The Invisible Polyjuice Potion: An Effective Physical Adversarial Attack Against Face Recognition

Ye Wang
The University of Kansas
Lawrence, KS, USA
yeah_wong@ku.edu

Zeyan Liu
The University of Kansas
Lawrence, KS, USA
zyliu@ku.edu

Bo Luo
The University of Kansas
Lawrence, KS, USA
bluo@ku.edu

Rongqing Hui
The University of Kansas
Lawrence, KS, USA
rhui@ku.edu

Fengjun Li
The University of Kansas
Lawrence, KS, USA
fli@ku.edu

ABSTRACT

Face recognition systems have been targeted by recent physical adversarial machine learning attacks, which attach or project visible patterns on adversaries' faces to trick backend FR models. While these attacks have demonstrated effectiveness in the literature, they often rely on visibly suspicious patterns, are susceptible to environmental noise, or exhibit limited success rates in practice. In this paper, we propose a novel physical adversarial attack against deep face recognition systems, namely Agile (adversarial glasses with infrared laser). It generates adjustable, invisible laser perturbations and emits them into the camera CMOS to launch dodging and impersonation attacks against facial biometrics systems. To do so, we first theoretically model physical adversarial perturbations and convert them to the digital domain. The generated synthesized attack signals are utilized to guide real-world laser settings. Our experiments with real-world attackers and a benchmark face database show that Agile is highly effective in DoS, dodging, and impersonation attacks. More importantly, the candidate impersonation target and optimal attack settings identified by Agile's attack synthesis approach are highly consistent with real-world physical attack results. The grey-box and black-box evaluation against commercial FR models also confirms the effectiveness of the Agile attack.

CCS CONCEPTS

• **Security and privacy → Biometrics**; • **Computing methodologies → Machine learning**.

KEYWORDS

Physical Adversarial Attacks; Face Recognition; Infrared Laser

ACM Reference Format:
Ye Wang, Zeyan Liu, Bo Luo, Rongqing Hui, and Fengjun Li. 2024. The Invisible Polyjuice Potion: An Effective Physical Adversarial Attack Against Face Recognition. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3670382

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA.

ACM ISBN 979-8-4007-0636-3/24/10.
https://doi.org/10.1145/3658644.3670382

1 INTRODUCTION

With recent advances, state-of-the-art face recognition (FR) systems that employ complex deep learning models trained with large-scale datasets of millions of faces [4] can achieve over 99% recognition accuracy [30, 38, 39, 41]. Face-recognition-based authentication systems, a.k.a. facial biometrics recognition, have been widely adopted in real-world applications, for example, in the travel industry [14] and consumer electronics [28]. U.S. Customs and Border Protection has deployed facial biometrics at all international airports [6].

Along with their growing adoption, many security concerns arise. One of them is adversarial machine learning attacks [5, 15, 21, 24], which aim to inject carefully-crafted *adversarial examples* into face images to trigger misclassification of the target neural models. However, most of these attacks are designed in the digital world. When applying them in the physical domain, a significant drop in attack practicality has been observed. To tackle this problem, physical adversarial attacks have been proposed to inject physical-domain perturbations to the object or the imaging component (e.g., cameras) of an FR system. The key challenge lies in generating appropriate physical adversarial artifacts that precisely achieve the attack effect of their counterpart generated in the digital domain.

Existing physical attacks against FR systems leverage printed patterns, projected visible patterns, and infrared signals, as summarized in Table 1. Attacks with *printed patterns* have been extensively studied in the literature, which generate adversarial perturbations in the digital domain and convert them to printable patterns in the physical domain. The patterns are then attached to attackers' eyeglass frames [34], hats [20], T-shirts [53], face masks [66], face stickers [25, 29, 49, 52, 56], and makeup [16, 65]. While not relying on special hardware, printed perturbations often involve a complex digital-to-physical transformation (e.g., non-printability score [34], expectation over transformation [3], total variation [20]) to handle printing noise. Besides, they are fixed once being printed out and thus cannot handle dynamic environmental noises. Another important drawback is that printed patterns may be visually suspicious [25], as demonstrated in Figure 1. Attacks with *projected patterns* emit light directly on the attacker's face [35] or on the key facial areas [22]. The adversarial patterns could be adjusted during the attack to achieve good robustness against environment noise.

However, they require special hardware or cooperation among multiple attackers for precise projection and have low visual stealthiness, which restricts the attack practicality in real-world settings. Finally, some recent attacks leverage *invisible infrared light* to generate adversarial perturbations and improve attack stealthiness. However, the perturbation patterns generated by infrared light are so limited that they can trigger only Denial of Service (DoS) [54] or untargeted impersonation [63] with low attack success rates.

We also observe that existing physical attacks cannot achieve high attack efficacy and high practicality (i.e., stealthiness) at the same time. For example, increasing the coverage of pixels from 0.5% to 2.5% (i.e., large stickers) in [29] can remarkably boost attack success rate from 10% to 90.32%. Similarly, in Adv-hat [20], attacks with 10% pixel coverage outperformed those with 6%. Meanwhile, to increase the success rate, existing attacks [29, 49] often take a brute-force approach to test a large number of physical perturbations, which is obviously impractical in real-world environments.

To tackle these challenges, we propose Agile (adversarial glasses with an infrared laser), a novel physical adversarial attack exploiting infrared laser interference against deep FR models. As shown in Figure 1, Agile utilizes ultra-small laser diodes at a low hardware cost to emit invisible infrared signal. It achieves superior stealthiness compared to existing attacks that rely on printed or projected patterns. The Agile attack consists of two steps: first, it theoretically models the interference of a selected laser diode, which is then used to generate synthesized attack face images with simulated laser perturbations using image fusion. Next, it uses a novel PSAS (Point-Specific Activation Similarity) filter and a similarity filter to identify candidate attack face images, which have high success rates to attack the target image under laser perturbations. With synthesized attack images and the corresponding laser parameter settings in the digital domain, Agile identifies a small range for laser parameter configuration in the physical domain, which can effectively activate dodging and impersonation attacks. Compared with existing physical adversarial attacks, Agile can continuously generate adversarial input within the identified optimal attack range to achieve a higher attack success rate (e.g., 80%) within a short period of time (e.g., 8s). We summarize our contributions as follows:

- We proposed a novel physical adversarial attack against facial biometrics using infrared lasers. It is the first theory-backed, controllable laser-enabled dodging and impersonation attacks with high effectiveness, stealthiness, robustness, and cost-efficiency.
- We presented a novel attack synthesis approach for physical attacks. We first developed a theoretical model to formalize laser interference patterns for the synthesis of attack images, which guides real-world attacks to achieve high success rates within a feasible time frame. Then, we designed a novel PSAS filter to accelerate the search for candidate attack face images.
- We evaluated the effectiveness of Agile against SOTA face embedding models in the white-box, grey-box, and black-box settings. The results showed that Agile is highly effective.
- We explored Agile's ability to bypass existing defenses and uncovered the new attack surfaces it introduces, to highlight the need for future study on proper defenses against this novel attack.

Ethical Considerations. We investigated the vulnerability of SOTA FR models under infrared laser-based physical adversarial attacks. All the experiments were conducted in controlled lab environments simulating real-world applications using open-source models and benchmark datasets. The experiments have been reviewed and approved by the Human Research Protection Program and the Laser Safety Committee at the University of Kansas.

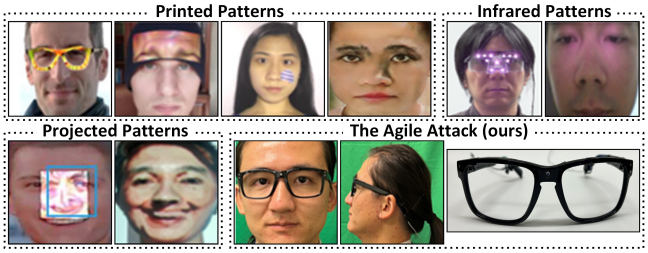

Figure 1: Physical adversarial attack examples and visual stealthiness: printed patterns [20, 34, 49, 65], projected patterns [22, 35], infrared patterns [54, 63], and Agile (ours).

The rest of the paper is organized as follows: we introduce deep face recognition and the attacks against FR systems in Section 2, and present the threat model in Section 3, followed by an attack overview and feasibility study in Section 4 and the design details of the Agile attack in Section 5. We report the evaluation settings and results in Section 6 and explore additional attacking surfaces enabled by Agile in Section 7. Existing defense strategies are discussed in Section 8. Finally, Section 9 concludes the paper.

2 BACKGROUND AND RELATED WORK

2.1 Face Recognition Systems

Face-based authentication systems can be categorized into *face recognition* and *face verification*. Face recognition answers the question "*who are you?*" by matching a candidate face against a list of known faces, i.e., one-to-many matching. Face verification answers "*Is that you?*" by matching the face against one known face to confirm the candidate's identity, i.e., one-to-one matching. In this work, we focus on deep-learning-based face recognition systems, which have demonstrated outstanding performance, e.g., [13] reported that more than 80 models have been built and tested on the *LFW* dataset [19] and 56 of them achieved 99.0% or higher accuracy.

