Closed-Box 3-D Face Reconstruction Attack on Face Recognition From a Single Image

Shizong Yan, Huixiang Wen, Shan Chang, *Member, IEEE*, Hongzi Zhu, *Senior Member, IEEE*, and Luo Zhou

***Abstract*—The 3-D face recognition systems are frequently susceptible to spoofing attacks, with 3-D face presentation attacks being particularly notorious. Attackers commonly exploit 3-D scanning and printing techniques to generate masks of target individuals, a method proven successful in various real-world scenarios. A defining characteristic of these attacks involves acquiring 3-D face models via 3-D scanning, a process that is notably more expensive and cumbersome compared to obtaining 2-D photographs. In this work, we introduce the depth recovery attack method (DREAM), a novel method for recovering 3-D face models from a single 2-D image. Specifically, our approach adopts a closed-box strategy, reconstructing sufficient depth information to compromise target recognition models—such as face identification and authentication systems—by merely accessing their output and the corresponding RGB photograph. Our key insight is that achieving successful attacks does not necessitate restoring the precise ground-truth depth values; instead, it only requires recovering the essential features that are salient to the target model's decision-making process. We evaluate DREAM's effectiveness using four public 3-D face datasets. Experimental results indicate that DREAM achieves a 94% success rate on face authentication models, even in cross-dataset testing. For face identification models, the success rate is 36%. Building upon DREAM, we further propose DREAM-3D, which leverages a 3-D GAN to reconstruct depth images for deceiving 3-D face recognition systems. In addition, we evaluate DREAM-3D's effectiveness on two datasets. Experimental results indicate that DREAM-3D achieves attack success rates (ASRs) exceeding 90% and approximately 50% against different models.**

***Index Terms*—3-D face recognition, closed-box attack, RGB-depth (RGB-D) images, twin deep networks.**

I. Introduction

FACE recognition, akin to other biometric recognition such as fingerprint and iris recognition, leverages inherent physiological traits of the human body to perform identity authentication (verification) or identification. Among them, face recognition, due to its convenience and noncontact nature, has been extensively applied in fields such as public security, payment, and access control. These systems typically require prior enrollment before recognition can occur. During the enrollment phase, a user's face images are captured and subsequently transformed into an embedding, which is ultimately stored in the system as a template for future comparison. In the recognition phase, the input face data are systematically compared against all preregistered templates in the database to ascertain whether a match exists. If a correspondence is found, the individual is authenticated and recognized as authorized. Although 2-D face recognition technologies have been extensively adopted, their inherent lack of spatial information or depth leads to incomplete face representations. Consequently, 2-D face recognition systems frequently exhibit reduced accuracy under suboptimal lighting conditions or when subject to variations in face pose and orientation. In recent years, with the significant reduction in the cost of 3-D cameras, there has been a notable shift from 2-D to 3-D face recognition technologies.

There have been a variety of spoofing or adversarial attacks against 3-D face recognition systems. Among these, 3-D presentation attacks have been extensively investigated. Such attacks involve leveraging 3-D scanning and printing technologies to create life-sized 3-D replicas of a victim's face [1], with the intention of deceiving 3-D face recognition systems. However, constructing accurate 3-D face models typically requires specialized 3-D scanning equipment, which can be prohibitively expensive (for instance, a portable optical 3-D scanner such as the GOM TRITOP is priced close to United States dollar (USD) 10 000 [2]). In comparison, acquiring 2-D face images is significantly more cost-effective, as these can be easily obtained from sources such as video conference or surveillance camera. Motivated by this, this work seeks to explore whether it is feasible to construct a 3-D face model from captured 2-D images for the purpose of circumventing 3-D face recognition systems.

A straightforward approach to carrying out 3-D face spoofing attacks using 2-D face images is to leverage 3D-from-2D reconstruction techniques to accurately recover the corresponding 3-D face structure. These techniques can be broadly categorized into three main strategies: *photometric stereo* [3], *statistical model fitting* [4], and *deep learning* [5]. Photometric stereo involves estimating the local surface orientations of the face by analyzing a series of images of the subject

Received 7 July 2025; revised 1 September 2025; accepted 1 October 2025. Date of publication 6 October 2025; date of current version 8 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62472083, Grant 62432008, and Grant 62502010; and in part by the AI-Enhanced Research Program of Shanghai Municipal Education Commission under Grant SMEC-AI-DHUZ-01. *(Corresponding author: Shan Chang.)*

Shizong Yan, Shan Chang, and Luo Zhou are with the School of Computer Science and Technology, Donghua University, Shanghai 201620, China (e-mail: 1222040@mail.dhu.edu.cn; changshan@dhu.edu.cn; luozhou@mail.dhu.edu.cn).

Huixiang Wen is with the School of Medical Imaging, Bengbu Medical University, Bengbu 233030, China (e-mail: huixiangwen@bbmu.edu.cn).

Hongzi Zhu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: hongzi@cs.sjtu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2025.3618402

captured from the same viewpoint but under varying lighting conditions; these local normals are then integrated to reconstruct the overall face geometry. Statistical model fitting, on the other hand, relies on a large dataset of 3-D face scans from diverse individuals to build a statistical 3-D face model. This model is subsequently adjusted based on the input 2-D face image. However, since the adaptation is controlled by a limited number of parameters, it often fails to capture fine-grained geometric details, which can compromise the attack's effectiveness. Deep neural network (DNN) approaches hold promise by learning a mapping from 2-D images to 3-D shape; some methods train DNNs to directly predict 3-D face models from 2-D inputs though such training requires a substantial number of 3-D face scans, which are not always readily available [6]. Other techniques attempt to infer scene depth from images [7], [8], but these typically yield only coarse depth maps rather than the detailed local geometries necessary for successful face spoofing. Consequently, if an adversary merely possesses a single 2-D image of a target and access to a small dataset of 3-D faces, there is currently no effective method to mount a successful spoofing attack against 3-D face recognition systems.

We introduce a novel 3-D face spoofing attack, termed the depth recovery attack method (DREAM). In our approach, an adversary possessing only a single 2-D face image of a target individual is able to generate a 3-D face, which can bypass the target face recognition system successfully. For instance, a potential attack scenario involves an attacker attempting to unlock a victim's mobile device within the permitted number of consecutive authentication attempts during the victim's temporary absence. The DREAM attack is motivated by the observation that face recognition systems typically rely on DNNs to extract and compare salient features between stored template faces and probe faces. This insight leads to two important considerations.

1) DREAM does not require the precise reconstruction of the entire 3-D face geometry but rather focuses on synthesizing a face that is sufficient to deceive the target face recognition system, emphasizing the key features most relevant to recognition performance.
2) It is feasible to infer and replicate the discriminative face features utilized by the recognition model through systematic interactions with the system.

Executing such an attack presents significant challenges for two primary reasons. First, the attacker's interaction with the target commercial device is strictly limited to a closed-box setting, with only a small number of permitted consecutive authentication attempts (e.g., *5*). This constraint severely restricts the amount of information about the device's 3-D face templates that can be inferred. Second, the closed-box nature of access prevents the attacker from directly obtaining the specific face features used by the system for matching, thereby complicating the identification and prioritization of local geometric features during the reconstruction process. To address these challenges, DREAM is designed as a two-stage approach comprising *zero-knowledge pretraining of a depth generator* and *depth fine-tuning using the target model*. In this framework, the attacker initially generates a coarse 3-D face model through pretraining, which is then iteratively refined by querying the target device. This two-step process enables the attacker to maximize the likelihood of a successful attack within the limited number of allowed authentication attempts.

In the initial stage, a straightforward method for estimating face depth from a 2-D image involves training a generative adversarial network (GAN) [9] using a 3-D face dataset available to the attacker. It is important to note that this dataset can be sourced from publicly accessible repositories and does not need to contain identities registered on the target device. However, relying exclusively on a GAN for depth estimation often results in limited attack efficacy, as the network alone may not capture the critical depth features required for successful spoofing. To address this limitation, we incorporate a pretrained agency face recognition model, registering the 3-D face to this agency system. The attacker can obtain such an agency model through various means, such as downloading similar face recognition software or utilizing a commercial device identical to the target. This agency model serves as an external supervisor, guiding the face depth reconstruction process. In addition, we integrate an attention mechanism within the generator to emphasize essential face features and introduce a dual-contrastive loss function [10] to enhance the generator's capability, particularly under small-sample settings. During the second stage, the attacker employs the 3-D face produced in the first stage to challenge the target device. Two outcomes are possible: the attack is either immediately successful, or it is not. In the latter case, the attacker utilizes feedback from the device—such as similarity scores—to refine the GAN's input (i.e., its random noise vector), thereby generating an updated 3-D face for subsequent attempts. This iterative process continues until either the attack succeeds or the maximum permitted number of queries is exhausted.

We evaluate the effectiveness of DREAM against both 3-D face authentication (1:1) and identification (1:N) scenarios. For the authentication task, we adopt the widely used Siamese network architecture [11] as the target model. For the identification task, we select Led3D [12] and the architecture proposed by Uppal et al. [13] as target models. Our experiments leverage three public 3-D face datasets—Pandora [14], the RGB-depth (RGB-D) Facial Dataset under Pose Variation [15], and the Texas Database [16]—for authentication evaluation, while Lock3DFace [17] is employed for identification evaluation. Comprehensive experimental results demonstrate that DREAM achieves attack success rates (ASRs) of 94.73%, 85.71%, and 88.50% within five attempts against the face authentication model and 36.36% and 89.32% within five attempts against the identification models.

