CONTROLLABLE EVALUATION AND GENERATION OF PHYSICAL ADVERSARIAL PATCH ON FACE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have revealed the vulnerability of face recognition models against physical adversarial patches, which raises security concerns about the deployed face recognition systems. However, it is still challenging to ensure the reproducibility for most attack algorithms under complex physical conditions, which leads to the lack of a systematic evaluation of the existing methods. It is therefore imperative to develop a framework that can readily and fairly evaluate the vulnerability of face recognition in the physical world. To this end, we propose to simulate the complex transformations of faces in the physical world via 3D face modeling, which serves as a digital counterpart of physical faces. The generic framework allows us to control different face variations and physical conditions to conduct reproducible evaluations conveniently. With this digital simulator, we further propose a **Face3DAdv** method considering the 3D face transformations and realistic physical variations. Extensive experiments validate that Face3DAdv can significantly improve the effectiveness of diverse physically realizable adversarial patches in both simulated and physical environments, against various white-box and black-box face recognition models.

1 INTRODUCTION

Face recognition, as a prevailing task in computer vision, has experienced substantial improvements thanks to the rapid development of deep neural networks (DNNs) (Deng et al., 2019a; Wen et al., 2016). DNNs facilitate the broad application of face recognition in various safety-critical fields, including finance/payment, public access, surveillance, etc. However, face recognition models based on DNNs are vulnerable to *adversarial examples* (Sharif et al., 2016; Dong et al., 2019; Yang et al., 2020b; Komkov & Petiushko, 2021; Tong et al., 2021) — maliciously generated inputs to mislead a target model, which may lead to serious consequences or security problems in real-world applications.

Extensive efforts have been devoted to studying the generation of adversarial examples on face recognition models, which can be conducive to investigating model robustness (Yang et al., 2020b; Tong et al., 2021). Some work (Dong et al., 2019; Yang et al., 2021) has proposed to apply minimal perturbations (measured by the ℓ_p norm) to face images in the *digital* world, aiming to evade being recognized or to impersonate another identity. However, practical face recognition systems usually process face photos taken in the *physical* world. Thus, it is of particular importance to explore physical adversarial attacks to identify the weaknesses of these models before they are deployed. To this end, some typical approaches generate various adversarial patches (Brown et al., 2017) that are wearable on faces, including eyeglass (Sharif et al., 2016; 2017), hats (Komkov & Petiushko, 2021), and stickers (Shen et al., 2021). They can take effect in deceiving the *unattended* payment system of vending machines (GeekPwn, 2020) and unlocking a mobile phone or car (Technologies, 2020).

Despite the success, the existing physical attack methods on face recognition still have several limitations. First, there is no systematic testing protocol for physical attacks. The evaluation is usually conducted by asking a few volunteers to attach the adversarial patches, followed by testing in a specific environment (e.g., printer, viewpoints, lighting conditions) (Sharif et al., 2016; Komkov & Petiushko, 2021; Shen et al., 2021; Zheng et al., 2021; Zolfi et al., 2021a), making it hard to evaluate and compare the effectiveness of different methods. The inconsistent experimental settings and potential bias by uncontrolled volunteers also limit the reproducibility of physical adversarial examples in different conditions. Second, most methods aim to craft adversarial examples robust to

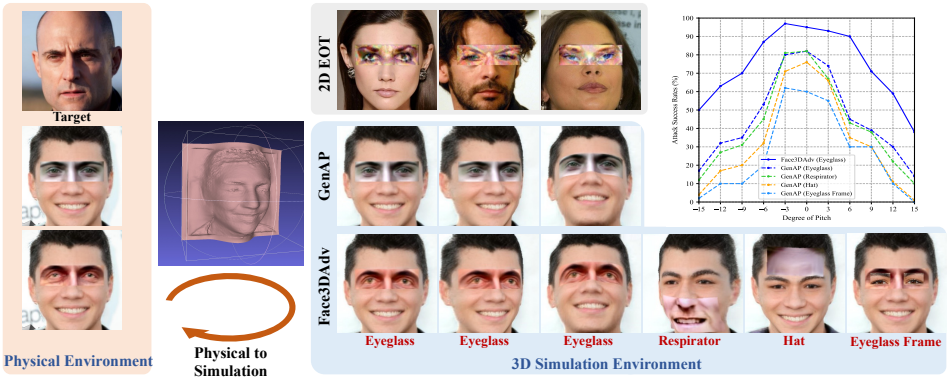


Figure 1: A controllable and high-fidelity simulator via 3D face modeling for evaluating the performance of different physical attacks, e.g., GenAP (Xiao et al., 2021) and our Face3DAdv. The simulation framework provides multiple physically realizable attacks, including Eyeglass (Xiao et al., 2021), Respirator (Tong et al., 2021), Hat (Komkov & Petiushko, 2021), and Eyeglass frame (Sharif et al., 2016; 2017), which have verified their practicality in many *unattended* recognition scenarios. We further demonstrate consistent effectiveness of our method, against ArcFace (Deng et al., 2019b) over $[-15, 15]$ degrees of the pose of *pitch* under a certain degree of *yaw*. More results regarding effectiveness and practicality are presented in Sec. 5.

varying physical conditions by optimizing over 2D image transformations (Athalye et al., 2018), such as rotation, translation and additive Gaussian noise, but they fail to consider other physical variations of 3D faces, such as viewpoint and lighting.

In this paper, we build a novel *simulation framework* that can reflect the characteristics of physical faces, enabling us to readily and fairly evaluate different physical attacks on face recognition. The framework fully leverages recent advances (Deng et al., 2020; Henderson et al., 2020; Shi et al., 2021) in 3D face modeling, which have demonstrated that they could generate controllable high-fidelity virtual face images that are even hard to distinguish from real ones. Based on this, the accessible and general framework has the ability to pair attackers acting in a simulated environment with counterparts acting in a realistic physical environment (as demonstrated in Sec. 5 experimentally). Specifically, we first embed a single-view testing face image onto the latent manifold of the pre-trained 3D generator (Shi et al., 2021) and reconstruct its 3D face information, including texture, shape, viewpoint, and lighting. Then, we propose a texture-based adversarial attack paradigm to generate a 3D adversarial face, which can naturally stitch a patch onto the face to make the adversarial patch more versatile and realistic. Finally, after introducing a differentiable renderer (Ravi et al., 2020), we can obtain 2D adversarial faces under diverse physical variations. Once informed with such a simulator (as shown in Fig. 1), future researchers can have priority to *conveniently* and *easily* conduct hundreds of experimenters in a controllable virtual environment for identifying problems and improving their physical attacks, whereas previous work (Komkov & Petiushko, 2021; Shen et al., 2021; Zheng et al., 2021) usually chooses no more than **10** volunteers for physical experiments.

Based on this simulation framework, the attacker has the ability to develop more reliable physical adversarial attacks by controlling the simulated environments, thus enabling the crafted adversarial patches to be more robust to physical transformations. To demonstrate this, we propose a **Face3DAdv** attack method to generate robust adversarial patches by optimizing over diverse physical transformations in adversarial scenarios based on the simulation framework. Moreover, since the physical variations are much more abundant in our method, we adopt a more effective strategy to focus on favorable transformations within a principled optimization framework. As a comparison, the previous methods (Athalye et al., 2018; Sharif et al., 2016) typically select physical transformations fully at random to optimize robust perturbations, without considering the different importance of physical variations. Extensive experiments demonstrate that our Face3DAdv achieves consistent improvements in both simulated and physical environments. As for imperceptibility, 3D adversarial patches crafted by Face3DAdv are also more conducive to steadily passing defensive mechanism (*commercial Face Anti-spoofing API*) in automatic face recognition systems.

To the best of our knowledge, this is the first attempt that conducts a reproducible physical-world adversarial attacks on face recognition, especially including 3D face recognition models. Our contributions can be summarized as: (i) We develop a reproducible simulation framework via 3D face modeling for readily and fairly evaluating the performance of physical attacks on face recognition, which can conveniently simulate the complex transformations of faces in the physical world; (ii) We

propose a 3D-aware attack method — Face3DAdv to generate robust adversarial patches, showing significant improvements over the previous methods with a particular focus on diverse physical conditions of 3D transformations, lighting variations, etc.