The face recognition pipeline roughly consists of three phases: pre-processing, representation (embedding), and recognition. (1) In pre-processing, faces are detected from the possibly complex background and their geometric structures (e.g., landmark points such as eyes, nose, and mouth) are identified. Popular pre-processing approaches include OpenCV [23]), RetinaFace [9], and MTCNN [59]). (2) In face embedding, face images are represented as vectors in a high-dimensional vector space, so that the similarity (or distance) between two images can be calculated in recognition. State-of-the-art (SOTA) face embedding models are complex DNNs trained with very large datasets, e.g., DeepFace [41], FaceNet and FaceNet512 [30], OpenFace [2], VGGFace2 [4], ArcFace [8], and SFace [61]. The models also generate decision thresholds based on a preset false acceptance rate. (3) In recognition, when the distance between a candidate face and an image in the identity dataset is smaller than the threshold, the face is matched to that identity. Last, continual learning has been adopted to improve recognition accuracy by

Table 1: Categorization of representative physical adversarial attacks against face recognition.

Attacks	Printed Patterns				Projected Patterns		Infrared Light		
	Adv-hat [20]	Adv-glasses [34]	Adv-sticker [29, 49]	Adv-makeup [16, 58, 65]	Small patterns [22]	Entire face [35]	LED glasses [54]	LED bulbs [63]	Agile (our attack)
Supported Attacks	**D**	**D**, $\mathbf{I}_t$	**D** , $\mathbf{I}_t$	**D** , $\mathbf{I}_t$	$\mathbf{I}_t$, $\mathbf{I}_u$	$\mathbf{I}_t$, $\mathbf{I}_u$	**DoS**	**DoS**, $\mathbf{I}_u$	**D**, **DoS**, $\mathbf{I}_t$, $\mathbf{I}_u$
Stealthiness	×	×	✓	✓	×	×	×	✓	✓
Noise Robustness	×	×	×	×	✓	✓	✓	×	✓
Hardware Costs	low	low	low	medium	high	high	low	low	low

$\mathbf{I}_t$=targeted impersonation, $\mathbf{I}_u$=untargeted impersonation, **D**=dodging, **DoS**=escape from face detection.

adding the newly identified face images to the training dataset and retraining the classifier.

Building a SOTA deep face recognition/embedding model requires very large training data and excessive computation. In practice, complex models are often trained and shared by large vendors. For example, FaceNet, DeepFace, and VGGFace are trained by Google, Facebook, and the VGG community, respectively.

The interpretability of deep networks is still an open and challenging problem [11]. Guided backpropagation (GBP) [36] generates gradients through backpropagation to visualize the contribution of each input pixel to classification. Recent work reported that GBP fails on the sanity check [1]. Grad-CAM [31] generalized the class activation map (CAM) [62] to generate activation maps for arbitrary CNN architectures by using the gradient from the last convolutional layer. To better explain the results, Grad-CAM++ [7] adopted more gradient heuristics such as the second derivative of the gradients and the decomposition framework. In deep face recognition, metric learning models are trained along with the classifier. The activation decomposition approach provides a visual explanation for deep metric learning [64]. In this paper, we use interpretation for deep metric learning to identify candidate attack images that are more likely to impersonate the target under laser interference, thereby substantially reducing the computation overhead.

2.2 Adversarial Attacks against FR Systems

Attacks in the Digital Domain. While some general-purpose DNN attacks may be utilized to compromise face recognition models, attacks specifically designed for deep face recognition models are proposed [47, 60]. They could be roughly grouped into two categories: *adversarial examples*: Deepfool [21], FGSM [15], C&W [5], etc, and *backdoors*: BppAttack [48], TaCT [42], etc.

Attacks in the Physical Domain. Existing physical-domain attacks utilize printed or projected adversarial patterns, e.g., face photos and masks [17], eyeglass frames with adversarial patterns [34], stickers with adversarial perturbations [20], adversarial makeups [16], etc. Earlier attacks design adversarial perturbations in the digital world and convert them to the physical world. However, it is difficult to precisely generate physical world perturbations that match the digital world attack performance. To increase the success rate, [22, 35] introduced dynamic perturbations, i.e., projected patterns that are adjustable during the attack. However, the size and complexity of the attacking equipment make it easily noticeable and highly suspicious. Recent attacks modeled the physical-digital conversion of the adversarial patterns to guide the design of physical-world perturbations, e.g., stickers [29, 49] and infrared LED [63]. Their performance is still limited for two reasons: first, modeling digital-physical conversion is highly challenging due to physical signal dynamics. Besides, physical-domain parameters are either fixed or can only be adjusted at an overly coarse granularity, which restricts the attack settings in the real world and results in low attack success rates.

2.3 Light Source-based Adversarial Attacks

Adversarial lasers and light have been used in other physical-world attacks. The AdvLB attack [12] generates visible laser beams as perturbations to trick object detection DNNs into misclassifying. The Rolling Colors [55] uses a visible laser beam on traffic lights to trigger the camera of the autonomous vehicles to misclassify the traffic light's color. GhostImage[57] uses visible light projection to tamper with object detection, leveraging the ghost effect to influence cameras even outside their field of vision. Compared with laser/light-based FR attacks, the attack scenarios, threat models, targeted systems, and approaches differ significantly, making it infeasible to directly apply these light/laser attacks in face recognition systems. In particular, object detection is vulnerable to changes in color and shape, whereas FR models show greater robustness to these changes, complicating efforts to tamper with FR systems. Impersonation attacks in FR require the face detection module to detect and align the face accurately but also need to deceive the face recognition module into misclassifying the altered face image. This process is more challenging than simply disrupting object recognition through color-based interference. Meanwhile, the image projection device used in [57] is large and complex, making the attack impractical in our scenario, where stealthiness is required.

Finally, the laser has been used in DoS and untargeted impersonation attacks against the FR systems. Privacy Visor [54] embedded IR LEDs in glasses to escape face detection in video surveillance (DoS), which is less challenging than dodging and impersonation. We extracted the attack images from their paper [54] and evaluated them with SOTA face detection models. Faces are correctly detected by MTCNN [59] and RetinaFace [9]. The Invisible Mask [63] uses three IR LEDs mounted on a cap to project invisible light on the face. They only achieve untargeted impersonation in small-scale experiments (one attacker with four successful targets). The synthesized images do not demonstrate strong similarity with the physical attack images. The large-scale attack simulation was not validated with corresponding physical attacks. [59] also acknowledges the health risk since the attacker's eyes are exposed to strong IR light sources (three 5 W LEDs) in a very short distance (approximately 3 inches). On the contrary, our design has been proven safe. Last, the dodging attack in [59] is essentially a DoS attack.

3 THE FR SYSTEM AND THREAT MODEL

We consider a typical facial biometrics system in which a camera continuously captures video streams of a user standing in front of it. For a full-HD camera with a prime lens, the detectable face

(e.g., 112x112 to 1920x1080 pixels) requires a subject-to-camera distance of 20-60 cm. We assume that the laser diode points directly at the camera, maintaining a stable distance within the required range, which is practical in real-world FR systems. The distance could be approximated while inaccuracies could be compensated by current adjustment (see Section 6.5.2). Additionally, angles can be aligned using the ghost effect, regardless of the availability of a face alignment guide on the user interface/screen (see Section 6.5.1). Most FR systems will continuously take videos/pictures until the subject is identified. Second-level delays are acceptable in practice.

The FR model recognizes an input face image as a known identity if its vector-space distance to that identity (class) is smaller than a preset threshold. When there are multiple matches, the system selects the class with the smallest distance (or highest similarity).

Threat Model. We adopt a threat model that is consistent with other physical-domain FR attacks in the literature. In particular, the attacker should not be able to tamper with the hardware and software components of the FR system. Besides, the FR systems are typically automated and often do not involve human observers.

We also assume the attacker has no knowledge of the face dataset for training the embedding model, but he may have (1) full knowledge of the face database for the recognition task (e.g., authorized or unauthorized identities), the DNNs used for face embedding and classification, and the camera parameters (i.e., ***white-box attacks***); (2) full knowledge about the face database but *not* the DNNs used in the FR system (i.e., ***grey-box attacks***); and (3) no prior knowledge about the FR system or the database (i.e., ***black-box attacks***).

Attacker's Goals. The terminologies for physical-world attacks against FR systems are inconsistent in the literature. Here, we categorize the attacks into three types: (1) *Impersonation attacks*, where the attacker who is *not* in the face database is recognized by the FR system as *a selected* authorized identity (i.e., targeted impersonation) or *any* authorized identity (i.e., untargeted impersonation) in the face database. (2) *Dodging attacks*, where the attacker who is *in* the face database is not recognized by the FR system as *himself* (i.e., identity dodging) or as *anyone in the block list* (i.e., database dodging). And (3) *Denial-of-Service (DoS) attacks*, in which the attacker aims to disable the face recognition functionality of the FR system so that no face is detected from the captured image. Note that dodging is more challenging than DoS since it requires a face to be detectable but misclassified. In practice, face recognition gates or turnstiles in public venues (exhibition halls, sports arenas, stadiums, government buildings) often use blacklists. The gate opens only when a person not on the blacklist is detected, e.g.,[46]. In this case, DoS is ineffective while dodging is desired.

