A Physical-World Adversarial Attack for 3D Face Recognition

Yanjie Li,¹ Yiquan Li,¹ Xuelong Dai,¹ Songtao Guo,² Bin Xiao¹

¹ Computing Department, The Hong Kong Polytechnic University
² College of Computer Science, Chongqing University

yanjie.li@connect.polyu.hk, yiquan.li@connect.polyu.hk, xuelong.dai@connect.polyu.hk, guosongtao@cqu.edu.cn, csbxiao@comp.polyu.edu.hk

Abstract

The 3D face recognition has long been considered secure for its resistance to current physical adversarial attacks, like adversarial patches. However, this paper shows that a 3D face recognition system can be easily attacked, leading to evading and impersonation attacks. We are the first to propose a physically realizable attack for the 3D face recognition system, named structured light imaging attack (SLIA), which exploits the weakness of structured-light-based 3D scanning devices. SLIA utilizes the projector in the structured light imaging system to create adversarial illuminations to contaminate the reconstructed point cloud. Firstly, we propose a 3D transform-invariant loss function (3D-TI) to generate adversarial perturbations that are more robust to head movements. Then we integrate the 3D imaging process into the attack optimization, which minimizes the total pixel shifting of fringe patterns. We realize both dodging and impersonation attacks on a real-world 3D face recognition system. Our methods need fewer modifications on projected patterns compared with Chamfer and Chamfer+kNN-based methods and achieve average attack success rates of 0.47 (impersonation) and 0.89 (dodging). This paper exposes the insecurity of present structured light imaging technology and sheds light on designing secure 3D face recognition authentication systems.

Introduction

3D face recognition systems use facial structural characteristics for user authentication and other tasks to compensate for the vulnerabilities of 2D face recognition. The common methods for acquiring real-time 3D face data mainly include structured light imaging, ToF camera, and laser. The structured light technology has high measurement precision and outperforms on uniform textures (Feng et al. 2021). The technology is popular to be adopted in 3D face measurement and is widely used in 3D scanners and smartphones such as Kinect v1 and iPhone X (Zhou and Xiao 2018).

The security of image-based face recognition systems has been thoroughly studied (Sharif et al. 2016, Dong et al. 2019, Xiao et al. 2021). However, the security of the 3D face recognition system is still in an early stage. This, to a large extent, is because the 3D face recognition system is resistant to existing physical attack methods, such as adversarial patches. Tsai et al. (Tsai et al. 2020) firstly proposed 3D printable adversarial examples for point cloud classification tasks. However, this method is difficult to be applied to 3D face recognition in the physical world because of liveness detection. In this paper, we propose a physically realizable adversarial attack on 3D face recognition named structured light imaging attack (SLIA), demonstrating that the structured-light-based 3D face recognition system can be easily attacked. We conceal the adversarial perturbations into the projected patterns, making them hard to be perceived by human eyes. The perturbed patterns will result in points shifting in the face point cloud and trigger dodging and impersonation attacks. This paper’s main contributions are:

- We are the first to realize a physically realizable adversarial attack on the 3D face recognition system. We camouflage the adversarial perturbations into the normal illuminations of the structured light system to make it hard to be discovered and can bypass the liveness detection.

- We propose a 3D transformation-invariant loss, which models the head’s movement as stochastic spatial rotation and translation transformations. Experiments show that our method significantly improves the physical attack’s robustness to the head’s movement.

- We propose efficient methods to map the point shifting to the pixel shifting in the projected patterns. We also integrate them into the optimization process by minimizing the total pixel-shifting. Our method needs fewer modifications on patterns than Chamfer and kNN-based methods.

- We successfully attack a real-world structured-light-based 3D face recognition system. We also conduct ablation studies of different modules and evaluate our attack’s robustness to random movements and transferability to black-box 3D models.

Fig. 1(a) shows the flowchart of the normal structured-light-based 3D face recognition process and SLIA. We suppose the adversary cannot directly modify the 3D classifier’s inputs. Therefore, we utilize the flaws of the structured-light-based 3D scanning device and modify the projected patterns instead. Fig. 1(b) shows an intuitive explanation of the attack principle, which reverses the adversarial point clouds to the projected patterns by mapping point-shifting to pixel-shifting. We suppose the 3D scanner is a white box so that we can design efficient reversing methods. We also propose some skills to satisfy the physical constraint and reduce the inverse
Figure 1: The flowchart and basic principle diagram of SLIA. (a) In the normal process, the scanner firstly sends specific patterns to the projector (①), then projects them on the face and captures the modulated pictures (②), and finally recovers the 3D data(③). In SLIA, the attacker modifies the projected patterns to pollute the 3D data indirectly. (b) The correspondence between point shifting $\Delta z$ and pixel shifting $\delta y_p$. We move the pixels in the projected patterns to result in point shifting in the point cloud.

