This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ProjAttacker: A Configurable Physical Adversarial Attack for Face Recognition via Projector

Yuanwei Liu[1,2*] Hui Wei[1,2*] Chengyu Jia[1,2] Ruqi Xiao[1,2]
Weijian Ruan[3,4] Xingxing Wei[5] Joey Tianyi Zhou[6] Zheng Wang[1,2†]

[1]National Engineering Research Center for Multimedia Software, Institute of Artifcial Intelligence, School of Computer Science, Wuhan University, Wuhan, China [2]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan, China
[3]Hangzhou Research Institute of Xidian University, Hangzhou, China [4]Smart City Research Institute of China Electronics Technology Group Corporation, Shenzhen, China [5]Institute of Artificial Intelligence, Beihang University, Beijing, China
[6]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore

Abstract

*Previous physical adversarial attacks have shown that carefully crafted perturbations can deceive face recognition systems, revealing critical security vulnerabilities. However, these attacks often struggle to impersonate multiple targets and frequently fail to bypass liveness detection. For example, attacks using human-skin masks [28] are challenging to fabricate, inconvenient to swap between users, and often fail liveness detection due to facial occlusions. A projector, however, can generate content-rich light without obstructing the face, making it ideal for non-intrusive attacks. Thus, we propose a novel physical adversarial attack using a **projector** and explore the superposition of projected and natural light to create adversarial facial images. This approach eliminates the need for physical artifacts on the face, effectively overcoming these limitations. Specifically, our proposed **ProjAttacker** generates adversarial 3D textures that are projected onto human faces. To ensure physical realizability, we introduce a light reflection function that models complex optical interactions between projected light and human skin, accounting for reflection and diffraction effects. Furthermore, we incorporate camera Image Signal Processing (ISP) simulation to maintain the robustness of adversarial perturbations across real-world diverse imaging conditions. Comprehensive evaluations conducted in both digital and physical scenarios validate the effectiveness of our method.*

1. Introduction

Face recognition has become a widely adopted technology for authentication and access control [2, 13]. Despite its prevalence, significant security vulnerabilities remain. Recent studies have shown that imperceptible, carefully crafted perturbations can deceive deep neural network based face recognition models, leading to incorrect predictions [5, 20, 25, 30]. These attacks remain effective in the physical world [24], exposing substantial security risks. Developing physical adversarial attack methods to identify weaknesses and assess the robustness of these systems is essential for secure real-world deployment.

Existing physical adversarial attacks on face recognition systems largely rely on physical artifacts like textured

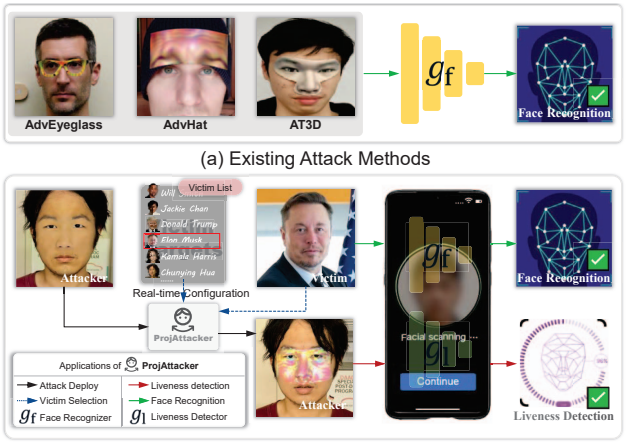


Figure 1. **Comparison of our ProjAttacker with existing attack methods.** While existing methods [10, 18, 28] can successfully attack face recognition, they are difficult to fabricate, inconvenient to swap between targets, and unable to bypass liveness detection due to facial occlusions. In contrast, our **ProjAttacker** projects light-based adversarial textures onto the face, enabling a non-intrusive attack via projector that overcomes these limitations, allowing real-time, configurable face swaps that simultaneously deceive both liveness detection and recognition models.

*Equal contribution †Corresponding author

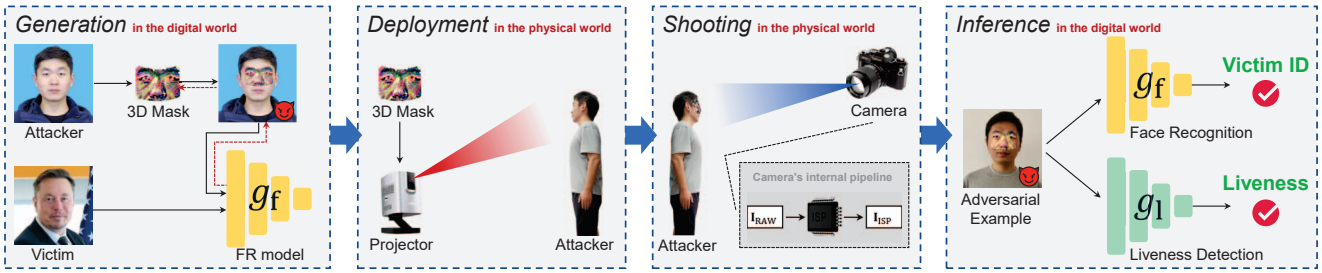


Figure 2. **Workflow of our ProjAttacker.** ProjAttacker works in four steps: **(1)** generating a light-based adversarial 3D mask; **(2)** deploying the attack in the real world using projection devices; **(3)** capturing the attacker's image with a camera; and **(4)** executing the attack during model inference. This workflow mirrors our attack pipeline in the digital domain.

eyeglasses [18, 19], human-skin masks [28], stickers [25], forehead-covering hats [10], and makeup [29]. While these approaches can successfully fool recognition models, they face notable limitations: 1) they are user-specific, difficult to adapt for multiple individuals, and often require complex fabrication processes (*e.g.*, 2D or 3D printing) and high-quality materials to accurately replicate designed adversarial textures, making them impractical for multi-user scenarios; and 2) their occlusive nature makes it difficult to bypass liveness detection, reducing their attack effectiveness in real-world settings.