2 Related Work

Adversarial attacks in the physical world. Recent work has shown that adversarial examples (Goodfellow et al., 2015; Szegedy et al., 2014) can exist in the physical world (Kurakin et al., 2017; Athalye et al., 2018), resulting in an emerging threat. In particular, adversarial patches (Brown et al., 2017) only perturb a small cluster of pixels, and can be applied to real objects in the physical world (Eykholt et al., 2018; Zhao & Stamm, 2020; Xu et al., 2020; Zolfi et al., 2021b; Yang et al., 2020a). Adversarial patches on face recognition have also been explored (Sharif et al., 2016; Brown et al., 2017). By attaching a carefully generated patch to the face, some studies (Pautov et al., 2019; Komkov & Petiushko, 2021) have shown success of physical attacks against the state-of-the-art face recognition models. However, these methods did not consider the face variations in the physical world, thus resulting in performance degeneration in real testing scenarios. Meanwhile, existing physical attacks commonly use EOT (Athalye et al., 2018) by randomly sampling the transformations during optimization without considering the different importance for the diverse physical variations.

3D face modeling. As one of the popular 3D face modeling mechanisms, 3D Morphable Model (3DMM) is commonly adopted to represent faces (Tuan Tran et al., 2017), which are parameterized by identity, expression, and illumination. Although 3DMM offers control over the semantic parameters, it suffers from photorealism and models only the essential parts of a portrait image (e.g., hair, mouth interior, background). More recent work reconstructs plausible 3D face shapes by exploiting knowledgeable parameter metrics of 3DMM (Tewari et al., 2020; Deng et al., 2020). On the other hand, some face representation methods leverage 3D position maps (Shi et al., 2021; Henderson et al., 2020) to represent and output the mesh of the target, and achieve the controllable parametric nature of existing face models. Therefore, we can construct a flexible environment that simulates the physical world with the aid of these blossoming techniques on 3D face modeling. The digital surrogate of a real face provides us a possible solution to conduct reliable and reproducible evaluation for facilitating physical attacks.

3 Robust Evaluation for Physical Attacks

The current physical attacks on face recognition (Sharif et al., 2016; Guo et al., 2021; Komkov & Petiushko, 2021) are usually evaluated by: 1) printing adversarial patches (e.g., eyeglass frames, hats, etc); 2) asking a few volunteers to attach them; and 3) testing the attack performance under a specific environment. However, the evaluation methodology is insufficient due to the lack of a systematic testing protocol. The experimental settings (including printer, chosen volunteers, and physical environment) are obviously inconsistent across different research, making it hard to compare and evaluate the effectiveness of existing methods.

To address this problem, it is imperative to develop a reproducible framework for readily and fairly evaluating the performance of physical adversarial attacks on face recognition. We advocate using a simulator rather than performing experiments in the physical world for the following reasons: 1) **completeness:** the simulator can provide a complete picture of the effectiveness of different attack methods given various controllable physical conditions; 2) **fair and replicable comparisons:** based on the same simulator, the comparisons between different attacks are fairer, and the evaluation results are replicable; 3) **cheap and easy to support large experiments:** conducting experiments on the simulation framework is much cheaper and easier, which supports larger-scale testing experiments.

As illustrated in Fig. 2, our simulation framework consists of four modules: 3D face modeling, adversarial generation, rendering, and evaluation. We introduce the details of these four steps in the following discussion.

3D face modeling. Given a 2D face image $\boldsymbol{x} \in \mathcal{X}$, we aim to generate the corresponding *realistic* 3D face representation that can be easily manipulated, by exploiting 3D parametric fitting in 3D face modeling. The 3D face is expected to approximate a real face in the physical world. In this paper, we leverage the state-of-the-art pre-trained 3D generator of $\mathcal{G}_{3D}$ (Shi et al., 2021) for 3D face modeling, which can disentangle the generation process of a 2D generator $\mathcal{G}_{2D}$ instantiated by StyleGAN (Karras et al., 2019) into different 3D modules for a 3D shape representation. Therefore, a

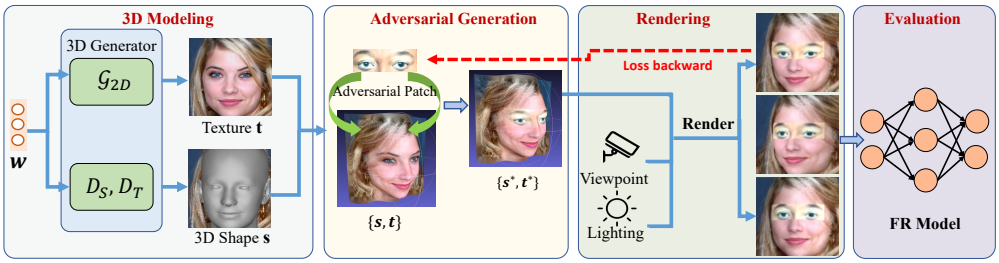


Figure 2: The overall simulation framework for evaluating and developing physical attacks include four modules of **3D Modeling**, **Adversarial Generation**, **Rendering**, and **Evaluation**. **3D Modeling** reconstructs a 3D face $\{\boldsymbol{s}, \boldsymbol{t}\}$ that can be manipulated, including a 2D texture face $\boldsymbol{t}$ from the generative model $\mathcal{G}_{2D}$ and a shape representation $\boldsymbol{s}$ from deep networks of shape D_S and transformation map D_T. **Adversarial Generation** adopts a texture-based adversarial method to apply a patch to a certain region to generate an adversarial 3D face $\{\boldsymbol{s}^*, \boldsymbol{t}^*\}$. **Rendering** adopts a renderer to produce a series of 2D rendered adversarial faces given $\{\boldsymbol{s}^*, \boldsymbol{t}^*\}$, different viewpoints and lighting conditions. **Evaluation** aims to test the performance by feeding the rendered images into the face recognition model.

3D face representation can be obtained given a style code $\boldsymbol{w}$, including a 3D shape representation of $\boldsymbol{s}$ and a 2D texture face of $\boldsymbol{t}$ from $\mathcal{G}_{2D}$.

Given a face recognition model $f(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^d$ and a random initialization parameter of $\boldsymbol{w}$, we propose to search for the optimal parameter of $\boldsymbol{w}$ for the generator by minimizing the distance between the original face image and the rendered image of $\boldsymbol{x}'$ as

$$\min_{\boldsymbol{w}} \mathcal{D}_f(\boldsymbol{x}', \boldsymbol{x}) + \lambda \|\boldsymbol{x}' - \boldsymbol{x}\|_1, \tag{1}$$

where $\boldsymbol{x}' := \mathrm{R}(\mathcal{G}_{3D}(\boldsymbol{w}); V_0, L_0)$ with R being a differentiable renderer, and V_0 and L_0 are corresponding parameters of neutralized viewpoint and lighting; and λ is a balancing hyperparameter. We adopt the ℓ_1 norm in the objective since the ℓ_2 norm can lead to blurry textures (Huh et al., 2020). $\mathcal{D}_f$ computes the distance of the feature representations of f as

$$\mathcal{D}_f(\boldsymbol{x}', \boldsymbol{x}) = \|f(\boldsymbol{x}') - f(\boldsymbol{x})\|_2^2. \tag{2}$$

By optimizing the objective function (1), we can obtain the optimal $\boldsymbol{w}^*$ and get the 3D face as $\{\boldsymbol{s}, \boldsymbol{t}\} = \mathcal{G}_{3D}(\boldsymbol{w}^*)$.

Adversarial generation. The next step is to apply the adversarial examples to the 3D face model. The existing attack methods usually adopt texture-based adversarial patches, i.e., for a face image $\boldsymbol{x}$, these methods can generate an adversarial face image $\boldsymbol{x}^*$ by applying a patch to a certain region. Since they do not modify the face shape, we directly replace the texture $\boldsymbol{t}$ of the original face image $\boldsymbol{x}$ as $\boldsymbol{t}^* = \boldsymbol{x}^*$. Notably, our framework can also perform adversarial attacks on face shapes if necessary for new attackers.