The Agile Attack and its Objectives. In this paper, we present a physical-domain attack against FR systems, called Agile, which can realize all three attacks, i.e., impersonation, dodging, and DoS. While being practical in real-world applications, physical-domain impersonation attacks often have unsurprisingly low attack success rates. To improve the attack performance, we propose novel *informed* attacks under the white-box setting: based on the knowledge about the face dataset and the DNN models, the attacker computes attack configurations (for impersonation, dodging, and DoS) and candidate attacker-target pairs (for targeted impersonation), with which the attack achieves a high success rate in the physical domain. In addition, we evaluate the DoS, dodging, and impersonation attacks under the grey-box and black-box settings to demonstrate the effectiveness of Agile in real-world applications.

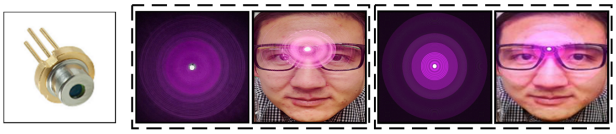

Figure 2: (A) laser diode, (B) annular interference pattern and image captured by a color CMOS camera, and (C) simulated laser signal and synthesized image.

4 ATTACK FEASIBILITY AND CHALLENGES

4.1 Attack Rationale and Feasibility

The Agile attack embeds a small laser diode into a pair of eyeglasses, which emits the laser beam directly into the imaging component of the FR system, e.g., a color CMOS camera. The narrowband infrared laser diode, as shown in Figure 2A, generates infrared laser signals with an annular interference pattern produced after passing through the lens. The camera captures the laser signal, revealing a physical adversarial perturbation affecting the facial image, as shown in Figure 2B. Intuitively, this results in a "blur" effect on the captured face, potentially saturating some pixels, e.g., the central bright dot in Figure 2B, or increasing the red-channel value of others. These alterations can significantly impact the feature representation of the captured image, subsequently misleading the classification model. Meanwhile, the infrared laser light is invisible to the human eyes, which makes the attack stealthy.

Next, we investigate the hardware, attack effects, and potential success rate to assess the feasibility of the Agile attack.

Hardware. Commodity infrared laser diodes come with varied wavelength and output power options. Due to safety and cost concerns, we prefer low-power laser diodes, such as HL6750MG [43] and L785H1 [44] from *Thorlabs* with 685 *nm* and 785 *nm* wavelengths, respectively. They are low-cost ($70-$90 for retail) and have maximum operating currents of 120 *mA* and 250 *mA*, respectively. The imaging component of the face recognition systems can be any commodity camera. In this work, we focus on CMOS cameras, which are more prevalent in real-world facial recognition systems compared to CCD cameras. The CMOS sensors are usually sensitive to optical signals with a wavelength in $[400\ nm, 1000\ nm]$, while the human eyes can only notice signals within $[400\ nm, 700\ nm]$. Therefore, the infrared light in $[700\ nm, 1000\ nm]$ can be accurately sensed by the camera sensors but invisible to human eyes.

Deep Face Recognition Models. We consider SOTA deep face recognition models such as FaceNet [30], VGGFace2 [4], DeepFace [41], ArcFace [8], and SFace [61]), which are open-source and extensively tested in the literature, as the target of the attack.

Attack Effects. A laser perturbation could change the feature representation of a face image. Even a small change may affect the recognition result, causing a misclassification (untargeted or targeted impersonation). As shown in Figure 3, along with the increasing of the laser signal, the blurred area increases, which potentially obscures key facial features and prevents the face from being correctly recognized, resulting in a dodging attack. As the blur effect intensifies, no face could be detected from the image, resulting in a DoS attack.

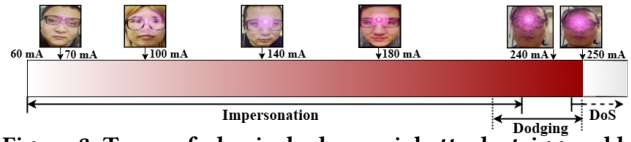

Figure 3: Types of physical adversarial attacks triggered by different laser signals and the potential power ranges.

Attack Success Probability. Laser perturbations have been shown effective in DoS against face recognition [63]. Similar results were observed in our experiments. As shown in Figure 3, the perturbations from strong laser signals (laser current $\geq 248mA$ in our settings) lead to excessive overexposure, rendering the faces undetectable. We also ran tests for untargeted impersonation attacks against a face database with 1,000 users to assess the attack success rates. For example, in a system using MTCNN for face detection/alignment and DeepFace for recognition, as shown in Table 6, all 15 attackers could impersonate one or multiple users in the database in real-world tests (i.e., the physical domain). In the digital domain, face images of 15 attackers under 15 different laser settings (power and distance) could "match" 722, 181, and 370 users using the Euclidean, Euclidean-L2, and cosine distance metrics, respectively.

4.2 Challenges and The Agile Attack Overview

A straightforward yet naive Agile physical adversarial attack is to place the attacker wearing the laser-embedded glasses in the field and test all possible perturbations. In a simple scenario where the laser setting is solely determined by the operating current of the laser diode, the total number of settings is $K(\delta) = (I_{max} - I_{min})/\delta$, where the current is in $[I_{min}, I_{max}]$ and δ denotes the step-width.

Challenges. The naive approach is impractical in dodging attacks, where the attacker should remain unrecognized all the time. However, exhaustively attempting all possible perturbations may cause the attacker to be identified under a weak or ineffective perturbation. For example, we have tested the attack success rate (ASR) of naive identity dodging in the physical domain, where 5 attackers were randomly selected from 15 available attackers to attack the FaceNet [30] model using the L785H1 laser diode [44]. We set $I_{min} = 60mA$, $I_{max} = 250mA$, and $\delta = 1mA$. On average, there were only 18 settings out of 190 possible settings that enabled a successful identity dodging (with an $ASR = 9.4\%$). The results also showed that the effective current for dodging attacks has a very narrow range, as shown in Figure 3. Furthermore, the attack effects on different attackers vary. The same perturbation enabling a dodging attack for one attacker may lead to an impersonation or a DoS attack when applied to another attacker.

Meanwhile, both dodging and impersonation attacks are susceptible to adversarial perturbations. To enhance the attack success probability, smaller step widths are preferred since more settings are likely to produce effective perturbations. However, it would significantly increase the attack time, which may cause the attack impractical in real-world applications. For instance, in a naive targeted impersonation, we randomly selected 5 targets from the LFW dataset and counted the number of successful impersonation attacks launched by 15 available attackers under 190 laser settings. The average ASR was between 0% and 0.77% as shown in Table 2, and the average attack time was 283.5 seconds.

Table 2: The average number of successful targeted impersonations of 15 attackers against 5 randomly selected targets.

Targets	T1	T2	T3	T4	T5
# of successes / ASR	0/0	22/0.77%	17/0.6%	6/2.1%	0/0

Finally, laser perturbations are constrained by laser power and interference patterns. Without an effective attack strategy, launching targeted impersonation using the naive approach may often fail. For example, targets T1 and T5 in Table 2 cannot be impersonated by any attacker under any laser perturbations. This is because the images of the attacker and the target are far away from each other in the embedding space, so that small perturbations are unable to push them close enough to be considered similar by a face recognition model. Recruiting more attackers would help to improve the attack success rate, however, it would also increase the attack time.

To address these challenges, we propose a novel framework for physical adversarial attacks. It identifies attack settings with high success probabilities in the digital domain and uses the information to guide physical-domain attacks, making the attacks more practical. The settings include laser settings in both dodging and impersonation, and additionally suggest optimal attackers among a candidate attacker set in impersonation. We refer to the Agile attacks as ***informed dodging*** or ***informed impersonation*** to distinguish them from conventional dodging or impersonation.

Agile Framework Overview. Attacks in the Agile framework have two phases, *digital attack synthesis* (denoted by blue lines in Figure 4) and *physical attack implementation* (denoted by red lines). Digital attack synthesis outputs suggested attackers and a small number of laser settings for controlling the laser diode to generate effective laser perturbations in physical attack implementation.

The key component of the Agile framework is digital attack synthesis. We will elaborate on its detailed design in Section 5. The main idea is to model the laser signal produced by a laser diode at a given power (step ①), and generate a set of synthesized attack images with simulated laser perturbations through laser interface fusion (step ②). With the knowledge of the face dataset, the victim FR model, and its camera parameters, the Agile framework computes the settings under which the simulated digital-domain attacks achieve high attack success rates (step ③). Finally, in physical-domain attacks (step ④), the laser diode generates laser signals using the identified power settings to tamper with the target FR model, for example, inducing it to misclassify the attacker as a particular user (targeted impersonation) or someone other than the attacker himself (identity dodging).