A projector can produce content-rich light without obstructing the face, making it ideal for non-intrusive attacks. Leveraging this advantage, we introduce **ProjAttacker**, a novel and configurable physical adversarial attack for face recognition. ProjAttacker explores the superposition of projected and natural light to create adversarial facial images. This approach effectively addresses both aforementioned limitations, simultaneously enabling multi-target impersonation and circumvention of liveness detection. Fig. 1 presents a comparison between our method and existing state-of-the-art (SOTA) approaches.

To ensure real-world attack effectiveness, we analyze the workflow of ProjAttacker, which comprises four steps (see Fig. 2). For the adversarial 3D mask generated in step **(1)** to remain effective through step **(4)**, it must successfully traverse two critical domain transitions: the projector-mediated transition from digital domain to physical space, and the camera-mediated transition from physical space back to digital images. *Maintaining resilience across these domain transitions is crucial for attack success.* Therefore, we first incorporate a **light reflection function** into the adversarial 3D mask generation pipeline, ensuring robustness against projection inconsistencies. To mitigate camera-induced color shifts in real-world scenarios, we develop a **differentiable proxy network** that simulates camera Image Signal Processing (ISP) modules. Specifically, this proxy network transforms the camera's internal denoising and color correction units into differentiable modules, and through stochastic sampling of conditional input parameters during optimization, generates adversarial patterns that remain effective under various camera capture conditions. Extensive experiments in both digital and physical scenarios demonstrate that our proposed method effectively compromises mainstream models and APIs. The main contributions are summarized as follows:

- **Real-world configurable face swap**. We present ProjAttacker, the first light-based attack that leverages projected 3D masks with adversarial patterns to simultaneously deceive face recognition models and liveness detection systems, enabling efficient and cost-effective face-swapping attacks.
- **Physically realizable attack pipeline**. We bridge the digital-physical gap through: (i) a light reflection simulation function that ensures robustness against projection inconsistencies, and (ii) a differentiable camera ISP proxy network that accurately models the image capture process, guaranteeing the real-world effectiveness of the adversarial 3D mask.
- **Exceptional attack effectiveness**. Our ProjAttacker achieves superior performance in both digital and physical domains. Comprehensive experimental results validate our approach and reveal significant security vulnerabilities in contemporary face recognition systems.

2. Related Works

Physical Adversarial Attack for Face Recognition. Face recognition systems have been extensively deployed in safety-critical applications, making the study of physical-world adversarial attacks against such systems increasingly crucial. Prior works have predominantly explored the integration of adversarial perturbations with physical accessories, including hats [10], glasses [18, 19, 27], and stickers [25, 26]. Additionally, alternative approaches such as Adv-Makeup [29] demonstrate successful attacks through subtle eye-area cosmetic modifications, while AT3D [28] employs 3D-printed masks with adversarial textures. However, these methodologies either suffer from diminished effectiveness when transitioning from digital to physical domains or require complex, time-intensive preparation processes. Furthermore, their reliance on material-based facial