Compared to 2-D GANs, 3-D GANs inherently incorporate substantial 3-D priors, which enables them to exhibit exceptional capabilities in synthesizing 3-D human faces. Consequently, we introduce a 3-D GAN-driven depth reconstruction approach, DREAM-3D, encompassing two distinct phases: *pretraining of the geometry-aware encoder* and *depth domain-adaptation with the target model*. In the first stage, we leverage synthetic data to perform supervised training of a geometry-aware encoder. This encoder is designed to

map an RGB face image to a latent vector within the 3-D GAN's latent space. From this latent vector, an identity-consistent 3-D face can be generated, and a corresponding depth map can be rendered, which is capable of deceiving 3-D face recognition systems. To ensure that the generated depth images can deceive the target system, we employ RGB pixel loss, perceptual loss, and the ID loss from the 3-D face recognition system. Since the depth images rendered by the 3-D GAN might exhibit a domain discrepancy compared to real 3-D face datasets, we introduce a second-stage domain-adaptation network. This network fine-tunes the depth images, enabling them to effectively deceive 3-D face recognition systems.

We evaluate the effectiveness of DREAM-3D against 3-D face authentication (1:1 matching) scenarios. For the authentication task, we modify the widely used RGB face recognition models to accept four-channel RGB-D input, designating them as our target models. We also leverage only the backbones from Led3D [12] and Uppal et al. [13], adapting them through retraining to serve as face authentication models. Our experiments leverage VIPL-MumoFace-2K [18] for training 3-D face recognition models and Texas 3-D and Lock3DFace for evaluation. Comprehensive experimental results demonstrate that DREAM-3D achieves an ASR exceeding 90% against RGB-D recognition models. The ASR is approximately 50% when targeting models that rely solely on depth images.

II. Related Work

This section briefly reviews 3-D face recognition systems and reconstruction, and GANs.

A. 3-D Face Recognition System

Face recognition, as a form of biometric recognition, has seen widespread adoption due to its convenience and non-intrusive nature. It has found numerous applications across domains such as security, commerce, healthcare, and robotics [19], [20], [21]. Current face recognition technologies are generally categorized into 2-D and 3-D approaches based on the modality of input data [22]. Owing to the affordability and accessibility of 2-D imaging devices, the majority of research and commercial solutions have historically focused on 2-D face recognition. Nevertheless, recent advancements in 3-D sensors (e.g., Intel RealSense and ORBBEC Gemini) and computational hardware (e.g., GPUs) have facilitated the increasing integration of 3-D face recognition into daily applications, such as Apple Face ID [23]. The 3-D face data can be represented in several formats, including RGB-D images, point clouds, and meshes. For instance, some studies [13], [15], [24] employ RGB-D image pairs to simultaneously extract RGB and depth features for robust face recognition. Several cloud-based platforms [25], [26] also utilize RGB-D image pairs, typically leveraging depth information for liveness detection while relying on RGB features for identity matching. Other approaches, such as Led3D [12], use solely depth information as input for face recognition. In addition, PointFace [27] directly processes face point clouds to extract discriminative features and measures similarity against template point clouds for recognition purposes.

Compared to their 2-D counterparts, 3-D face recognition systems offer several notable advantages [28], [29], [30]. First, 3-D face data inherently preserve rich geometric information, eliminating the need for projection from the 3-D physical space to a 2-D image plane. This enables the extraction of more discriminative features for recognition tasks. Second, 3-D data are inherently robust to variations in pose, illumination, and facial expression, enhancing the adaptability and reliability of face recognition systems in dynamic, real-world environments. Third, 3-D face recognition provides improved security; by leveraging depth information for liveness detection, these systems are naturally resistant to typical 2-D spoofing techniques such as printing and replay attacks [25], [26], [31].

Face spoofing attacks are generally categorized into 2-D and 3-D attacks based on the techniques employed [32]. The 2-D attack commonly utilizes printed photographs or digital displays to imitate a legitimate user, whereas the 3-D attack often involves the use of fabricated face masks. While basic 2-D spoofing methods are typically ineffective against 3-D face recognition systems, their effectiveness can be enhanced when combined with advanced optical attack. For instance, DepthFake [33] infers the target individual's face depth from their RGB face image and projects specially designed scatter patterns containing depth cues onto the image, thereby imbuing the 2-D photograph with 3-D characteristics suitable for authentication. Experimental results have demonstrated that DepthFake can successfully deceive several commercial SDKs and devices. More sophisticated attacks, such as 3-D spoofing and morphing, usually employ meticulously crafted face masks that either closely replicate the victim's features or are generated using adversarial techniques. For example, Singh and Ramachandra [34] demonstrated that morphing the 3-D features of two individuals can effectively bypass PointNet-based 3-D face recognition systems in digital attack settings. Similarly, Li et al. [35] developed end-to-end attack algorithms that utilize either inherent or auxiliary projectors to generate adversarial illuminations, producing adversarial points at arbitrary locations on 3-D faces. These methods successfully compromised a 3-D face recognition system that utilizes point clouds and depth images in both digital and physical environments.

B. 3-D Face Reconstruction

In recent years, the integration of 3-D data into face analysis and its related applications has garnered significant attention. Although 3-D face data offer a more precise representation of face geometry, and the availability of 3-D sensors has increased, acquiring 3-D face images remains more challenging than capturing 2-D images. Moreover, publicly available 3-D face datasets are relatively limited in scale, both in terms of the number of images and the diversity of subjects. This scarcity has motivated considerable research efforts toward developing methods for reconstructing 3-D face structures from single, uncalibrated 2-D images. However, this 3D-from-2D reconstruction task is inherently ill-posed, necessitating

the incorporation of prior knowledge to constrain the solution space. According to [6], 3-D face reconstruction techniques can be classified into three main categories based on how prior knowledge is incorporated: statistical model fitting, photometric methods, and deep learning approaches. Among these, deep learning-based methods have emerged as the state-of-the-art (SOTA) in recent years, driven by advancements in network architectures and learning algorithms. These methods are capable of recovering intricate face details, including those necessary for face animation. For example, Lei et al. [36] introduced a hierarchical representation network (HRN) that achieves precise and highly detailed 3-D face reconstructions from a single image by disentangling geometry and leveraging hierarchical representations for detailed modeling. Similarly, Feng et al. [37] proposed the first approach capable of regressing both 3-D face shape and animatable details. This was achieved through a novel detail-consistency loss function that effectively isolates individual-specific features from expression-induced wrinkles. The ability to recover high-fidelity face details through 3-D face reconstruction also introduces new security challenges. For instance, attackers can exploit these techniques to fabricate 3-D masks of victims' faces for use in 3-D presentation attacks [38]. In addition, Shahreza and Marcel [39] first applied 3-D face reconstruction to template inversion attacks, enabling the recovery of 3-D faces from 2-D face recognition system templates.

C. GAN

GAN is a class of generative models capable of learning to generate data in both semisupervised and unsupervised settings. A typical GAN framework is composed of two components: a generator and a discriminator. The generator aims to capture the distribution of real data to produce highly realistic synthetic samples, while the discriminator's task is to distinguish between genuine and generated samples. This interaction can be formulated as a two-player minimax game, in which the generator and discriminator iteratively improve their performance until reaching a dynamic equilibrium. Recent years have witnessed substantial progress in GAN research, resulting in numerous variants designed to address the limitations of the original GAN architecture. For example, some works [40], [41] focus on overcoming issues such as mode collapse and gradient vanishing, while others [42], [43], and [44] seek to enhance the fidelity of synthesized images and broaden GAN applications to areas such as text-to-image synthesis, cross-domain translation, and image enhancement. Furthermore, GANs have been employed in security-related topics; Yuan et al. [45] and Khosravy et al. [46] utilized GAN-based priors for model inversion attacks, enabling the reconstruction of high-dimensional data such as face images. In addition, Wang et al. [47] proposed a GAN-based framework for synthesizing dense depth images with detailed textures in indoor scenes by integrating RGB images and depth data. The capacity of GANs to effectively capture the spatial distribution of real-world data allows them to contribute prior knowledge to a wide array of tasks. It is, thus, plausible that the introduction of GAN-derived priors could also benefit 3-D face-related applications.

III. Threat Model

To evaluate the effectiveness of DREAM and the vulnerability of a specific 3-D face recognition system that employs a DNN-based recognition model, it is essential to first establish the threat model that delineates the capabilities and objectives of the adversary.

A. Attack Scenario

As shown in Fig. 1, we consider two attack scenarios.

1) *Face Identification:* The system stores multiple 3-D template faces corresponding to registered users. When a 3-D face is presented to the target identification model, the model computes similarity scores between the input face and each of the stored templates. Based on these similarity scores and a predefined threshold, the system determines whether the input face matches any of the registered templates.
2) *Face Authentication:* The system stores one or more 3-D template faces associated with each registered user. Upon receiving an input face, the model computes the similarity score between the input and template faces, and determines whether the input face corresponds to the same individual as the template based on the calculated similarity.

In this article, we denote 3-D faces in terms of RGB-D image pairs, and these face recognition models use either an RGB-D image pair or a depth image as input.

B. Properties for Adversary

Adversary's Goal: The goal of an attacker is to successfully bypass identification or authentication to gain unauthorized access to the device. For instance, an attacker may covertly acquire the victim's device and attempt to unlock it during brief periods of the victim's absence.

Adversary's Knowledge: It is assumed that the attacker possesses the following information.

1) The attacker holds an RGB photograph of the victim, which can be obtained from videos captured by cameras or downloaded from the Internet, and so on.
2) The attacker possesses a limited auxiliary dataset consisting of 3-D faces, for example, stored as RGB-D images. These data may originate from publicly available datasets or be constructed using 3-D scanning. It is worth highlighting that this auxiliary dataset does not overlap with the target model's training or template data.
3) The attacker operates under a closed-box scenario regarding the target model, meaning that they have no information about the internal architecture or parameters of the recognition model and have access only to its output, specifically the similarity scores.

Adversary's Capability: The attacker is presumed to have the following capabilities.