5 THE AGILE INFORMED ATTACKS

In this section, we theoretically model the infrared laser interference to the imaging systems and present a method to fuse attack images. Then, we present Agile informed attack design to compose physical-domain *informed* dodging and impersonation attacks using results from simulated attacks in the digital domain.

5.1 Laser Interference Modeling

A laser beam with a high temporal coherence emits light with a very narrow spectrum. This results in annular interference patterns when propagating through multiple lenses of the camera, as shown

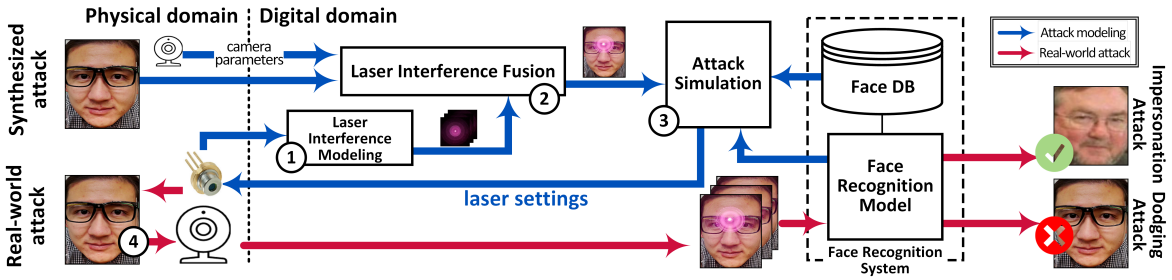

Figure 4: Agile attack workflow: blue lines denote attack synthesis process and red lines denote real-world attack process.

in Figure 2B, After propagating through, each cell of the CMOS sensor behind the lens catches the annular interference light shining on its surface and converts it into electrons. Then, an analog-to-digital converter translates the differing charges of individual cells into pixel values of various colors. As a result, physical adversarial perturbations are injected into digital images.

It is very challenging to accurately measure the perturbation at each pixel of the digital image, because laser interference modeling not only depends on the hardware (laser and camera) but also is highly sensitive to physical settings and environment dynamics. When an ideal model is not available, we speculate that an imperfectly accurate model (as shown in Figure 2C) could still provide useful information to guide a robust attack, because the perturbations in our attack are strong but mono-featured. In this work, we explore this hypothesis and demonstrate its practicality. Table 3 summarizes the notations used in the laser interference modeling.

The shape of a laser spot is typically defined by its *irradiance distribution*, which forms a beam profile, and the *phase*, which determines the beam profile over the propagation distance. Laser diodes used in this work have a Gaussian beam profile, but our model can be extended to other beam profiles with small modifications. The Gaussian irradiance profile $P(r)$ [10] can be described as follows:

$$P(r) = \frac{2P}{\pi\omega(z)^2} \exp\left(\frac{-2r^2}{\omega(z)^2}\right) \tag{1}$$

As illustrated in Figure 5A, the power P of the laser is dispersed across the wavefront during propagation. The camera captures only the portion of the signal P_a that passes through its aperture r_a. For the detailed P_a derivation approach. Once the laser signal enters the aperture, the lenses of the camera create annular interference patterns that can be modeled based on their specific optical characteristics as shown in Equation 2.

$$P_I = P_a \cdot \cos\left(\frac{2\pi nt \times \cos(\arcsin(n\sin(\arctan(d/f))))}{\lambda}\right)^2 \tag{2}$$

5.2 Laser Interference Fusion

Next, we quantify the adversarial perturbation generated by the laser by measuring the effect of a laser signal on the RGB channels of the image according to its *quantum efficiency*. It is a sensitivity measure to describe the probability that a photon landing on a pixel gives off an electron and causes a change in brightness. When a laser with λ and P_I is shed into the CMOS, the color filters on each photodiode filter the signal and output P_I^R, P_I^G, and P_I^B, where $P_I^R = P_I \cdot f_R(\lambda)$, $P_I^G = P_I \cdot f_G(\lambda)$, and $P_I^B = P_I \cdot f_B(\lambda)$, and $f_R(\lambda)$, $f_G(\lambda)$, and $f_B(\lambda)$ are the quantum efficiency of three color channels.

When the light strength is sufficiently high, the Red channel with a larger quantum efficiency will be saturated first. Meanwhile, the Green and Blue channels with smaller quantum efficiency will keep receiving light until they are overflowed. When all three color channels are saturated, the corresponding pixels become white. Therefore, we need to capture the color-filtering effect and calibrate the RGB values when some pixels are saturated. We define a relu-like function $\mathcal{F}(x) = min(x, \alpha)$ following [55], where α stands for the overflow value of each channel. It is typically set to 255. The converted pixel value of the laser signal is computed as $Y = (Y_R, Y_G, Y_B)$, where $Y_R = \mathcal{F}(P_I^R)$, $Y_G = \mathcal{F}(P_I^G)$, and $Y_B = \mathcal{F}(P_I^B)$.

Next, we measure the pixel values of the laser interference in the digital domain and then apply it to an input image by adding the perturbation on three channels. Since some pixels may get saturated, we need to calibrate the color channel overflow again. Therefore, for an input image X and a laser perturbation Y, we fuse them to generate an output RGB image as $\mathcal{F}(X + Y) = min(X + Y, \alpha)$.

However, in real-world attacks, when the laser intensity is large, the camera will adjust its exposure time to reduce the light shed on the CMOS sensor. The synthetic face image with laser perturbation is brighter than real-world images captured by the camera. So, we further adjust its brightness by a ratio $r_b = P_o/(P_o + P_a)$, where P_a is the laser intensity and P_o is the power of the input figure X. Finally, we obtain the simulated attack image following Equation 3 and compute the simulated perturbation as $X_o - X$. Figure 2B shows a laser interference pattern and the corresponding synthesized face image.

$$X_o = \mathcal{F}(r_b(X + Y)) = min(r_b(X + Y), \alpha) \tag{3}$$

5.3 Attack Simulation and Informed Attacks

After modeling the laser interference and fusing it into an image, we simulate an attack image by applying a synthetic laser perturbation to any input image. Let x_a denote the face image of attacker a and $z_{a,r}$ denote the simulated attack image under a laser perturbation r. $z_{a,r}$ can be computed following Equation 3. This enables us to run simulated attacks in the digital domain and search for optimal attack settings to guide physical-domain attacks. So, we call this new attack technique the Agile informed attack. The informed attacks are expected to be: (i) *effective*, where physical-domain attacks informed by digital-domain simulations should achieve high attack success rates; (ii) *practical* such that the identified settings should be easy-to-implement and remain stealthy in the physical domain; and (iii) *efficient*, which requires the computation cost for searching for optimal settings remain reasonably low.

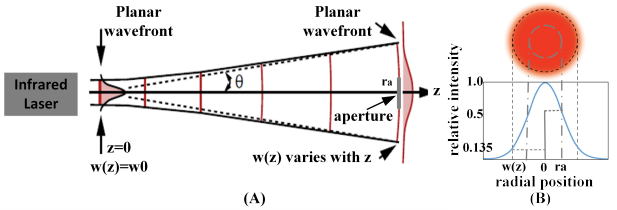

Figure 5: Laser power distribution during propagation: (A) the curvature of the wavefront of a Gaussian beam; (B) the energy passing through the aperture.

5.3.1 Attack Objectives. We propose two types of informed attacks, i.e., informed dodging and informed impersonation, which include identity dodging, database dodging, untargeted impersonation, and targeted impersonation. Moreover, we consider DoS attacks a special instance of dodging, which has a simpler objective function since any overly strong laser signal (e.g., generated by a current $\geq 248mA$ in our experiments) could trigger DoS.

Let $f(\cdot)$ denote a face embedding model with a decision threshold τ based on $\mathbb{D}$, and $\mathcal{L}(\cdot)$ represents a distance function measuring the similarity of two feature embeddings, such as Euclidean-L2. In dodging attacks, we aim to find a small perturbation r, when applied to the image of an attacker a, who is in the face database $\mathbb{D}$, the simulated attack image $z_{a,r}$ would not be recognized as a (i.e., *identity dodging*) or anyone in a pre-defined deny-list $\mathbb{L}$ (i.e., *database dodging*). The objectives of the two dodging attacks can be respectively described by Equations 4 and 5 as follows.

$$\mathcal{L}(f(z_{a,r}), f(x_a)) > \tau, \quad x_a \in \mathbb{D} \tag{4}$$

$$\mathcal{L}(f(z_{a,r}), f(x)) > \tau, \quad x_a \in \mathbb{D} \text{ and } \forall x \in \mathbb{L} \tag{5}$$

Next, let us denote the face detection and alignment model in the target FR system as $\mathcal{M}(\cdot)$, where $\mathcal{M}(x) = 1$ means a face is detected in an image x. Then, the objective of DoS attacks can be simplified as $\mathcal{M}(z_{a,r}) = 0$, indicating no face detected in the attack image $z_{a,r}$.