error, e.g. single-direction constraint and projector undistortion. We realize both impersonations (targeted) and dodging (untargeted) attacks a real-world 3D face recognition system. Experiments show that our methods need fewer perturbations compared with Chamfer and kNN-based methods and achieve average attack success rates of 0.47 on impersonation attacks and 0.89 on dodging attacks.

**Related Work**

**3D Adversarial Attack** 3D adversarial attacks focus on generating adversarial examples for 3D deep learning models. Most of them aim for point cloud data. Xiang et al. (Xiang, Qi, and Li 2019) first proposed an adversarial attack on the point cloud by point perturbation and point generation. Zheng et al. (Zheng et al. 2019) and Wicker et al. (Wicker and Kwiatkowska 2019) spoofed the deep learning model by dropping some critical points based on the saliency map. To improve the transferability of the attack, Hamdi et al. (Hamdi et al. 2020) proposed advPC which adds an auto-encoder module into the adversarial point cloud generation process. Some other studies focused on improving the imperceptibility of the adversarial examples based on the geometric properties of the adversarial point cloud (Wen et al. 2022; Huang et al. 2022). However, these attacks modify the classifier’s input directly and are hard to be realized in the real world.

For physical realizable attacks, Cao et al. (Cao et al. 2019) successfully fooled the Lidar sensor on autonomous vehicles by adding fake front-near obstacles through a time-lapse module and a laser emitter. However, the sensors used for face recognition are significantly different from the autopilot. Tsai et al. (Tsai et al. 2020) proposed kNN loss to generate 3D printable adversarial examples. Tu et al. (Tu et al. 2020) proposed utilizing Laplacian loss to improve the mesh’s smoothness and the 3D printability. However, because of liveness detection, the 3D-printing-based attacks cannot be directly applied to the 3D face recognition. To the best of our knowledge, no one had realized a physical adversarial attack on the 3D face recognition system, whereas the security of the face recognition system is crucial in reality.

**Optical Adversarial Attack** The optical adversarial attacks change the illumination of the target objects to spoof the classifiers. Compared with printing-based attacks, they can bypass the liveness detection and have better camouflage. The first work was proposed by Nicoles et al. (Nichols and Jasper 2018), which generates adversarial illuminations through iteratively capturing and optimizing. Zhou et al. (Zhou et al. 2018) utilized infrared LEDs to attack the face recognition system. But their method has restricted perturbation space. Worzyk et al. (Worzyk, Kahlen, and Kramer 2019) projected perturbations onto the road stop signs. Nguyen et al. (Nguyen et al. 2020) applied the optical adversarial attack to the 2D face recognition and proposed transformation-invariant adversarial loss to improve the physical robustness. Gnanasambandam et al. (Gnanasambandam, Sherman, and Chan 2021) improved attack success rate by considering spectral non-linear. However, these attacks currently only consider the 2D image classification tasks and cannot be directly applied to 3D scenarios because of the large difference in their imaging principles. In this work, we propose SLIA attack to apply optical adversarial attacks to the 3D face recognition scenario.

**Methodology**

Fig 2 shows the pipeline of SLIA. The black block diagram is the normal process of SLI-based 3D face recognition. The red block diagram is our attack, which consists of the adversarial point cloud generation and the inverse function to map the adversarial point cloud to the input pattern of the projector. We firstly propose a 3D transform-invariant loss function to generate adversarial point clouds with higher physical robustness. Then we briefly introduce the SLI-based face recognition system and design high-efficiency inverse methods to map the adversarial point clouds back to the projector’s inputs.
Generating adversarial point cloud
In this subsection, we introduce our method of generating robust adversarial perturbations on the face point clouds. We firstly introduce the perturbation’s form. Then we introduce the loss function for generating the adversarial point cloud.

