm{d}}(f_{2\mathrm{d}}(\boldsymbol{t}_{1_c}), f_{2\mathrm{d}}(\boldsymbol{t}_{2_c}), \theta_{2\mathrm{d}}) \rightarrow \{0, 1\}, \tag{2}$$

and

$$m_{3\mathrm{d}}(f_{3\mathrm{d}}(\boldsymbol{t}_{1_d}), f_{3\mathrm{d}}(\boldsymbol{t}_{2_d}), \theta_{3\mathrm{d}}) \rightarrow \{0, 1\}. \tag{3}$$

Based on the above notation, our objective in master face generation is to produce a forged sample $\boldsymbol{x}$ that can match the highest number of enrolled templates in the training set and compromise both 2D and 3D FR systems with the most false matches.

$$\boldsymbol{x} = \arg\max_{\boldsymbol{x}} \sum_{\boldsymbol{t} \in \mathcal{T}_h} M(\boldsymbol{x}, \boldsymbol{t}, \theta_{2\mathrm{d}}, \theta_{3\mathrm{d}}) \tag{4}$$

Since our objective is to generate a master face that can simultaneously compromise both the 2D and 3D FR systems, focusing solely on maximizing cases where $m_{2\mathrm{d}} = m_{3\mathrm{d}} = 1$ is both sufficient and effective. This design ensures that the optimization process concentrates on satisfying the shared constraints of both systems without being distracted by the edge cases of inconsistency.

To this end, we use a 3DMM-based face generator G to synthesize a 3D face mesh, conditioned on a set of latent codes, which are the camera code $\boldsymbol{c}$, albedo code $\boldsymbol{\alpha}$, light code $\boldsymbol{l}$, shape code $\boldsymbol{\beta}$, pose code $\boldsymbol{\varphi}$, and expression codes $\boldsymbol{\psi}$ [37]. Human face templates in the FR systems are typically front-facing and expressionless, so we optimize only the albedo code $\boldsymbol{\alpha}$ and shape code $\boldsymbol{\beta}$ and freeze the other codes to simplify the training procedure. We then utilize the same data preprocessor P to produce the RGB-D image pair of this synthesized face. We therefore re-formulate the master face generation problem as finding an optimal pair of latent vectors $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ that results in the highest FMR:

$$(\boldsymbol{\alpha}, \boldsymbol{\beta})_{opt} = \arg\max_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} \frac{\sum_{\boldsymbol{t} \in \mathcal{T}_h} M(P(G(\boldsymbol{\alpha}, \boldsymbol{\beta})), \boldsymbol{t}, \theta_{2\mathrm{d}}, \theta_{3\mathrm{d}})}{\|\mathcal{T}_h\|} \tag{5}$$

In particular, our maximization objective deliberately ignores cases where $m_{2\mathrm{d}} \neq m_{3\mathrm{d}}$, as these inconsistencies fail to provide clear guidance on how to update the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Specifically, changes in the albedo code primarily affect the facial appearance, influencing the 2D FR system but having little impact on the 3D FR system. In contrast, changes in the shape code alter the facial geometry, significantly affecting both the 2D and 3D FR systems. As a result, in cases where inconsistencies occur, it is ambiguous whether the they come from the appearance variation or the geometric variation.

Maximizing the count of matches requires an iterative process to refine $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. For this purpose, we introduce an LVE strategy in the following Section 3.2.

3.2 Latent Variable Evolution Algorithm

We formalized the process for refining an initial latent vector $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as outlined in Algorithm 1. To address the optimization challenges inherent in generating master faces, which involve non-differentiable thresholding operations, we used the covariance matrix adaptation evolution strategy (CMA-ES) [26] as our optimizer.

Our implementation of the LVE algorithm leverages the ask-and-tell interface of CMA-ES. First, we initialize the CMA-ES solver with random latent codes. When we "ask" the solver for solutions, it generates potential candidate solutions by sampling from a multivariate normal distribution with parameters determined during initialization. We execute the complete generation and matching procedure using these candidate solutions to obtain fitness scores from our objective function. These scores are subsequently "told" to the CMA-ES optimizer. The optimizer utilizes this feedback to update its distribution parameters, including the distribution mean vector and covariance matrix, for the subsequent iterations of the ask-and-tell process. This

Algorithm 1 Latent variable evolution pseudo code

$m \leftarrow 22$ ▷ Population size
$\mathcal{F}, \mathcal{S} \leftarrow \{\}$ ▷ Empty master face & score set
$(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leftarrow \mathbf{0}$ ▷ Initialization
for n iterations **do** ▷ Run LVE n times
  $\mathcal{X} \leftarrow P(G((\boldsymbol{\alpha}, \boldsymbol{\beta})))$ ▷ Generate m faces
  $\boldsymbol{s} \leftarrow \mathbf{0}$ ▷ Initialize scores $\boldsymbol{s} \in \mathbb{R}^m$
  for face $\boldsymbol{x}_i$ in faces $\mathcal{X}$ **do**
    for face $\boldsymbol{t}_j$ in faces $\mathcal{T}$ **do**
      $\boldsymbol{s}_i \leftarrow \boldsymbol{s}_i + M(\boldsymbol{x}_i, \boldsymbol{t}_j, \boldsymbol{\theta})$ ▷ See Section 3.2
    end for
  end for
  $\boldsymbol{s} \leftarrow \frac{\boldsymbol{s}}{|\mathcal{T}|}$ ▷ Calcuating FMR
  $\boldsymbol{x}_b, \boldsymbol{s}_b \leftarrow \text{GetBestFace}(\mathcal{X}, \boldsymbol{s}, \boldsymbol{\theta})$
  $\mathcal{F} \leftarrow \mathcal{F} \cup \{\boldsymbol{x}_b\}$ ▷ Append best master face
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{s}_b\}$ ▷ Append best score
  $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leftarrow \text{CMA-ES}(\mathbf{1} - \boldsymbol{s})$
end for
return $\mathcal{F}, \mathcal{S}$
$\boldsymbol{f}_b, \boldsymbol{s}_b \leftarrow \text{GetBestFace}(\mathcal{F}, \mathcal{S}, \boldsymbol{\theta})$ ▷ Find best master faces

iterative approach enables the optimizer to progressively explore the search space, ultimately converging towards an optimal solution.

The key challenge lies in defining an appropriate objective function that guides the CMA-ES algorithm effectively toward improved solutions. In prior studies on 2D master face generation [42], the optimization process used scores of the similarity between two faces, aiming at increasing these scores. In contrast, our work introduces complexity by incorporating both 2D and 3D FR systems, emphasizing simultaneous matches. Our experiments in Section 4.8 show that enhancing similarity scores for both 2D and 3D FR systems might not yield the desired outcomes. This is because a face sample with high average similarity scores in both 2D and 3D FR systems could be matched to different individuals across modalities due to distinct feature space distributions. To address this challenge, we use the matching function described in Section 3.1, which quantifies the count of concurrent 2D and 3D matches for the same individual. The final objective function tends to maximize the joint FMR on both 2D and 3D FR systems:

$$\mathcal{L}_{joint} = 1 - \text{FMR}_{\text{joint}} + \omega \|\boldsymbol{\beta}\|_2, \tag{6}$$

where the joint FMR is defined as:

$$\mathrm{FMR}_{\mathrm{joint}} = \frac{\sum_{\mathbf{t}\in\mathcal{T}_h} M(\mathbf{x}, \mathbf{t}, \theta_{2\mathrm{d}}, \theta_{3\mathrm{d}})}{\|\mathcal{T}_h\|}. \tag{7}$$

Here the $\omega\|\boldsymbol{\beta}\|_2$ defines a regularization term of the shape vector $\boldsymbol{\beta}$ in a 3DMM. This regularization penalizes extreme deviations in the shape vector that could result in unrealistic or anatomically implausible face shapes, ensuring that the generated shapes remain within a reasonable range of natural human facial geometry.

