Physical-World Optical Adversarial Attacks on 3D Face Recognition
Session TAG: THU-PM-392

Yanjie Li,[1] Yiquan Li,[1] Xuelong Dai,[1] Songtao Guo,[2] Bin Xiao[1]
[1] The Hong Kong Polytechnic University [2] Chongqing University

May 30, 2023

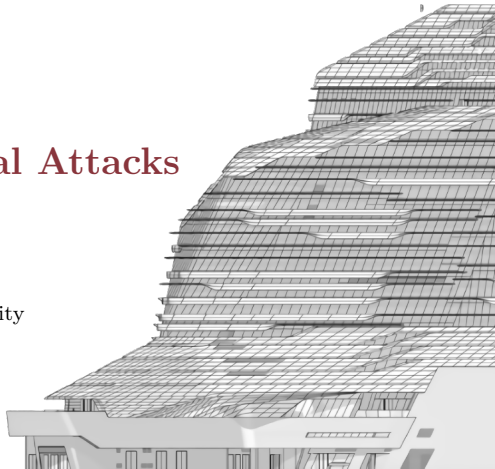


Table of Contents

Motivations

- A lack of study on the security of 3D face recognition systems against physical-realizable adversarial attacks.
- Present physical 3D adversarial attacks can only generate adversarial points adjacent to the surface, limiting the attack success rate.
- Present physical 3D adversarial examples are not resistant to random rotation and transition, which are not practical in the real world.

Overview

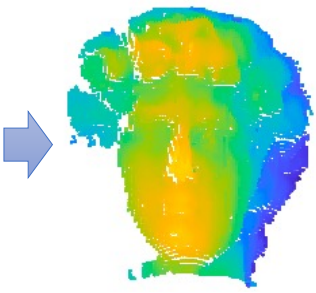
1 Introduction

Normal Modulated Image

David

Adversarial Modulated Image

Athena



Adversarial Point Cloud

Figure: A demonstration of our attack. We project optical noises on the 3D faces to generate adversarial point clouds. Our attack modifies fewer points than previous attacks and does not need the adversarial points to be adjacent to the 3D surface.

Contributions

- We are the first to realize the end-to-end physical adversarial attack against 3D face recognition through adversarial illuminations.
- Our attack can generate adversarial points at any position by utilizing the 3D reconstruction principle.
- Our attack involves random 3D transformations in the attack pipeline, which significantly improve adversarial examples' robustness to random movements.
- We attack both point-cloud-based and depth-image-based 3D face recognition models. Compared with previous attacks, our method needs fewer perturbations with a high success rate in experiments.

Table of Contents

Structured-light-based 3D reconstruction

- Structured light imaging is a popular method of acquiring 3D face data due to its high precision and superiority in uniform textures.
- According to the 3D reconstruction principles, it can be classified as the phase-shifting-based algorithms and the DNN-based algorithms.

Figure: Acquiring the phase shift image

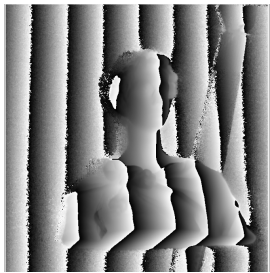


Figure: Calculating the phase map

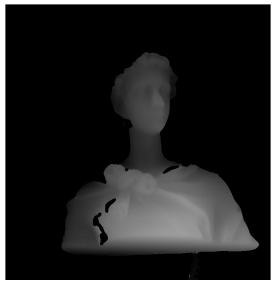


Figure: Getting the depth image

Table of Contents

3 Methodology

Overview of attacks

- We propose **phase shifting attack** and **phase superpositon attack** for phase-shifting-based and DNN-based 3D reconstruction algorithms respectively.

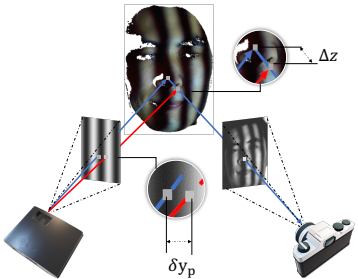


Figure: Phase shifting attack

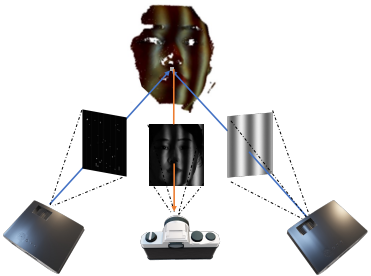


Figure: Phase superposition attack

Phase shifting attack

3 Methodology

- Directly modify the projected patterns.
- The depth changes are mapped to pixel transition in the projected images through an opposite process of 3D reconstruction.
- We project the adversarial point displacements onto the normal vector of the camera imaging plane.
- This projection can make the 3D pixel shifting only change the depth in the camera view.