Rendering 2D images with transformations. Given a 3D adversarial face $\{\boldsymbol{s}, \boldsymbol{t}^*\}$, we can adopt a renderer (Ravi et al., 2020) to produce 2D rendered adversarial faces given different viewpoints and lighting conditions. Specifically, we choose a set of viewpoints $V = \{V_i\}_{i=1}^{N_v}$ and lighting variations $L = \{L_j\}_{j=1}^{N_l}$, and then render an adversarial image as

$$\boldsymbol{r}^*_{(i,j)} = \mathrm{R}(\boldsymbol{s}, \boldsymbol{t}^*; V_i, L_j). \tag{3}$$

We can also apply some 2D image transformations (e.g., rotation, translation, scaling, etc) to $\boldsymbol{r}^*_{(i,j)}$.

Evaluation. The final step is to evaluate the performance of the attack by feeding the rendered adversarial images into the face recognition model. For different tasks, we can evaluate the attack performance in different ways.

Discussion on the printer. Previous research (Thys et al., 2019; Xu et al., 2020; Zheng et al., 2021) has studied the color deviation between the digital image and its printed version by mapping a digital color spectrum to printed counterpart or adopting non-printability losses. However, we mainly focus on constructing a simulation framework for evaluating physical adversarial attacks. The previous approaches are generally compatible with our framework in the physically realizable procedure.

4 FACE3DADV

In this section, we propose a **Face3DAdv** method to exploit the various physical transformations.

4.1 PRELIMINARY

Face recognition usually has two sub-tasks: face verification and face identification (Huang et al., 2007). We mainly consider face verification in this paper, while the proposed approach can be naturally extended to face identification. In face verification, the feature distance between a pair of images $\{\boldsymbol{x}^a, \boldsymbol{x}^b\} \subset \mathcal{X}$ is first calculated as $\mathcal{D}_f(\boldsymbol{x}^a, \boldsymbol{x}^b)$. Then the prediction of face verification can be formulated as

$$\mathcal{C}(\boldsymbol{x}^a, \boldsymbol{x}^b) = \mathbb{I}(\mathcal{D}_f(\boldsymbol{x}^a, \boldsymbol{x}^b) < \delta), \tag{4}$$

where $\mathbb{I}$ is the indicator function, and δ is a threshold. When $\mathcal{C}(\boldsymbol{x}^a, \boldsymbol{x}^b) = 1$, the two images are recognized as the same identity, otherwise different identities. Note that this definition is consistent with the commonly used cosine similarity metric, since f outputs a normalized feature.

Given the original face images $\boldsymbol{x}^a$ and $\boldsymbol{x}^b$, we aim to generate an adversarial image $\boldsymbol{x}^*$ by adding a perturbation to $\boldsymbol{x}^a$ to mislead the face recognition model when recognizing $\boldsymbol{x}^*$ and $\boldsymbol{x}^b$. There are generally two types of adversarial attacks on face recognition: *dodging* and *impersonation*. A dodging attack aims to make the face recognition model fail to recognize the identity of $\boldsymbol{x}^*$, i.e., to make $\mathcal{C}(\boldsymbol{x}^*, \boldsymbol{x}^b) = 0$ while $\mathcal{C}(\boldsymbol{x}^a, \boldsymbol{x}^b) = 1$; an impersonation attack aims to make the face recognition model recognize $\boldsymbol{x}^*$ as a specific identity, i.e., to make $\mathcal{C}(\boldsymbol{x}^*, \boldsymbol{x}^b) = 1$ while $\mathcal{C}(\boldsymbol{x}^a, \boldsymbol{x}^b) = 0$.

4.2 3D-AWARE ADVERSARIAL ATTACK

To facilitate the physical realizability of the adversarial examples, we study adversarial patches that are restricted to a specifically designed region. Although some elaborate adversarial patches (Komkov & Petiushko, 2021; Xiao et al., 2021) consider 2D image transformations, they do not take into account other realistic 3D physical transformations, thus leading to inevitable degeneration of their effectiveness. To make the crafted adversarial patch more versatile and effective in the real world, we optimize the adversarial patch over both the common 2D transformations and the newly considered 3D transformations. Based on our simulation framework, we can readily optimize the adversarial patches over 3D transformations. Therefore, the attack objective function of crafting adversarial examples can be formulated as

$$\begin{aligned} &\min_{\boldsymbol{s}^*, \boldsymbol{t}^*} \ \mathbb{E}_{V_i \sim V, L_j \sim L}[\mathcal{J}_f(\mathrm{R}(\boldsymbol{s}^*, \boldsymbol{t}^*; V_i, L_j), \boldsymbol{x}^b)], \\ &\text{s.t.} (\mathbf{1} - \mathrm{M}) \odot \mathrm{R}(\boldsymbol{s}^*, \boldsymbol{t}^*; V_i, L_j) = (\mathbf{1} - \mathrm{M}) \odot \mathrm{R}(\boldsymbol{s}^a, \boldsymbol{t}^a; V_i, L_j), \end{aligned} \tag{5}$$

where $M \in \{0, 1\}^n$ is a binary mask to apply the perturbations to pixels where the value of the mask is 1, $\odot$ is the element-wise multiplication operation, $\{\boldsymbol{s}^a, \boldsymbol{t}^a\}$ is the 3D face obtained by optimizing Eq. (1) given a 2D face image $\boldsymbol{x}^a$ and $\mathcal{J}_f$ is the attack loss. In this paper, we adopt $\mathcal{J}_f = -\mathcal{D}_f$ for a dodging attack and $\mathcal{J}_f = \mathcal{D}_f$ for an impersonation attack. Since 2D transformations (Xie et al., 2019) are generally compatible with the objective (5), we can craft a 3D adversarial face to fool the face recognition systems for diverse 2D and 3D face transformations.

Mapping shape representation. Note that the optimization problem (5) is constrained, which must ensure that the shape representation $\boldsymbol{s}^*$ of the 3D adversarial face is only modified in the designed region in every optimization step. However, this can give rise to the inconsistency of $\boldsymbol{s}^*$ in the designed mask region and original face representation $\boldsymbol{s}^a$ after a long optimization trajectory, consequently leading to inevitable performance degradation due to shape disharmony of the whole 3D face. To address this issue, we directly reduce the optimization space of the 3D adversarial face by replacing $\boldsymbol{s}^*$ with $\boldsymbol{s}^a$ in the optimization. In this way, we can entirely restrict the 3D adversarial face in a prior fixed shape, and only optimize the texture map $\boldsymbol{t}^*$.

Given the objective function in Eq. (5), we can iteratively apply the fast gradient method (Kurakin et al., 2017) with a small step size α to generate adversarial examples. In particular, we optimize the adversarial texture image $\boldsymbol{t}^*$ via

$$\boldsymbol{t}^*_{k+1} = \Pi_{D^t}\left(\boldsymbol{t}^*_k - \alpha \cdot \mathrm{M} \odot \operatorname{sign}\left(g_{k+1}\right)\right), \tag{6}$$

where g_{k+1} is the updated gradient at the $(k+1)$-th iteration, and Π_{D^t} is the projection function that projects the adversarial images onto the $D^t = \{\boldsymbol{t} : \|\mathrm{M} \odot \boldsymbol{t} - \mathrm{M} \odot \boldsymbol{t}^a\|_\infty \leq \epsilon\}$. We call it **Face3DAdv ($\boldsymbol{x}$)**. Besides, $\boldsymbol{t}^*$ can be optimized in the latent space $\boldsymbol{w}^*$ in $\mathcal{G}_{2D}$ by following a state-of-the-art transferable adversarial method (Xiao et al., 2021) on face recognition, which can be formulated as $\boldsymbol{t}^* = \mathcal{G}_{2D}(\boldsymbol{w}^*)$. And $\boldsymbol{w}^*$ can be optimized by adopting a popular optimizer, such as Adam (Kingma & Ba, 2015), which is called **Face3DAdv ($\boldsymbol{w}$)**.