1) The attacker is capable of introducing arbitrary 3-D information into a 3-D face recognition system, for instance, by means of an optical adversarial attack [35]. For simplicity, we operate under the assumption that the

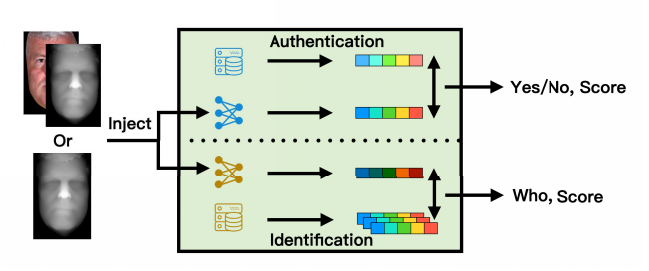

Fig. 1. Attack scenarios: 3-D face authentication versus identification.

attacker can input any RGB-D images into a closed-box 3-D face recognition system.

2) The attacker is permitted to submit a finite number of successive queries to the target device. Following a series of unsuccessful attempts, the device will temporarily lock or power off as a security precaution.
3) The attacker is able to access an agency model that permits an unlimited number of queries. This agency model can be acquired through various means, such as by downloading identical or similar recognition applications, or by utilizing a device that is the same type as the target device.

IV. Design of Dream

DREAM operates in two main stages: *zero-knowledge pretraining of the depth generator* and *depth fine-tuning with the target model*. In the initial stage, the attacker pretrains a GAN-based generator capable of extracting and reconstructing salient depth features from 2-D face images, with the goal of producing depth information sufficient to evade a 3-D face recognition system. Next, the attacker inputs the victim's image into the generator to generate a preliminary 3-D face. In the subsequent stage, this preliminary 3-D reconstruction is iteratively refined by querying the target model multiple times. After each query, the target model response is used to optimize the reconstructed 3-D face, incrementally improving its quality. The process terminates either upon a successful attack or when the maximum allowed number of queries is reached. For clarity and without loss of generality, we focus on the reconstruction of a depth image from its corresponding RGB input, i.e., restoring the depth (D) channel, as RGB-D images can be readily converted into other 3-D data representations such as point clouds.

A. Zero-Knowledge Pretraining of Depth Generator

To facilitate depth image reconstruction, we employ a GAN due to its ability to learn the distribution of depth images from publicly available datasets. Leveraging this prior knowledge enables the generation of plausible face depth images. For training the generator and discriminator, we utilize a public 3-D dataset that shares no overlapping identities with those in the target network's dataset. The pipeline of the offline training process is illustrated in Fig. 2(a).

1) Architecture of Generator: An auxiliary RGB image is incorporated as an additional input to the generator to ensure that the synthesized depth image preserves identity-specific information and remains aligned with the corresponding RGB image, which inherently contains rich identity cues. To further enhance feature extraction, a convolutional block attention module (CBAM) is integrated after extracting the auxiliary information. CBAM allows the network to automatically determine both the salient features and their spatial locations within the image by sequentially applying channel and spatial attention mechanisms, thereby emphasizing relevant characteristics in both dimensions. The combined use of auxiliary RGB input and the attention module enables the GAN to more effectively focus on identity-related features, leveraging prior knowledge of 3-D face structure. As a result, the generated depth images are better suited to mislead the target recognition model.

2) Dual Contrastive Loss: Traditional GAN often encounters issues such as mode collapse and gradient vanishing during training. Several improved approaches [40], [41] have addressed these challenges by modifying the loss function and network architecture. Herein, we replace the conventional GAN loss with the dual contrastive loss introduced in [10] to train the GAN. This loss function enhances the discriminator's ability to identify bona fide versus artificial samples, particularly on small datasets. Simultaneously, it encourages the generator to produce more realistic images by increasing the synthesis quality under the discriminator's guidance.

Adversarial training fundamentally depends on the discriminator's capability to differentiate between authentic and synthetic data. Discriminators, similar to other classification models, can experience overfitting, especially if the available dataset is limited. In our scenario, the auxiliary dataset leveraged by the attacker to train the GAN is also constrained in size, and the training data within each batch may not be fully utilized during the training process. Drawing inspiration from contrastive learning, we treat the ground-truth depth images within the same batch as positive examples and the generated depth images as negative examples. This approach encourages the discriminator to learn more robust and discriminative representations by explicitly comparing similarities and differences between samples. Accordingly, we adopt a dual contrastive loss to replace the original GAN loss, as formulated in the following equations:

$$L^{+}_{\mathrm{GAN}} = \frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{e^{D(\mathrm{depth}_i)}}{e^{D(\mathrm{depth}_i)}+\sum_{j=1}^{N}e^{D(G(\mathrm{rgb}_j,z_j))}}\right] \tag{1}$$

$$L^{-}_{\mathrm{GAN}} = \frac{1}{N}\sum_{j=1}^{N}\left[\log\frac{e^{-D(G(\mathrm{rgb}_j,z_j))}}{e^{-D(G(\mathrm{rgb}_j,z_j))}+\sum_{i=1}^{N}e^{-D(\mathrm{depth}_i)}}\right] \tag{2}$$

where N denotes the batch size, $\mathrm{depth}_i \sim p_{\mathrm{data}}$ represents the ground-truth depth image sampled from the data distribution, $z_j \sim p_{\mathrm{noise}}$ is a noise vector, and $G(\mathrm{rgb}_j, z_j)$ denotes the generated depth image corresponding to the input RGB image and noise. In (1), the objective is to enable the discriminator to identify the real depth image from a batch of synthesized depth maps. Conversely, in (2), the discriminator is trained to distinguish a synthesized depth from a batch of real depth

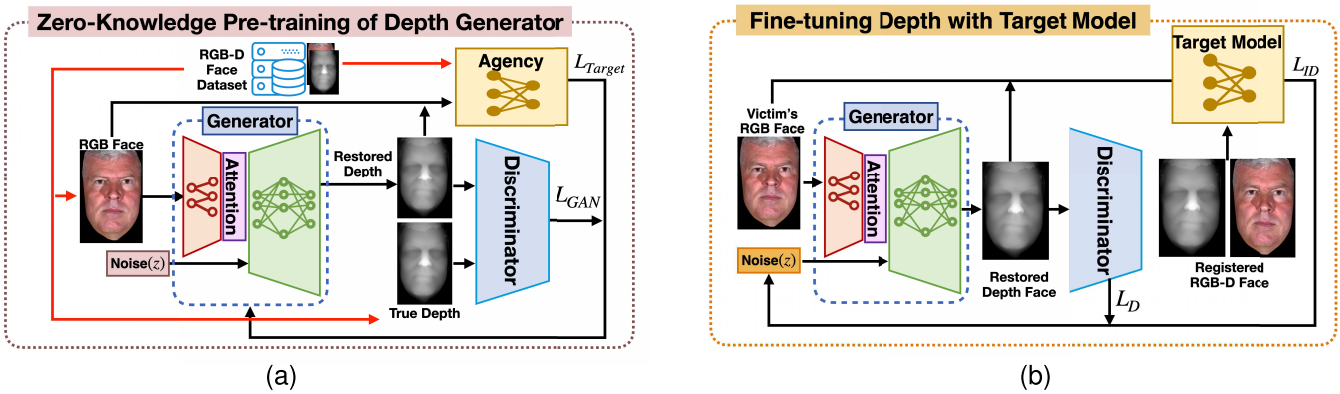

Fig. 2. Overview of DREAM: DREAM consists of two stages. (a) Zero-knowledge pretraining of the depth generator. (b) Fine-tuning depth with target model.

images. Thus, the overall GAN loss function is, thus, formulated as shown in the following equation:

$$\min_G \max_D L(G,D) = L^{+}_{\text{GAN}} + L^{-}_{\text{GAN}}. \tag{3}$$

3) Target Loss: In addition, we incorporate distinct target loss functions tailored to various target models, aiming to make the generated depth more similar to the template face depth in the target system, enough to deceive the target recognition system. Before introducing the target loss, the generator was limited to producing a depth image that appeared visually realistic. However, to effectively deceive the target recognition system, it is essential for the generated depth to be sufficiently close to the template face depth within the feature space of the target system. Therefore, we design and apply specific target loss functions corresponding to different target models.

For 3-D **face verification**, we introduce the target loss as in (4), which is commonly used in face verification model training

$$\begin{aligned}\min_G L^{\text{ver}}_{\text{Target}}(G) = (1-Y)\frac{1}{2}\left(\text{Dis}\left(x, G\left(\text{rgb}, z\right)\right)\right)^2 \\ + (Y)\frac{1}{2}\left\{\max\left(0, m - \text{Dis}\left(x, G\left(\text{rgb}, z\right)\right)\right)\right\}^2\end{aligned} \tag{4}$$

where Y indicates whether the two input image pairs correspond to the same identity, with 0 representing a match and 1 indicating a nonmatch. x denotes the ground-truth depth image or RGB-D image pair, and $G(\text{rgb}, z)$ refers to the generated image pair. $\text{Dis}(x, G(\text{rgb}, z))$ represents the Euclidean distance between the two input image pairs as computed by the target network, and m is a margin. This formulation is designed to guide the GAN in learning which features are critical for successful system verification.

For 3-D **face identification**, to accommodate different identification models, we introduce both the cross-entropy loss and (5) as target loss functions. Typically, 3-D face recognition systems are trained using a softmax layer in conjunction with cross-entropy loss. However, in certain systems, the softmax layer is omitted, and identity recognition is performed by calculating the similarity score (or distance) between feature vectors to determine the most likely identity. Therefore, we propose using (5) as the target loss function in such cases

$$\begin{aligned}&\min_G L^{\text{ide}}_{\text{Target}}(G) \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{Sim}\left(x_i, G\left(\text{rgb}_i, z_i\right)\right)\right)}{\sum_{j=1}^{N}\exp\left(\text{Sim}\left(x_j, G\left(\text{rgb}_i, z_i\right)\right)\right)}\end{aligned} \tag{5}$$

where N represents the batch size, x_i denotes the ith ground-truth depth image or RGB-D image pair within the batch, and $G(\text{rgb}_i, z_i)$ corresponds to the generated images. $\text{Sim}(\cdot,\cdot)$ indicates the similarity between the two input image pairs as computed by the target network. During training, pairs of auxiliary RGB images and real depth images are utilized. The generator receives an RGB image and noise z as input to produce a corresponding depth image. The similarity between the real image and its generated counterpart, as measured by the target model, is treated as a positive example. In contrast, the similarity of real images to other generated images within the same batch is regarded as a negative example. This approach encourages similar data samples to be mapped closely together in the feature space of the target system, while dissimilar samples are pushed further apart.