Algorithm 1 Phase Shifting Attack Algorithm

Input: $I_c = \{I_1^c, ..., I_N^c\}, I_p, A_c = K[R\ T]$, target label t
Output: $I_p^{adv} = \{I_1^p, ..., I_N^p\}$

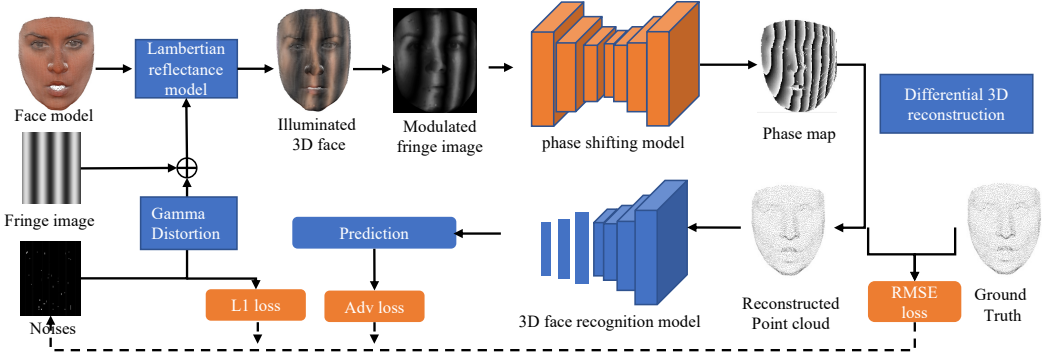
$\phi_a \leftarrow f(I_c)$
for $i = 0$ to N **do**
  Reconstruct the point cloud $\mathcal{P} \leftarrow h(\phi_a)$
  $\mathcal{P}' \leftarrow K \cdot R \cdot \mathcal{P}$
  Compute the gradient $\nabla_{\mathcal{P}'} l_{adv}(\mathcal{P}', t)$
  $\nabla_{\mathcal{P}'} l_{adv}(\mathcal{P}')[:, 0:2] \leftarrow 0$
  $\nabla_{\phi_a} l_{adv}(\phi_a) \leftarrow \nabla_{\mathcal{P}'} l_{adv}(\mathcal{P}') \cdot \frac{\partial \mathcal{P}'}{\partial \phi_a}$
  Compute the total gradient $\triangle = \nabla_{\phi_a} l_{total}(\phi_a)$
  $\phi_a \leftarrow clip(\phi_a + \alpha \cdot \frac{\triangle}{\|\triangle\|_2})$
end for
$u_p' \leftarrow round(\frac{w\phi_a}{2\pi n_s})$
$I_p^{adv}(u_p) \leftarrow I_p(u_p')$
return $I_p^{adv} = \{I_1^p, ..., I_N^p\}$

Figure: Phase shifting attack algorithm

Phase superposition attack

3 Methodology

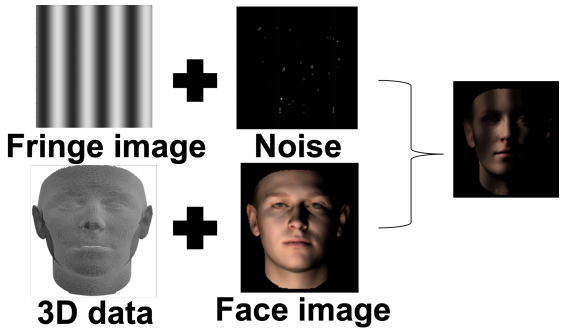
- In the real world, the adversary usually cannot directly modify the projected image.
- The adversary uses an additional projector to project additive noises on the faces, resulting in dodging or impersonation attacks.



Lambertian reflection model

- We use the Lambertian reflection model to simulate the real-world face-lighting process.
- The process can be formulated as

$$I(\boldsymbol{x}) = a(\boldsymbol{x})\boldsymbol{n}(\boldsymbol{x}) \cdot (\boldsymbol{s}_{p_1}(\boldsymbol{x}) + \boldsymbol{s}_{p_2}(\boldsymbol{x})). \tag{1}$$



3D transform invariant loss

- When attacking a physical system, the distance and head may move unexpectedly.
- We propose 3D transform invariant loss to make the adversarial point clouds can generate consistent results when the head rotates or moves.

$$\mathcal{T}(\mathcal{P}) = (\mathbf{R}_{(\theta_x,\theta_y,\theta_z)}\mathcal{P}^T)^T + \mathbf{M}_{(\eta_x,\eta_y)} \tag{2}$$

$$l_{adv}(\phi'_a) = \mathbb{E}_{\mathcal{T}} l_{logits}\left(\mathcal{M}\left(\mathcal{N}\left(\mathcal{T}(h(\phi'_a))\right)\right)\right), \tag{3}$$