Algorithm 1 Face3DAdv

Require: A pre-trained 3D generative model $\mathcal{G}_{3D}$, a FR model f, a real face image $\boldsymbol{x}^a$, a target face image $\boldsymbol{x}^b$, 2D transformation function T.
Ensure: Adversarial image t^*.
for iter in MaxIterations N_1 **do** ▷ *Stage I: Obtain a 3D face*
  Initialize latent code $\boldsymbol{w} = \boldsymbol{w}_0$;
  Obtain $\mathcal{J}$ from Eq. (1);
  $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla_w \mathcal{J}$;
end for
Forward pass the optimal $\boldsymbol{w}^*$ into $\mathcal{G}_{3D}$ to the 3D face $\{\boldsymbol{s}^a, \boldsymbol{t}^a\}$;
Initializing $\boldsymbol{t}_0^* = \boldsymbol{x}^b$; ▷ *Stage II: Optimize t^**
for k in MaxIterations N_2 **do**
  $\boldsymbol{t}_k^* = \boldsymbol{t}^a \odot (\mathbf{1} - \mathrm{M}) + \boldsymbol{t}_k^* \odot \mathrm{M}$;
  Construct 3D adversarial face $\{\boldsymbol{s}^a, \boldsymbol{t}_k^*\}$;
  Get importance probability $\hat{P}_{i,j}$ from Eq. (7);
  Draw M rendered images $\{\boldsymbol{r}_{k,m}^*\}_{m=1}^M$ according to $\hat{P}$;
  Obtain the gradient $\boldsymbol{g}_{k+1} = \nabla_{\boldsymbol{t}} \mathcal{J}_f(\sum_m T_m(\boldsymbol{r}_{k,m}^*), \boldsymbol{x}^b)$;
  Update $\boldsymbol{t}_{k+1}^*$ via Eq. (6);
end for

4.3 Optimization by Importance Sampling

The typical EOT (Athalye et al., 2018) randomly selects transformations to craft adversarial examples during optimization, without considering the importance among different transformations. As illustrated in Fig. 1, we show the heatmaps of impersonation attacks under different face variations, which motivates us to conduct a more effective sampling strategy to learn the more difficult or critical transformations.

Given an adversarial patch, a larger loss $\mathcal{J}_f$ on the condition $\{V_i, L_j\}$ represents a greater attack difficulty. Thus, we can utilize $\mathcal{J}_f$ as a surrogate for evaluating the usefulness of the condition $\{V_i, L_j\}$. Those transformations with larger losses should be selected more frequent in the optimization phase, yielding a more effective learning strategy.

To achieve this, we define a flexible importance sampling strategy in every iteration through a probability distribution P, where $P_{i,j}$ indicates sampling probability on the condition $\{V_i, L_j\}$, which can be represented by a softmax function based on Eq. (5) as

$$P_{i,j} = \frac{1}{Z} e^{\mathcal{J}_f(\mathrm{R}(\boldsymbol{s}^a, \boldsymbol{t}^*; V_i, L_j), \boldsymbol{x}^b)}, \tag{7}$$

where $Z = \sum_{i,j} e^{\mathcal{J}_f(\mathrm{R}(\boldsymbol{s}^a, \boldsymbol{t}^*; V_i, L_j), \boldsymbol{x}^b)}$ is the normalization factor. Therefore, if a loss value is larger, we assign a larger value to $P_{i,j}$ such that the transformation $\{V_i, L_j\}$ will be selected with higher probabilities in the optimization trajectory. In each iteration, we sample the points with a batch size of k according to P. The detailed optimization procedure of **Face3DAdv ($\boldsymbol{x}$)** is summarized in Algorithm 1, which can be easily extended to **Face3DAdv ($\boldsymbol{w}$)** by optimizing the latent code $\boldsymbol{w}^*$ for obtaining $\boldsymbol{t}^*$ in Stage II.

5 Experiments

In this section, we first present a simulation-based evaluation framework, and present the experimental results to demonstrate the effectiveness of our proposed Face3DAdv. Finally, we also validate that the evaluation results in our simulator can obtain a consistent tendency with ones in the real world.

5.1 Experimental Settings

Testing protocol. To facilitate the fair and convenient evaluation of physical attacks on face recognition, we aim to construct a comprehensive testing protocol. Although previous methods (Zheng et al., 2021) have considered the significance of evaluating physical variations, e.g., certain poses and lighting in practical scenarios, it is still difficult to conduct a fair comparison between different methods due to poor reproducibility. To tackle this problem, we first customize realistic transformation conditions based on the simulation framework to reduce the potential bias by an uncontrolled

Table 1: The attack success rates (%) of the different face recognition models against impersonation attacks on LFW with adversarial glasses. * indicates white-box attacks.

	Method	*Pitch*			*Yaw*			*Lighting*			*Mixture*		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	MIM	75.65*	8.97	7.84	89.63*	11.10	8.00	94.81*	11.33	5.76	48.45*	3.05	5.65
	EOT	86.58*	16.16	17.48	99.63*	17.53	16.83	99.29*	17.67	12.62	73.73*	6.78	12.46
	GenAP	86.39*	27.87	31.68	99.03*	37.80	31.17	99.33*	41.10	29.19	68.68*	14.89	27.18
	Face3DAdv($\boldsymbol{x}$)	**94.42***	17.23	17.65	99.63*	21.33	17.03	99.29*	22.86	16.81	80.88*	7.73	15.14
	Face3DAdv($\boldsymbol{w}$)	94.39*	**32.29**	**31.81**	**99.90***	**42.27**	**32.00**	**99.95***	**47.10**	**31.33**	**84.08***	**19.69**	**31.75**
CosFace	MIM	9.61	61.32*	13.94	12.60	83.8*	13.97	13.48	95.52*	11.71	5.24	31.08*	11.31
	EOT	21.71	75.71*	29.45	27.77	97.53*	32.40	28.38	96.19*	28.19	14.20	59.71*	26.84
	GenAP	28.74	69.45*	36.90	37.67	94.73*	36.87	40.38	98.14*	34.38	20.95	48.34*	32.72
	Face3DAdv($\boldsymbol{x}$)	23.23	85.16*	29.55	28.90	97.83*	32.80	30.71	96.90*	28.76	15.18	71.03*	28.37
	Face3DAdv($\boldsymbol{w}$)	**40.06**	**87.19***	**46.65**	**51.40**	**98.13***	**46.20**	**54.00**	**98.62***	**45.52**	**31.30**	**72.15***	**43.03**
FaceNet	MIM	3.77	7.42	66.81*	7.50	10.13	67.50*	5.00	10.29	70.52*	2.34	2.88	34.80*
	EOT	9.87	17.97	**98.10***	13.53	20.73	**98.87***	12.62	22.86	96.14*	5.70	8.19	86.73*
	GenAP	22.35	23.29	89.61*	29.63	31.67	94.47*	26.33	31.90	92.57*	14.56	12.77	78.72*
	Face3DAdv($\boldsymbol{x}$)	15.13	21.55	98.19*	20.27	26.67	98.57*	20.62	33.43	**98.95***	9.56	12.59	95.94*
	Face3DAdv($\boldsymbol{w}$)	**28.94**	**33.23**	**98.10***	**38.17**	**44.47**	98.70*	**38.29**	**46.48**	98.29*	**21.54**	**20.54**	**96.45***