3.3 Baseline

We compare our methods with the first 3D master face generation method [20], which reconstructs the 3D geometry from images generated through StyleGAN2.

One limitation of the baseline derives from the instability of the unconditional GAN-based generator. Randomly sampling the latent vector could yield human faces with varying poses and expressions. Faces with exaggerated expressions or excessively deviated poses are difficult to optimize, which degrades performance. The authors, therefore, ran the LVE algorithm five times and selected the optimal outcome for evaluation. Although effective, this method is computationally expensive.

Another limitation is that optimizing within the latent space of 2D GAN during the optimization stage compromises the information available from 3D FR. 3D face reconstruction from a single image is an ill-posed problem. Therefore, the reconstruction process typically introduces inaccuracies and uncertainties, leading to a loss of information related to the characteristics of the 3D master face. Additionally, since the 3D geometry is estimated from 2D images, controlling the 3D domain without affecting 2D appearance is challenging, resulting in reduced controllability.

In contrast, our method stably generates highly controllable 3D master faces and effectively utilizes information from both modalities. For comparison, we re-implemented the baseline method using StyleGAN2 and the DECA 3D face reconstruction model instead of the original reconstruction network [12]. The reason is that DECA uses the FLAME topology for 3D face reconstruction, enhancing fairness in comparisons. Furthermore, DECA achieved better reconstruction performance than the work mentioned above on the NoW benchmark [48].

4 Experiments

4.1 Experimental Setup

Datasets

In our experiments, we used four 3D face datasets and four FR systems, enabling us to explore various configurations and assess the generalizability of master faces. The details of the datasets involved are presented in Table 1. We extracted data for 60 individuals, comprising a total of 1,500 scans, from the BU-3DFE dataset [64] to form the training set for master face generation. To ensure an extensive evaluation, the remaining 40 identities were randomly shuffled and allocated to the development (dev) and evaluation (eval) sets. The Headspace [10] and Texas3D [25] datasets are used as targets in the attacking phase and split into dev and eval sets too. Specifically, the dev set of each dataset was used for conducting a grid search to identify an optimal threshold that effectively balances the false acceptance rate (FAR) and false rejection rate (FRR), ultimately minimizing the equal error rate (EER), as shown in Table 2. As Headspace provides only one sample image per individual, we manually selected thresholds to ensure that both 2D and 3D FR systems achieved an EER of less than 2%.

Table 1: Details of 3D facial datasets used in our experiments.

Database	Data type	IDs	Scans	Exps
BU-3DFE [64]	Mesh	100	2,500	25
Texas3D [25]	Range Images	118	1,149	Various
Headspace [10]	Mesh	1,519	1,519	1
FaceScape [62]	Mesh	847	16,940	20

Table 2: Equal error rates (%) computed on each dataset-FR system pair.

	FaceNet [50]	AdaFace [36]	IResnet100 [13]	Led3D [40]
BU-3DFE	1.17	10.35	9.27	11.70
Texas3D	0.08	6.64	4.69	4.00
Headspace	1.95	1.70	1.79	1.29

Although the FaceScape dataset [62] has the largest number of samples, its facial topology does not include eyes and mouth, making it unsuitable for training the master face. We thus used its released bilinear model to generate 300 different samples, each having 52 different expression meshes rendered in 9 different poses. Inspired by Kim *et al.* [35], we used these rendered depth

maps to fine-tune a pre-trained 2D FR system [13], resulting in a workable 3D FR system.

Data Preprocessing

Our experiments required two rounds of data preprocessing. First, for datasets with inconsistent topologies and varying facial poses as raw data, i.e., BU-3DFE and Headspace, we selected one facial scan as a template. We then conducted a Procrustes analysis based on the landmark data for each facial scan to align them. This enabled us to further use the selected intrinsic parameters to render the entire mesh dataset into an RGB-D dataset.

During preprocessing, we used face detection and cropping to transform the rendered datasets into valid input data for the FR systems. We used the same parameters settings for the MTCNN face detector [68] used for FaceNet and AdaFace. During the training process, we used a face parser based on the bilateral segmentation network (BiSeNet) [65] to filter out irrelevant information, such as background and neck regions, from the intermediate results.

For the rendered 3D depth maps based on the FLAME topology, we first used a pre-defined vertex mask to retain only the depth information for the facial region. We then carried out preprocessing relevant to the target FR system. The preprocessing pipeline corresponds to that for Led3D, which includes nose tip calibration, outliers removal, and depth normalization.

Face Recoginition Systems

From among the many open-source 2D FR systems, we selected FaceNet and AdaFace. FaceNet [50] is based on the GoogLeNet (InceptionNet) [56] architecture and trained with triplet loss. As a highly regarded 2D FR model widely used to this day, FaceNet has demonstrated high efficiency and accuracy. We used a FaceNet model pre-trained on the VGGFace2 [4] dataset for the experiments. AdaFace [36] features a novel loss function based on adjustable image quality. We used an AdaFace model, which used ResNet18 [27] as the backbone, pre-trained on the CASIA-WebFace dataset [63].

There are relatively few open-source models for 3D FR systems, primarily due to the scarcity of public available databases. Hence, we used a fine-tuned IResnet100 model originally trained on the MS1MV2 dataset [11]. We also used a 3D FR system based on an open-source lightweight CNN model named Led3D [40], which incorporates a spatial attention vectorization module for multi-level feature fusion. Initially pre-trained on a combination of the Face Recognition Grand Challenge (FRGC) v2 dataset [45] and Bosphorus dataset [49], it was further fine-tuned using the Lock3DFace dataset [67], which consists of Kinect-captured low-quality 3D face images. Notably, for fair experiments, we carefully selected the pre-trained 2D and 3D FR systems to ensure that their training sets did not overlap with the dataset we used for training and evaluating master faces.

Setting

We simulate and evaluate different attack scenarios as shown in Figure 2. In **Master Face Generation Phase**, we use the BU-3DFE training dataset and FaceNet/IResnet100 FR systems pair to generate a set of master faces. The evaluation is done in the **Attacking Phase**, where we use the generated master faces to attack specific settings of a face authentication system. If the targeting system shares the same dataset and FR systems with those used for the generation phase, we consider this a **white-box attack**. If the only partial settings are overlapped, we consider it a **gray-box attack**. The most difficult case is the **black-box attack**, where both the dataset and the FR systems of the target is completely different from the training setting.

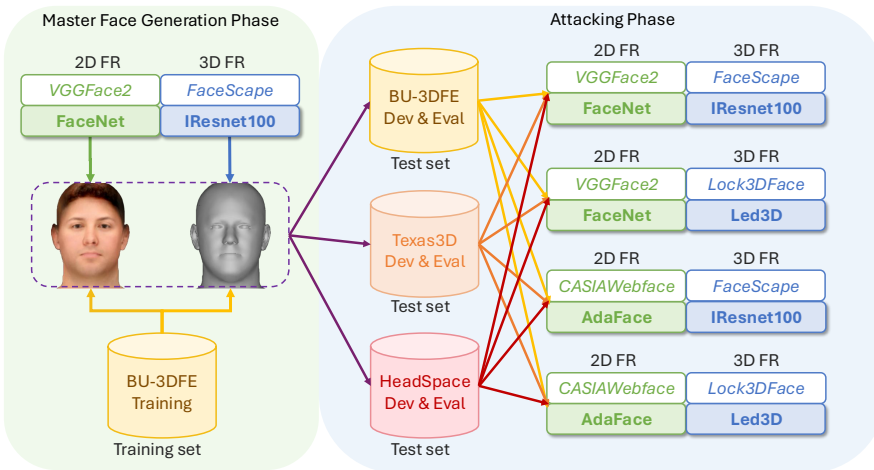


Figure 2: **Master face attack scenarios**. Master faces were created during the generation phase on a fixed dataset and FR systems and then used for attacking. A combination of 3 test datasets (further divided into dev and eval sets) and 4 FR pairs resulted in a total of 12 attack settings, categorized as white/gray/black-box attacks depending on the extent of overlap with the generation phase.