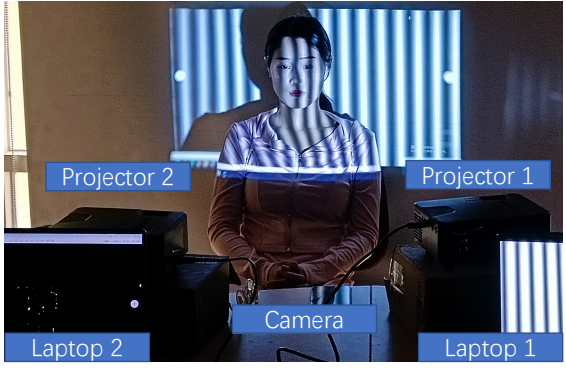
Rotation Transition

Table of Contents

4 Experiment results

Experiment setup

- Datasets: Bosphrus, Eurecom, SIAT-3DFE
- 3D face recognition model: Pointnet, Pointnet++, DGCNN, CurveNet
- Compared attacks: C&W, 3DAdv, KNNadv, GeoA3
- Physical settings: one industry camera, two home projectors



Digital attack results

4 Experiment results

- We evaluate the attack performance by the attack success rate (higher is better) and RMSE (lower is better). RMSE is used to measure the noise size.
- Compared with the state-of-art geometry-ware attack GeoA3, our attacks have fewer 3D reconstruction errors while maintaining a high attack success rate.

	PointNet		PointNet++(SSG)		Point++(MSG)		DGCNN	
Metrics	ASR (%)	RMSE	ASR (%)	RMSE	ASR (%)	RMSE	ASR (%)	RMSE
C&W	**0.98**	0.35	0.92	0.45	0.88	0.36	0.94	0.26
3Dadv	0.89	0.27	0.86	0.34	0.77	0.25	0.89	0.16
KNNadv	0.86	0.15	0.85	0.22	0.79	0.23	0.85	0.08
GeoA3	0.75	0.18	0.91	**0.14**	0.85	0.13	0.84	0.09
Ours	0.95	**0.13**	**0.93**	0.15	**0.99**	**0.11**	**0.96**	**0.05**

Table: The untargeted attack performance. We evaluate the attack performance by the attack success rate (higher is better) and RMSE (lower is better).

Physical attack results

4 Experiment results

- We attack 10 people in physical attacks and achieved 90% ASR for untargeted attacks and 40% ASR for targeted attacks for both phase shifting and superposition attacks.
- Compared with the state-of-art geometry-ware attack GeoA3, our attacks have fewer 3D reconstruction errors while maintaining a high attack success rate.

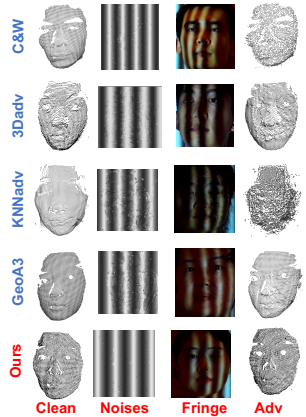
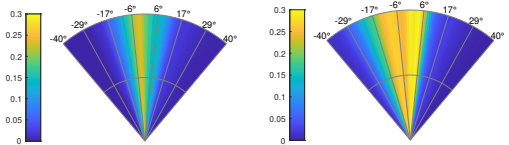


Figure: The physical phase shifting attack results.

Ablation study

4 Experiment results



(a) Without 3D-TI (b) With 3D-TI

Figure: The adversarial examples' robustness against face rotations. We generate perturbations on Pointnet and then add them to the rotated point cloud. We plot the predictions on the target label without (left) and with (right) 3D-TI module.

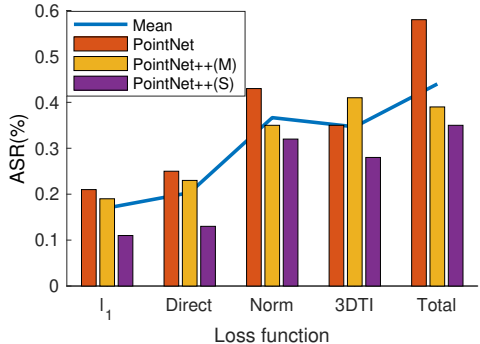


Figure: The ASR results of ablation study. *Direct* means the direction constraint. *Norm* is the renormalization. $3DTI$ means 3D transform invariant loss.

Q&A

Thank you for listening!
Your feedback will be highly appreciated!