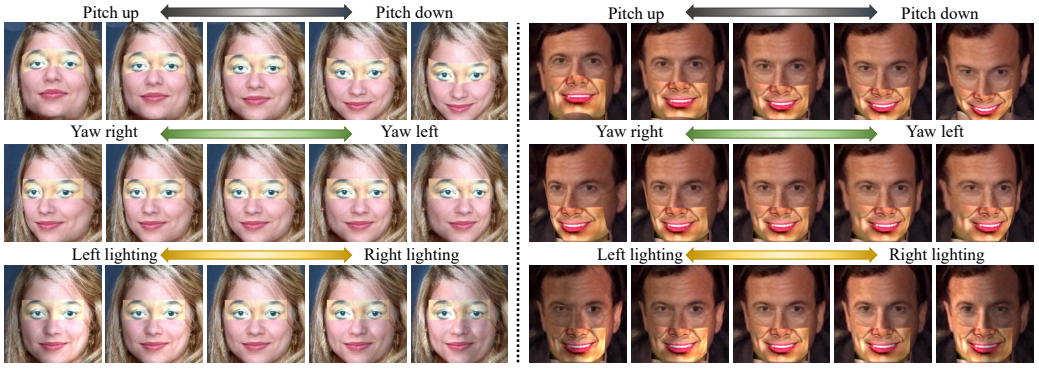


Figure 3: Sample results in simulation framework for physical attacks of **Eyeglass** and **Respirator**, which realizes 3D control of the adversarial examples. Thus, the framework can be used as a surrogate for implementing physical adversarial attacks due to cheap and easy implementation. More examples are presented in Appendix C.

experimenter. In our framework, we conduct a total of 200 experimenters from LFW (Huang et al., 2007) and CelebA-HQ (Karras et al., 2017), which are two of the most widely used benchmark datasets on both low-quality and high-quality face images. For every experimenter, we introduce controllable variations, including different poses and lightings, for reliable evaluation of adversarial attacks. As for poses, we choose 3D face variations, i.e., *yaw* and *pitch*. These two variations of all experimental faces are required to have specific movement ranges of the cruciform rail from -15 to 15 angles, respectively. Meanwhile, we also create a series of relighted testing images by creating a shading map of *lighting* from left to right. Furthermore, we have linearly combined these thee conditions to constitute a new type, named *mixture*. The detailed testing protocol and results of 2D transformations (Xie et al., 2019; Komkov & Petiushko, 2021) are provided in Appendix A.

Networks. We use three face recognition models with different model architectures and training objectives for evaluation, i.e., ArcFace (Deng et al., 2019b), CosFace (Wang et al., 2018), and FaceNet (Schroff et al., 2015). Each model obtained over 99% benign recognition accuracy on LFW by following its corresponding optimal threshold. If the distance of two images that are fed into the model exceeds the threshold, we regard them as different identities; otherwise, as the same identities.

Compared methods. We compare with **MIM** (Dong et al., 2018) that integrates a momentum for improving the transferability of adversarial examples, **EOT** (Athalye et al., 2018) that synthesizes examples over a distribution of transformations, and **GenAP** (Xiao et al., 2021) that is a state-of-the-art transferable adversarial method on face recognition based on generative models. We also take **AdvHat** (Komkov & Petiushko, 2021) as another baseline by wearing hats, which is also blended into EOT (Athalye et al., 2018) to boost the black-box transferability.

Attack types. We consider three types of physically realizable attacks in the simulation environment, i.e., **Eyeglass** (Xiao et al., 2021), **Respirator** (Tong et al., 2021; Zhu et al., 2022), and **Hat** (Komkov & Petiushko, 2021) in 3D pasting ways. Then, we mainly adopt Eyeglass for evaluating the vulnerability of face recognition system in the physical world due to its overall excellent black-box performance, which is also consistently observed in (Xiao et al., 2021). Besides, we further verify the better practicality of the attack mechanism of 3D Eyeglass than 2D ones w.r.t. imperceptibility in Sec. 5.3.

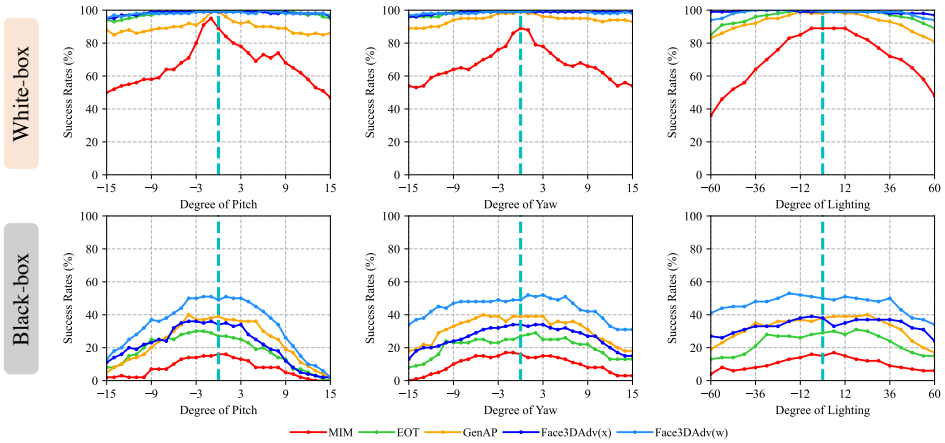
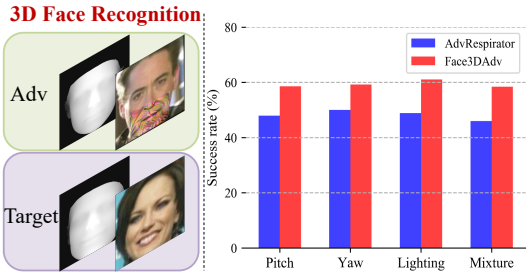


Figure 4: Attack success rates (%) of different attacks under various variations, including *pitch*, *yaw*, and *lighting*. FaceNet is chosen as a white-box model, and test black-box performance by CosFace.

Table 2: Comparison of AdvHat and ours by two types. CosFace is a white-box model.

Type	Testing	Method	Face variations			
			Pitch	*Yaw*	*Lighting*	*Mixture*
Hat	ArcFace	AdvHat	2.97	4.37	4.43	2.34
		Face3DAdv	11.13	11.83	12.24	8.63
	CosFace	AdvHat	63.13	84.50	89.76	39.44
		Face3DAdv	75.29	81.03	88.54	56.45
	FaceNet	AdvHat	3.35	3.80	4.86	4.60
		Face3DAdv	9.45	9.57	9.29	9.89
Respirator	ArcFace	AdvRespirator	19.97	24.83	26.71	12.71
		Face3DAdv	36.26	48.43	49.10	29.02
	CosFace	AdvRespirator	74.77	94.83	95.67	47.44
		Face3DAdv	89.58	96.30	96.29	67.31
	FaceNet	AdvRespirator	16.45	18.53	18.67	12.47
		Face3DAdv	30.65	31.67	32.95	26.52

Figure 5: The attack success rate (%) of two methods against black-box 3D face recognition model by using the attack type of Respirator.



Implementation details. We mainly perform impersonation attacks based on the pairs with different identities in this paper, considering the more difficult and practical property than dodging attacks. The hyperparameters and the evaluation results of dodging attacks are presented in Appendix B.

5.2 Benchmarking on Simulation Framework

In this section, we compare the performance of different physical attacks on face recognition, based on the proposed simulation framework. Fig. 3 shows examples for a physical attack in the simulation framework, which effectively achieves 3D control of the adversarial examples.

Effectiveness of the proposed method. To verify the effects of different face variations, we compare the performance of different methods. Table 1 show the attack success rates (%) of the different face recognition models on LFW, respectively. We can see that different face variations weaken the attack performance of the methods in varying degrees, especially for the effect of *mixture* type. Despite this, Face3DAdv with two variations leads to higher white-box attack success of face recognition models. The results also demonstrate that Face3DAdv can achieve more robust and effective testing performance, benefitting from various physical variations in the optimization phase.

Transferability of the proposed method. We then feed the crafted adversarial images against one face model into other models for testing the transferablity. The results indicate that Face3DAdv can obtain better black-box transferability in the simulation framework. Meanwhile, Fig. 4 shows the detailed performance of the different face variations based on white-box FaceNet. Note that Face3DAdv ($\boldsymbol{w}$) performs best where the axis of face conditions belongs to zero, revealing that our method can consistently enhance the black-box performance even in *without* variations.