4.2 Metrics and Reference Anchor

Given $\mathbf{x}$ as the generated master face sample and given the context of the target, we typically used the joint FMR on both 2D and 3D FR systems as the evaluation metric, as defined in Equation 7.

Apparently, the FMR is affected by the choice of the training dataset and the performance of the FR systems selected. Due to variations in the assessments of different FR systems in previous research, **there is currently no unified benchmark for evaluating the success rate of master face**

attacks. To the best of our knowledge, our research is the first attempt to simultaneously assess this success rate for both 2D and 3D systems in terms of generalization. Therefore, besides the reconstruction-based baseline [20] detailed in Section 3.3, we set two reference anchors, which are the FMRs of natural master faces obtained on the training set and the test set.

A natural master face is a **bona fide face sample** that possesses master face capability. Given an arbitrary dataset and 2D/3D FR systems pair, for each bona fide face data within the dataset, we can calculate the number of genuine templates in the dataset that it could falsely match with, conditioned on the matching function of the given FR systems. The one with the highest FMR is identified as the natural master face under that specific setting. Therefore, we can compute the natural master face on the training set using the generation phase setting. In addition, for each of the twelve settings in the attacking phase, as shown in Figure 2, we can obtain the natural master face on the test set.

To be specific, for each attacking scenario out of the twelve settings, we evaluate the FMR with the following baseline:

1. **Attack with the natural master face based on the test set**: We assume the attacker already knows the gallery and the FR systems of the targeted face authentication system, making the attack **white-box**. While generally impossible in real-world scenarios, it serves as an anchor for evaluating the **"best ideal" performance of a master face attack**.

2. **Attack with the natural master face based on the training set**: In this attack setting, the natural master faces calculated with the settings from the generation phase are used. This means that they are generated under the same conditions as our synthesized master faces. This anchor supports the comparison of the attack success rates **between genuine and synthesized master face samples**.

3. **Attack with the synthesized master face from Friedlander *et al.* [20]**: We use the same setting in the generation phase to get master faces from the baseline [20]. We try both attacks with a single master face or multiple master faces using a greedy strategy.

The FMR resulting from the above baselines is compared to the FMR achieved using our synthesized master face approach to evaluate effectiveness. We present results in in Tables 3 and 4, with further analysis in Section 4.6.

4.3 Master Face Generation and Attack

Master face generation refers to the generation phase depicted in Figure 2, in which we ran the LVE algorithm (Algorithm 1) for 1,000 iterations on a

Table 3: Success rates for master face attacks simulated with different settings (in total 12 settings, each setting on dev and eval set), divided into two sub-tables.

(a)

FRs	Strategy	BU-3DFE dev (%)			BU-3DFE eval (%)			Headspace dev(%)		
		2D	3D	Joint	2D	3D	Joint	2D	3D	Joint
FaceNet IResNet	Avg	1.09	9.06	0.01	1.39	13.99	0.35	3.89	3.56	0.34
	Best	**1.20**	**1.60**	**0.80**	**10.60**	**34.40**	**7.40**	**17.56**	**11.78**	**4.19**
	Single[20]	0.00	6.80	0.00	3.20	5.20	0.00	0.20	0.20	0.00
	Greedy[20]	0.20	23.80	0.00	3.20	28.60	0.00	7.78	1.40	0.00
	Single	0.80	40.00	0.80	4.20	56.60	4.20	5.99	6.79	1.00
	Greedy	3.00	48.40	2.80	15.40	64.60	14.00	15.97	16.37	2.59
	Morph	***4.40***	***51.80***	***4.40***	***19.60***	***67.00***	***19.20***	***20.96***	***22.75***	***4.59***
FaceNet Led3D	Avg	1.09	11.74	0.06	1.39	22.74	0.84	3.89	2.58	0.27
	est	**5.20**	**2.20**	**2.20**	**10.60**	**46.80**	**9.40**	**17.56**	**10.78**	**4.19**
	Single[20]	0.00	6.80	0.00	3.20	0.80	0.00	0.20	0.00	0.00
	Greedy[20]	0.20	15.20	0.00	3.20	14.20	0.00	7.78	0.00	0.00
	Single	0.80	35.80	0.60	4.20	46.80	4.00	5.99	6.99	0.40
	Greedy	3.00	49.80	2.00	15.40	53.40	11.60	15.97	10.78	0.40
	Morph	***4.40***	***55.40***	***4.20***	***19.60***	***60.60***	***18.20***	***20.96***	***13.77***	***2.20***
AdaFace IResNet	Avg	9.88	9.06	1.91	9.96	13.99	3.04	3.39	3.56	0.31
	Best	**36.40**	**40.40**	**16.60**	**31.60**	**47.80**	**22.00**	**18.56**	**13.97**	**4.99**
	Single[20]	0.60	6.80	0.00	4.20	5.20	0.00	0.00	0.20	0.00
	Greedy[20]	2.60	23.80	0.00	6.20	28.60	1.00	0.00	1.40	0.00
	Single	5.20	40.00	5.20	4.80	56.60	4.80	0.00	6.79	0.00
	Greedy	8.40	48.40	7.00	8.20	64.60	7.40	0.40	16.37	0.00
	Morph	***19.40***	***51.80***	***15.00***	***25.00***	***67.00***	***22.60***	***1.80***	***22.75***	***0.60***
AdaFace Led3D	Avg	9.88	11.74	3.08	9.96	22.74	4.51	3.39	2.58	0.25
	Best	**26.60**	**51.20**	**20.20**	**34.20**	**48.80**	**25.20**	**14.97**	**7.98**	**3.39**
	Single[20]	0.60	6.80	0.00	4.20	0.80	0.00	0.00	0.00	0.00
	Greedy[20]	2.60	15.20	0.00	6.20	14.20	0.00	0.00	0.00	0.00
	Single	5.20	35.80	4.60	4.80	46.80	3.80	0.00	6.99	0.00
	Greedy	8.40	49.80	6.80	8.20	53.40	6.80	0.40	10.78	0.00
	Morph	***19.40***	***55.40***	***17.00***	***25.00***	***60.60***	***20.60***	***1.80***	***13.77***	***0.40***

(b)