The final loss function can be written as the following equation:

$$\min_G \max_D L(G,D) = L_{\text{GAN}}(G,D) + \lambda L_{\text{Target}}(G). \tag{6}$$

Overall, in the first stage, we use (6) to train our proposed depth image generator. The L_{GAN} term in this equation is included to address the limitation of a small-scale training dataset, and the L_{Target} term is designed to ensure that the generated depth is sufficient to deceive 3-D face recognition systems.

B. Depth Fine-Tuning With Target Model

As illustrated in Fig. 2(b), after the training, our goal is to identify latent vector $\hat{z}$ that allows the generated depth image or image pair (comprising the auxiliary RGB image and the generated depth image) to attain maximal similarity with the template face as evaluated by the target network. To this end,

we formulate the optimization problem in (8) to derive the optimal vector $\hat{z}$

$$L_D(z) = -\log\left(D\left(G\left(\text{rgb}, z\right)\right)\right) \tag{7}$$

$$\hat{z} = \arg\min_z L_D(z) + \alpha L_{\text{ID}}(z) \tag{8}$$

where $L_D(z)$ serves to constrain atypical face features, while the target loss $L_{\text{ID}}(z)$ drives the generated depth images toward optimal resemblance to the template face as evaluated by the target network.

For 3-D face verification, we use $L_{\text{ID}}^{\text{ver}}$ [see (9)] as the L_{ID}. Since the target model in face verification outputs the distance $\text{Dis}(x, G(\text{rgb}, z))$ between the template and the input, the attacker's objective is simply to minimize this distance to approach the template as closely as possible

$$L_{\text{ID}}^{\text{ver}}(z) = (1-Y)\frac{1}{2}\left(\text{Dis}\left(x, G\left(\text{rgb}, z\right)\right)\right)^2 + (Y)\frac{1}{2}\left\{\max\left(0, m - \text{Dis}\left(x, G\left(\text{rgb}, z\right)\right)\right)\right\}^2. \tag{9}$$

For 3-D face identification, we use cross-entropy loss and cosine embedding loss [see (10)] as L_{ID} for different identification models

$$L_{\text{ID}}^{\text{ide}}(z) = \begin{cases} 1 - \cos\left(x, G\left(\text{rgb}, z\right)\right), & \text{label} = 1 \\ \max\left(0, \cos\left(x, G\left(\text{rgb}, z\right)\right) - m\right), & \text{label} = -1. \end{cases} \tag{10}$$

If the target system identifies the depth image recovered by DREAM or the image pair as the intended identity, the label is assigned as 1; otherwise, it is set to −1. The image pair denotes the auxiliary RGB image and the depth image recovered by DREAM. Here, $\cos(x, G(\text{rgb}, z)$ denotes the cosine similarity, which is the output of the target model. DREAM iteratively optimizes the losses until either a successful attack is achieved or the maximum number of attempts is reached.

In this stage, the constraint is that we can only obtain the target system's output and are limited to five queries. Using the generator and discriminator trained in the previous stage, we combine them with queries to the target system to find the optimal $\hat{z}$.

V. Experiments of DREAM

In this section, we will first detail the datasets used and the implementation details. Then, we will present the experimental results to demonstrate the effectiveness of DREAM against different FR systems under the closed-box scenario.

A. Datasets

We use four datasets, samples shown in Fig. 3, to evaluate DREAM.

1) *Pandora* [14] utilizes the Microsoft Kinect One sensor to collect depth data, comprising over 250 000 RGB-D images from 20 subjects (ten males and ten females).
2) *RGB-D Facial Dataset under Pose Variation* [15] utilizes the PrimeSense camera to capture RGB-D images in a consistent indoor environment. It contains over 24k RGB-D images from 952 individuals. The dataset consists exclusively of young Asian subjects, with a gender distribution of 30% male and 70% female. In this article, this dataset is referred to as USTC.
3) *Texas 3-D* [16] is acquired with a stereo imaging system developed by 3Q Technologies, providing x, y, and z measurements at a high resolution of 0.32 mm. Each session simultaneously captured color and range images, resulting in 1149 image pairs from 105 adult participants.
4) *Lock3DFace* [17] is collected using the second-generation Kinect sensor and comprises a total of 5671 RGB-D face video clips from 509 individuals. It includes diverse variations in expression, pose, time, and occlusion, with 377 male and 122 female subjects.

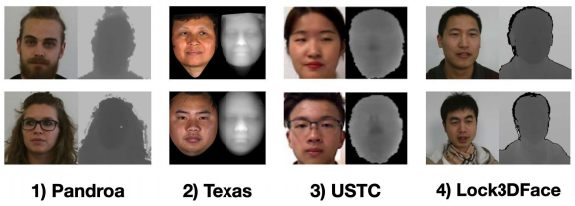

Fig. 3. RGB-D samples from datasets: Pandora, Texas 3-D, USTC, and Lock3Dface.

1) Preprocessing: For the first three datasets, images are centrally cropped and resized to 64×64 pixels. Of these, 50% are allocated for training the face authentication network, 40% for training the DREAM framework, and 10% is reserved for evaluation. For the Lock3DFace dataset, images are similarly cropped at the center and resized to 224×224 pixels. Images from 340 and 100 individuals are used to train the face identification network and the DREAM framework, respectively, while the remaining images from 69 subjects are designated for attack evaluation. It is ensured that the data used to train the target network have no shared images with those used to train the DREAM.

B. Target Model Implementation

We employ three distinct 3-D face recognition models to evaluate the effectiveness of DREAM.

1) Face Authentication: We use a Siamese network [48] implemented in PyTorch and train the RGB-D face authentication model with a contrastive loss function as defined in (11), employing the Adam optimizer lr = 0.001 and a batch size of 32

$$L(W, Y, X_1, X_2) = (1-Y)\frac{1}{2}\left(F_W(X_1, X_2)\right)^2 + (Y)\frac{1}{2}\left\{\max\left(0, m - F_W(X_1, X_2)\right)\right\}^2 \tag{11}$$

where X_1 and X_2 are the two inputs (RGB-D image pair) to the authentication network and F and W are the parameters of the authentication network. The output F_W corresponds to the Euclidean distance between the feature representations of the two inputs as computed by the face authentication network, and m is the margin.

We evaluate the authentication model using the false acceptance rate (FAR) and the false rejection rate (FRR), selecting

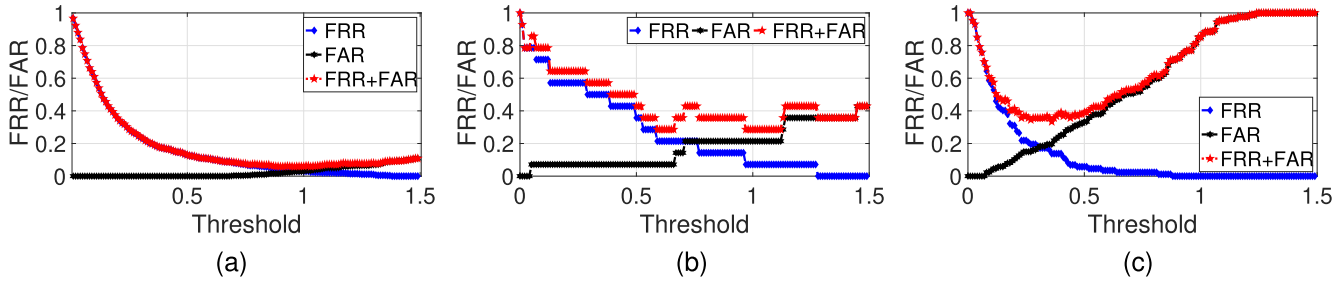

Fig. 4. Relationship between FAR/FRR and threshold on (a) Pandora, (b) Texas, and (c) USTC datasets.

TABLE I
THRESHOLD SELECTION OF FACE AUTHENTICATION MODEL

	Pandora	Texas	USTC
Threshold	1.00	0.60	0.23
FAR/FRR	0.03/0.03	0.07/0.21	0.14/0.21

TABLE II
PERFORMANCE OF TWO IDENTIFICATION MODELS

Model	Input	Accuracy
Led3D [12]	Depth	80.12%
Uppal [13]	RGB+Depth	89.61%

TABLE III
IMPACT OF λ ON DREAM'S ATTACK PERFORMANCE

	λ=0.8	λ=1.0	λ=1.2	λ=1.5
Pandora	73.17%	81.57%	94.73%	94.73%
Texas	70.52%	78.57%	85.71%	71.42%
USTC	86.20%	88.50%	89.65%	89.65%
Lock3DFace(Led3D)	28.78%	36.36%	31.81%	30.31%
Lock3DFace(Uppal)	89.32%	89.32%	89.32%	89.32%

the decision threshold empirically. The FRR represents the probability of incorrectly classifying images of the same individual as belonging to different individuals, whereas the FAR denotes the probability of incorrectly classifying images of different individuals as belonging to the same individual. As shown in Fig. 4, both FAR and FRR vary with the choice of threshold. The subset of Texas used to test the face authentication model performance contains very few identities and images; this causes the curves to look abnormal. We determine the optimal threshold for each metric, as shown in Table I, corresponding to the point at which the sum of FAR and FRR is minimized.