Similarly, in targeted impersonation, we aim to find appropriate r such that under the laser perturbation, the attack image of attacker a, who is in a set of known attacker $\mathbb{A}$, would be misidentified as an image of the target user t who is in the face database $\mathbb{D}$. In untargeted impersonation, t can be any user in $\mathbb{D}$. Therefore, the objectives of targeted and untargeted impersonation can be described by Equations 6 and 7, respectively.

$$\mathcal{L}(f(z_{a,r}), f(t)) < \tau, \quad x_a \in \mathbb{A}, t \in \mathbb{D} \tag{6}$$

$$\mathcal{L}(f(z_{a,r}), f(x)) < \tau, \quad x_a \in \mathbb{A}, \exists x \in \mathbb{D} \tag{7}$$

5.3.2 Informed Attack Design. In the Agile framework, an informed attack consists of two steps.

Candidate Setting Search: In this step, we search for candidate laser settings under which the laser signals lead to successful adversarial attacks in the digital domain. A laser setting typically includes parameter values about the laser's wavelength, optical power, as well as the laser-to-camera distance and angle. Among them, the wavelength is determined once a laser is selected and the laser-to-camera angle should be 0° since we require the laser diode to point right at the camera (we will discuss laser-camera alignment in Section 6.5.1). Therefore, we only need to explore all possible combinations of laser power and laser-to-camera distance.

Table 3: Notations used in laser interference modeling.

Notation	Definition
P	Laser optical output power
$P(r)$	Representing power intensity at a given location r
λ	Wavelength of the laser
θ	Divergence angle of the laser
r	Radial distance from the axis
z	Propagation distance
r_a	Aperture radius
t	Lens thickness
f	Focal length
n	Refractive index
d	Distance to the CMOS imaging plane center
$\omega(z)$	Laser beam radius where the irradiance is $1/e^2$ of the peak

The effective laser power (i.e., laser light captured by the camera's aperture) is proportional to the electrical current supplied to the laser diode. It is also inversely proportional to the square of the distance. In theory, we can fix the laser-camera distance while adjusting the laser current to produce any laser signal captured at another laser-camera distance. For example, the laser signal produced by a current of 100 mA at a laser-to-camera distance of 35 cm can approximate the signal generated by the same laser operating with a current of 204.1 mA and at a distance of 50 *cm*. Therefore, by maintaining a constant laser-camera distance (e.g., 35 *cm*) and varying laser current, we can simplify the search process.

The search starts from the minimal operating current with a step-width δ and generates simulated perturbations. The corresponding synthesized attack images $z_{a,r}$ for the attacker a are used to identify candidate currents satisfying the objectives described in Equations 4 - 7 in different attack cases. We employ a basic grid search for dodging attacks (Section 5.3.3). To improve search efficiency, we propose an enhanced search scheme comprising two filters based on the embedding similarity and a novel point-specific activation similarity for impersonation attacks (Section 5.3.4).

Optimal Setting Identification: Among all candidate settings, we aim to identify the optimal ones to inform real-world attacks. It is a non-trivial task because laser settings producing successful digital-domain perturbations may not necessarily result in successful physical attacks, due to inaccuracies in laser signal modeling related to the susceptibility of laser signals to physical-domain noise, such as measurement inaccuracies, environmental lighting, scattering during light propagation, etc. To tackle this challenge, we choose to determine candidate current ranges, rather than individual values, such that a current within each range is highly likely to yield a successful physical-domain attack. For example, small distance inaccuracies can be compensated by small adjustments near the candidate current. This strategy has been shown effective for addressing uncertainties in digital-to-physical transformation and mitigating the influence of modeling inaccuracies [55].

In particular, we employ a sliding window of length k to identify consecutive laser settings that lead to successful digital-domain attacks, e.g., $\{I_1, ..., I_k\}$. Based on empirical analysis, we set k to 3. Then, we merge adjacent ranges to form longer ranges and output a set of candidate ranges ranked by the range size. The first (also the longest) range, denoted as $[R_{min}, R_{max}]$, will be used to inform

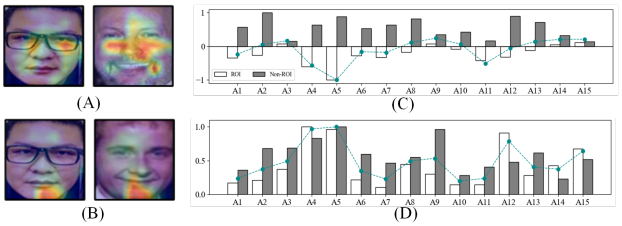

Figure 6: (A) and (B): the heatmaps between the same attacker and two different targets. (C) and (D): the average difference of the similarity between the attack and target images before and after applying laser perturbations, for each of 15 attackers and two different groups of targets.

the physical attack, which will start from the range center (i.e., $I = (R_{min} + R_{max})/2$) and increase or decrease the current by δ in each search step. In our experiments, we set δ to 1 mA.

5.3.3 Informed Dodging Attacks. To inform dodging attacks in the physical domain, we need to find if there exists a laser perturbation enabling digital-domain identity or database dodging, and the candidate range to produce that laser signal. Following the design in Section 5.3.2, we can directly apply the grid search to identify the optimal range satisfying the objectives defined by Equations 4 and 5 and launch physical Agile dodging attacks. Experiment results in Section 6.2 show that the informed identity dodging and database dodging attacks achieve close to 100% average attack success rates.

5.3.4 Informed Impersonation Attacks. To inform physical impersonation attacks, we expect to find if there exists a laser perturbation making the synthesized attack image of any known attacker similar to the selected target (targeted impersonation) or making the synthesized attack image of a selected attacker similar to any user in the database (untargeted impersonation). Besides, we need to determine the candidate range to produce that laser signal.

We could apply the grid search, however, it yields a high computation cost, since each search step needs to generate the synthesized attack image for each attacker and perform a classification. The computation is dominated by the embedding-calculation operation, denoted as O_e. Therefore, the grid search cost for an informed targeted impersonation is proportional to the number of known attackers N_A and possible laser settings $K(\delta)$, which can be approximated as $N_A \cdot K(\delta) \cdot O_e$. For example, $N_A = 15$ and $\delta = 1mA$, and $K(\delta)$ is 150 in our tests. The cost for untargeted impersonation increases significantly by a factor of N_T, i.e., the number of users in the face database, which is set to 1,000 in our experiments.

Enhanced Search: We propose a novel enhanced search scheme utilizing two filters based on *embedding similarity* (ES) and *point-specific activation similarity* (PSAS). It identifies a small set of optimal attackers for a selected target, or a small set of targets among all users in the database for a selected attacker, to improve the attack success rates while reducing the number of searches.

For example, in informed targeted impersonation, the ES filter identifies the attackers close to the target in the embedding space while the PSAS filter finds the ones holding certain characteristics that make them susceptible to laser perturbations. Let $\mathbb{A}$ denote the set of known attackers. We represent the attacker sets after the ES filter and PSAS filter as $\mathbb{A}_{ES}$ and $\mathbb{A}_{PSAS}$, respectively, where

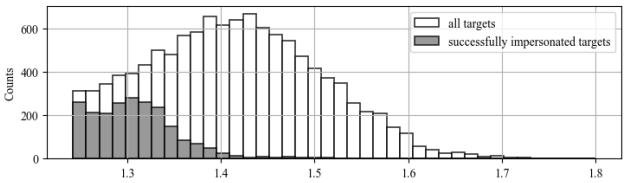

Figure 7: The effectiveness of ES filter.

$\mathbb{A}_{PSAS} \subseteq \mathbb{A}_{ES} \subseteq \mathbb{A}$. Accordingly, the objectives of targeted and untargeted impersonation defined in Equations 6 and 7 are revised as follows:

$$\mathcal{L}(f(z_{a,r}), f(t)) < \tau, \quad x_a \in \mathbb{A}_{PSAS}, t \in \mathbb{D} \tag{8}$$

$$\mathcal{L}(f(z_{a,r}), f(x)) < \tau, \quad x_a \in \mathbb{A}_{PSAS}, \exists x \in \mathbb{D} \tag{9}$$

We return the top candidate attacker in $\mathbb{A}_{PSAS}$ or the entire set to inform physical-domain attacks, making targeted impersonation more practical. In untargeted impersonation, the filters will output candidate target sets for a given attacker, denoted as $\mathbb{T}_{ES}$ and $\mathbb{T}_{PSAS}$, to predict attack results without testing with the entire target set.

Next, we present the design for targeted impersonation. However, the scheme can be applied to untargeted impersonation by simply swapping the attacker and target.

(1) Embedding Similarity Filter: Low-power laser diodes are used due to cost and safety concerns, resulting in small laser perturbations, which may be constrained to manipulate pairs of attack and target images that are widely separated in the embedding space. On the other hand, if two images are already close (near the decision boundary), small perturbations have a good chance to push them closer. Therefore, we calculate the similarity between the embeddings of each attacker in $\mathbb{A}$ and the target, ranked the attackers by the similarity scores, and output $\mathbb{A}_{ES}$ with the top-3 attackers.