FRs	Strategy	Headspace eval (%)			Texas3d dev (%)			Texas3d eval (%)		
		2D	3D	Joint	2D	3D	Joint	2D	3D	Joint
FaceNet IResNet	Avg	3.48	2.90	0.31	0.08	4.31	0.01	0.17	2.80	0.01
	Best	**9.18**	**11.18**	**3.79**	**3.24**	**23.73**	**1.69**	**6.60**	**9.40**	**2.60**
	Single[20]	0.20	0.00	0.00	0.00	0.62	0.00	0.00	1.80	0.00
	Greedy[20]	6.99	1.60	0.20	0.00	0.62	0.00	0.00	1.80	0.00
	Single	5.59	4.39	0.60	0.00	0.00	0.00	0.00	0.00	0.00
	Greedy	14.17	13.17	1.20	0.00	0.46	0.00	0.00	4.40	0.00
	Morph	***20.76***	***18.76***	***4.19***	***0.00***	***0.46***	***0.00***	***0.00***	***4.40***	***0.00***
FaceNet Led3D	Avg	3.48	2.06	0.20	0.08	3.81	0.05	0.17	12.25	0.05
	Best	**11.18**	**11.18**	**2.40**	**8.32**	**20.18**	**7.55**	**11.40**	**8.00**	**5.00**
	Single[20]	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Greedy[20]	6.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Single	5.59	6.99	0.40	0.00	0.00	0.00	0.00	0.00	0.00
	Greedy	14.17	10.18	0.60	0.00	0.62	0.00	0.00	0.00	0.00
	Morph	***20.76***	***12.18***	***2.20***	***0.00***	***0.62***	***0.00***	***0.00***	***0.00***	***0.00***
AdaFace IResNet	Avg	3.75	2.90	0.30	6.18	4.31	0.49	6.40	2.80	0.31
	Best	**16.97**	**11.18**	**4.79**	**31.28**	**18.34**	**9.71**	**22.00**	**9.40**	**4.60**
	Single[20]	0.00	0.00	0.00	8.94	0.62	0.00	4.20	1.80	0.00
	Greedy[20]	0.00	1.60	0.00	9.86	0.62	0.00	4.60	1.80	0.00
	Single	0.00	4.39	0.00	0.31	0.00	0.00	0.20	0.00	0.00
	Greedy	0.00	13.17	0.00	1.69	0.46	0.00	5.80	4.40	0.40
	Morph	***1.60***	***18.76***	***0.20***	***4.01***	***0.46***	***0.00***	***18.20***	***4.40***	***0.60***
AdaFace Led3D	Avg	3.75	2.06	0.18	6.18	3.81	0.68	6.40	12.25	1.50
	Best	**23.75**	**10.98**	**4.39**	**28.51**	**20.18**	**14.79**	**27.00**	**28.20**	**13.60**
	Single[20]	0.00	0.00	0.00	8.94	0.00	0.00	4.20	0.00	0.00
	Greedy[20]	0.00	0.00	0.00	9.86	0.00	0.00	4.60	0.00	0.00
	Single	0.00	6.99	0.00	0.31	0.00	0.00	0.20	0.00	0.00
	Greedy	0.00	10.18	0.00	1.69	0.62	0.00	5.80	0.00	0.00
	Morph	***1.60***	***12.18***	***0.20***	***4.01***	***0.62***	***0.00***	***18.20***	***0.00***	***0.00***

BU-3DFE training set consisting of 1,500 facial data samples to train our master faces. The FR systems used in our experiments were *FaceNet* and fine-tuned *IResNet100* as mentioned above. Notably, the training set for the FR systems(*VGGFace2*, *FaceScape*) was distinct from the training set for the LVE algorithm(*BU-3DFE*).

Table 4: Results for using a selected natural master face to attack face authentication systems for 12 settings. Natural master face was computed using the BU-3DFE training set, and FR systems used were FaceNet and IResNet. The computation setting matched that for our master face generation. The FMR for each attacking setting is shown in column **Natural**. Our best results of the master face morphing attacks are shown in column **Morph** in comparison with such kind of natural master face attack.

(a)

FRs		FaceNet IResNet				FaceNet Led3D			
Strategy		Avg	Best	**Natural**	***Morph***	Avg	Best	**Natural**	***Morph***
BU-3DFE dev (%)	2D	1.09	1.20	**0.00**	***4.40***	1.09	5.20	**0.00**	***4.40***
	3D	9.06	31.60	**39.60**	***51.80***	11.74	2.20	**38.20**	***55.40***
	Joint	0.01	0.80	**0.00**	***4.40***	0.06	2.20	**0.00**	***4.20***
BU-3DFE eval (%)	2D	1.39	10.60	**0.00**	***19.60***	1.39	10.60	**0.00**	***19.60***
	3D	13.99	34.40	**46.20**	***67.00***	22.74	46.80	**39.00**	***60.60***
	Joint	0.35	7.40	**0.00**	***19.20***	0.84	9.40	**0.00**	***18.20***
Headsapce dev (%)	2D	3.89	17.56	**0.20**	***20.96***	3.89	17.56	**0.20**	***20.96***
	3D	3.56	11.78	**3.79**	***22.75***	2.58	10.78	**1.60**	***13.77***
	Joint	0.34	4.19	**0.00**	***4.59***	0.27	4.19	**0.00**	***2.20***
Headspace eval (%)	2D	3.48	9.18	**0.80**	***20.76***	3.48	11.18	**0.80**	***20.76***
	3D	2.90	11.18	**2.40**	***18.76***	2.06	11.18	**1.40**	***12.18***
	Joint	0.31	3.79	**0.00**	***4.19***	0.20	2.40	**0.00**	***2.20***
Texas3D dev (%)	2D	0.08	3.24	**0.00**	***0.00***	0.08	8.32	**0.00**	***0.00***
	3D	4.31	23.73	**0.00**	***0.46***	3.81	20.18	**0.00**	***0.62***
	Joint	0.01	1.69	**0.00**	***0.00***	0.05	7.55	**0.00**	***0.00***
Texas3D eval (%)	2D	0.17	6.60	**0.00**	***0.00***	0.17	11.40	**0.00**	***0.00***
	3D	2.80	9.40	**0.00**	***4.40***	12.25	8.00	**0.00**	***0.00***
	Joint	0.01	2.60	**0.00**	***0.00***	0.05	5.00	**0.00**	***0.00***

(b)

FRs		AdaFace IResNet				AdaFace Led3D			
Strategy		Avg	Best	**Natural**	***Morph***	Avg	Best	**Natural**	***Morph***
BU-3DFE dev (%)	2D	9.88	36.40	**18.00**	***19.40***	9.88	26.60	**18.00**	***19.40***
	3D	9.06	40.40	**39.60**	***51.80***	11.74	51.20	**38.20**	***55.40***
	Joint	1.91	16.60	**9.80**	***15.00***	3.08	20.20	**12.20**	***17.00***
BU-3DFE eval (%)	2D	9.96	31.60	**17.80**	***25.00***	9.96	34.20	**17.80**	***25.00***
	3D	13.99	47.80	**46.20**	***67.00***	22.74	48.80	**39.00**	***60.60***
	Joint	3.04	22.00	**12.00**	***22.60***	4.51	25.20	**10.40**	***20.60***
Headsapce dev (%)	2D	3.39	18.56	**0.00**	***1.80***	3.39	14.97	**0.00**	***1.80***
	3D	3.56	13.97	**3.79**	***22.75***	2.58	7.98	**1.60**	***13.77***
	Joint	0.31	4.99	**0.00**	***0.60***	0.25	3.39	**0.00**	***0.40***
Headspace eval (%)	2D	3.75	16.97	**1.00**	***1.60***	3.75	23.75	**1.00**	***1.60***
	3D	2.90	11.18	**2.40**	***18.76***	2.06	10.98	**1.40**	***12.18***
	Joint	0.30	4.79	**0.00**	***0.20***	0.18	4.39	**0.00**	***0.20***
Texas3D dev (%)	2D	6.18	31.28	**7.70**	***4.01***	6.18	28.51	**7.70**	***4.01***
	3D	4.31	18.34	**0.00**	***0.46***	3.81	20.18	**0.00**	***0.62***
	Joint	0.49	9.71	**0.00**	***0.00***	0.68	14.79	**0.00**	***0.00***
Texas3D eval (%)	2D	6.40	22.00	**6.00**	***18.20***	6.40	27.00	**6.00**	***18.20***
	3D	2.80	9.40	**0.00**	***4.40***	12.25	28.20	**0.00**	***0.00***
	Joint	0.31	4.60	**0.00**	***0.60***	1.50	13.60	**0.00**	***0.00***

To compare our master face generation method with the reconstruction-based baseline, we ran the baseline multiple times using the same FR systems, dataset, and iteration number, each time with a different initialization latent code. We then selected the output with a realistic visual appearance and the highest FMR.