Single-direction perturbation In the SLI technology, there is a one-to-one correspondence between the point in the point cloud and the pixel in the modulated images. Therefore we define the adversarial perturbations as additive noises, which shift some points of the original point cloud, rather than adding or dropping points. In addition, to satisfy the physical constraint of the 3D reconstruction algorithm, we firstly rotate the point cloud to make its z-axis coincide with the camera coordinate system’s z-axis. Then we restrict the perturbation on the z-direction. This constraint can make the point shifting correspond to the depth change of the matching pixel, therefore reducing the influence on other pixels and reducing the reconstruction error (we prove this later). Therefore, we have \( P' = P + \Delta = P(x, y, z + \Delta_z) \), where \( P, \Delta \in \mathbb{R}^{N \times 3} \), \( N \) is the number of points in the point cloud.

Loss function We choose the C&W attack (Carlini and Wagner 2017) as our basic attack. We make some important modifications to make it fit the 3D face data and physical attack context. For the impersonation attack, The aim is to make the classifier \( M(\cdot) \) identify the 3D face data as a specific person. The loss function is defined as

\[
\arg \min_{\Delta} \left( f_{t'}(M(P + \Delta)) + \lambda \cdot D(P, P') \right)
\]

where \( t' \) is the target label, \( D(P, P') \) is the distance metric, and \( \lambda \) is a hyperparameter to balance these two terms. The \( f_{t'}(\cdot) \) is the logits loss function, which is defined as

\[
f_{t'}(P) = \max(\max(Z(P)_{ij \neq t'}) - Z(P)_{t'}, \kappa)
\]

where \( Z(P) \) is the output of the logits layer and \( \kappa \) controls the minimal margin. However, in the real world, the head may have small movements when the attack is in progress. To solve this problem, we propose a 3D transform-invariant loss (3D-TI) to make the adversarial point clouds more robust in physical attacks. We model the small movements of the human head as random rotating around the \( X, Y, \) and \( Z \) axes and translating along the \( XY \) plane and involve these random 3D transformations in the optimization. The spatial transformation function can be modeled as

\[
T(P) = (R(\theta_x, \theta_y, \theta_z)P)^T + M(\eta_x, \eta_y),
\]

where \( M \) and \( R \) are spatial translation and rotation matrix, \( \eta_x, \eta_y \) are random displacements sampling from a normal distribution \( \mathcal{N}(0, 0.01) \) and \( \theta_x, \theta_y, \theta_z \) are random rotation angles sampling from \( \mathcal{N}(0, 10^\circ) \). Moreover, to simulate the real-world preprocessing process, we involve a resample and renormalization function in the optimization. The final loss function is defined as

\[
\arg \min_{\Delta} \frac{1}{k} \sum_{i=0}^{k-1} f_{t'}(M(N(T_i(P) + \Delta))) + \lambda \cdot D(P, P')
\]

where \( T_i \) is \( i \)th random 3D transformation matrix. \( N(\cdot) \) is a resample and renormalization function, which resamples \( N \) points from \( P' \) and then renormalizes it into zero mean and unit variance. \( M \) is the classification model. \( k \) is the 3D transformations’ number in one iteration. The hyper-parameter \( \lambda \) is determined using binary search. We compared several different distance metric in our experiments, including \( l_2 \) loss, Chamfer loss (Xiang, Qi, and Li 2019), Chamfer+kNN loss (Tsai et al. 2020) and total pixel shifting of adversarial patterns, which is illustrated in the next subsection. We optimize this loss function using Adam algorithm.

Moreover, through experiments, we find that even if the same adversarial sample is fed into the PointNet++, different classification results may be produced, which may be because the PointNet++’s down-sampling layers improve its robustness to adversarial samples. Experiments show that the re-normalization function (with or without resampling) can greatly improve the models’ classification consistency to adversarial examples. The details are shown in the ablation study. Last but not least, at the end of each iteration, we clip the perturbation to make \( \| \Delta \|_{\infty} < \xi \). This is because when the surfaces have too large jumps, the multi-step phase-shift algorithm may unwrap the corresponding pixel’s phases into false periods.

Inversing the SLI system
The SLI system firstly projects and captures a group of phase-shift images in sequential order then reconstructs the 3D shape and sends it to the face recognition system. Therefore,
We model the SLI system as a series system, as shown in Fig[2] which consists of a projector, a face, a camera, a 3D reconstruction algorithm, and a classifier. We first briefly introduce the 3D reconstruction algorithm we use and then introduce our inverse method. Finally, we discuss some significant tips to improve the reconstruction accuracy.

The 3D reconstruction algorithm We reference Piccirilli’s work [Piccirilli et al. 2016] to build the real-time 3D face data acquisition system, which uses fringe projection techniques and the representative multi-phase shift-masking (MSPS) algorithm. This structured light imaging system consists of one camera and one projector. The projector projects a group of phase-shift patterns to encode every unitary position of a surface. Then the camera reconstructs the 3D face data from modulated pictures. We choose the phase-shift step number $N$ as 12 in this paper.