(2) Point-Specific Activation Similarity (PSAS) Filter: In synthesized attack images, the laser signals have a stronger influence on certain pixels than others. So, we partition the image into two areas, denoting the area under stronger influence as the *region of interest* (ROI) and the remaining area as the non-ROI. Besides, we can calculate the contributions of the ROI and non-ROI regions towards the overall similarity between two images, using the point-specific activation similarity (PSAS) [64].

In particular, let A^a and A^t denote the feature maps in the last convolution layer of the deep FR model for two images, for example, an attack and a target image. For each point (i, j) in the attack image, we compute its overall activation in the target image:

$$s = (W^a_{i,j} A^a_{i,j}) \cdot \sum_{x,y} (W^t_{x,y} A^t_{x,y}) \tag{10}$$

where (x, y) is a point in the target image, $W_{i,j}$ and $W_{x,y}$ are the weight matrix for point (i, j) and (x, y), respectively. It represents the cosine similarity between the point (i, j) and the feature vector of the target image in the embedding space, i.e., the contribution of the point (i, j) towards the overall similarity. If we visualize the point-specific activation similarity using heatmaps, we can see the regions in one image contributing more to its overall similarity with the other image. For example, Figure 6A shows the upper face region of both the attack and target images contributes more to the overall similarity between them, while in Figure 6B, the bottom region contributes more.

We observed that the laser perturbations, which are effective in the digital domain, often exhibit a "blur" effect in the ROI. Therefore, under its interference, the similarity in the non-ROI region will have a greater impact on the overall similarity between two images than the similarity in the ROI. Intuitively, if an attack image is close to the target image and has a higher similarity in the non-ROI compared to the ROI, the laser perturbation is likely to enhance the overall similarity between the two images, thus increasing the success rates of the attack. Therefore, we design the PSAS filter to select attackers from $\mathbb{A}_{ES}$, where the similarity in the non-ROI region between their attack images and the target image is higher than the similarity in the ROI.

The PSAS filter consists of three steps. First, we use the nose and mouth landmarks generated during face detection and alignment to partition the image into two parts. The upper partition is the ROI region. Second, we average the overall activation for each point in the ROI and non-ROI regions and compare their aggregate contributions to the overall similarity. Finally, we select the attack images with higher non-ROI similarity to form the set $\mathbb{A}_{PSAS}$.

Filter Effectiveness and Efficiency: We empirically validate the effectiveness of ES and PSAS filters using 15 attackers and 1,000 targets (in the I-1K set in Section 6.3). For the ES filter, we demonstrate the distribution of the embedding space distances between each attacker a (without adversarial laser) and each target in I-1K in Figure 7 (white bars). The distribution of the distances between a and targets who can be impersonated (denoted as $\mathbb{T}_a$) are shown in the gray bars. As shown, 98% of the impersonation cases happened between attacker-target pairs whose original distances are among the top 25% of all pairwise distances. That is, if we employ the least expensive ES filter to eliminate 70% to 75% of the targets in I-1K who are less similar to a's benign image, the remaining set $\mathbb{T}_{ES}$ contains the majority of $\mathbb{T}_a$.

For the PSAS filter, we divide the targets into two groups for each attacker, one with a higher ROI similarity and the other with a higher non-ROI similarity. Then, we apply laser signals generated under 15 settings (i.e., the current in $[60mA,200mA]$ with $\delta = 10mA$) to each attack image, and calculate the average similarity difference between the attack and target images before and after applying laser perturbations. Figure 6C and 6D report the results for two groups with a higher ROI similarity and a higher non-ROI similarity, respectively. The light and dark bars represent the similarity changes in the ROI and non-ROI regions, while the green dots denote the overall similarity changes. Apparently, for the group in Figure 6D, the laser perturbations on average improve the overall similarity, improving the attack success rate.

The ES and PSAS filters involve both embedding-calculation and embedding multiplication operations, denoted as O_e and O_m, respectively ($O_e \gg O_m$). In this process, we calculate the embedding vectors for N_A attackers in $\mathbb{A}$ and the target, the feature maps for the attacker in $\mathbb{A}_{ES}$ and the target, and the embedding vectors for the attacker in $\mathbb{A}_{PSAS}$ under each of $K(\Delta)$ settings, where Δ is a coarse step-width, e.g., $\Delta = 10mA$. Therefore, the estimated computation is $(N_A+1+K(\Delta)\cdot|\mathbb{A}_{PSAS}|)O_e$. Compared to the computation of the grid search, the enhanced search reduces computation of O_e by 97.3% at least, with $N_A = 15$, $K(\delta) = 150$, $K(\Delta) = 15$, and $|\mathbb{A}_{PSAS}| = 3$ at most.

Table 4: Average ASR of identity-dodging (ID) and database-dodging (DD) within the optimal current ranges.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
Range Length in ID	13	12	15	13	10	15	20	14	12	10	19	12	11	13	14
Range ASR in ID (%)	92.3	83.3	86.7	76.9	80.0	80.0	85.0	78.5	83.3	90.0	84.2	83.3	81.8	84.6	92.8
Range ASR in D-50 (%)	92.3	83.3	86.7	76.9	80.0	80.0	85.0	78.5	83.3	90.0	84.2	83.3	81.8	84.6	92.8
Range ASR in D-200 (%)	92.3	83.3	86.7	76.9	80.0	73.3	80.0	78.5	83.3	90.0	78.9	75.0	81.8	84.6	92.8
Range ASR in D-500 (%)	84.6	83.3	80.0	75.0	77.7	73.3	78.9	71.4	75.0	88.8	77.7	75.0	80.0	76.9	92.3
Range ASR in D-1K (%)	84.6	81.8	80.0	75.0	77.7	71.4	77.7	71.4	75.0	77.8	77.7	70.0	80.0	76.9	83.3

6 IMPLEMENTATION AND EVALUATIONS

6.1 Experiment Settings

Hardware. In the experiments, we used the L785H1 laser diode [44]. It has a narrow line width ($< 0.1\ nm$) to produce reliable interference fringes and small beam divergence angles (5° lateral and 10° vertical angles) to cover a substantial area while remaining safe for cameras. We drilled a small hole in the nose bridge of the frame of the eyeglasses to embed the laser diode. This position allows the laser perturbation pattern to reliably influence critical regions on the face (i.e., eyes and nose). Besides, when the attacker wears the glasses and looks at the camera, the incidence angle of the laser beam will be close to 90° for better laser alignment.

The diode is powered by a current source, the *Kungber SPS305* DC power supply, and operates within a voltage range between $2V$ and $3V$. It could be powered by two AAA alkaline batteries using a simple circuit to convert the voltage source into an adjustable current source. We used an iPhone 13 ultra-wide camera with the Sony IMX772 CMOS to capture face images in most experiments. Besides, we tested the effectiveness of Agile on a different camera (an iPhone 12 camera with the Sony IMX372 CMOS). In all experiments, we set the subject-to-camera distance to 35 cm to capture the faces with effective sizes for different FR models.

FR Models and Parameters. We adopted the *LFW* dataset [18] in the experiments, which contains 13,233 face images of about 5,749 users (identities). In each experiment, we randomly sampled a set of individuals to create the identity database or the denylist.

In most experiments, we used MTCNN [59] for face detection and alignment, a pre-trained FaceNet model[1] for face embedding, and the normalized Euclidean distance. In impersonation attacks, we also explored 4 state-of-the-art face recognition models (VGGFace2 [4], DeepFace [41], ArcFace [8], and SFace [61]) and other face detection models and distance metrics. To calculate the similarity between two images, we adopted the optimal L_2-distance threshold of 1.242 for the LFW dataset [30], which is widely accepted in the literature for unrestricted, labeled data. All the experiments were conducted using PyTorch 1.7.1 on a desktop computer with Intel Core i7 8700 CPU and Nvidia RTX 2080 GPU.

Attackers and Data Collection. We recruited 15 volunteers (denoted as A1 to A15) from the University of Kansas to act as attackers, including: (1) undergraduate and graduate students, and faculty; (2) both females and males; and (3) individuals from different race and ethnicity groups (white, black, and Asian). For each individual, we took photos and videos without the infrared laser signal (i.e., *benign* images/videos) and with the laser interference generated under different current settings increasing from 60 mA to 250 mA in 40 seconds (i.e., *attack* images/videos).

[1]Model 20180402-114759 is an Inception ResNet (v1) [40] trained with VGGFace2, achieving 0.9965 accuracy on LFW. https://github.com/timesler/facenet-pytorch.