Figure 3 presents the joint FMR (the rate of the master face being falsely matched as the same individual by 2D and 3D FR systems) on the *BU-3DFE* training set, constituting a white-box scenario. As shown by the RGB-D avatars, StyleGAN2 tightly entangled shape, appearance, head pose, and expression attributes, leading to joint adjustments during the optimization process. In contrast, our 3DMM disentangled these attributes, enabling optimization with fewer degrees of freedom and resulting in better FMR results(6.60%) than the baseline result(2.87%).

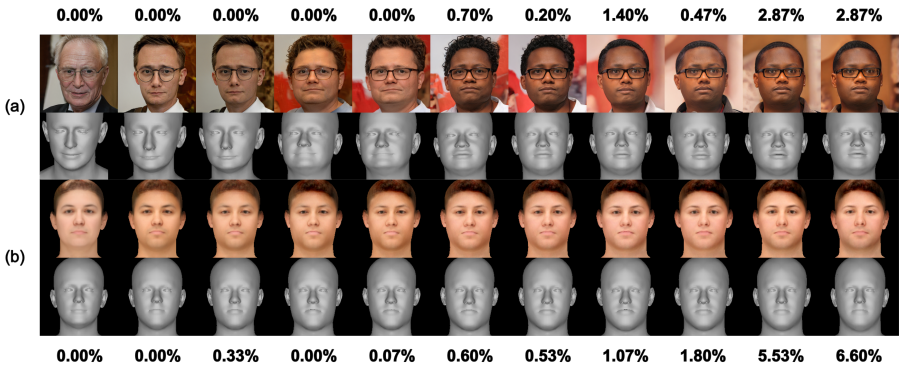


Figure 3: Intermediate faces and their **joint FMRs** on the **training set**. Row (a) was generated by the baseline, and (b) was generated by our method. The leftmost column is the initialized face sample, and the rightmost column is the master face sample obtained after 1,000 iterations.

4.4 Master Face Attack in a Greedy Manner

The comparison described above was conducted for a white-box attack scenario, which is seldom the case in reality. However, Friedlander *et al.*, in their original work, only trained and evaluated the reconstruction-based baseline method on a single 3D facial dataset named *Texas3D*. It is still unclear whether 3D face masters generated from a training set can be successfully generalized to real-world face authentication systems with unknown FR architectures or dataset distributions. In our evaluation, however, we test the generalizability of the master face generated by both the baseline [20] and our methods.

3D master face generalization has proven challenging due to the potential misalignment or conflict between the densest clusters in the feature space distributions of 2D and 3D FR systems. Even in the simplest scenario of a white-box attack, the FMRs on the dev and test sets can be zero when attacking with only a single master face generated from the training set. To address this limitation, we use a greedy strategy, which starts by generating one master face from the training set. Subsequently, individuals that have already been matched are removed, and another face is generated repeatedly. This strategy enables the exploration of more possible clusters of master faces in the feature space of the training set, **with no overlap in individuals matched by each master face**. We use this set, rather than a single master face, to conduct a master face attack.

4.5 Master Face Morphing

While the greedy strategy has proven effective in improving the master face attacks, it comes with a high time cost when generating a larger number of

master faces. The inherent nature of the LVE algorithm dictates that each training run results in only one master face sample. For 1,000 iterations, the baseline method running on a system with an NVIDIA Tesla V100 card takes approximately 14 hours to create a single master face. Our approach reduces this time cost by 1 hour as it omits the StyleGAN generation steps, but the time cost remains relatively high.

However, our approach enables the quick generation of new master faces through interpolation between existing master faces, supported by the interpolation control capabilities of 3DMMs. These morphs effectively preserve both shape and appearance, as shown in Figure 4. By smoothly bridging between the "densest cluster" within which the source/target master face falls, these morphs not only cover a subset of mismatched identities from the source master faces but also introduce new mismatches that are not covered by the input master faces. This enhances the master face attack in terms of efficiency and effectiveness.

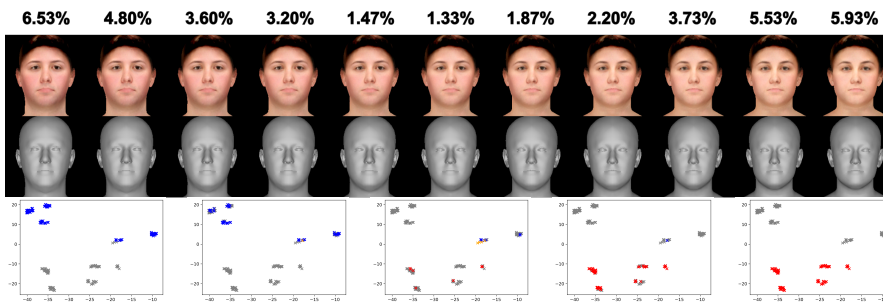


Figure 4: Effect of master face morphing. Columns show generated face samples with their **joint FMR** on top. From left to right, interpolation weight increases from 0.1 to 0.9. T-SNE visualization displays matching results for the left source face, morph with weights 0.2, 0.5, 0.8, and right source face, respectively. Orange points represent newly matched samples that were not covered before.

For instance, the baseline takes around 17 day to generate 30 master faces. Our method, however, allows us to train 3 master faces in 1.5 days only and to create 27 morphs from pairs of 3 master faces in less than a minute. Using these 30 samples in an attack greatly improves the attack success rate, as illustrated by the results in Table 3. In this example, we save 10x time than the baseline. The time saving is even more significant when generating a large number of master faces for brute-force attack.

4.6 Master Face Attack Simulation Analysis

We present the complete results for our comprehensive experimental settings, as illustrated in Figure 2, in Tables 3 and 4.

We conducted evaluations across combinations of four 2D and 3D FR system pairs with three 3D facial datasets, simulating a total of twelve master face attack scenarios. These scenarios include one white-box attack, two black-box attacks, and nine gray-box attack cases, which are shaded respectively from white to dark gray in the table. Notably, for each setting, we present results computed using seven different strategies. For each strategy, we report the results with the highest joint FMRs, along with the corresponding 2D and 3D FMRs.

The results for **Avg** and **Best** were computed using the natural faces belonging to the corresponding targeted face authentication system setting in a **white-box** manner. They were used as references to evaluate whether our **Single, Greedy, and Morph** results can surpass the natural best result in white-box cases.

In Table 3, the **third and fourth rows** for each setting represent the evaluation results for a single master face instance and for a set of three master faces generated greedily with the reconstruction-based baseline, respectively. The **fifth and sixth rows** present the evaluation results for master faces generated with our 3DMM-based method instead. The final row, labeled as **Morph**, highlights our key results, which are computed using the combination of the three master faces generated by the greedy mechanism and their intermediate morphs, resulting in a total of thirty samples used for the attack.

In Table 4, the **Avg**, **Best**, and **Morph** columns are the same as described above while the **Natural** column shows values for the second anchora natural master face attack based on the training setting equivalent to the one used in the generation phase.

The experimental results demonstrate that master faces generated by our method achieve high FMRs across various attack settings. This underscores the effectiveness of our 3D master face attack approach in real-world scenarios. In contrast, while the baseline demonstrates success in attacking individual 2D or 3D FR systems, it fails to target both 2D and 3D FR systems simultaneously. Compared to the natural master face on the test set, our morph attack method achieves significantly better joint FMR in the white-box attack scenario. In gray-box attack scenarios, when the dataset distribution is unknown (e.g., attacks on Headspace and Texas3D), and the FR system architecture is known (i.e., the target FR systems are FaceNet and IResNet, the same architectures used to generate the master face), our method outperforms the natural master face on the Headspace dataset. When the FR system architecture is only partially known, our FMR shows some decline but still remains significantly higher than the average FMR of bona fide samples. Moreover, when the dataset distribution is known (e.g., attacks on BU-3DFE), regardless of whether the FR system architecture is partially known or unknown, our results either surpass or are on par with the FMR of the natural master face. Even in the most difficult black-box attack scenario, our method can attain a

joint FMR higher than the average bona fide face's FMR on Headspace.