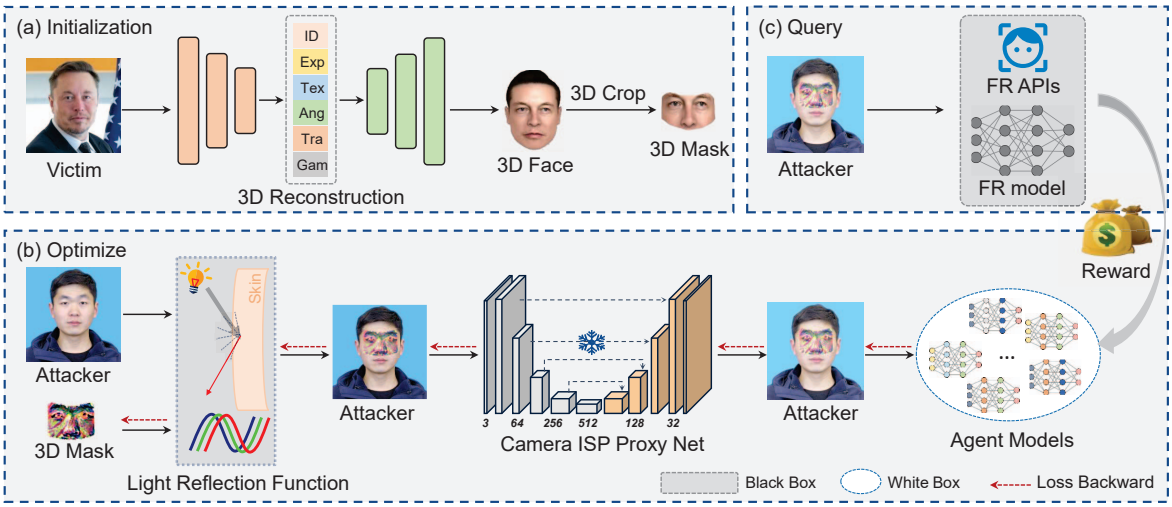


Figure 3. **Overview of ProjAttacker.** **(a)** Initially, we employ CNN to extract features and perform parameter fitting on the image, precisely regressing the 3D Morphable Model (3DMM) coefficients. These coefficients are then rendered and cropped to produce the initial 3D adversarial pattern. **(b)** Subsequently, we simulate the realistic projection of the pattern in the physical space using the Light Reflection Function (LRF), and employ an camera ISP proxy model to emulate adversarial face images under various camera imaging systems. The adversarial pattern is dynamically updated through multiple iterations of optimization using an ensemble learning model. **(c)** Finally, we utilize the optimized adversarial face image for black-box attacks. If the attack fails, we dynamically adjust the weight distribution of the ensemble model through a reinforcement learning-based reward function and repeat step (b) to continually refine the attack strategy and enhance the success rate.

artifacts often fails to circumvent liveness detection mechanisms. In this work, we propose a light-based attack that addresses these limitations by simultaneously compromising both face recognition and liveness detection systems.

Light-Based Attack. Light has emerged as a promising adversarial medium for projecting attack patterns onto human faces [24]. Current light-based attacks against face recognition systems are limited to two primary approaches: Zhou et al. [31]'s infrared light projection method and Nguyen et al. [14]'s 2D RGB adversarial pattern projection. While these approaches represent significant advances, they fail to adequately account for light distortion and reflection on facial surfaces, resulting in suboptimal attack success rates and performance. We systematically re-examine the light-based attack pipeline, as illustrated in Fig. 2, and introduce novel light reflection and camera capture simulation modules. These innovations enable the generation of robust light-based adversarial patterns that maintain effectiveness throughout both projection and capture processes, ensuring high attack success rates in real-world scenarios.

Domain Transition. Since Sharif et al. [18] first demonstrated adversarial attacks in the physical world, the research community has increasingly focused on exploring the effectiveness of adversarial examples under real-world conditions [7, 21, 23]. This process involves two domain transformations. As shown in Fig. 2, the first domain transformation occurs during the attack deployment stage, where the adversarial pattern transitions from the digital space to the physical world. The second domain transformation takes place during the camera capture stage, where the adversarial pattern transitions back from the physical world to the digital space. Existing methods have primarily addressed the former, such as non-printability score (NPS) [18] and D2P [8], which aim to mitigate the impact of the digital-to-physical transformation on attacks. To the best of our knowledge, no method has yet addressed the latter transformation, i.e., from physical to digital. We have designed a comprehensive attack pipeline that incorporates a light reflection function and a camera ISP proxy network to accurately simulate the real-world workflow, thereby enhancing the attack effectiveness in the physical world.

3. Methodology

3.1. Problem Formulation

Threat Model. Let $\mathcal{X}$ denote the space of all possible facial images. Given a black-box facial recognition system $f : \mathcal{X} \to \mathbb{R}^d$ that maps an input image to a d-dimensional embedding space, our goal is to generate physical adversarial patterns that can be projected onto an attacker's face to impersonate a target victim. Formally, for an attacker's face image $x^a \in \mathcal{X}$ and a victim's face image $x^v \in \mathcal{X}$, we aim to find an adversarial perturbation function $F(\cdot)$ that generates a physically realizable adversarial example x^{adv} such that:

$$x^{adv} = F(x^a, x^v), \tag{1}$$

where $F(\cdot)$ consists of a 3D mask M and a light-based perturbation pattern $\tilde{x}$.

Attack Objective. The attack is considered successful if the following conditions are satisfied: (1) Impersonation Success: The adversarial example should be verified as the victim: $Ver(x^{adv}, x^v) = 1$ where $Ver(x^1, x^2) = \mathbb{I}\left(\frac{\langle f(x^1), f(x^2)\rangle}{|f(x^1)|\cdot|f(x^2)|} > \tau\right)$. (2) Physical Realizability: The adversarial perturbation must be physically realizable through light projection:

$$x^{adv} = x^a \odot (1 - M) + T_l(x^a + \tilde{x}) \odot M, \tag{2}$$

where $T_l(\cdot)$ represents the Light Reflection Function that models real-world lighting effects. The complete optimization can be formulated as:

$$\min_{M,\tilde{x}} L_F[T_c(x^{adv}), x^v]$$
$$\text{s.t. } x^{adv} = x^a \odot (1 - M) + T_l(x^a + \tilde{x}) \odot M \quad M \in \mathcal{M}, \tag{3}$$

where $L_F[\cdot,\cdot]$ is the objective function measuring the dissimilarity between the adversarial example and victim, $T_c(\cdot)$ represents the camera capture simulation, $\mathcal{M}$ is the space of valid 3D masks obtained through 3D face reconstruction, $\odot$ denotes element-wise multiplication.

This optimization problem is solved through a reinforcement learning framework [11], where the agent learns an optimal perturbation policy through interactions with an environment constructed from white-box face recognition model ensembles, ultimately generalizing to attack the black-box target model.

3.2. Overall Framework

Given a victim's facial image x^v, we first construct a 3D face model using a pre-trained 3D face reconstruction network. This network decomposes the facial attributes into multiple components, and then used them to reconstruct a complete 3D face model of the victim. To obtain the projection area for the adversarial pattern, we perform 3D cropping on the reconstructed face model to generate the initial 3D mask $M \in \mathcal{M}$. This mask defines the spatial region where the light-based perturbation will be applied.