2) Face Identification: We employ Led3D [12] and the model proposed by Uppal et al. [13] as face identification models. During training, the output of Led3D is a softmax layer, optimized using cross-entropy loss; however, the softmax layer is removed during inference, and cosine similarity is computed between the input sample and all gallery templates, assigning the identity corresponding to the template with the highest similarity. We utilize the pretrained Led3D model to assess the effectiveness of our attack. Uppal's model is implemented following the parameter settings described in [13]. During training, the model produces one main output and two auxiliary outputs, all supervised with softmax and cross-entropy loss, while only the main output is retained for testing. Both Led3D and Uppal are evaluated in terms of rank-one accuracy on the Lock3DFace dataset, with results presented in Table II.

C. DREAM Implementation

1) GAN: The GAN is implemented using PyTorch. The generator and the discriminator are optimized with the same Adam optimizer ($\beta_1 = 0.5$ and $\beta_2 = 0.999$) and a learning rate(lr) of 0.001. Batch sizes are set to 32 for spoofing identification and 64 for authentication tasks. The choice of λ is further discussed in Section V-D.

2) Attention Block: We implement the CBAM attention module in PyTorch. CBAM comprises two sequentially connected submodules: the channel attention module (CAM) and the spatial attention module (SAM). It sequentially infers a channel attention map and a spatial attention map when processing feature maps. These attention maps are then multiplied back into the original feature maps.

3) Depth Fine-Tuning: In the second stage, we set $\alpha = 100$ and employ the SGD optimizer to update the latent vector z, using (lr = 0.02) and a momentum of 0.9. The vector z is initialized randomly from a standard Gaussian distribution and is optimized for *five* iterations.

4) Comparisons: We employ three 3-D face reconstruction methods, HRN [36], DECA [37], and the Face++ API [49], as comparative baselines. For HRN and DECA, we utilize the publicly available pretrained models to perform face reconstruction and generate corresponding depth images. In the case of the Face++ API [49], one or more RGB images are uploaded as input to obtain a 3-D face model. Subsequently, both RGB and depth images are rendered from the reconstructed model. The 3-D rendering process is carried out using a Python script in Blender [50], a free and cross-platform open-source 3-D graphics software.

D. Determining Key Parameter λ

The parameter λ is introduced to balance the target loss and the intrinsic loss of the GAN, thereby ensuring that the reconstructed depth image can successfully bypass authentication or identification systems while preserving the geometric structure of the face. We systematically evaluate the ASRs across various values of λ. As shown in Table III, the attacks on face authentication and identification models achieve

TABLE IV
ATTACK PERFORMANCE COMPARISON OF FOUR METHODS

Method	Datasets				
	Pandora	Texas	USTC	Lock3DFace	
				Led3D	Uppal
Face++	55.26%	3.57%	38.50%	0%	89.32%
HRN	60.52%	17.85%	44.82%	0%	89.32%
DECA	63.15%	21.42%	48.27%	0%	89.32%
DREAM	**94.73%**	**85.71%**	**88.50%**	**36.36%**	**89.32%**

optimal effectiveness when λ is set to 1.2 and 1.0, respectively. Accordingly, in our experiments, we set λ to 1.2 for attacks targeting face authentication models and to 1.0 for those targeting face identification models.

E. Experimental Results

1) Performance Comparison: The overall performances of DREAM are shown in Table IV. To represent various attack scenarios, we use different recognition models and loss functions. For the face authentication scenario, we used a Siamese network-based model, which is trained and tested on the Pandora, Texas 3-D, and USTC datasets. For the face identification scenario, we use Led3D and Uppal, which are trained and tested on the Lock3DFace dataset. Our method achieves effective attacks across various attack scenarios. It is evident that DREAM consistently outperforms conventional 3-D face reconstruction attacks, achieving significantly higher ASR on both the Texas and Lock3DFace (Led3D) datasets. This improvement can be attributed to the limitations of traditional 3-D face reconstruction attacks, which are one-shot in nature and primarily depend on the quality of the reconstructed face, without leveraging vulnerabilities in the authentication system. In contrast, DREAM incorporates the target loss from the target network during training, enabling the generator to iteratively minimize the distance between the depth image generated by DREAM and the template. As a result, the system is more likely to misclassify the generated samples.

The ineffectiveness of the 3-D face reconstruction attack on Led3D can be attributed to the fact that Led3D exclusively uses depth images as input. Moreover, the 3-D reconstruction approach does not incorporate feedback from the target model during optimization, leading to reconstructed depth images that lack adequate identity-specific features. In contrast, DREAM attains a maximum ASR of 36%, which is substantially higher than the 0% ASR achieved by the 3-D face reconstruction attack.

From the rightmost column of Table IV, the attack performance of the 3-D face reconstruction attack is comparable to that of DREAM when evaluated against the Uppal [13] model. Such an observation can be linked to the identification model's inherent characteristics, wherein the depth image is utilized to generate an attention map, and the final decision is based on RGB features multiplied by this attention map.

Fig. 5 presents the recovered depth images by DREAM and reconstructed depth images by SOTA 3D face reconstruction methods for several representative successful attack cases.

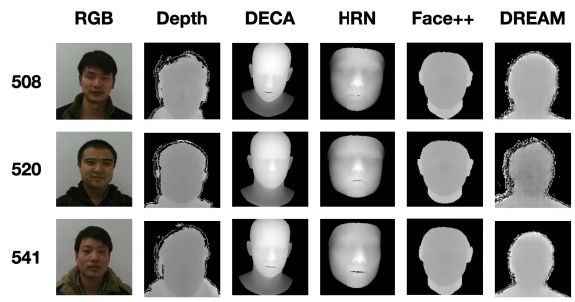

Fig. 5. Successful attack cases: ground-truth RGB-D images and depth images restored by DECA, HRN, Face++, and DREAM.

TABLE V
DREAM'S ATTACK PERFORMANCE CROSS-DATASETS

	Pandora	Texas	USTC
Pandora	94.73%	14.28%	71.26%
Texas	44.73%	85.71%	83.90%
USTC	78.94%	71.42%	88.50%

The leftmost number indicates the case ID, RGB denotes the available auxiliary data, depth represents the ground-truth depth image, and the remaining columns correspond to four comparison methods. The HRN, DECA, and Face++ 3-D face reconstruction algorithms utilize face models such as FLAME [51] and BFM [52] for reconstruction, resulting in depth images that are visually more similar to natural human faces. However, 3-D face reconstruction is a one-time process and cannot leverage feedback from the face recognition model to iteratively refine the depth image. In contrast, DREAM is designed to interact directly with the target recognition models, learning which features are most effective for spoofing and continuously optimizing them throughout the online attack process.

2) Cross-Dataset Testing: We conduct cross-dataset attacks on the face authentication model, with the results summarized in Table V. For instance, a value of 71.26% indicates the ASR when DREAM is trained using the Pandora dataset and tested against a target model trained on the USTC dataset. Notably, the DREAM trained using the Texas dataset achieves a relatively low ASR when attacking the target model trained on Pandora. This may be attributed to significant differences in the distributions of the two datasets, such as variations in the number of subjects or the total number of images; the reverse scenario exhibits similar trends. Overall, these results demonstrate the strong generalizability of DREAM, as it focuses on minimizing the distance between the input and the template, thus enabling effective transfer across datasets.

3) Ablation Experiment: To verify the effectiveness of the architecture of GAN, dual contrastive loss, and target loss, we conduct ablation studies across four datasets. As presented in Table VI, the ASR increases with the inclusion of each component, demonstrating that each element plays a significant role in enhancing the overall ASR.

A standard GAN generates face depth images in a random manner; however, incorporating a target loss introduces an additional objective, guiding the generated depth images to

TABLE VI
ABLATION EXPERIMENT OF DREAM

Method	Datasets				
	Pandora	Texas	USTC	Lock3DFace	
				Led3D	Uppal
G	65.78%	42.85%	57.47%	0%	89.32%
G + L_T	81.57%	71.42%	78.16%	21.23%	89.32%
G + A + L_T	86.84%	78.57%	83.90%	27.28%	89.32%
G + A + L_T + L_{DCL}	94.73%	85.71%	88.50%	36.37%	89.32%

G denotes GAN.
L_T denotes target loss.
A denotes attention block.
L_{DCL} denotes dual contrastive loss.

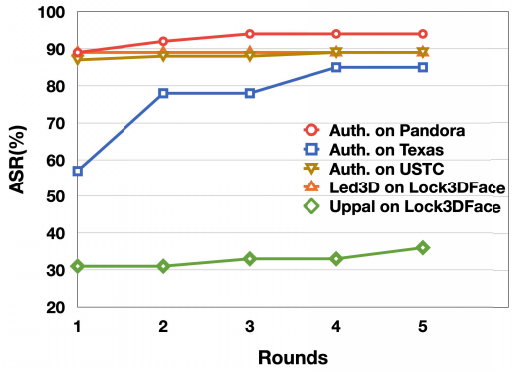

Fig. 6. DREAM's attack performance at different rounds.

better deceive the target recognition system. By integrating an auxiliary input and an attention block, DREAM is able to more precisely extract and generate salient features from relevant regions, rather than indiscriminately producing key features. Furthermore, the dual contrastive loss enhances the discriminator's capability while simultaneously compelling the generator to produce images that more closely align with the distribution of authentic depth images. This, in turn, contributes to a higher ASR.

When the available auxiliary dataset is limited, the modified target loss (5), which is based on contrastive learning, ensures that the depth image generated from the auxiliary RGB remains sufficiently close to the identity of the corresponding RGB image while being adequately separated from other identities in Led3D. In addition, we train the attack model on Led3D using both cross-entropy loss and cosine embedding loss. The results show that the cosine similarity score between the recovered depth image and all identities, including the target identity, exceeds 0.8. However, the similarity score to the target identity is not the highest among all identities, indicating that these loss functions fail to enforce sufficient separation between the recovered depth image and nontarget identities.