We observed that the attack success rate of master faces is constrained by dataset distribution differences, particularly in gray-box or black-box attacks where the target dataset distribution cannot be accurately estimated. The performance gap between HeadSpace and BU-3DFE further supports the conclusion that mismatched dataset distributions can significantly reduce attack success rates on the target dataset.

However, our results still demonstrate the potential threat posed by morphable master faces to the joint 2D and 3D face recognition systems. By integrating research on neural network architecture estimation [58, 44] and dataset distribution inference [54, 5, 8, 30], the difficulty of using master face attacks against face authentication systems can be further reduced, thereby amplifying the associated risks.

4.7 Master Face Reenactment and Presentation Attack

Although most methods rely on static master face samples to attack FR systems, our method enables dynamic facial reenactment by manipulating the pose and expression codes in the FLAME model. Specifically, the FLAME model learns both pose code $\boldsymbol{\varphi}$ and expression code $\boldsymbol{\psi}$ distributions from 4D facial sequences. By sampling expression codes within chosen standard deviations of these learned distributions, we ensure natural facial deformations and can generate a diverse range of realistic expressions. Similarly, we control pose variations by sampling head pose and jaw articulation parameters within appropriate angular ranges, enabling natural head movements and mouth articulations. While baseline methods fail to attack FR systems with liveness detection due to their lack of semantic control over the generated output, our method's high controllability demonstrates significant advantages.

As shown in Figure 5, due to the sensitivity of 2D FR systems to pose variations, the success rate of attacks targeting specific poses may be relatively low. Nonetheless, our results still highlight the potential of utilizing a controllable 3D master face to strengthen presentation attacks against 2D face authentication systems, particularly against systems that require users to exhibit specific facial expressions. However, current active presentation attack detection systems often require users to perform specific facial expressions or movements based on text instructions. While our method enables facial reenactment by manipulating latent variables for expressions and poses, it falls short of addressing such dynamic, real-time interactions. Incorporating a large language model (LLM) agent could be a promising direction for enhancing adaptability and achieving more sophisticated attacks in the future.

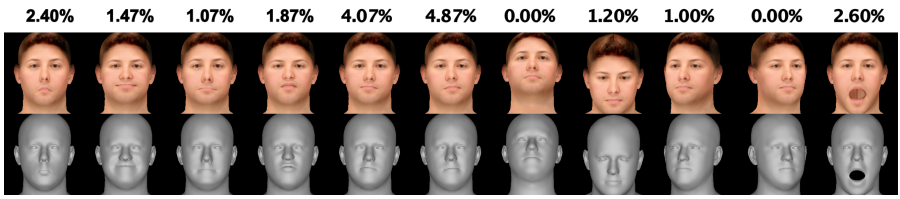


Figure 5: Effect of master face reenactment. Columns show generated face samples with their **joint FMR** on top. The first to sixth columns show variations in the first three principal components of the expression. The others show visualizations of changes w.r.t poses.

4.8 Ablation Study

Attacking 3D FR systems only

We conduct an ablation study to validate our hypothesis that a master face generation method based on 3DMM can better learn from the shape information within the 3D facial dataset, resulting in a higher rate of false matching. In contrast, the baseline method based on 3D face reconstruction has limited abilities to preserve and utilize 3D shape information. This is due to various factors such as optimization within the 2D latent variable space, unstable latent variable initialization, and errors in the 3D face reconstruction process. In this experiment, we used only the FMR computed from the 3D FR system as the objective function for the optimizer. The training curve obtained, shown in Figure 6, demonstrates that the 3DMM-based method is better at learning crucial features for a 3D master face, resulting in higher 3D FMRs.

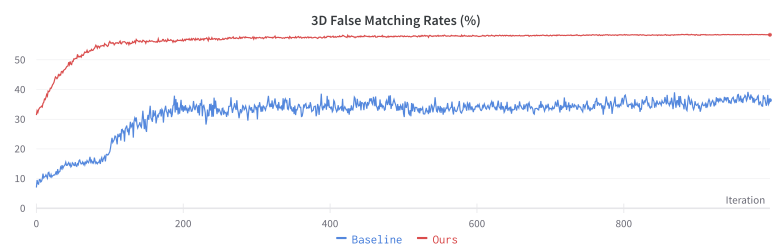


Figure 6: Training curves for two master faces, one generated using the baseline method and one generated using our method, guided only by feedback from the 3D FR system. Our method shows better initialization and higher FMRs.

Attacking 2D FR systems only

One of the criticisms of 3DMM is its tendency to blur textures. To assess whether this affects our method's 2D FMR, we used feedback from only the 2D FR system to optimize the master face. The design aim was to compare

the final 2D FMRs between our 3DMM-based method and the reconstruction-based baseline. Since FaceNet performs exceptionally well, to avoid having the CMA-ES optimizer fail due to an initially close-to-zero FMR, we used a relatively low threshold starting point and gradually increased its matching threshold every 200 iterations. We found that the 3DMM-based method also outperformed the baseline method in terms of 2D FMRs, as shown in Figure 7. We hypothesize that the 2D FR results are affected by pose and expression. In our training dataset, all facial data corresponded to a frontal pose, which aligns with the use case in real life. This pose is modeled with the fixed pose parameters of our 3DMM-base method. In contrast, in StyleGAN, facial pose and expression are uncontrollable during training, which may degrade the final 2D error matching rates.

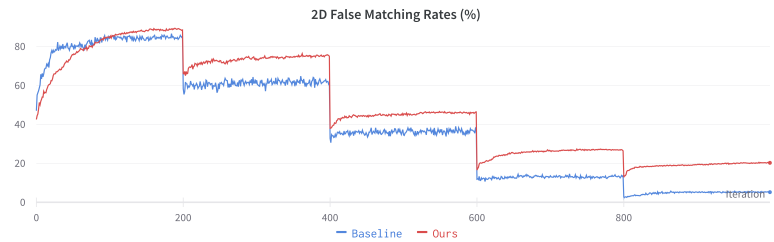
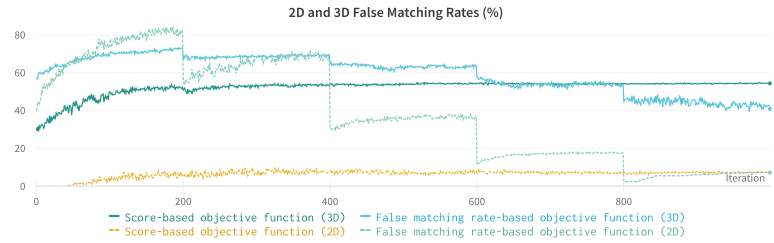


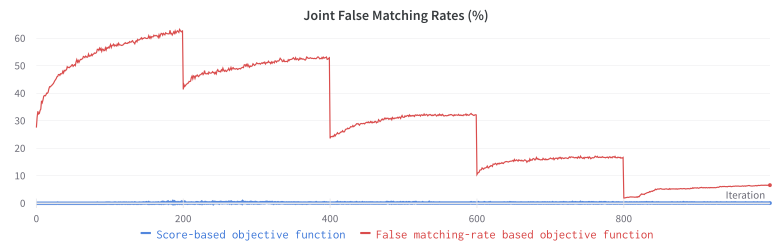
Figure 7: Training curves for two master faces, one generated using the baseline method and one generated using our method, guided only by feedback from the 2D FR system. Our method exhibited better robustness when the threshold was increased.