The core of our framework lies in the optimization stage, where we iteratively refine the adversarial pattern through a reinforcement learning approach. This stage involves several key components: **Light Reflection Function** $T_l(\cdot)$: We design a differentiable light reflection function that simulates how projected patterns interact with the attacker's face under various physical conditions. This function models the transformation of the initial perturbation $\tilde{x}$ when projected onto the 3D mask M and combines it with the attacker's face image x^a according to Eq. 2. **Camera ISP Proxy Network** $T_c(\cdot)$: To bridge the gap between digital and physical domains, we incorporate a Camera ISP (Image Signal Processing) proxy network that simulates real-world image capture effects. This network ensures that the generated patterns remain effective under real camera conditions. **Agent Models**: We construct an environment using an ensemble of white-box face recognition models. The agent interacts with these models to learn an optimal perturbation strategy that can generalize to black-box scenarios. The learning process is guided by a reward signal derived from the cosine similarity scores between the adversarial example and the victim's embedding.

In the final stage, we evaluate the effectiveness of our attack against the target black-box face recognition model. The attacker's face, augmented with the optimized adversarial pattern projected through the 3D mask, is presented to the target system. The success of the attack is measured by the verification result $Ver(x^{adv}, x^v)$. This stage validates whether our generated adversarial pattern can successfully fool the target system while maintaining physical realizability. The entire framework operates in an end-to-end manner, with loss gradients propagating backward through the optimization stage to continuously refine the adversarial pattern.

3.3. 3D Mask Initialization

Face recognition systems capture the 3D structure of facial features, such as surface protrusions and recesses. However, most existing attack methods [10, 18, 19, 25] rely on 2D perturbation simulations, which struggle to effectively map onto 3D facial features in real-world applications. This dimensional mismatch reduces the effectiveness of physical-space attacks. To address this limitation, we introduce a high-dimensional, geometric parameter-driven 3D facial projection construction. Starting from facial images, we employ a convolutional neural network (CNN) regression model to extract 3D Morphable Model (3DMM) coefficients that encode individual-specific facial geometry. These coefficients then guide a facial mesh reconstruction module, applying 3D deformation and topological mapping to accurately generate a 3D mesh model aligned with the individual's facial structure. Spatial coordinate transformations are then applied to map the mesh vertices from object coordinates to camera coordinates, and an intrinsic camera matrix enables perspective projection, which normalizes depth information into 2D pixel coordinates for precise 3D-to-2D mapping on the image plane.

$$x = \mathcal{T}\left(K \cdot \mathcal{R}\left(B \cdot [\Theta_{id}, \Theta_{exp}, \Theta_{tex}, ...]^T + \overline{S}\right)\right). \tag{4}$$

In this formulation, Θ represents various 3DMM coefficients. B is the matrix containing the basis shape vectors, which is used to map the coefficients to the corresponding shape space, while $\overline{S}$ serves as the mean facial shape. $\mathcal{R}$ is the spatial transformation function mapping 3D mesh vertices from object to camera coordinates, while the intrinsic camera matrix K enables perspective projection, and $\mathcal{T}$ normalizes depth information to 2D pixel coordinates.

Through this process, we effectively simulate digital-space attack patterns in the physical domain by projecting refined 2D adversarial patterns onto 3D faces, substantially enhancing attack success rates.

3.4. Light Reflection Function

The efficacy of light-based adversarial attacks hinges on the faithful conversion of digital adversarial patterns into physical light projections. Two primary factors influence this process: (1) the alteration of received light wavelengths due to specular reflection, diffuse reflection, and absorption of adversarial light on facial surfaces, and (2) the interaction between projected light and ambient illumination, resulting in the reception of reflected ambient light. These phenomena modulate the pixel values of the digital pattern. Consequently, formulating these transformations is crucial for mitigating their impacts and preserving attack effectiveness.

We propose a Light Reflection Function (LRF) that models the process of pattern projection, reflection on skin, and camera capture, formulated as the transformation $T_l(\cdot)$:

$$x^{mask} = T_l(x^a + \tilde{x}) \odot M. \tag{5}$$

This process primarily involves modeling the following physical phenomena: the projector emits light patterns corresponding to the RGB values of each pixel in the image. Both projected and ambient light strike the human face and, after absorption and reflection, produce new light patterns. The camera then captures these patterns and converts them into RGB format. This process is principally determined by three factors (λ denotes the wavelength of light):

- $S(\lambda)$: the spectral power distribution of the projector.
- $\rho(\lambda)$: the spectral reflectance of the skin.
- $\sigma(\lambda)$: the sensitivity of the camera's photodetectors, with $\sigma_x(\lambda)$, $\sigma_y(\lambda)$, and $\sigma_z(\lambda)$ representing the impact on red, green, and blue components respectively.