To reconstruct the 3D data, we firstly calibrate the projector and camera to get their projection matrices $A_p$ and $A_c$. Then we project the sinusoidal fringe patterns on faces and reconstruct the 3D coordinates by calculating the absolute phase. The projected sinusoidal fringe images have a phase difference of $\frac{2\pi}{N}$. The modulation process is

$$I'_i(u, v) = a(u, v) + r(u, v)A\cos(\phi_{w}(u, v) - \frac{2\pi i}{N}).$$

(5)

where $i \in [0, N - 1]$ is the phase-shift step number. The $(u, v)$ is the pixel coordinate in $I_c$. The $a(u, v)$ and $r(u, v)$ are respectively the background light and reflectivity. The $A$ is the amplitude of sinusoidal fringe patterns. The $\phi_w$ is the wrapped phase. Therefore, after getting the modulated pictures $I'_0, ..., I'_{N-1}$, $\phi_w$ can be derived from Eq[5]

$$\varphi_w(u, v) = \tan^{-1}\left(\frac{\sum_{n=0}^{N-1} I_i(u, v)\sin(2\pi n/N)}{\sum_{n=0}^{N-1} I_i(u, v)\cos(2\pi n/N)}\right).$$

(6)

After getting the wrapped phase, we unwrap it to get the absolute phase, $\phi_a(u, v) = \varphi_w(u, v) + 2\pi K(u, v)$, where $K(u, v)$ is the stripe’s order. We use cyclic complementary gray code [Wu et al. 2019] to compute $K(u, v)$. Then we match pixels in $I_p$ and $I_c$ according to their absolute phases should be equal. Finally, we use Eq[32] in Feng’s paper [Feng et al. 2021] to recover the 3D coordinate, which, for the sake of simplicity, can be expressed as the following function,

$$[x, y, z] = h(A_c, A_p, u, v, c, v_p, c_p, v_p),$$

(7)

where $[x, y, z]$ is the 3D coordinate. The $u, v$ are the pixel’s horizontal and vertical coordinates in $I_c$. $u_p$ is the matching pixel’s horizontal coordinate in $I_p$. We exclude $v_p$ in Eq[7] because $v_p = v_p$ when the projector and camera are horizontally placed.

Inverse method After shifting the point from $(x, y, z)$ to $(x, y, z + \Delta z)$, we inverse the point shifting to pixel shifting in the projected patterns. We firstly calibrate the projector and camera to get projection matrices $A_p$ and $A_c$. Then we get the matching pixels’ horizontal coordinates in $I_p$ by solving the inverse function of Eq[7]

$$u'_p = h^{-1}(x, y, z') = A_{i,j}^P x + A_{i,j}^P y + A_{i,j}^P z' + A_{i,j}^P$$

(8)

where $A_{i,j}$ is the element at $(i, j)$ in $A$. To accelerate the calculation. Then we change the color of the matching pixel $(u'_p, v_p)$ in projected patterns to make SLI system think the light comes from $(u'_p, v_p)$.

$$I'_p(u'_p, v_p) \leftarrow I_p(u'_p, v_p).$$

(9)

Fig[1][b] shows an intuitive explanation of Eq[9]. This inversion method can be integrated into the optimization. But because the pixel coordinate is not differential, we optimize $\phi_a$ instead of $u_p$ directly. Then we get $u_p$ though $u_p = \text{round}(\frac{w\phi_a}{2\pi n} - w)$, where $w$ is the fringes’ number, $w$ is the width of projected images. We also add a $l_2$ loss term between the original and revised absolute phase to minimize the total pixel-shifting. The new loss function is defined as

$$\arg\min_{\phi'_a} \frac{1}{k} \sum_{i=0}^{k-1} f_{\psi} (I(M(h(\phi'_a)) + h(\frac{w\phi_a}{2\pi n} - \phi'_a))$$

$$+ \lambda_1 \cdot D(P, P') + \lambda_2 \cdot \|\phi'_a - \phi_a\|_2)$$

(10)

Where calculate $\frac{\partial f_{\psi}}{\partial \phi_a}$ in the optimization, we fix the values of $x$ and $y$ and calculate $\frac{\partial f_{\psi}}{\partial \phi_a}$ by Eq[8]

The influence of single-direction perturbation We assume that the camera directly faces the face. In this case, the $z$-axes of the world and camera coordinate system coincide (otherwise, we rotate the point cloud to make their $z$-axes coincide). Then we have the following theorem.