We further evaluate the ASRs under varying numbers of queries. As illustrated in Fig. 6, "Auth." refers to the face authentication model. Notably, a high ASR is achieved even in the initial round of queries, and the ASR continues to rise with additional query rounds, particularly in the case of Texas. These results indicate that both stages of DREAM play a substantial role in enhancing the effectiveness of the attack.

VI. DESIGN AND EXPERIMENT OF DREAM-3D

As illustrated in the Fig. 7, the overarching framework of DREAM-3D is structured into two sequential stages: *pretraining of the geometry-aware encoder* and *depth domain-adaptation with the target model*. In the initial stage, the attacker trains a geometry-aware encoder to transform an RGB image into a latent space vector, $\hat{w}$. Subsequently, this vector $\hat{w}$ is fed into a 3-D GAN, which then synthesizes an RGB image $\hat{I}^{\text{RGB}}_{\text{render}}$ and its associated depth image $\hat{I}^{D}_{\text{render}}$. During the subsequent second stage, the attacker trains a domain adaptation network to refine the depth image generated in the first stage. The ultimate goal is to yield an adapted depth image that can effectively deceive 3-D face recognition systems.

A. Geometry-Aware Encoder

Thanks to recent advancements in 3-D GANs, high-quality and identity-diverse 3-D faces can be synthesized by 3-D GANs, so we can now directly acquire high-resolution RGB images of faces and their corresponding depth images. Among these, EG3D [53] stands out as a typical representative; it comprises a mapping network, a StyleGAN2-based generator, a triplane 3-D representation with a decoder, a neural volume renderer, and a super-resolution module. The simplified pipeline of EG3D is visually represented in the top-left section of Fig. 7; we primarily focus on the 3-D GAN mapping network, the generator, and the renderer. First, the mapping network M_{3DGAN} receives two inputs: a random vector z sampled from a standard normal distribution and a pose parameter p, which it then maps to an intermediate vector w; we denote the mapping process as $w = M(z, p)$. Subsequently, this intermediate code w is then fed into the 3-D GAN generator and renderer to obtain the rendered RGB $I^{\text{RGB}}_{\text{render}}$ and depth I^{D}_{render}. Consequently, w determines the identity and other attributes of the generated 3-D face. Given an RGB face image I^{RGB}, can we recover the corresponding latent vector $\hat{w}$ such that the reconstructed 3-D face maintains identity consistency with the RGB face I^{RGB}?

To derive $\hat{w}$ from an RGB face image I^{RGB}, we designed a geometry-aware encoder, $\hat{w} = E(I^{\text{RGB}})$. The recovered latent code $\hat{w}$ is progressively derived by leveraging the pyramid features from the backbone network. The network architecture of the geometry-aware encoder is based on Swin transformer [54], with attention blocks incorporated at multiscale feature layers to generate corresponding level latent codes. Once the input RGB face image is projected to the latent code $\hat{w}$, we input $\hat{w}$ into the generator and render of 3-D GAN to get the corresponding rendered RGB $\hat{I}^{\text{RGB}}_{\text{render}}$ and depth $\hat{I}^{D}_{\text{render}}$.

The training process of the geometry-aware encoder E is illustrated in the upper portion of Fig. 7; synthetic data were leveraged for the supervised training. Training with synthetic data not only mitigates the challenge of limited available RGB-D datasets for attackers but also enables direct supervision of the latent code w learning. Therefore, w loss is a crucial component of L_E. We also incorporate the commonly used L_2 pixel loss, perceptual loss L_{LPIPS} [55], and ID loss L_{ID} calculated by a pretrained RGB-D face recognition network into L_E. Consequently, L_E can be formulated as follows:

$$L_E = L_{\text{ID}}\left(I^{\text{RGB}}_{\text{render}}, \hat{I}^{D}_{\text{render}}, I^{\text{RGB}}_{\text{render}}, I^{D}_{\text{render}}\right) + \|\hat{w} - w\|_2 + \left\|\hat{I}^{\text{RGB}}_{\text{render}} - I^{\text{RGB}}_{\text{render}}\right\|_2 + L_{\text{LPIPS}}\left(\hat{I}^{\text{render}}_{\text{RGB}}, I^{\text{render}}_{\text{RGB}}\right). \quad (12)$$

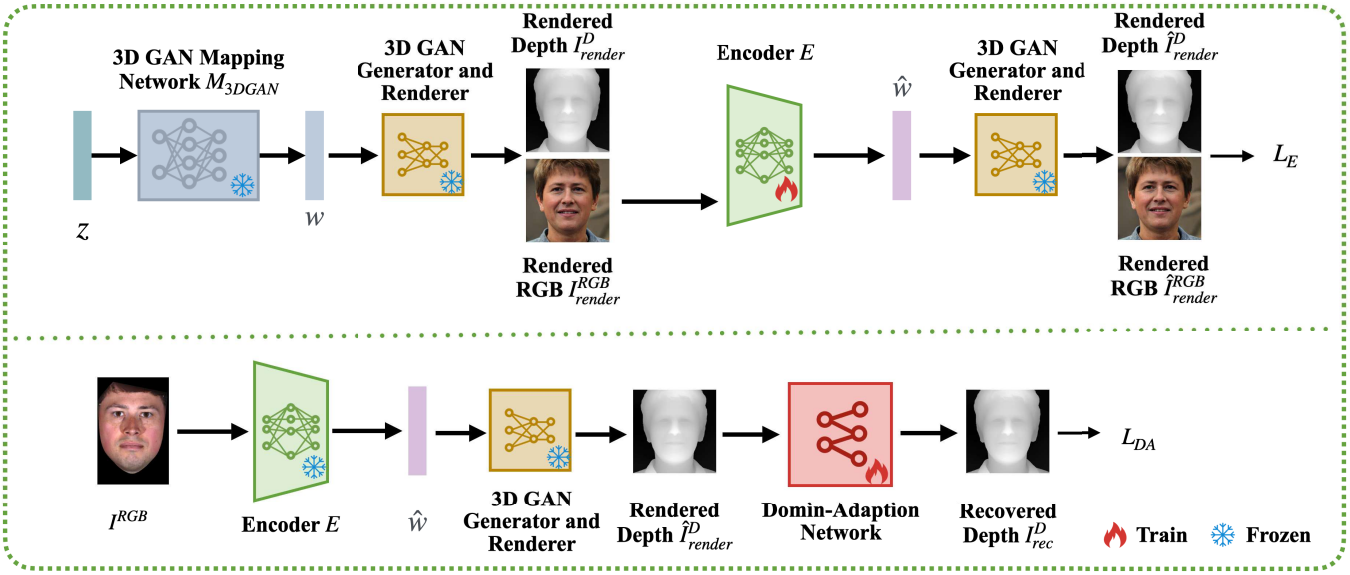

Fig. 7. Overview of DREAM-3D: pretraining of the geometry-aware encoder and depth domain-adaptation with the target model.

Specifically, the pixelwise L_2 loss and perceptual loss L_{LPIPS} are employed to ensure that the reconstructed 3-D faces exhibit greater similarity in texture and visual appearance. Concurrently, the w loss enables the geometry-aware encoder to directly learn the projection of the input RGB image into the W-space. Furthermore, the ID loss guarantees that the generated depth image can effectively deceive 3-D face recognition systems.

To ensure that the reconstructed latent code $\hat{w}$ aligns with the distribution of W-space, we also employ a discriminator to differentiate between the reconstructed $\hat{w}$ and authentic w. We train our geometry-aware encoder E and the discriminator D within a GAN-based framework. D is trained using the typical nonsaturating GAN loss [9] while also incorporating R_1 regularization [56]. Hence, we train E and D using the following loss functions:

$$L_D = -\mathbb{E}\left[\log D(w)\right] - \mathbb{E}\left[\log\left(1 - D(\hat{w})\right)\right] + \frac{\alpha}{2}\mathbb{E}\left[\|\nabla D(w)\|_2\right] \tag{13}$$

$$L_E^{\text{adv}} = -\mathbb{E}\left[\log D(\hat{w})\right] \tag{14}$$

where L_D is the loss for the discriminator D and α represents the coefficient for R_1 regularization. L_E^{adv} is an additional loss term for encoder E.

B. Domain-Adaptation Network

With the geometry-aware encoder effectively trained, it becomes possible to map an RGB image to a latent code $\hat{w}$, simultaneously yielding an identity-unified 3-D face. Nevertheless, a discernible domain discrepancy might exist between the generated depth image and its true counterpart. We design a U-Net-based domain-adaptation network to reduce the discrepancy between depth images rendered by the 3-D GAN and those from a target dataset while simultaneously enabling the adapted depth image to deceive 3-D face recognition systems. We train the domain-adaptation network using a limited, real RGB-D dataset. The loss function L_{DA} for training the domain-adaptation network is given as follows:

$$L_{\text{DA}} = (1 - Y)\frac{1}{2}\left(\text{Dis}\left(I^{\text{RGB}}, I^D, I^{\text{RGB}}, I^D_{\text{rec}}\right)\right)^2 + (Y)\frac{1}{2}\left\{\max\left(0, m - Dis\left(I^{\text{RGB}}, I^D, I^{\text{RGB}}, I^D_{\text{rec}}\right)\right\}^2 \tag{15}$$

where Y indicates whether the two input image pairs correspond to the same identity, with 0 representing a match and 1 indicating a nonmatch. I^{RGB}, I^D denotes the ground-truth RGB-D image pair, and $I^{\text{RGB}}, I^D_{\text{rec}}$ refers to the generated image pair. Dis() represents the Euclidean distance between the two input image pairs as computed by the target network, and m is a margin. This formulation is designed to guide the domain-adaptation network in learning which features are critical for successful system verification.