Objective Function Selection for CMA-ES Solver

As described in Section 3.2, after the CMA-ES solver samples and provides possible candidate answers, the fitness scores corresponding to these answers are returned to CMA-ES to aid it in further optimization. The score function thus plays a decisive role in the efficiency of optimization. Previous research on master faces has proposed two approaches to optimize based on similarity scores or FMRs. We leverage the FMR-based objective function for its better performance when attacking joint FR systems. As shown in Figure 8a, when we optimize with a single-modal FR system, both objective functions yield similar results and efficiency. However, for cross-modal optimization, using a score-based objective function causes the optimizer to focus on improving individual performance while ignoring the need to find a "cross-modal space." As a result, the FMR of the master face generated by the score-based function is much lower than the one generated by the FMR-based function, as shown in Figure 8b.



(a)



(b)

Figure 8: Training curves with score-based and false matching rate-based objective functions. Figure 8a shows training curves for four master faces, two generated using different objective functions, guided only by feedback from the 2D FR system, and the other two guided only by the 3D FR system. As shown in Figure 8a, these two different objective functions achieved similar FMRs in the 2D-only scenario. For 3D FMR, the score-based function performed better. However, Figure 8b shows that the score-based function failed to jointly attack the 2D and 3D FR systems. After 1,000 iterations, the FMR-based function has an FMR of 6.6%, while the score-based function holds only 0.06%.

3D Morphable Face Model Regularization

One crucial point to note in the implementation of our method is that with 3DMM, its parameters are assumed to follow a Gaussian distribution with a mean of zero. This assumption is violated during the optimization process of the CMA-ES solver, and the objective function we use leads the optimizer to focus only on improving the FMR without regard for whether the generated shapes are anatomically plausible. To address this problem, we introduce a regularization term into the objective function to penalize shape codes that deviate too far from the zero vector, as depicted in Section 3.2.

However, this regularization term to some extent limits the ability of the CMA-ES solver to optimize shape variables, as shown in Figure 9. Therefore, choosing an appropriate weight is important to balance between a high FMR and an anatomically plausible shape.

Shape images of two master faces generated with the same settings except for the weight for the regularization term are shown in Figure 10.

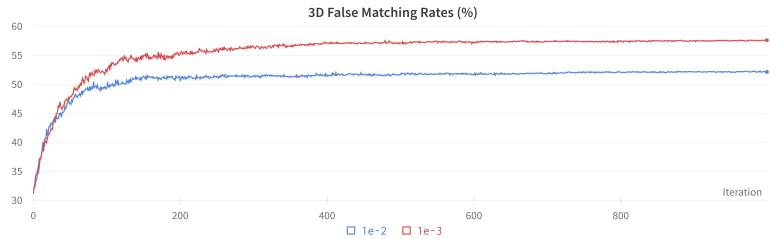
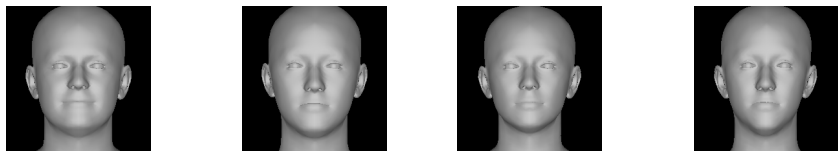


Figure 9: Training curves for two master faces generated using our methods with different weights of regularization term, guided only by feedback from the 3D FR system. It is evident that the larger regularization term limited the ability to further craft the shape code, resulting in a lower 3D FMR.



(a) (b) (c) (d)

Figure 10: Shape images generated with different settings for reference. That in Figure 10a is from the initialized face with a zero vector as shape code. That in Figure 10c is from the master face generated using the FMR-based objective function and a larger weight of regularization term (1e-2). That in Figure 10b is from the master face generated using the score-based objective function and a smaller weight of regularization term (1e-3). That in Figure 10d is **from the master face with the best 3D FMR**, generated using the FMR-based objective function and a smaller weight of regularization term (1e-3).

5 Defense Against 3D Master Face Attack

Our research has identified significant concerns regarding the vulnerability of 2D and 3D FR systems against controllable 3D master face attacks. Despite extensive research on security for 2D FR systems in the past decade, these findings do not seamlessly extend to 3D FR systems. For example, presentation attack detection [29, 16] and deepfake attack detection [47, 31, 38, 66, 28] can be readily adapted to counter physical and digital 2D morphing face attacks, respectively. However, similar work has not yet been done for 3D FR systems, which highlights the urgent need for research and development in this area. Another concern is the generalizability of detectors for both 2D and 3D FR systems, which remains an active research topic in biometric security.

6 Limitation

Our 3DMM-based method for 3D master face generation has below limitations: 1) Most 3DMM models have limited texture resolution and therefore cannot generate high-fidelity 2D faces that would convincingly deceive human eyes. This means that if the 2D FMR can be increased by improving the texture quality, it may be possible to increase the joint FMR. 2) The LVE algorithm is less efficient as it can optimize only one latent vector at a time. 3) Black-box master face attacks do not succeed when the distribution of the training dataset is dissimilar to that of the attack dataset.

Future work includes exploring potential countermeasures against 3D controllable and morphable master face attacks as our evaluation results revealed that these attacks are significant threats. It also includes enhancing the quality of 2D facial appearance generated by 3DMM to further improve joint FMRs, or utilizing the differentiable properties of 3DMM to learn distributions of master faces, rather than individual latent vectors, to reduce the time cost of the master face generation.

7 Discussion and Conclusion

Existing methods cannot be effectively applied to real-world attack scenarios due to the following limitations: **1) Ill-posed 3D face reconstruction from a single 2D image:** Current approaches that generate 2D master faces and then reconstruct 3D master faces from them suffer from significant information loss. **2) High computational cost:** Existing methods are extremely costly, requiring weeks of computation to generate a large number of master faces for achieving relatively effective attacks in a greedy manner. **3) Lack of flexibility and controllability**: Current methods lack the adaptability needed to bypass face authentication systems equipped with liveness detection techniques, such as active presentation attack detection systems, which demand dynamic user interactions such as facial expressions or specific movements.

We propose, for the first time, a method to generate **deformable, controllable, and morphable** master faces using a 3D Morphable Face Model, allowing the production of master faces capable of effectively compromising both 2D and 3D face recognition systems in real-world scenarios. Our approach directly generates and optimizes 3D faces without a lossy reconstruction procedure to improve the FMR. We further generate a large number of master face morphs that also possess master face capability to improve the generalizability of the master face when performing gray-box and black-box attacks. Compared to the reconstruction-based baseline [20], our method is over ten times faster in generating more master faces. Furthermore, the controlla-

bility of our master face represents a significant advancement in overcoming limitations posed by liveness detection technologies.

We employ multiple 3D face datasets and 2D/3D face recognition systems to simulate real-world gray-box/black-box attacks. As the first study to evaluate master face attacks across various attack scenarios, our greedy generation and morph creation method demonstrated the potential to compromise face authentication systems even when the architectures of the face recognition systems or face gallery distributions are unknown. In addition, by using disentangled parameters, we can easily change the facial expressions and poses of the master faces while retaining the ability to achieve false matching. Our findings have revealed significant security risks associated with controllable and morphable master face attacks and emphasize the need for research on defense strategies.

In conclusion, we propose a novel master face attack method that leverages 3D morphable face models for generating morphable and controllable master faces and evaluate its performance on various attacking scenarios simulating real-world gray-box and black-box attacks. Our results demonstrate the potential threat posed by such master face attacks to existing active face authentication systems, highlighting the necessity for further research into effective defense mechanisms.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grants JP21H04907 and JP24H00732, by JST CREST Grant JPMJCR20D3 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, and by JST K Program Grant JPMJKP24C2 Japan.

References

[1] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows", *ACM Transactions on Graphics (ToG)*, 40(3), 2021, 1–21.

[2] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces", in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, 187–94.

[3] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross, "Deepmasterprints: Generating masterprints for dictionary attacks via latent variable evolution", in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2018, 1–9.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age", in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, IEEE, 2018, 67–74.

[5] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models", in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, 5253–70.