Theorem 1. When the camera and world coordinate systems’ $z$-axes coincide, only perturbing on the $z$-axis has a very small influence on the corresponding pixel coordinate.

Proof. Because the $z$-axis of the world coordinate system and camera coordinate system coincides, the point’s coordinate in the camera coordinate system can be deduced by

$$\begin{bmatrix} x_c \ y_c \ z_c' \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \ y \ z + \Delta z \end{bmatrix} + T$$

(11)

According to the camera imaging model, the transition matrix from camera coordinate to image coordinate is

$$z_c' \begin{bmatrix} u_c' \\ v_c' \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & 0 \\ 0 & \frac{1}{dy} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_c & 0 & 0 \\ 0 & f_c & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c' \end{bmatrix}.$$  

(12)

where $(u_c', v_c')$ is the point’s corresponding pixel coordinate in $I_c$ after the perturbation. $f_c$ is the focal length, $dx$ and $dy$ are the pixel’s physical size. Therefore, when $|\Delta z| < 10^{-2}|z|, u_c < 100, v_c < 100$, we have

$$u_c' = \left(\frac{f_c x_c}{dx}\right)/(\Delta z) + u_0 \approx u_c + 1$$

(13)

$$v_c' = \left(\frac{f_c y_c}{dy}\right)/(\Delta z) + v_0 \approx v_c + 1$$

(14)

Therefore, when we only add small perturbations on the $z$-axis, the pixels’ coordinates in modulated images are almost uninfluenced.
The influence of ambient light and reflectance  In SLIA, the fringe intensity captured by the camera becomes

\[ I'_c(u_c, v_c) = a(u_c, v_c) + r(u_c, v_c) I'_p(u_p, v_p) \]  

where \( I'_c \) is the modulated images and \( I'_p \) is the adversarial patterns. This equation is very similar to the normal modulating process (see Eq.5), just replacing the original sinusoidal stripes with adversarial patterns. From Eq.6, we can see that the background illumination and reflectance can be canceled out in the multi-step phase-shifting algorithm because they are the bias term and first-order coefficient. Therefore, we don’t consider their influence in this work.

Linearizing the projector distortion  Although the ambient light and reflectance influence can be counteracted, the projector’s intrinsic nonlinear distortions can still cause reconstruction error and reduce the attack success rate. We model the projector’s distortion as a cubic polynomial function and rectify the adversarial pattern’s grayscale before it is projected. Briefly speaking, suppose \( I_s \) is the source image of the projector, \( I_p \) is the projector’s output, and \( f \) is the projector’s distortion function. We find an distortion linearization function \( f^{-1} \) to make

\[ f(f^{-1}(I_s)) = \frac{\max(I_p) - \min(I_p)}{\max(I_s) - \min(I_s)} (I_s - \min(I_s)) + \min(I_p) \]

through cubic spline fitting. The details can be found in the appendix.

Experiment

Experiment Setup

Dataset and structured light system  We use Bosphorus database (Savran et al. 2008) and Eurecom face database (Min, Kose, and Dugelay 2014) to evaluate the attack performance. Bosphorus database is collected by Inspeck Mega Capturor II 3D, which is a commercial structured light-based 3D acquisition equipment and used in many 3D face recognition works (Zhou and Xiao 2018). The sensor resolution in \( x, y, z \) directions are 0.3–0.4mm. It consists of 105 different faces with a rich set of expressions and occlusions. Each face consists of 35K points. Eurecom face database consists of 52 faces with nine facial expressions. It is acquired by Kinect, which is also a commercial structured light-based device. Each face has about 62K points. We also collect ten people’s faces using our own structured light system and add them to the above datasets. We downsample all these 3D face data to 4K points to train 3D classifiers.

Our own structured light system refers to Piccirilli’s design (Piccirilli et al. 2016), which includes an industry camera and a home projector and captures images at a distance of about 1.5m. The projector’s resolution is 1600×1200, and the camera’s resolution is 640×480. We linearize the color distortion before the images are projected. Then we crop the face in the modulated images using Viola-Jones algorithm and resize it to 65 × 65 centered by the face. We reconstruct the 3D data using the 12-step phase-shift algorithm. We also implement the physical attacks on our structured light system.