To simulate and quantify human visual response to colors, we utilize the CIE XYZ color space model, integrating the color matching functions to calculate an image's coordinates within this space: $X = \int S(\lambda)\rho(\lambda)\sigma_x(\lambda)d\lambda, Y = \int S(\lambda)\rho(\lambda)\sigma_y(\lambda)d\lambda, Z = \int S(\lambda)\rho(\lambda)\sigma_z(\lambda)d\lambda$. The interaction between ambient and projected light is considered through linear superposition of their spectral intensities:

$$S(\lambda) = S_{natural}(\lambda) + S_{projector}(\lambda). \tag{6}$$

Finally, XYZ coordinates are converted to RGB coordinates. Using the PAL standard RGB system as an example, the transformation is:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} X_r & Y_r & Z_r \\ X_g & Y_g & Z_g \\ X_b & Y_b & Z_b \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}. \tag{7}$$

$S(\lambda)$, $\rho(\lambda)$, $\sigma(\lambda)$, and PAL color conversion matrix(X_r etc.) can all be obtained from technical documentation and official standards. As shown in Fig. 4, our LRF enables the transformed adversarial 3D mask to closely align with its real-world counterpart. Integrating the LRF within our optimization pipeline ensures that the adversarial 3D mask, when projected into the physical world, maintains its attack effectiveness.

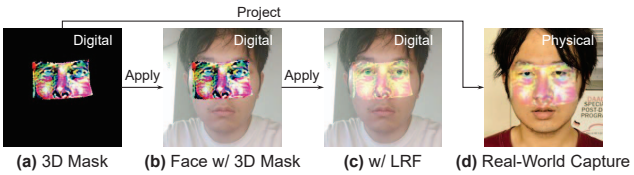


Figure 4. **Light Reflection Function.** (c) is obtained by applying our light reflection function to (b), while (d), obtained by projecting (a) in the real world, demonstrates consistency with (c).

3.5. Camera Capture Simulation

The camera's internal Image Signal Processing (ISP) pipeline plays a crucial role in transforming physical scenes into digital RGB images. Variations in ISP configurations can lead to divergent visual characteristics of the same adversarial texture across devices, potentially affecting attack performance. Therefore, we introduce ISP simulation $T_c(\cdot)$ to model the camera capturing process. To address the non-differentiable nature of traditional ISP processes, we build a differentiable camera ISP proxy network that enables gradient propagation during optimization. The camera ISP proxy network simulates various ISP processes through conditional inputs, focusing on representative parameters from color and tone correction modules, as well as the denoising module, which significantly impact image quality.

Our camera ISP proxy network is based on the U-Net architecture [16], comprising 3×3 convolutions and 2×2 max pooling operations in its encoder and decoder. The feature channels expand progressively through {32, 64, 128, 256, 512}. This structure enables the ISP proxy network to maintain differentiability while simulating diverse camera ISP parameter configurations, thus adapting to various hardware environments.

We normalize ISP parameters (e.g., brightness, contrast, gamma correction) to [0,1] and concatenate them with feature maps, enabling the ISP proxy network to dynamically adjust image tone, contrast, color balance, and denoising characteristics to flexibly simulate diverse device imaging profiles. To train our camera ISP proxy network, we utilize an undifferentiable camera ISP simulator [15] along with the COCO dataset [12], generating a total of 2,270 data pairs. During training, we employ a composite loss function—including pixel-level MSE loss, perceptual loss, and regularization loss—to ensure high image fidelity and sensitivity to conditional ISP parameters, while maintaining full differentiability for gradient optimization. Our framework leverages the ISP network simulating various lighting ef-

Source	LFW dataset					CelebA-HQ dataset				
	ArcFace	MoblieNet	ResNet	IR	API	ArcFace	MoblieNet	ResNet	IR	API
ArcFace [1]	–	40.40%	25.37%	39.22%	12.37%	–	60.80%	64.32%	62.12%	16.49%
MobileNet [4]	13.28%	–	46.98%	24.71%	12.05%	49.50%	–	66.33%	40.80%	12.86%
ResNet [3]	25.37%	54.31%	–	33.14%	18.33%	63.49%	75.50%	–	56.35%	12.35%
IR [3]	22.32%	41.44%	42.37%	–	17.86%	78.79%	78.39%	58.79%	–	14.29%
Ensemble	**59.09%**	**72.91%**	**75.28%**	**53.52%**	**38.78%**	**90.50%**	**88.29%**	**88.37%**	**88.50%**	**32.84%**

Table 1. **Evaluation of black-box impersonation attack.** We report the attack success rates (ASR) on ArcFace, MobileNet, ResNet, IR, and a commercial API. Each row represents results using a different source model for the attack. In the ensemble setting, the target model corresponding to each column is excluded from the ensemble used to generate the attack.

Dataset	Type	Method	ArcFace	FaceNet	MobileNet	ResNet	CosFace	API	Average
LFW	Transfer-based	TAP [27] CVPR'21	53.50%	46.00%	-	-	50.00%	10.00%	39.89%
		AT3D [28] CVPR'23	86.25%	-	91.50%	86.50%	63.00%	45.25%	74.50%
	Query-based	AdvSticker [25] TPAMI'22	-	51.11%	-	-	48.06%	-	49.59%
		SOPP [26] TPAMI'22	50.28%	72.83%	-	-	40.08%	37.56%	50.19%
		ProjAttacker Ours	78.07%	52.17%	79.49%	82.98%	78.17%	41.43%	68.72%
CelebA-HQ	Transfer-based	TAP [27] CVPR'21	65.75%	52.50%	-	-	58.25%	20.00%	49.13%
		AT3D [28] CVPR'23	92.50%	-	96.25%	93.75%	87.50%	50.25%	84.05%
	Query-based	AdvSticker [25] TPAMI'22	-	48.96%	-	-	47.73%	-	48.35%
		SOPP [26] TPAMI'22	50.28%	72.83%	-	-	40.08%	37.56%	50.19%
		ProjAttacker Ours	95.15%	84.47%	98.13%	97.20%	93.20%	44.03%	85.36%

Table 2. **Quantitative comparison with existing methods.** All methods are evaluated under a black-box setting, where TAP [27] and AT3D [28] are transfer-based attacks, while AdvSticker [25], SOPP [26], and ProjAttacker are query-based attacks. The highest and second-highest Attack Success Rates (ASRs) across both datasets are highlighted in red and blue, respectively.

fects in the physical world to enhance the robustness of adversarial patterns and improve attack efficacy.