[6] K. Carta, A. Huynh, S. Mouille, N. El Mrabet, C. Barral, and S. Brangoulo, "How video injection attacks can even challenge state-of-the-art Face Presentation Attack Detection Systems", in *Proceedings IMCIC-International Multi-Conference on Complexity, Informatics and Cybernetics*, 2023, 105–12.

[7] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, "Meshgan: Non-linear 3d morphable models of faces", *arXiv preprint arXiv:1903.10384*, 2019.

[8] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks", in *International conference on machine learning*, PMLR, 2021, 1964–74.

[9] V. Ciancaglini, C. Gibson, D. Sancho, O. McCarthy, M. Eira, P. Amann, and A. Klayn, "Malicious uses and abuses of artificial intelligence", *Trend Micro Research*, 2020, 4–79.

[10] H. Dai, N. Pears, W. Smith, and C. Duncan, "Statistical modeling of craniofacial shape and texture", *International Journal of Computer Vision*, 128, 2020, 547–71.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 4690–9.

[12] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, 0–0.

[13] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition", in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, 9415–22.

[14] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, *et al.*, "3d morphable face modelspast, present, and future", *ACM Transactions on Graphics (ToG)*, 39(5), 2020, 1–38.

[15] N. Erdogmus and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect", in *2013 IEEE sixth international conference on biometrics: theory, applications and systems (BTAS)*, IEEE, 2013, 1–6.

[16] European Union Agency for Cybersecurity (ENISA), "Remote Identity Proofing - Attacks & Countermeasures", *tech. rep.*, Accessed: Feb. 3, 2025, European Union Agency for Cybersecurity (ENISA), 2023, https://www.enisa.europa.eu/publications/remote-identity-proofing-attacks-countermeasures.

[17] Europol, "Facing Reality? Law Enforcement and the Challenge of Deepfakes", 2022.

[18] H. Felouat, H. H. Nguyen, T.-N. Le, J. Yamagishi, and I. Echizen, "eKYC-DF: A Large-Scale Deepfake Dataset for Developing and Evaluating eKYC Systems", *IEEE Access*, 2024.

[19] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images", *ACM Transactions on Graphics (ToG)*, 40(4), 2021, 1–13.

[20] T. Friedlander, R. Shmelkin, and L. Wolf, "Generating 2D and 3D Master Faces for Dictionary Attacks with a Network-Assisted Latent Space Evolution", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022.

[21] S. Giebenhain, T. Kirschstein, M. Georgopoulos, M. Rünz, L. Agapito, and M. NieSSner, "Learning neural parametric head models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 21003–12.

[22] S. Giebenhain, T. Kirschstein, M. Georgopoulos, M. Rünz, L. Agapito, and M. NieSSner, "Mononphm: Dynamic head reconstruction from monocular videos", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 10747–58.

[23] S. Z. Gilani and A. Mian, "Learning from millions of 3D scans for large-scale 3D face recognition", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 1896–905.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Advances in neural information processing systems*, 27, 2014.

[25] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik, "Texas 3D face recognition database", in *2010 IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI)*, IEEE, 2010, 97–100.

[26] N. Hansen, Y. Akimoto, and P. Baudis, "CMA-ES/pycma on Github", Zenodo, DOI:10.5281/zenodo.2559634, February 2019, DOI: 10.5281/zenodo.2559634, https://doi.org/10.5281/zenodo.2559634.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–8.

[28] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2), 2024, e1520.

[29] J. Hernandez-Ortega, J. Fierrez, A. Morales, and J. Galbally, "Introduction to presentation attack detection in face biometrics and recent advances", *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, 2023, 203–30.

[30] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey", *ACM Computing Surveys (CSUR)*, 54(11s), 2022, 1–37.

[31] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network", *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3), 2021, 1089–102.

[32] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments", in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[33] 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*, 2019, 4401–10.

[34] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, 8110–9.

[35] D. Kim, M. Hernandez, J. Choi, and G. Medioni, "Deep 3D face identification", in *2017 IEEE international joint conference on biometrics (IJCB)*, IEEE, 2017, 133–42.

[36] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality Adaptive Margin for Face Recognition", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[37] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans", *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017, 194:1–194:17, https://doi.org/10.1145/3130800.3130813.

[38] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, "FAMM: facial muscle motions for detecting compressed deepfake videos over social networks", *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12), 2023, 7236–51.

[39] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 212–20.

[40] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang, "Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 5773–82.

[41] H. H. Nguyen, S. Marcel, J. Yamagishi, and I. Echizen, "Master face attacks on face recognition systems", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3), 2022, 398–411.

[42] H. H. Nguyen, J. Yamagishi, I. Echizen, and S. Marcel, "Generating master faces for use in performing wolf attacks on face recognition systems", in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2020, 1–10.

[43] G. Pan, Z. Wu, and L. Sun, "Liveness detection for face recognition", *Recent advances in face recognition*, 109, 2008, 124.

[44] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning", in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, 506–19.

[45] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge", in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1, IEEE, 2005, 947–54.

[46] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders", in *Proceedings of the European conference on computer vision (ECCV)*, 2018, 704–20.

[47] C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, *Handbook of digital face manipulation and detection: from DeepFakes to morphing attacks*, Springer Nature, 2022.

[48] S. Sanyal, T. Bolkart, H. Feng, and M. Black, "Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision", in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019, 7763–72.

[49] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis", in *Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers 1*, Springer, 2008, 47–56.

[50] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 815–23.

[51] G. Shamai, R. Slossberg, and R. Kimmel, "Synthesizing facial photometries and corresponding geometries using generative adversarial networks", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s), 2019, 1–24.

[52] Z. Shao, Z. Wang, Z. Li, D. Wang, X. Lin, Y. Zhang, M. Fan, and Z. Wang, "Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 1606–16.

[53] R. Shmelkin, T. Friedlander, and L. Wolf, "Generating master faces for dictionary attacks with a network-assisted latent space evolution", in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, IEEE, 2021, 1–8.

[54] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models", in *2017 IEEE symposium on security and privacy (SP)*, IEEE, 2017, 3–18.

[55] R. Slossberg, G. Shamai, and R. Kimmel, "High quality facial surface and texture synthesis via generative adversarial networks", in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, 0–0.

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 1–9.

[57] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, 1701–8.

[58] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}", in *25th USENIX security symposium (USENIX Security 16)*, 2016, 601–18.

[59] M. Une, A. Otsuka, and H. Imai, "Wolf attack probability: A new security measure in biometric authentication systems", in *ICB*, 2007, 396–406.

[60] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, 5265–74.

[61] Y. Xu, L. Wang, Z. Zheng, Z. Su, and Y. Liu, "3d gaussian parametric head model", in *European Conference on Computer Vision*, Springer, 2024, 129–47.

[62] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 601–10.

[63] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch", *arXiv preprint arXiv:1411.7923*, 2014.

[64] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research", in *7th international conference on automatic face and gesture recognition (FGR06)*, IEEE, 2006, 211–6.

[65] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation", in *Proceedings of the European conference on computer vision (ECCV)*, 2018, 325–41.

[66] D. Zhang, J. Chen, X. Liao, F. Li, J. Chen, and G. Yang, "Face forgery detection via multi-feature fusion and local enhancement", *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[67] J. Zhang, D. Huang, Y. Wang, and J. Sun, "Lock3DFace: A large-scale database of low-cost kinect 3d faces", in *2016 International Conference on Biometrics (ICB)*, IEEE, 2016, 1–8.

[68] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks", *IEEE signal processing letters*, 23(10), 2016, 1499–503.

[69] M. Zheng, H. Yang, D. Huang, and L. Chen, "Imface: A nonlinear 3d morphable face model with implicit neural representations", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, 20343–52.

[70] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou, "Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, 1097–106.