Applied model  We evaluate our attack on several state-of-the-art 3D classifier models, including Pointnet (Qi et al. 2017a), single-scale and multi-scale Pointnet++ (Qi et al. 2017b), DGCNN (Phan et al. 2018) and CurveNet (Xiang et al. 2021). We implement dodging and impersonation attacks on these models. The dodging attack aims to make these 3D classifiers classify these acquired faces into any classes except the ground truth. The impersonation attack aims to classify these faces into randomly chosen labels.

Evaluation metrics  We set the search space of \( \lambda \) in Eq.4 as \([0, 10]\) and use binary search to narrow down \( \lambda \). We set the binary search step as 5, the iteration of each step as 100, and \( \kappa \) as 30. We terminate the optimization process after the binary search step is reached and return the adversarial examples with minimum distance loss. We evaluate \( l_2 \) loss, Chamfer loss, Chamfer+kNN loss, and total pixel shifting in the projected patterns in our experiments and use attack success rate and \( l_2 \) distance between the adversarial and original patterns as evaluation metrics.

Experiment Results

In this section, we first evaluate the SLIA on several state-of-the-art 3D classifiers with different distance metrics. Then we conduct ablation studies of different components, like \( z \)-direction perturbation, 3D transformation, and renormalization. Next, we evaluate the effect of 3D transformation-invariant loss on real-world disturbance. Last but not least, we evaluate our attack’s transferability to black-box models and the influence factors of the attack success rate.

Adversarial examples with different distance metrics  Fig.3 shows the generated adversarial point clouds and patterns using different distance metrics. We implement physical impersonation attacks on ten people. We firstly generate digital adversarial point clouds on PointNet with randomly chosen targets, then we inverse them to the adversarial patterns \( I'_p \). Only one of the multi-step phase-shifting fringes is shown here. Other fringes are generated in the same way. We also show the modulated pictures \( I'_c \) and reconstructed adversarial face point cloud \( P'_o \) in Fig.3. From the figure, we can see that with the 3D-TI module, the perturbations on the projected patterns tend to be uniform and global.

We evaluate the perturbation size of the projected patterns under different distance metrics by \( \delta = \sum_{i=0}^{N-1} \frac{\| p_i - p'_i \|}{\| p_i \|} \) and plots their interquartile ranges. As shown in Fig.3(a), our attack only needs a small modification of projected patterns, especially for \( l_2 \) loss on the absolute phase. We also evaluate the perturbation size of adversarial point clouds with and without the 3D-TI module. As shown in Fig.3(b), with the 3D-TI module, the perturbation size slightly improves, which we think is the necessary cost to improve the physical attack’s robustness.

To evaluate the physical targeted and dodging attack success rate of SLIA on the whole datasets, we simulate the physical modulation process through Eq.5. The irradiance and reflectance of the scene are estimated through Gnanasambandam’s method (Gnanasambandam, Sherman, and Chan 2021). Then we attack this simulation system using our methods. The attack results are shown in Table1. All these distance metrics achieve high ASRs on dodging attacks. For the targeted attack, Chamfer+kNN distance outperforms other distances on the PointNet++. We think this is because Cham-
Figure 3: The digital and physically reconstructed adversarial point clouds using original and 3D-TI loss functions with different distance metrics. The first three distance metrics are measured on the point cloud, and the last distance metrics are measured on the absolute phase map, which is equivalent to the total pixel coordinate shifting.

| Attack Method | Pointnet | Pointnet++(SSG) | Pointnet++(MSG) | DGCNN | CurveNet |
|---------------|----------|-----------------|-----------------|-------|----------|
| SLIA($l_2$)   | 0.57     | 0.98            | 0.42            | 0.92  | 0.32     |
| SLIA(Chamfer) | 0.58     | **1.00**        | 0.39            | **1.00** | 0.35   |
| SLIA(Chamfer+kNN) | 0.52 | 0.95             | **0.45**        | **1.00** | **0.38** |
| SLIA($l_2$ on $\varphi_a$) | **0.62** | 0.98            | 0.37            | 0.75  | 0.26     |

Table 1: The attack performance of SLIA. We evaluate the attack performance by the attack success rate. We conduct attacks on five different models with both impersonation and dodging attacks.