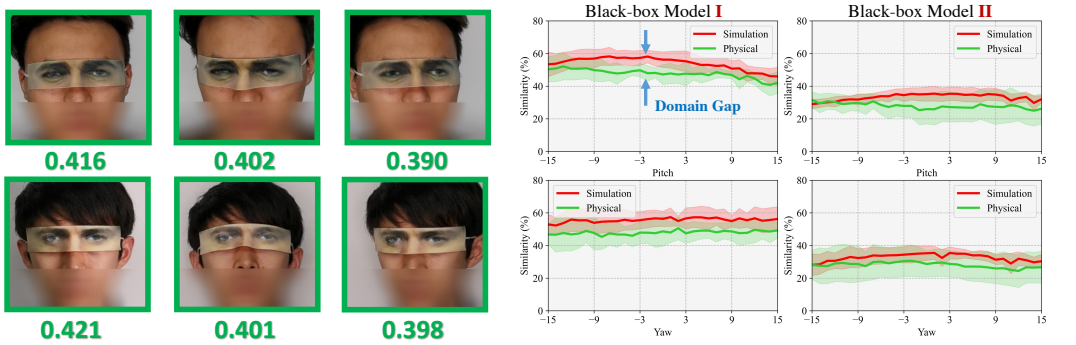
Comparison with AdvHat (Komkov & Petiushko, 2021). We compare the performance of our method with AdvHat by adopting the attack type of *Hat* in Table 2. Besides, we introduce its variation of attack type based on *Respirator* (Zhu et al., 2022), named as AdvRespirator. We found that the optimized region of Hat is not very prominent in the whole face region, making it hard to fully utilize the information of 3D variations in the white-box optimization phase. Nevertheless, our method consistently obtains better results in terms of effectiveness and transferability in these two types.

Table 3: Ablation study of the importance sampling strategy. 'w/o IS' indicates equally sampling. CosFace is a white-box model.

Test	*Pitch*		*Yaw*		*Lighting*	
	w/o IS	with IS	w/o IS	with IS	w/o IS	with IS
Arc.	39.12	40.06	50.07	51.40	52.62	54.00
Cos.	84.90	87.19	98.37	98.13	98.10	98.62
Fac.	45.16	46.65	46.03	46.20	45.67	45.52

Table 4: The attack success rates (%) of physical experiments with 3D adversarial glasses against CosFace. We also adopt the popular methods of face anti-spoofing to test the imperceptibility, including a commercial API.

Method	**Effectiveness**			**Imperceptibility**	
	Pitch	*Yaw*	*Lighting*	*DBMnet*	*API*
GenAP	42.43	53.50	57.70	90.62	5.11
Face3DAdv	64.62	69.37	72.40	96.77	85.52



(a) (b)

Figure 6: (a) shows black-box results of physical attacks by wearing the 3D adversarial glasses. Green box refers to successful attacks, and similarity scores are marked. (b) presents the mean similarity (%) in the simulator and real world by 10 volunteers. The domain gap derives from printer, uncontrolled bias by volunteers, etc.

Effectiveness on 3D face recognition model. Since the proposed method lies in textured-based attacks almost without changing the depth map of a face, it should be able to attack 3D face recognition by leveraging the black-box transferability. To verify this, we introduce typical RGBD-FR (Xiong et al., 2019) that utilizes depth images to explore the global facial layout. Fig. 5 shows the results against RGBD-FR based on the attack type of Respirator. Our method by texture-based attacks can achieve effective attacks against 3D face recognition based on black-box transferability.

Ablation study of the importance sampling strategy. We conduct an ablation study to investigate the effects of the sampling strategy introduced in Sec. 4.3. Table 3 shows the attack success rates *with* and *without* importance sampling. After introducing this strategy, ours can better exploit profitable transformations in the optimizing phase, making it more effective during the testing phase.

5.3 Experiments in the Physical World

In this section, we invited 10 volunteers to be the attackers, and assigned another random identity as the victim for this experiment. The main steps were as follows: First, we took **one** face photo of a volunteer with a fixed camera under natural light. Then, we used the simulation framework for adversarial attacks under different variations and get adversarial glasses for each volunteer. The adversarial glasses were 3D-printed and pasted on real faces. Finally, after wearing glasses, the volunteers tried to reproduce different poses and lighting via a stabilized environment source. Fig. 6 and Table 4 illustrate the effectiveness of our method under varying face variations over the baseline in the real world. The main reason is that Face3DAdv benefits from various simulated physical transformations, and presents the consistent performance in the real world. Besides, the curves also verify that simulator can obtain a consistent tendency of attacks in the physical world. Furthermore, 3D texture-based attack is also more conducive to passing **Face Anti-Spoofing** steadily, e.g., popular DBMnet (Jia et al., 2021) and commercial API, since 3D texture-based attack does not almost change the depth. We also provide details in Appendix D and video demos in the supplementary material.

6 Conclusion

In this paper, we introduce a simulation framework based on 3D face modeling, which can control different face variations and physical conditions to conduct reproducible evaluations. Based on this, we also propose Face3DAdv to craft more robust adversarial patches by considering the 3D face transformations. Extensive experiments verify the consistent improvements over the previous methods in both simulated and physical environments, against diverse face recognition models.

ETHICS STATEMENT

Face recognition models based on DNNs are vulnerable to adversarial examples, which may lead to serious security problems in real-world applications. It is very imperative to understand the actual progress of the field. This paper proposes a novel framework to explore the security vulnerabilities of face recognition models, which can facilitate to evaluate and develop more robust models. One current limitation of our method is that the cost of 3D printing for physical attacks is more expensive than 2D printing.

REPRODUCIBILITY STATEMENT

We provide the code to reproduce our results in the supplementary material.

REFERENCES

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019b.

Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5154–5163, 2020.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

AI Face Mask Challenge GeekPwn, 2020. http://2020.geekpwn.org/zh/index.html Accessed: 2020-10-24.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Ying Guo, Xingxing Wei, Guoqiu Wang, and Bo Zhang. Meaningful adversarial stickers for face recognition in physical world. *arXiv preprint arXiv:2104.06728*, 2021.

Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7498–7507, 2020.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Technical report*, 2007.

Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *European Conference on Computer Vision*, pp. 17–34. Springer, 2020.

Yunpei Jia, Jie Zhang, and Shiguang Shan. Dual-branch meta-learning network with distribution alignment for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17: 138–151, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 819–826. IEEE, 2021.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshops*, 2017.

Mikhail Pautov, Grigorii Melnikov, Edgar Kaziakhmedov, Klim Kireev, and Aleksandr Petiushko. On adversarial patches: real-world attack on arcface-100 face recognition system. In *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pp. 0391–0396. IEEE, 2019.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 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*, pp. 1528–1540. ACM, 2016.

Mahmood Sharif, Sruti Bhagavatula, and Bauer. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *arXiv preprint arXiv:1801.00349*, 2017.

Meng Shen, Hao Yu, Liehuang Zhu, Ke Xu, Qi Li, and Jiankun Hu. Effective and robust physical-world attacks on deep learning face recognition systems. *IEEE Transactions on Information Forensics and Security*, 16:4063–4077, 2021.

Yichun Shi, Divyansh Aggarwal, and Anil K Jain. Lifting 2d stylegan for 3d-aware face generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6258–6266, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Vee Technologies, 2020. http://visagetechnologies.com/face-recognition-in-cars/ Accessed: 2020-10-9.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6142–6151, 2020.

Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Liang Tong, Zhengzhang Chen, Jingchao Ni, Wei Cheng, Dongjin Song, Haifeng Chen, and Yevgeniy Vorobeychik. Facesec: A fine-grained robustness evaluation framework for face recognition systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13254–13263, 2021.

Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5163–5172, 2017.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11845–11854, 2021.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Xingwang Xiong, Xu Wen, and Cheng Huang. Improving rgb-d face recognition via transfer learning from a pretrained 2d network. In *International Symposium on Benchmarking, Measuring and Optimization*, pp. 141–148. Springer, 2019.

Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, pp. 665–681. Springer, 2020.

Xiao Yang, Fangyun Wei, Hongyang Zhang, and Jun Zhu. Design and interpretation of universal adversarial patches in face detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pp. 174–191. Springer, 2020a.

Xiao Yang, Dingcheng Yang, Yinpeng Dong, Wenjian Yu, Hang Su, and Jun Zhu. Robfr: Benchmarking adversarial robustness on face recognition. *arXiv preprint arXiv:2007.04118*, 2020b.

Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3897–3907, 2021.

Xinwei Zhao and Matthew C Stamm. Defenses against multi-sticker physical domain attacks on classifiers. In *European Conference on Computer Vision*, pp. 202–219. Springer, 2020.

Xin Zheng, Yanbo Fan, Baoyuan Wu, Yong Zhang, Jue Wang, and Shirui Pan. Robust physical-world attacks on face recognition. *arXiv preprint arXiv:2109.09320*, 2021.

Jiayi Zhu, Qing Guo, Felix Juefei-Xu, Yihao Huang, Yang Liu, and Geguang Pu. Masked faces with faced masks. *arXiv preprint arXiv:2201.06427*, 2022.

Alon Zolfi, Shai Avidan, Yuval Elovici, and Asaf Shabtai. Adversarial mask: Real-world adversarial attack against face recognition models. *arXiv preprint arXiv:2111.10759*, 2021a.

Alon Zolfi, Moshe Kravchik, Yuval Elovici, and Asaf Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15232–15241, 2021b.

A Detailed Testing Protocol

In the simulation framework, we choose a total of 200 experimenters from LFW and CelebAHQ, which are near-front angles. For every experimenter, we introduce a controllable environmental testing protocol including different poses and lightings as follows.

1) **Pitch:** based on the proposed simulation framework, we control specific movement ranges of the cruciform rail from -15 to 15 angles, and evaluate the performance of attack methods by using the obtained image of every angle. Thus there are a total of 30 images for every experimenter.
2) **Yaw:** we similarly control movement ranges of the cruciform rail from -15 to 15 angles, and evaluate the performance of attack methods for images for every angle. Therefore, there are a total of 30 images for every experimenter in this type.
3) **Lighting:** we obtain relighted testing images by creating a shading map of lighting from -60 to 60 degrees. There are a total of 20 images for every experimenter in this type when the sampling interval is set to 6.
4) **Mixture:** We have linearly combined these three conditions to constitute a new type, named mixture. Specifically, we sample uniformly at intervals of 6 under $[-15, 15]$ degrees of yaw and pitch, respectively, meanwhile setting three different degrees of lighting as -40, 0 and 40. Thus there are a total of 108 images for every experimenter in this type.

In total, our testing protocol in the simulation framework consists of 200 experimenters and a total of 37,600 testing faces. Therefore, a wide range of different physical types in the evaluation, far ahead of the previous datasets, makes our testing protocol challenging and realistic for the existing attack methods.

Evaluation of 2D transformations. We consider three types of 2D physical transformations, which are rotation, projective transformation and their mixture as follows.

1) **Rotation:** the angle of the rotation is sampled from $\mathcal{N}(0, \sigma_1)$.
2) **Projective transformation:** it has eight parameters including $[a_0, a_1, a_2, b_0, b_1, b_2, c_0, c_1]$. Given a point (x, y), we can calculate the mapping point $(x', y') = ((a_0x + a_1y + a_2)/k, (b_0x + b_1y + b_2)/k,)$, where $k = c_0x + c_1y + 1$. a_0 and b_0 are sampled from $\mathcal{N}(1, \sigma_1)$, and other parameters are sampled from $\mathcal{N}(0, \sigma_1)$.
3) **Mixture-2D:** We orderly combine these two conditions to constitute a new type, named Mixture-2D.

In the evaluation of 2D transformation, we set the fixed random seed and sample σ uniformly from $\mathcal{U}(0, 0.1)$. Table 5 shows comparison of EOT and Face3DAdv by 2D variation types. We can see that the performance of white-box attack between the two methods is close to 100%, indicating that the methods can resist the effect of 2D variations in certain varying degrees. The main reason is that the 2D variations can be easily integrated into the optimization phase. Meanwhile, Face3DAdv can obtain better black-box transferability due to the involvement of various 3D physical conditions. Therefore, 3D transformations can be regarded as more difficult and practical than 2D transformations, which also further encourages us to evaluate the performance of different attack methods in varying 3D physical transformations.

B More Experiments

B.1 Implementation Details

Note that MIM and EOT select optimal parameters as report for black-box performance by following Xiao et al. (2021). We thus set the number of iterations as $N = 400$, the learning rate $\alpha = 1.5$, the decay factor $\mu = 1$, and the size of perturbation $\epsilon = 40$ for impersonation and $\epsilon = 255$ for dodging under the ℓ_∞ norm bound, which are identical for all the experiments. The sampling number of EOT is set as $M = 10$. And GenAP adopts original public hyperparameters. As for Face3DAdv, We set the number of iterations $N_1 = 300$, $N_2 = 100$, and the learning rate of Adam optimizer $\eta = 0.01$. Besides, we sample 10 transformations from 20 candidates for Ours in every optimization step.

Table 5: Comparison of EOT and ours by 2D variation types. CosFace is a white-box model.

Testing	Method	Face variations		
		Rotation	*Projection*	*Mixture-2D*
ArcFace	EOT	16.0	10.0	12.0
	Face3DAdv	18.0	12.0	16.0
CosFace	EOT	99.0	99.0	99.0
	Face3DAdv	98.0	98.0	98.0
FaceNet	EOT	13.0	13.0	13.0
	Face3DAdv	18.0	18.0	18.0

Table 6: The attack success rates (%) of the different models against impersonation attacks on CelebA-HQ with adversarial glasses. * indicates white-box attacks.

	Method	*Pitch*			*Yaw*			*Lighting*			*Mixture*		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	MIM	72.71*	9.00	11.68	90.63*	10.33	11.17	94.57*	13.05	9.71	45.58*	4.26	8.55
	EOT	81.13*	12.94	16.81	97.33*	16.93	18.23	98.57*	16.86	15.38	57.92*	5.93	12.98
	GenAP	85.90*	30.16	39.48	99.10*	40.33	40.23	99.10*	48.48	36.81	69.05*	19.19	35.44
	Face3DAdv($\boldsymbol{x}$)	92.19*	15.32	21.23	98.77*	20.47	23.37	99.33*	25.86	22.71	79.74*	8.76	18.58
	Face3DAdv($\boldsymbol{w}$)	**93.84***	**34.23**	**46.35**	**99.97***	**47.27**	**48.70**	**99.71***	**53.71**	**44.19**	**83.04***	**24.29**	**42.15**
CosFace	MIM	16.10	54.48*	19.90	21.07	77.27*	22.83	21.33	89.76*	19.86	9.58	25.94*	16.63
	EOT	19.32	60.06*	24.97	24.30	85.23*	29.90	26.38	94.81*	23.76	11.69	31.81*	20.32
	GenAP	44.29	68.13*	52.03	55.67	95.07*	55.30	56.14	98.10*	51.57	32.31	48.62*	46.12
	Face3DAdv($\boldsymbol{x}$)	28.97	83.19*	36.06	37.50	95.40*	39.53	40.57	99.05*	36.76	21.75	65.38*	32.23
	Face3DAdv($\boldsymbol{w}$)	**49.71**	**83.68***	**55.29**	**59.40**	**95.43***	**58.30**	**60.95**	**98.86***	**56.29**	**39.79**	**67.31***	**52.92**
FaceNet	MIM	8.39	6.16	68.26*	9.97	7.70	70.47*	10.95	8.71	65.14*	5.79	3.26	34.83*
	EOT	10.17	9.77	88.55*	13.40	11.97	92.93*	15.24	13.38	87.76*	7.07	5.14	56.93*
	GenAP	32.23	25.81	94.45*	40.77	36.67	99.13*	40.05	40.38	94.14*	23.33	15.91	82.79*
	Face3DAdv($\boldsymbol{x}$)	20.26	18.74	98.74*	26.73	23.67	99.87*	30.43	31.62	99.67*	15.38	11.95	95.86*
	Face3DAdv($\boldsymbol{w}$)	**42.29**	**36.61**	**99.77***	**54.20**	**51.57**	**100.0***	**52.90**	**56.76**	**99.76***	**32.44**	**25.63**	**98.47***

B.2 Training Efficiency

We set the number of iterations as N and the sampling number of EOT as M in baselines, thus the adversarial patch is generated by $N * M$ forward and backward propagations. As a comparison, our method requires sampling M times from M_L candidates ($M_L = 2 * M$ in our setting) at every iteration, thus needs to perform $N * M_L$ forward propagations and $N * M$ backward propagations. Overall, we only use acceptable overhead on running complexity in the inference phase, and obtain a better performance.