C. Experiment

1) Datasets: Texas 3-D, Lock3DFace, and VIPL-MumoFace-2K [18] are used in DREAM-3D. The first two datasets are detailed in Section V-A. There are more than 163k RGB-NIR-D image pairs in the VIPL-MumoFace-2K dataset, and each pair is aligned via face landmarks and cropped to a resolution of 256 × 256 pixels.

a) Preprocessing: Texas 3-D and Lock3DFace datasets are used to construct the evaluation datasets and train the domain-adaptation network. Following the benchmarks for RGB face recognition, all images for evaluation were aligned according to ArcFace's face coordinates and then cropped to 112 × 112 pixels. The remaining portion of the two datasets can be utilized for training the domain-adaptation network, ensuring no identity overlap with the evaluation dataset. The VIPL-MumoFace-2K dataset is used to train the 3-D face

TABLE VII
PERFORMANCE OF RGB-D MODELS AT FAR = 0.01 ON TEXAS 3D AND LOCK3DFACE DATASETS

Model	Input	Datasets	
		Texas 3D	Lock3DFace
ArcFace-D	RGB+Depth	99.02%	99.55%
ElaFace-D	RGB+Depth	97.08%	98.96%
Led3D	Depth	92.23%	86.37%
Uppal [13]	RGB+Depth	98.05%	99.70%

recognition model, with alignment and cropping procedures identical to those described previously.

b) Evaluation protocol: For the Texas 3-D dataset, we follow the LFW [57] evaluation protocol to select 103 positive pairs and 103 negative pairs for performance evaluation. For the Lock3DFace dataset, since the depth images were captured at different distances, we normalize the face region to a similar depth range and then select 675 positive pairs and 675 negative pairs for performance evaluation. We evaluate the recognition performance of a 3-D face recognition model by the metric TAR@FAR = 10^{-2}, and TAR@FAR means true acceptance rate achieved at a given FAR. We evaluate the attack performance of DREAM-3D by the metric ASR@FAR = 10^{-2}. The ASR herein denotes the proportion of successful attacks when a generated RGB-D image pair replaces the probe in the positive pairs of the evaluation protocol, given a threshold corresponding to FAR = 10^{-2}.

2) Target Model Implementation: Thanks to advancements in depth foundational models, monocular depth estimation models now exhibit considerably improved performance. We leverage the common RGB dataset Webface42M [58] and DepthAnything V2 [59] to create a synthetic RGB-D dataset, encompassing 10 000 unique identities. We use the synthetic RGB-D dataset and VIPL-MumoFace-2K to train the four 3-D face recognition models: ArcFace-D, ElasticFace-D, Led3D, and Uppal. The ArcFace-D and ElasticFace-D represent modified versions of ArcFace [60] and ElasticFace [61], respectively. These adapted models accept four-channel input, effectively performing signal-level fusion for RGB-D face recognition. For Led3D and Uppal, we exclusively utilized their backbone, training them with the methodology employed in ArcFace. The performance is shown in Table VII. Led3D demonstrated degraded performance when confronted with large datasets.

3) DREAM-3D Implementation: The network architecture of our geometry-aware encoder was designed based on the Swin transformer [54]. We also utilize ArcFace-D to compute the ID loss and employ a pretrained EG3D as both the 3DGAN generator and renderer. We use the Adam optimizer (lr = 2×10^{-5}) and the ranger optimizer (lr = 10^{-4}) for discriminator and encoder, respectively. We choose U-Net as the backbone of the domain-adaptation network and the Adam optimizer (lr = 10^{-3}), and the learning rate is decayed progressively during training.

4) Performance Comparison: We use DREAM trained on Texas 3-D and Lock3DFace as our comparison baseline. The overall performance of DREAM-3D is shown in Tables VIII and IX. We can see that our proposed DREAM and DREAM-3D maintain a high ASR even against high-performing RGB-D face recognition models. DREAM-3D's enhanced performance stems from several key factors. First, the 3D GAN is exceptionally well-trained and incorporates substantial 3-D human face priors. This foundational strength provides a rich understanding of face structures. Second, we integrated an ID loss during the training process of the encoder, which significantly boosts the attack's effectiveness. Furthermore, we have observed that while the depth images generated before domain-adaptation visually resemble real depth images, their attack efficacy on Led3D is notably poor. However, once domain-adaptation is applied, the attack performance improves dramatically. This highlights the critical role of domain-adaptation in bridging the gap between synthetically generated depth images and those capable of deceiving 3D face recognition systems.

When attacking the Led3D model, the ASR is comparatively lower than with other models. This is because Led3D relies solely on depth images for recognition, which imposes higher requirements on the quality and specificity of these depth images. Conversely, when attacking other models that utilize RGB-D images, the ASR is higher. This can be attributed to the fact that RGB-D models place greater emphasis on RGB face images, which are more readily available and easier to acquire.

Fig. 8 presents the recovered depth images by DREAM and DREAM-3D for several representative successful attack

TABLE VIII
ATTACK PERFORMANCE COMPARISON FOR DIFFERENT RGB-D MODELS AT FAR = 0.01 ON TEXAS 3D

Method	Texas 3D			
	ArcFace-D	ElasticFace-D	Led3D	Uppal
Dream	80.58%	77.66%	29.12%	70.87%
Dream-3D	92.23%	90.29%	58.25%	87.37%

TABLE IX
ATTACK PERFORMANCE COMPARISON FOR DIFFERENT RGB-D MODELS AT FAR = 0.01 ON LOCK3DFACE

Method	Lock3DFace			
	ArcFace-D	ElasticFace-D	Led3D	Uppal
Dream	84.44%	81.48%	31.11%	87.11%
Dream-3D	96.88%	93.33%	47.40%	92.14%

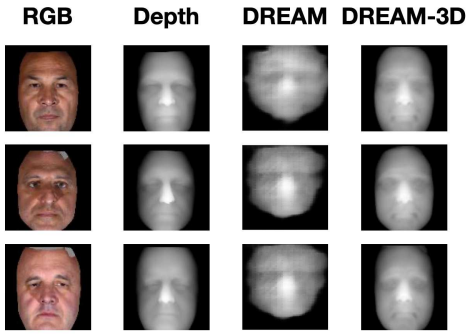

Fig. 8. Successful attack cases: ground-truth RGB-D images and depth images restored by DREAM and DREAM-3D.

TABLE X
COMPUTATIONAL COMPLEXITY COMPARISON

Method	HRN	DECA	DREAM	DREAM-3D
ms/sample	340	26.31	10	40

TABLE XI
ABLATION EXPERIMENT OF DREAM-3D ON TEXAS 3-D

Method	Texas 3D			
	ArcFace-D	ElasticFace-D	Led3D	Uppal
E	87.37%	86.40%	19.41%	83.49%
E+DA	92.23%	90.29%	58.25%	87.37%

TABLE XII
ABLATION EXPERIMENT OF DREAM-3D ON LOCK3DFACE

Method	Lock3DFace			
	ArcFace-D	ElasticFace-D	Led3D	Uppal
E	90.37%	86.67%	26.67%	85.03%
E+DA	96.88%	93.33%	47.40%	92.14%

E denotes the geometry-aware encoder.
DA denotes the domin-adaptation network.

cases in the Texas 3-D dataset. As shown in the figure, the depth recovered by DREAM-3D looks more like a real human face. We also further calculate the Fréchet inception distance (FID), with DREAM and DREAM-3D being 154.9 and 98.6, respectively.

5) Computational Complexity: We evaluate the computational complexity of various methods using a single RTX 3090, with the time taken per sample as the metric. As shown in Table X, DREAM is more time-efficient, and DREAM-3D is in the same order of magnitude.

6) Ablation Experiment: To verify the effectiveness of the geometry-aware and domain-adaptation network, we performed ablation experiments on Texas 3-D and Lock3DFace. Tables XI and XII show the change in ASR with the use of each component, which proves that each component contributes to the improvement of the ASR.

When targeting RGB-D face recognition models, their ASR is high even without a domain-adaptation network, indicating that RGB-D models heavily rely on RGB images. When facing models like Led3D, which rely solely on depth images, the attack effectiveness significantly improves after incorporating a domain-adaptation network. This is because there is a domain gap between the depth images rendered by 3-D GANs and those acquired by sensors, while the domain-adaptation network bridges the domain gap and significantly improves the attack effectiveness by learning the mapping from the 3-D GAN rendered depth images domain to the depth images domain of the specified dataset and supervising this process with an ID loss.

VII. DISCUSSION

Our main contribution is how to generate a depth image that can deceive 3-D face recognition systems based on a person's RGB image in the digital domain. However, it is important to discuss how our design can be adapted for use in practical settings. We can exploit an optical adversarial attack [35] to feed desired depth information directly into the camera in the physical world. Likewise, in the demo of [62], the depth image can be used to create a printable 3-D model. However, when using a mesh converted from a low-resolution depth map to make a mask, the surface of the mask may be smooth, which leads to inaccurate depth capture by the sensor. Even with a mask made from a high-resolution depth map, it is hard to ensure that the depth value captured by the sensor is the depth image that we generate due to the inherent errors of the sensor and the mask making. Addressing these limitations and challenges of executing the attack in physical settings may require super-resolution and end-to-end optimization.

VIII. CONCLUSION

In this article, we propose a closed-box attack method that can fool the 3-D face recognition system with depth information recovered from 2-D images. The attacker exploits the output of the target model and the 2-D images of the victim to recover the depth, which is convenient compared to previous methods of recovering depth information. We add a target loss of the target network and an attention block to the GAN to recover the depth of able to pass the authentication, rather than approaching the ground truth. Dual contrastive loss also contributes to the ASR. Finally, we evaluate the effectiveness of DREAM on four public 3-D face datasets. Our experimental results demonstrate that DREAM can successfully deceive 3-D face recognition systems and performs well on different datasets, and some 3-D face recognition systems still rely on RGB features, which are vulnerable. We further explore depth image reconstruction using 3-D GANs. Extensive experiments have demonstrated the effectiveness of DREAM-3D and highlighted the vulnerabilities of 3-D face recognition systems.