Table 5: Attack performance of untargeted impersonation.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
I-100	23	22	11	21	9	21	7	17	25	4	5	7	7	8	12
I-500	146	132	67	85	44	103	49	56	90	43	32	43	42	45	59
I-1K	248	236	157	227	75	217	109	121	207	95	74	96	77	88	134

6.2 Informed Dodging Attacks

Identity Dodging. In a successful identity-dodging attack, the attacker should not be recognized as himself. Therefore, given an attacker, we first searched within the power range of [60 mA, 250 mA] for candidate current settings that lead to successful simulated dodging attacks under synthesized laser interference. We used the step-width $\delta = 1$ mA and the decision threshold $\tau = 1.242$. Then, we identified the ranges of at least three consecutive candidate settings (i.e., $k = 3$), from which the optimal current range $[R_{min}, R_{max}]$ was determined for each attacker. For instance, the optimal current ranges for attackers A3 and A7 are [235 mA, 250 mA] and [228 mA, 248 mA], respectively. Finally, we generated laser perturbations using the current within each attacker's optimal range and ran the identity-dodging attack in the physical domain. Starting from the range center, we increased and decreased the current by $\delta = 1$ mA each time until reaching the range boundary.

"Informed" with the optimal range, the Agile attack becomes effective and efficient. For all 15 attackers, the physical identity-dodging attacks succeeded in the initial current setting (i.e., optimal range center), achieving a 100% average attack success rate (ASR). We also tested all the settings in the optimal current ranges with $\delta = 1$ mA interval and recorded the ones under which the physical dodging attacks were successful. This allows us to calculate the attack success rate in the optimal range (called range ASR). As shown in Table 4, the attacks succeeded in most settings in the optimal range, indicating high robustness against small measurement inaccuracy and environment noise.

The optimal range helps to determine the boundary between dodging and DoS attacks. For example, the laser perturbation with $I = 250$ mA launched successful identity dodging for most attackers, however, it would cause DoS for some attackers (e.g., A7). In general, an input current larger than R_{max} would generate a sufficiently large laser perturbation and cause image overexposure. The saturation of pixels over most of the face image renders the detection model ineffective. Therefore, the optimal range can guide Agile DoS attacks. We did experiments for attackers whose $R_{max} < 250$ mA and the DoS attacks were successful in the physical domain. However, we did not conduct full-scale DoS experiments in this work, since our infrared diodes have a maximum power of 250 mA. As sufficiently large laser signals would inevitably blind all imaging systems, it is reasonable to conclude that increasing laser power could achieve a high or 100% DoS success rate [37].

Database Dodging. In a successful database-dodging attack, the attacker should not be recognized as himself or anyone on the denylist. In this experiment, we randomly sampled 50, 200, 500, and 1,000 individuals from the *LFW* dataset and added the attacker's benign images to mimic the denylists of 4 different sizes, denoted as D-50, D-200, D-500, and D-1K, respectively.

Similar to the identity-dodging attack, we searched for candidate current settings and the optimal current range that led to successful simulated attacks in which the synthesized attack images could escape the entire denylist. Finally, we ran the physical Agile attacks with the laser signals generated within the optimal range and evaluated if the attacker was recognized as any identity on the corresponding denylist or not. The results showed that the database dodging attacks against all four denylists achieved 100% success rates with the initial current setting. We also calculated the range ASRs for 15 attackers using the settings in the optimal range. As shown in Table 4, the database dodging attack is more robust to succeed against smaller denylists with substantial distance between two identities than larger denylists where the distance between two identities is small.

Table 6: Attack performance in different models/settings.

	VGGFace2			DeepFace		
	MTCNN	Retina	OpenCV	MTCNN	Retina	OpenCV
Cosine	674|0|0	631|0|0	622|0|0	370|0|0	389|0|0	308|0|0
Euclidean	922|0|0	912|0|0	908|0|0	722|0|0	736|0|0	718|0|0
Euclidean-L2	555|0|0	401|0|0	389|0|0	181|0|0	184|0|0	106|0|0
	ArcFace			**SFace**		
	MTCNN	Retina	OpenCV	MTCNN	Retina	OpenCV
Cosine	104|0|0	95|0|0	97|0|0	88|1|1	60|3|3	74|1|1
Euclidean	356|1|1	370|0|0	373|1|1	654|0|0	718|0|0	582|0|0
Euclidean-L2	58|1|1	56|2|2	57|1|1	34|5|5	25|4|4	38|1|2

In each model and under each pre-processing setting: (left) the total number of identities impersonated by any of the 15 attackers; (middle) the number of attackers who failed to impersonate any target in the I-1K dataset; (right) the number of failed attackers who were identified by Agile's simulated attacks.

6.3 Untargeted Impersonation Attacks

Laser perturbations generated by low-cost laser diodes are typically small. Being applied to the attacker, they could cause small shifts of the feature vectors of the attack images in the embedding space. In this section, we investigated the effectiveness of such small shifts towards successful untargeted impersonation. The results would lay the foundation for Agile attacks for targeted impersonation.

In particular, we randomly selected 100, 500, and 1,000 individuals from the *LFW* dataset to construct three identity datasets, denoted as I-100, I-500, and I-1K, respectively. Since the face recognition model is not 100% accurate, the benign images of some attackers could be misidentified as individuals in *LFW*. We identified these individuals and excluded them from the identity datasets. Besides, in each new experiment, we resampled all the I-n datasets.

Next, we applied laser perturbations generated under 15 current settings, i.e., from 60 mA to 200 mA with 10 mA interval, to each attacker and captured the attack images. Finally, we calculated the distances between the feature representations of an attack image and each image in the identity datasets to identify successful impersonation cases.

Attacks against FaceNet. We used MTCNN [59] to detect faces from images and the pre-trained FaceNet model to compute the feature vectors of the face images. If the embedding distance between an attacker and a target (in an identity dataset) is smaller than the decision threshold (i.e., 1.242 for LFW), we consider it a successful untargeted impersonation. Finally, we counted the number of success cases for each attacker. Note that if the distances between an attacker and two different targets are both smaller than the threshold, we consider them as two success cases.

Table 7: Targeted impersonation against 5 targets.

	$\mathbb{A}_{ES}$	$\mathbb{A}_{PSAS}$	A_i	$[R_{min}, R_{max}]$	ASR_{OR}	ASR_{ER}
T1	{A9, A4, A11}	{A9, A4, A11}	A9	[192-198]	0	100%
T2	{A3, A15, A14}	{A3}	A3	[155-180]	100%	100%
T3	{A2, A12, A1}	{A2, A12}	A2	[137-154]	100%	100%
T4	{A10, A4, A13}	{A10}	A10	[80-99]	100%	100%
T5	{A1, A7, A6}	{A1, A7}	A1	[125-138]	100%	100%

Table 5 reports the total number of targets in each identity dataset that were impersonated by each attacker. The attack is more likely to succeed in larger identity datasets than smaller ones. For example, the average pair-wise distance between 15 attackers' benign images and the targets in the I-1K dataset is 1.279, which was reduced to 1.176 under laser perturbations. This indicates that the adversarial perturbations on average pushed the attackers towards the targets in the embedding space.

Attacks against SOTA FR Models. Next, we tested the performance of the Agile untargeted impersonation attack against four state-of-the-art FR models: VGGFace2 [4], DeepFace [41], ArcFace [8], and SFace [61]. Besides MTCNN, we also considered two other commonly used face detection and alignment modules, RetinaFace [9] and OpenCV [23]. Besides the normalized Euclidean distance (Euclidean-L2), we also tested the cosine similarity and Euclidean distance metrics, which are popular in FR applications [33]. Therefore, for each FR model, we tested a total of 9 settings.

We utilized the deepface platform [33] to implement 4 FR models. In each model, the decision threshold was calculated from a false acceptance rate (FAR) of 0.001 on the LFW dataset and coded in the *find_threshhold* function in deepface's *modules/verification.py* file. In each model and under each setting, we computed the embedding distances between 15 attackers and the targets in the I-1K dataset under 15 current settings to identify successful impersonation cases.

Table 6 reports an aggregated number of success cases across 15 attackers, showing the effectiveness of the Agile attack against deep face recognition models. For example, 555 identities in the I-1K dataset could be impersonated by at least one of the 15 attackers under one of the 15 current settings in the VGGFace2 model using MTCNN and the normalized Euclidean distance. Even in the latest models with very high recognition accuracy (e.g., SFace and ArcFace with 99.65% and 99.41% accuracy on *LFW*), the physical perturbations have a non-negligible chance to enable impersonation. Compared to the raw Euclidean distance and cosine similarity, the normalized Euclidean distance appears to be a robust distance metric against the Agile attacks. Meanwhile, the face detection and alignment models have a modest effect on attack performance.

While untargeted impersonation appears to be feasible with Agile attacks, we also found that a few attackers failed to impersonate any target in the I-1K dataset. For example, 5 attackers failed in the attack against SFace using MTCNN and the normalized Euclidean distance, as shown in Table 6. Obviously, they would fail in the targeted impersonation attack under the same setting. Using Agile's simulated attacks in the digital domain, we can identify these attackers effectively, as shown in Table 6, to reduce the attack time in the real world.

6.4 Informed Targeted Impersonation Attacks

Results in untargeted impersonation show that with a given set of attackers and small laser perturbations, the Agile attack cannot impersonate all the targets. For FaceNet (with MTCNN and normalized Euclidean distance) and a randomly selected I-1K dataset, only 364 identities could be impersonated by the 15 attackers recruited for our experiments. If we increase the size of the attacker set $\mathbb{A}$, we could impersonate more targets.