B.3 Evaluation of Dodging Attacks

We perform dodging attacks based on the pairs of images with the same identities on LFW. Table 7 shows the attack success rates (%) of the different face recognition models against dodging attacks on LFW with adversarial glasses. We can see that the overall success rates of dodging attacks are very high, which illustrate that impersonation attacks are more difficult than dodging attacks. Despite this, Face3DAdv with two variations leads to higher white-box and black-box success rates of face recognition models. Similar to the conclusion in impersonation attacks, the results of dodging attacks also demonstrate that Face3DAdv can achieve more robust and effective testing performance. The main reason is that Face3DAdv benefits from various physical variations in the optimization phase.

B.4 More Evaluation of Impersonation Attack

Table 6 show the attack success rates (%) of the different face recognition models on CASIA. The results also demonstrate that Face3DAdv can achieve more robust and effective testing performance, benefitting from various physical variations in the optimization phase.

C More Examples

We show the four mentioned types of physically realizable adversarial attacks in this paper in Fig. 7. In Fig. 8, Fig. 9 and Fig. 10, we show more results of Eyeglass in simulation framework for physical attack on different datasets, which effectively realize 3D control of the adversarial

Table 7: The attack success rates (%) of the different face recognition models against dodging attacks on LFW with adversarial glasses. * indicates white-box attacks.

	Method	*Pitch*			*Yaw*			*Lighting*			*Mixture*		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	MIM	**100.0***	65.19	65.19	99.93*	60.23	62.17	100.0*	69.81	70.81	99.79*	85.10	75.32
	EOT	**100.0***	76.32	76.61	**100.0***	74.40	74.43	**100.0***	82.57	81.67	99.98*	90.10	85.67
	GenAP	**100.0***	95.29	97.45	**100.0***	96.23	97.73	**100.0***	97.43	97.81	**100.0***	98.44	98.38
	Face3DAdv($\boldsymbol{x}$)	**100.0***	86.94	87.32	**100.0***	86.20	85.93	**100.0***	90.00	87.48	**100.0***	94.87	91.23
	Face3DAdv($\boldsymbol{w}$)	**100.0***	**96.52**	**98.39**	**100.0***	**96.33**	**98.13**	**100.0***	**97.69**	**97.97**	**100.0***	**98.91**	**98.77**
CosFace	MIM	45.03	99.77*	57.32	27.37	**100.0***	57.10	39.71	**100.0***	63.76	68.32	99.31*	70.08
	EOT	49.39	99.97*	61.13	33.20	**100.0***	60.87	45.81	**100.0***	64.71	71.78	99.78*	73.66
	GenAP	86.42	99.94*	98.29	81.27	**100.0***	97.90	85.71	**100.0***	97.71	93.87	99.99*	**98.94**
	Face3DAdv($\boldsymbol{x}$)	67.77	**100.0***	88.81	54.63	**100.0***	90.53	68.62	**100.0***	91.29	84.34	**100.0***	94.23
	Face3DAdv($\boldsymbol{w}$)	**89.32**	**100.0***	**98.65**	**85.80**	**100.0***	**98.33**	**87.76**	**100.0***	**98.19**	**95.42**	**100.0***	98.79
FaceNet	MIM	45.84	61.71	98.52*	29.07	57.37	97.60*	39.38	62.14	99.76*	67.92	83.03	97.56*
	EOT	52.00	78.26	**100.0***	37.07	76.77	**100.0***	47.10	79.81	**100.0***	73.36	91.01	**100.0***
	GenAP	92.42	96.03	**100.0***	88.83	98.13	**100.0***	90.05	98.38	**100.0***	96.31	98.38	**100.0***
	Face3DAdv($\boldsymbol{x}$)	66.74	93.84	**100.0***	52.20	92.80	**100.0***	67.90	94.10	**100.0***	82.47	97.06	**100.0***
	Face3DAdv($\boldsymbol{w}$)	**96.06**	**98.71**	**100.0***	**96.23**	**99.03**	**100.0***	**97.10**	**99.00**	**100.0***	**98.17**	**99.60**	**100.0***

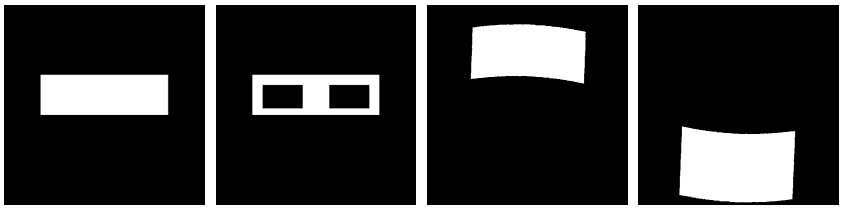


Figure 7: The used four types of physically realizable adversarial attacks in this paper.

examples, including pitch, yaw, lighting, and mixture. Thus, the framework can be reliably used as a surrogate for implementing physical adversarial attacks on face recognition due to cheap and easy implementation.

D PHYSICAL EVALUATIONS

In physical experiments, we mainly present the following steps. First, we took a face photo of a volunteer with a fixed camera under natural light. Then, we used the simulation framework for adversarial attacks under different variations and get adversarial glasses for each volunteer. The adversarial glasses were 3D-printed and pasted on real faces. Finally, after wearing adversarial glasses, the volunteers tried to reproduce different conditions, including some specific yaw, pitch and lighting via a stabilized environment source. We craft 3D adversarial eyeglasses against CosFace by proposed Face3DAdv, and perform attacks against white-box and two *unknown* black-box models (commercial face recognition API or other recognition models). We provide the whole video demos in the **supplementary material** for evaluation under varying physical conditions, which also show the output similarity and prediction results of every frame. By wearing 3D adversarial glasses, the attacker can effectively and steadily impersonate the target identity under different variations against white-box and black-box models, as predicted by these models.

Face Anti-Spoofing. To demonstrate the effectiveness of Face3DAdv in face anti-spoofing, we choose a powerful commercial face anti-spoofing API service. The working mechanism and training data are completely unknown for us. We then feed the crafted adversarial images into the black-box API for evaluating the effectiveness. We obtain a satisfying performance on passing the face anti-spoofing API with a success rate of 85.52% under diverse variations, which outperforms 2D methods by a margin. Since 3D texture-based attack does not almost change the depth map of a face, it is also more conducive to passing commercial Face Anti-Spoofing API steadily.

The practical usage of Eyeglass. The attack mechanism of Eyeglass can be regarded as a practical choice due to the following two aspects. 1) **Imperceptibility**. In a practical FR system, the main defensive module of detecting the abnormality or perceptibility is a Face Anti-Spoofing API, which aims to distinguish whether an image belongs to a real face or not. And we demonstrated that the adversarial 3D Eyeglass can steadily pass a commercial Face Anti-Spoofing API. 2) **Black-box**

effectiveness. The overall performance of Eyeglass is best while considering Eyeglass, Respirator and Hat in the main paper. Although the adversarial patches of Eyeglass Frame are more unsuspicious than ones of Eyeglass, they also lead to very undesirable black-box attack performance. Furthermore, the physical experiments of Eyeglass also demonstrated the practicality and effectiveness.

Figure 8: Sample results in simulation framework for physical attack on CelebA-HQ, which realizes 3D control of the adversarial examples, including *pitch*, *yaw*, *lighting* and *mixture*.

Figure 9: Sample results in simulation framework for physical attack on LFW, which realizes 3D control of the adversarial examples, including *pitch*, *yaw*, *lighting* and *mixture*.

Figure 10: Sample results in simulation framework for physical attack on LFW, which realizes 3D control of the adversarial examples, including *pitch*, *yaw*, *lighting* and *mixture*.