4. Experiments

4.1. Experiment Settings

Targeted Models. We selected seven face recognition models as our target models, including six representative open-source models (ArcFace [1], MobileNet [4], ResNet [3], FaceNet [17], CosFace [22], IR [3]) and a commercial face recognition API*. For attacking the API, we employed other open-source models as an ensemble of surrogate models. When performing an ensemble attack on one of the models, the remaining models served as surrogate models. In our experiments, we evaluate the attack effectiveness by calculating the cosine similarity of the vectors obtained from the input image pairs. Each model adheres to its respective threshold: if the similarity between two images exceeds the threshold, they are considered the same identity; otherwise, they are considered different identities. Notably, each model is well-trained, achieving an accuracy exceeding 99% on the LFW dataset.

Datasets. We conducted experiments in the digital domain on two extensively employed benchmark datasets: the Labeled Faces in the Wild (LFW) [6] and the CelebFaces Attributes High Quality (CelebA-HQ) [9]. For evasion attacks, we randomly selected several pairs of images with the same identity from the dataset, aiming to push the attacker's features away from their true identity. For impersonation attacks, we randomly chose an attacker from the dataset and paired them with several randomly selected victims, creating multiple pairs of images with different identities.

Comparison methods. We compare our proposed ProjAttacker with the following four methods: TAP [27], AdvSticker [25], SOPP [26], and AT3D [28]. All methods are physically realizable.

Implementation details. All experiments were conducted on a Linux server equipped with two NVIDIA GeForce RTX 3090 GPUs, with all code and algorithms implemented in PyTorch. We set the number of iterations per round to 300 and the learning rate to 0.45. Within the overall framework, the reward step size is set to 0.1. The total number of queries is 25, and if no result is obtained within 25 queries, the attempt is deemed unsuccessful.

4.2. Attacks in the Digital World

Black-box Attack. To assess the effectiveness of the attacks, we employ the Attack Success Rate (ASR), a metric that measures the efficacy of attack methodologies by indicating the proportion of successful attacks relative to the

*https://intl.cloud.tencent.com/product/facerecognition

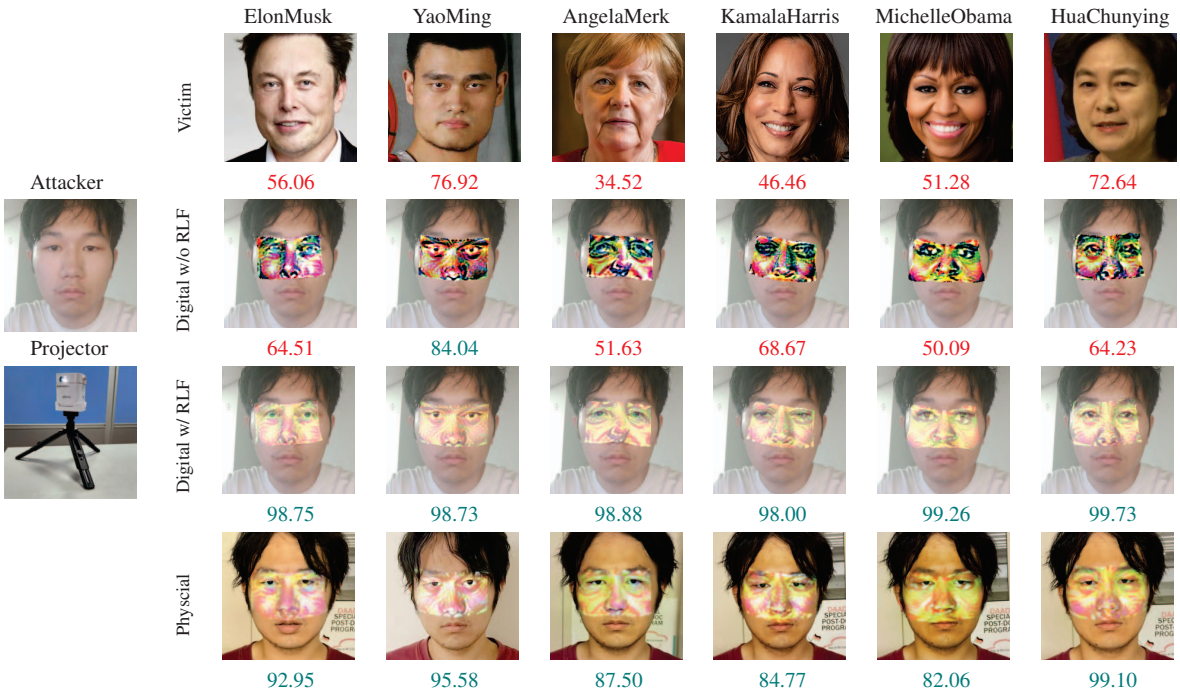


Figure 5. **Attack results of our ProjAttacker in both digital and physical domains.** By projecting diverse adversarial patterns onto facial regions, an attacker can induce the face recognition model (CosFace [22]) to misidentify the subject as various victims. Our method exhibits superior flexibility and cost-effectiveness compared to approaches based on wearable material artifacts. Green indicates successful attacks, while red denotes unsuccessful attacks.

total samples. Tab. 1 displays the quantitative results of impersonation attacks under black-box conditions against five different target models.