Ablation study To evaluate the effects of different modules, we conduct an ablation study of different modules. We simulate the physical attacks by involving the random rotations and translations of human faces, and necessary preprocessing steps (normalization and down-sampling) and then evaluate the ASR on five different models. We use Chamfer distance metrics as a basic method to find digital adversarial examples. The results are shown in Table 2. Through the experiments, we find that the original attack suffers a low ASR in the physical attack context, as shown in the 1st line. We think this is because it’s vulnerable to physical changes and preprocessing steps. The 2nd line shows that the $z$-direction constraint can slightly improve ASR, which we think is because it reduces the reconstruction error of adversarial point clouds. The 3rd line shows that the renormalization module can improve ASR by about 20%, even without integrating the resampling module. We think this is because the normalization can change the overall distribution of adversarial point clouds and make the original attack fail. The renormalization module can consider this change beforehand to advance ASR. The 4th and 5th line shows that the 3D-TI module
can improve the physical attack’s success rate when there are small physical changes like random rotations and translations. With all the above modules, the targeted attack success rate increases 34% on Pointnet, 20% on Pointnet++ SSG, 24% on Pointnet++ MSG, 20% on DGCNN, and 33% on CurveNet.

### Physical attack robustness evaluation
To evaluate the effectiveness of the 3D-TI module, we rotate the original point cloud along the z-axis at different angles and add the same perturbation to the rotated point clouds. Then we compare the output logits on the target label after Softmax. As shown in Fig 5, with the 3D-TI module, the prediction is more robust to the rotation transformation.

**Table 2:** The result of ablation study. We conduct simulated physical attacks on five different models with both impersonation and dodging attacks. \( \Delta z \) means the \( z \)-direction constraint. \( \mathcal{N} \) means the renormalization module. \( T \) means the 3D-TI module.

| Attack Method | targeted | dodging | targeted | dodging | targeted | dodging | targeted | dodging |
|---------------|----------|---------|----------|---------|----------|---------|----------|---------|
| Cf.           | 0.21     | 0.72    | 0.19     | 0.75    | 0.11     | 0.68    | 0.17     | 0.74    |
| Cf. + \( \Delta z \) | 0.25     | 0.78    | 0.23     | 0.81    | 0.13     | 0.77    | 0.23     | 0.79    |
| Cf. + \( \mathcal{N} \) | 0.43     | 0.86    | 0.35     | 0.88    | 0.32     | 0.79    | 0.41     | 0.85    |
| Cf. + \( T \) | 0.35     | 0.88    | \( \mathbb{0.41} \) | 0.91    | 0.28     | 0.85    | \( \mathbb{0.43} \) | 0.84    |
| Cf. + \( \Delta z + \mathcal{N} + T \) | 0.58     | \( \mathbb{1.00} \) | 0.39     | 1.00    | \( \mathbb{0.35} \) | 0.99    | 0.37     | 0.96    | 0.53    | 0.97    |

Table 3: The transferability of the adversarial point clouds by original Chamfer+kNN loss. The horizontal column is the substitute model, the same below.

| subs. | victim | PointNet | PN++(SSG) | PN++(MSG) | DGCNN | CurveNet |
|-------|--------|----------|-----------|-----------|-------|----------|
| PointNet | 1.00   | 0.07     | 0.10      | 0.12      | 0.11  |
| PN++(SSG) | 0.26   | 1.00     | 0.18      | 0.19      | 0.21  |
| PN++(MSG) | 0.21   | 0.17     | 1.00      | 0.19      | 0.09  |
| DGCNN | 0.27   | 0.21     | 0.30      | 1.00      | 0.13  |
| CurveNet | 0.37   | 0.36     | 0.45      | 0.43      | 1.00  |

Table 4: The transferability of the adversarial point clouds by 3D-TI Chamfer+kNN loss.

### Conclusion
This paper proposes a physical-world adversarial attack against the 3D face recognition system. We propose to use random 3D transformation to improve the physical attack’s robustness and a series of techniques to inverse the SLI system. The limitation is that we suppose the SLI algorithm is a white box, which is a strong assumption for real-world attacks. In experiments, we first evaluate our attack on several state-of-the-art 3D deep learning models and show that our attack can successfully attack the real-world system with very few modifications to projected patterns. Then we prove the effectiveness of the 3D-TI module to the physical attack’s
robustness. Last but not least, we show that our attack can improve the adversarial point clouds’ transferability.

References

Cao, Y.; Xiao, C.; Cyr, B.; Zhou, Y.; Park, W.; Rampazzi, S.; Chen, Q. A.; Fu, K.; and Mao, Z. M. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In Proceedings of the ACM SIGSAC conference on computer and communications security, 2267–2281.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp), 39–57. Ieee.

Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; and Zhu, J. 2019. Efficient decision-based black-box adversarial attacks on face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7714–7722.