REFERENCES

[1] S. Saunders. (2017). *Cyber Security Firm Uses a 3D Printed Mask to Fool Iphone X's Facial Recognition Software*. [Online]. Available: https://3dprint.com/194079/3d-printed-mask-iphone-x-face-id/
[2] L. Morovič and P. Pokorný, "Optical 3D scanning of small parts," in *Proc. Adv. Mater. Res.*, vols. 468–471, 2012, pp. 2269–2273.
[3] X. Cao, Z. Chen, A. Chen, X. Chen, S. Li, and J. Yu, "Sparse photometric 3D face reconstruction guided by morphable models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4635–4644.
[4] M. Hernandez, T. Hassner, J. Choi, and G. Medioni, "Accurate 3D face reconstruction via prior constrained structure from motion," *Comput. Graph.*, vol. 66, pp. 14–22, Aug. 2017.
[5] J. Lin, Y. Yuan, T. Shao, and K. Zhou, "Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5890–5899.
[6] A. Morales, G. Piella, and F. M. Sukno, "Survey on 3D face reconstruction from uncalibrated images," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100400. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157401372100040X
[7] G. Shen et al., "Learning regularizer for monocular depth estimation with adversarial guidance," in *Proc. 29th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 5222–5230, doi: 10.1145/3474085.3475639.
[8] J. Zhang et al., "Heuristic depth estimation with progressive depth reconstruction and confidence-aware loss," in *Proc. 29th ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 2252–2261, doi: 10.1145/3474085.3475386.
[9] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[10] N. Yu et al., "Dual contrastive loss and attention for GANs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6711–6722.

[11] J. Bromley et al., "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 6, 1994, pp. 25–44.

[12] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5766–5775.

[13] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad, "Depth as attention for face representation learning," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2461–2476, 2021.

[14] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-from-depth for driver pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4661–4670.

[15] L. Jiang, J. Zhang, and B. Deng, "Robust RGB-D face recognition using attribute-aware loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2552–2566, Oct. 2020.

[16] S. Gupta, K. R. Castleman, M. K. Markey, and A. C. Bovik, "Texas 3D face recognition database," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation (SSIAI)*, May 2010, pp. 97–100.

[17] J. Zhang, D. Huang, Y. Wang, and J. Sun, "Lock3DFace: A large-scale database of low-cost Kinect 3D faces," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[18] S. Yu, H. Han, S. Shan, and X. Chen, "CMOS-GAN: Semi-supervised generative adversarial model for cross-modality face image synthesis," *IEEE Trans. Image Process.*, vol. 32, pp. 144–158, 2023.

[19] J. Kittler, A. Hilton, M. Hamouz, and J. Illingworth, "3D assisted face recognition: A survey of 3D imaging, modelling and recognition approachest," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 3, Jun. 2005, p. 114.

[20] D. Smeets, P. Claes, D. Vandermeulen, and J. G. Clement, "Objective 3D face recognition: Evolution, approaches and challenges," *Forensic Sci. Int.*, vol. 201, nos. 1–3, pp. 125–132, Sep. 2010.

[21] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003, doi: 10.1145/954339.954342.

[22] Y. Guo, H. Wang, L. Wang, Y. Lei, L. Liu, and M. Bennamoun, "3D face recognition: Two decades of progress and prospects," *ACM Comput. Surv.*, vol. 56, no. 3, pp. 1–39, Oct. 2023, doi: 10.1145/3615863.

[23] Apple.(2024). *About Face ID Advanced Technology*. [Online]. Available: https://support.apple.com/en-us/HT208108

[24] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad, "Two-level attention-based fusion learning for RGB-D face recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Los Alamitos, CA, USA, Jan. 2021, pp. 10120–10127, doi: 10.1109/ICPR48806.2021.9412514.

[25] Baidu.(2024). *Baidu AI Cloud Face Recognition Platform*. [Online]. Available: https://intl.cloud.baidu.com/product/face

[26] Tencent.(2024). *Tencent Cloud Face Recognition Platform*. [Online]. Available: https://intl.cloud.tencent.com/

[27] C. Jiang, S. Lin, W. Chen, F. Liu, and L. Shen, "PointFace: Point set based feature learning for 3D face recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.

[28] M. Li, B. Huang, and G. Tian, "A comprehensive survey on 3D face recognition methods," *Eng. Appl. Artif. Intell.*, vol. 110, Apr. 2022, Art. no. 104669. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622000057

[29] Y. Jing, X. Lu, and S. Gao, "3D face recognition: A comprehensive survey in 2022," *Comput. Vis. Media*, vol. 9, no. 4, pp. 657–685, Dec. 2023.

[30] Y. Jing, X. Lu, and S. Gao, "3D face recognition: A survey," 2021, *arXiv:2108.11082*.

[31] L. Souza, L. Oliveira, M. Pamplona, and J. Papa, "How far did we get in face spoofing detection?," *Eng. Appl. Artif. Intell.*, vol. 72, pp. 368–381, Jun. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197618300927

[32] J. Galbally and R. Satta, "Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models," *IET Biometrics*, vol. 5, no. 2, pp. 83–91, Jun. 2016. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2014.0075

[33] Z. Wu, Y. Cheng, J. Yang, X. Ji, and W. Xu, "DepthFake: Spoofing 3D face authentication with a 2D photo," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2023, pp. 917–933.

[34] J. M. Singh and R. Ramachandra, "3-D face morphing attacks: Generation, vulnerability and detection," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 6, no. 1, pp. 103–117, Jan. 2024.

[35] Y. Li, Y. Li, X. Dai, S. Guo, and B. Xiao, "Physical-world optical adversarial attacks on 3D face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24699–24708.

[36] B. Lei, J. Ren, M. Feng, M. Cui, and X. Xie, "A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 394–403.

[37] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Trans. Graph.*, vol. 40, no. 8, pp. 1–13, Aug. 2021, doi: 10.1145/3450626.3459936.

[38] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3D masks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1084–1097, Jul. 2014.

[39] H. O. Shahreza and S. Marcel, "Template inversion attack against face recognition systems using 3D face reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19605–19615.

[40] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5769–5779.

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[43] T. Karras et al., "Alias-free generative adversarial networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 852–863.

[44] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[45] X. Yuan, K. Chen, J. Zhang, W. Zhang, N. Yu, and Y. Zhang, "Pseudo label-guided model inversion attack via conditional generative adversarial network," in *Proc. 37th AAAI Conf. Artif. Intell. 35th Conf. Innov. Appl. Artif. Intell. 13th Symp. Educ. Adv. Artif. Intell.*, vol. 37, 2023, pp. 3349–3357, doi: 10.1609/aaai.v37i3.25442.

[46] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, "Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 357–372, 2022.

[47] H. Wang et al., "RGB-depth fusion GAN for indoor depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6209–6218.

[48] Z. Tang et al., "VVSec: Securing volumetric video streaming via benign use of adversarial perturbation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3614–3623.

[49] (2024). *Face++ 3D Face Reconstruction Interface API*. [Online]. Available: https://www.faceplusplus.com.cn/3dface/

[50] R. Ton. (1995). *Blender*. [Online]. Available: https://www.blender.org/

[51] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–17, Dec. 2017, doi: 10.1145/3130800.3130813.

[52] T. Gerig et al., "Morphable face models—An open framework," 2017, *arXiv:1709.08398*.

[53] E. R. Chan et al., "Efficient geometry-aware 3D generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16123–16133.

[54] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[56] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually converge," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3481–3490.

[57] 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 *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, Jul. 2008, pp. 1–14.

[58] Z. Zhu et al., "WebFace260M: A benchmark unveiling the power of million-scale deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2021, pp. 10492–10502.

[59] L. Yang et al., "Depth anything V2," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 21875–21911.

[60] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[61] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 1578–1587.

[62] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 9492–9502.

Shizong Yan received the bachelor's degree from the School of Automation, Nanjing Institute of Technology, Nanjing, China, in 2014. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Donghua University, Shanghai, China.

His research interests include 3-D face recognition and artificial intelligence (AI) system security.

Huixiang Wen received the Ph.D. degree in computer science from Donghua University, Shanghai, China, in 2024.

He is now a Lecturer with the Department of Medical Imaging, Bengbu Medical University, Bengbu, China. His research interests include the Internet of Things (IoT) security, mobile computing, and computer vision.

Shan Chang (Member, IEEE) received the Ph.D. degree in computer software and theory from Xi'an Jiaotong University, Xi'an, China, in 2012.

From 2009 to 2010, she was a Visiting Scholar with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong. She was also a Visiting Scholar with BBCR Research Laboratory, University of Waterloo, Waterloo, ON, Canada, from 2010 to 2011. She is currently a Professor with the Department of Computer Science and Technology, Donghua University, Shanghai, China. Her research interests include security and privacy in mobile networks and sensor networks.

Dr. Chang is a member of IEEE Computer Society, Communication Society, and Vehicular Technology Society.

Hongzi Zhu (Senior Member, IEEE) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2009.

He was a Post-Doctoral Fellow with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, and the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, in 2009 and 2010, respectively. He is a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include mobile sensing and computing, and the Internet of Things.

Dr. Zhu received the Best Paper Award from IEEE Globecom 2016. He was a Leading Guest Editor of *IEEE Network Magazine*. He is an Associate Editor of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and IEEE INTERNET OF THINGS JOURNAL. He is a Senior Member of the IEEE Computer Society, IEEE Communication Society, and IEEE Vehicular Technology Society. For more information, please visit http://lion.sjtu.edu.cn

Luo Zhou received the M.S. degree from the Department of Electronics and Communication Engineering, Jiangsu University of Science and Technology, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Donghua University, Shanghai, China.

His research interests include ubiquitous and pervasive computing, mobile computing/sensing, and IoT security.