As shown in Tab. 1, we can observe: (1) The proposed model ensemble method demonstrates significant efficacy against impersonation attacks on both the LFW and CelebA-HQ datasets. Notably, the highest attack success rate achieved on the LFW dataset is 75.28%, while on the CelebA-HQ dataset, it reaches 90.50%. (2) The attack success rates on the CelebA-HQ dataset are generally higher than those on the LFW dataset. This can be attributed to the higher quality and resolution of facial images in the CelebA-HQ dataset, which facilitates more effective feature extraction by the models. (3) The ensemble method consistently outperforms models trained individually, indicating that our approach can adaptively adjust model weights to achieve optimal performance.

Comparisons with SOTA Methods. We compare ProjAttacker attack with five existing physically realizable black-box attack methods, as shown in Tab. 2. We conducted experiments on a set of images from the CelebA-HQ dataset. Each target model in the columns is excluded from our ensemble model, strictly adhering to the principles of black-box attacks. Our method outperforms the two baseline query-based approaches, AdvSticker and SOPP, across both datasets. Regarding transfer-based methods, our approach surpasses TAP on the LFW dataset but falls short of AT3D. However, on the CelebA-HQ dataset, our method outperforms both TAP and AT3D. These results demonstrate the excellent attack performance of our approach in black-box settings.

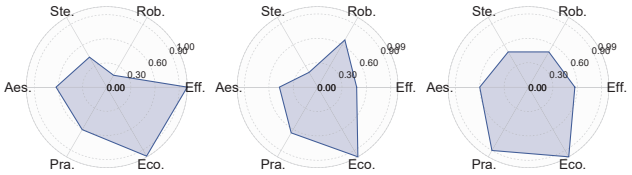
3D	LRF	ISP	ArcFace	MobileNet	ResNet	API
	✓	✓	56%	54%	52%	34%
✓		✓	38%	32%	32%	12%
✓	✓		44%	42%	48%	22%
✓	✓	✓	**68%**	**60%**	**62%**	**38%**

- "3D" represents the 3D mask form, as opposed to the 2D mask.
- "LRF" means the light reflection function in our pipeline.
- "ISP" means the camera ISP proxy network in our pipeline.

Table 3. **Ablation study in the physical world.** We report the ASR of our ProjAttacker attack.

4.3. Attacks in the Physical World

Qualitative Analysis. As shown in Fig. 5, we present an attacker, six victims, and their corresponding digital and physical space attacks. Without the light reflection function, the optimized adversarial patterns exhibit diverse colors with high saturation, impeding physical space reproduction. Incorporating the light reflection function facilitates the transfer of digital space adversarial patterns to the physical domain. Physical space attacks demonstrate our method's efficacy in misleading CosFace to misiden-



(a) TAP [27] (b) AT3D [28] (c) ProjAttacker

Figure 6. **Comparison with existing methods.** We compare the hiPAA [24] of three physical adversarial attack methods.

tify the attacker as different victims with high confidence. Notably, transitioning between victims merely requires altering the projected pattern, offering significant operational convenience. These results highlight our method's flexibility in executing effective attacks in real-world scenarios.

Ablation Study. To quantify the contribution of each module within our proposed framework to the success rate of physical space attacks, we conducted ablation experiments. We systematically removed the 3D, RLF, and ISP modules from a set of attacker-victim image pairs, generating four distinct adversarial patterns for physical testing, including those from the complete framework. The experiments encompassed 50 attacker-victim pairs, with statistical results presented in Tab. 3. We observe that the removal of any single module results in a decrease in Attack Success Rate (ASR). These findings validate the rationale and necessity of each module within our method.

Evaluation using *hiPAA* Metrics. To comprehensively evaluate the performance of physical adversarial attacks, Wei et al. [24] proposed the *hiPAA* metric, which includes six dimensions: Effectiveness, Stealthiness, Robustness, Practicability, Aesthetics, and Economics. We assess and compare the *hiPAA* scores of various methods, as illustrated in Fig. 6. Note that Stealthiness, Practicability, and Aesthetics are evaluated through a user study; refer to the *supplementary material* for more details of our evaluation. We observe that our ProjAttacker attack outperforms TAP in robustness and surpasses AT3D in effectiveness. Additionally, our method significantly outperforms in the *Practicability* dimension, as it does not require manually crafted 3D-printed masks or patches. For an attacker to mislead the facial recognition system into identifying them as different victims, only the adversarial pattern needs to be switched. This not only reduces costs but also eliminates the need for complex manufacturing processes, enhancing the practicality of real-world applications.