Feng, S.; Zu, C.; Zhang, L.; Tao, T.; Hu, Y.; Yin, W.; Qian, J.; and Chen, Q. 2021. Calibration of fringe projection profilometry: A comparative review. Optics and Lasers in Engineering, 143: 106622.

Gnanasambandam, A.; Sherman, A. M.; and Chan, S. H. 2021. Optical Adversarial Attack. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 92–101.

Hamdi, A.; Rojas, S.; Thabet, A.; and Ghanem, B. 2020. Advpc: Transferable adversarial perturbations on 3d point clouds. In European Conference on Computer Vision, 241–257. Springer.

Huang, Q.; Dong, X.; Chen, D.; Zhou, H.; Zhang, W.; and Yu, N. 2022. Shape-invariant 3D Adversarial Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15335–15344.

Min, R.; Kose, N.; and Dugelay, J.-L. 2014. Kinectfaceb: A kinect database for face recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44(11): 1534–1548.

Nguyen, D.-L.; Arora, S. S.; Wu, Y.; and Yang, H. 2020. Adversarial light projection attacks on face recognition systems: A feasibility study. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 814–815.

Nichols, N.; and Jasper, R. 2018. Projecting trouble: Light based adversarial attacks on deep learning classifiers. arXiv preprint arXiv:1810.10337.

Phan, A. V.; Le Nguyen, M.; Nguyen, Y. L. H.; and Bui, L. T. 2018. Dgcn: A convolutional neural network over large-scale labeled graphs. Neural Networks, 108: 533–543.

Piccirilli, M.; Doretto, G.; Ross, A.; and Adjeroh, D. 2016. A Mobile Structured Light System for 3D Face Acquisition. IEEE Sensors Journal, 16(7): 1854–1855.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, 30.

Savran, A.; Alyüz, N.; Dibeklioğlu, H.; Çeliktutan, O.; Gökbek, B.; Sankur, B.; and Akarun, L. 2008. Bosphorus database for 3D face analysis. In European workshop on biometrics and identity management, 47–56. Springer.

Sharif, M.; Bhagavatula, S.; Bauer, L.; and Reiter, M. K. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the 2016 acm sigsac conference on computer and communications security, 1528–1540.

Tsai, T.; Yang, K.; Ho, T.-Y.; and Jin, Y. 2020. Robust adversarial objects against deep learning models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 954–962.

Tu, J.; Ren, M.; Manivasagam, S.; Liang, M.; Yang, B.; Du, R.; Cheng, F.; and Urtasun, R. 2020. Physically realizable adversarial examples for lidar object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13716–13725.

Wen, Y.; Lin, J.; Chen, K.; Chen, C. L. P.; and Jia, K. 2022. Geometry-Aware Generation of Adversarial Point Clouds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2984–2999.

Wicker, M.; and Kwiatkowska, M. 2019. Robustness of 3D Deep Learning in an Adversarial Setting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Worzyk, N.; Kahlen, H.; and Kramer, O. 2019. Physical adversarial attacks by projecting perturbations. In International Conference on Artificial Neural Networks, 649–659. Springer.

Wu, Z.; Zu, C.; Guo, W.; Tao, T.; and Zhang, Q. 2019. High-speed three-dimensional shape measurement based on cyclic complementary Gray-code light. Optics express, 27(2): 1283–1297.

Xiang, C.; Qi, C. R.; and Li, B. 2019. Generating 3d adversarial point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9136–9144.

Xiang, T.; Zhang, C.; Song, Y.; Yu, J.; and Cai, W. 2021. Walk in the Cloud: Learning Curves for Point Clouds Shape Analysis. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 895–904.

Xiao, Z.; Gao, X.; Fu, C.; Dong, Y.; Gao, W.; Zhang, X.; Zhou, J.; and Zhu, J. 2021. Improving transferability of adversarial patches on face recognition with generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11845–11854.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2730–2739.

Zheng, T.; Chen, C.; Yuan, J.; Li, B.; and Ren, K. 2019. Pointcloud saliency maps. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 1598–1606.
Zhou, S.; and Xiao, S. 2018. 3D face recognition: a survey. *Human-centric Computing and Information Sciences*, 8(1): 1–27.

Zhou, Z.; Tang, D.; Wang, X.; Han, W.; Liu, X.; and Zhang, K. 2018. Invisible mask: Practical attacks on face recognition with infrared. *arXiv preprint arXiv:1803.04683*. 