4.4. Liveness Detection

Liveness detection technology primarily encompasses four methods: (1) **Transform**: Analysis of image Fourier transforms or Local Binary Patterns to discern whether a presented face is genuine or merely a static image. (2) **Depth**: Utilization of depth sensors or stereo cameras to assess the three-dimensional structure of a face, thereby verifying its authenticity. (3) **Motion**: Examination and tracking of eye movements, lip motions, and blink frequency to validate the presence of a live subject. (4) **MultiBio**: Integration of facial vascular flow detection with other biometric identification technologies, such as iris scanning, to provide a multi-tiered security validation approach.

Method	Transform	Depth	Motion	MultiBio
TAP [27] CVPR'21	✓	✗	✗	✗
AdvSticker [25] TPAMI'22	✓	✓	✓	✗
SOPP [26] TPAMI'22	✓	✓	✓	✗
AT3D [28] CVPR'23	✓	✓	✗	✗
ProjAttacker Ours	✓	✓	✓	✓

Table 4. **Liveness detection analysis.** Our light-based attack outperforms existing methods by successfully deceiving all liveness detection strategies.

	CTyun	**Baidu**	**Tencent**	**Aliyun**
Physical Evaluation	55/55	55/55	48/51	52/55

Table 5. **Bypass rate.** The number of successful bypasses achieved by our method across four commercial liveness detection APIs.

Tab. 4 shows that existing physical attack modalities typically require a tangible medium and often rely on partially obstructing the facial features to create adversarial perturbations. In contrast, our method utilizes light as the medium and adopts a non-intrusive attack. This approach avoids obscuring facial biometric features, thus demonstrating higher effectiveness against various liveness detection technologies. Tab. 5 shows that our method achieved successful bypasses across all four commercial liveness detection APIs, with a 100% success rate on CTyun and Baidu in particular.

5. Conclusion

We introduced ProjAttacker, a novel adversarial light attack that effectively compromises both face recognition systems and liveness detection models. Extensive experiments in both digital and physical space demonstrate that ProjAttacker outperforms existing methods, highlighting the vulnerability of current face recognition systems to sophisticated physical attacks. Our findings underscore the pressing need for more robust defense mechanisms to secure biometric systems against potential threats.

Ethics Statement. Our research investigates the vulnerabilities of face recognition systems through physical adversarial attacks. Aware of the dual-use potential, we conducted all experiments under strict ethical guidelines, ensuring no compromise of personal data or real-world systems. Our goal is to contribute to the development of more secure face recognition systems. We urge the community to consider both defensive and offensive aspects, promoting responsible application.

Acknowledgements. This work was supported by Hubei Key R&D Project (2022BAA033) and National Natural Science Foundation of China (62171325, 62402365). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. 6

[2] Hang Du, Hailin Shi, Dan Zeng, Xiao-Ping Zhang, and Tao Mei. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Computing Surveys (CSUR)*, 54(10s):1–42, 2022. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 6

[4] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. 6

[5] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15014–15023, 2022. 1

[6] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 6

[7] Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20514–20523, 2023. 3

[8] Steve TK Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 962–969, 2019. 3

[9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 6

[10] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th international conference on pattern recognition (ICPR)*, pages 819–826. IEEE, 2021. 1, 2, 4

[11] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017. 4

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014. 5

[13] Yuxi Mi, Zhizhou Zhong, Yuge Huang, Jiazhen Ji, Jianqing Xu, Jun Wang, Shaoming Wang, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using trainable feature subtraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 297–307, 2024. 1

[14] Dinh-Luan Nguyen, Sunpreet S Arora, Yuhang Wu, and Hao Yang. 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*, pages 814–815, 2020. 3

[15] Jueqin Qiu. Fast open image signal processor (fast-openisp). `https://github.com/QiuJueqin/fast-openISP`, 2021. 5

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015. 6

[18] 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*, pages 1528–1540, 2016. 1, 2, 3, 4

[19] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019. 2, 4

[20] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1

[21] 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*, pages 0–0, 2019. 3

[22] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018. 6, 7

[23] Hui Wei, Zhixiang Wang, Xuemei Jia, Yinqiang Zheng, Hao Tang, Shin'ichi Satoh, and Zheng Wang. Hotcold block: Fooling thermal infrared detectors with a novel wearable design. In *Proceedings of the AAAI conference on artificial intelligence*, pages 15233–15241, 2023. 3

[24] Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin'ichi Satoh, Luc Van Gool, and Zheng

Wang. Physical adversarial attack meets computer vision: A decade survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 3, 8

[25] Xingxing Wei, Ying Guo, and Jie Yu. Adversarial sticker: A stealthy attack method in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 2711–2725, 2022. 1, 2, 4, 6, 8

[26] Xingxing Wei, Ying Guo, Jie Yu, and Bo Zhang. Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2, 6, 8

[27] 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*, pages 11845–11854, 2021. 2, 6, 8

[28] Xiao Yang, Chang Liu, Longlong Xu, Yikai Wang, Yinpeng Dong, Ning Chen, Hang Su, and Jun Zhu. Towards effective adversarial textured 3d meshes on physical face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2023. 1, 2, 6, 8

[29] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021. 2

[30] Fengfan Zhou, Qianyu Zhou, Bangjie Yin, Hui Zheng, Xuequan Lu, Lizhuang Ma, and Hefei Ling. Rethinking impersonation and dodging attacks on face recognition systems. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2487–2496, 2024. 1

[31] Zhe Zhou, Di Tang, Xiaofeng Wang, Weili Han, Xiangyu Liu, and Kehuan Zhang. Invisible mask: Practical attacks on face recognition with infrared. *arXiv preprint arXiv:1803.04683*, 2018. 3