Table 8: Targeted impersonation against SOTA FR models.

FR Models	$\lvert\mathbb{T}_{candidate}\rvert$	$\lvert\mathbb{T}_{filtered}\rvert$	$\lvert\mathbb{T}_{impersonated}\rvert$	FE	ASR
VGGFace2	561	547	319	97.5%	56.8%
DeepFace	165	159	143	96.4%	86.7%
ArcFace	66	66	63	100%	95.5%
SFace	45	45	44	100%	97.8%

To evaluate the effectiveness of the Agile targeted impersonation attack, we randomly selected 5 targets from 364 candidates, denoted as T1 to T5. Based on the Euclidean-L2 distance between each target and the attackers, the ES filter output the $\mathbb{A}_{ES}$ set with top-3 attackers mostly close to each target in the embedding space, as shown in Table 7. Next, we used the PSAS filter to identify the attackers in $\mathbb{A}_{ES}$ who are more similar to the target in the non-ROI region and constructed the set $\mathbb{A}_{ASPS}$ for the Agile targeted impersonation attack in the physical domain.

$\mathbb{A}_{ASPS}$ and the optimal current range $[R_{min}, R_{max}]$ are used to guide the physical-domain attacks. For example, we selected the first attacker in $\mathbb{A}_{ASPS}$ for each target and calculated the corresponding optimal current range. In the physical Agile attack, we set the initial current at the center of the optimal range and increased and decreased the laser current by $\delta = 1\ mA$ each time until reaching the range boundary. The attack is considered successful if the attacker could impersonate the target within the optimal range. Table 7 shows that four targets were successfully impersonated while one failed, achieving an average attack success rate of 80%. This is because targeted impersonation requires accurate perturbations and therefore is less tolerant to measurement inaccuracies (e.g., laser-to-camera distance and angle misalignment) and environment noises (e.g., lighting conditions). As discussed in Section 6.5.2, such inaccuracies and noises could be compensated by adjusting the laser current. So, we could extend the current range to improve the attack success rate. Table 7 shows the attack success rates in the optimal range (denoted as ASR_{OR}) and in an extended range of $\pm 10\ mA$ from the optimal range center (denoted as ASR_{ER}).

Next, we evaluated Agile targeted impersonation against four SOTA FR models (with MTCNN and Euclidean-L2). In the I-1K dataset, we first identified the targets who could be impersonated (denoted as $\mathbb{T}_{candidate}$). Similar as discussed above, for each target in $\mathbb{T}_{candidate}$, we found the set of candidate attackers ($\mathbb{A}_{PSAS}$) and the corresponding optimal and extended ranges for laser current. The simulated attacks in the digital domain showed that the identified attackers could impersonate most of the targets (denoted as $\mathbb{T}_{filtered}$), showing that the ES and PSAS filters achieved high filter efficiency (FE). Finally, we launched the Agile targeted impersonation attacks in the physical domain with the attackers in $\mathbb{A}_{PSAS}$ and the extended ranges. Table 8 reports the set of the target ($\mathbb{T}_{impersonated}$) that were successfully impersonated in the real-world attacks and the average ASRs for each FR model. The ASR on VGGFace2 is lower than that of the other models since the attackers were misidentified as individuals other than the original targets (i.e., the attacker is close to more than one target in the embedding space). In the latest models that maximize the separation

Table 9: Informed targeted impersonation at different distances.

	Laser-to-Camera Distance	32 *cm*	35 *cm*	38*cm*
(A1, T1)	informed setting (mA)	[74-94]	[90-110]	[108-128]
	# of successful attacks	11	10	13
(A2, T2)	informed setting (mA)	[56-76]	[66-86]	[80-100]
	# of successful attacks	9	8	6
(A3, T3)	informed setting (mA)	[114-134]	[138-158]	[166-186]
	# of successful attacks	13	15	16

among identities, such as ArcFace and SFace, the Agile attack could achieve over 95% attack success rates.

6.5 Discussions

6.5.1 Attack Time. Agile consists of an offline synthesis phase and a real-time attack phase. Our experiments demonstrate that applying filters reduces over 90% of the offline synthesis time. For example, in a targeted impersonation attack against the I-1K database with 15 attackers, the attack synthesis process using grid search took 268.75 hours to test all 150 laser settings (using the FaceNet model on the DeepFace platform). However, applying filters reduced the simulation time to 2.97 hours.

The real-world attack phase involves camera alignment and current adjustment steps. During camera alignment, attackers can visually align the laser with the camera using screen feedback, leveraging the point-symmetric relationship between the light source and the ghost image as proposed by GhostImage [57]. We conducted an experiment with 7 participants, each repeating the alignment process 10 times. On average, participants completed the alignment in 2.35 seconds. Over repeated tests, alignment times improved significantly; the slowest participant initially took 4.5 seconds but reduced to 1.6 seconds after multiple attempts.

Participants without screen feedback could observe a red spot reflected in the camera lens and attempt to center it manually. In experiments involving 7 participants, the alignment time ranged from 1.41 seconds to 6.64 seconds, with an average of 2.81 seconds. Similar to the screen feedback method, participants showed improvement in alignment time with repeated attempts. Alignment without screen feedback may be less precise, but it still maintains accuracy within ±3°, which does not significantly impact the attack's performance.

In current adjustment, Agile selects the optimal current range based on informed information from the attack synthesis to compensate for inaccuracies in distance and ambient light. Since these factors are unpredictable, the adjustment time is empirically evaluated. Experimental results from Sections 6.3 and 6.4 indicate that all pairs can achieve the desired attack effect within 20 settings. The current adjustment interval is set to match the frame time, allowing up to 4 seconds at 5 frames per second, with an average of 1.6 seconds. This rate corresponds to the minimum refresh rate used by DeepFace [32] to ensure compatibility across all models. However, this rate can be significantly increased. For example, FaceNet recommends 25 frames per second in its demo [45], which could potentially reduce the time for current adjustment by five-fold.

6.5.2 Attack Robustness. Next, we discuss and empirically assess the tolerance of Agile against inaccuracies in laser-to-camera distance, angle misalignment, and varying lighting conditions.

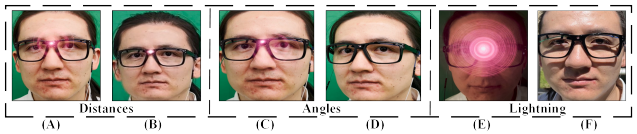

Figure 8: Images captured at (A) normal distance (35cm), (B) long distance (2m), (C) 6° angle, (D) 12° angle, (E) dark environment (no external lighting), and (F) extremely bright environment (under sunlight).

Attack Effectiveness with Distance Inaccuracies. The Agile attack operates effectively within a specific distance range due to the constraints of the selected laser diode's operating power. Based on this power, the effective laser-to-camera distance ranges between 32 *cm* and 38 *cm*, within which Agile can achieve targeted impersonation with a high success rate.

We empirically validated the effectiveness of Agile in targeted information attacks under different distances. We randomly selected 3 pairs of attackers and targets from 15 attackers using the I-1K dataset, denoted as (A1, T1), (A2, T2), and (A3, T3). Each attacker could successfully impersonate the target in the digital domain. Through informed impersonation, a current range was determined to guide the physical-domain attacks, as shown in Table 9, where all $K = 20$ settings within each range were tested.

Across three different distances, the attacks achieved high attack success rates (ASRs), which remained consistent. At distances below 30 *cm*, the laser signal was too powerful, often causing dodging attacks. Conversely, at distances longer than 40 *cm*, dodging attacks rarely succeeded. In extreme cases, such as distances over 2 meters, the attack failed due to the extremely weak laser signal, as illustrated in Figure 8B. Within the effective distance range, inaccuracies in distance measurement can be compensated by adjusting the current. For instance, at a distance of 35 cm, varying the distance between 1 *cm* and 3 *cm* in both forward and backward directions consistently resulted in successful Agile targeted impersonation attacks.

Attack Effectiveness with Angle Misalignment. While Agile ideally expects the laser-to-camera angle to be 0°, we empirically tested attack images with ±1, ±2, and ±3 degrees in targeted impersonation. These angles were calculated based on the distance and ghost pixel coordinates, following the method proposed in GhostImage [57]. The attacks generally succeeded with a slight decrease in attack success rates as the angle deviated from 0°. For instance, at angles up to 6° (Figure 8C), the attacks showed inconsistent results compared to perfectly aligned angles. When the angle increased to 12°, exceeding the laser's divergence angle, the attack failed most of the time. This failure occurred because the camera could not fully capture the laser signals beyond the divergence angle (Figure 8D).

Agile under Different Lighting Conditions. We conducted experiments under two ambient light conditions in an indoor setting: full lighting (500 lux) and partial lighting (300 lux). We repeated experiments for dodging, targeted, and untargeted impersonation, achieving above 90% attack success rates (ASRs). Dodging and untargeted impersonation attacks yielded similar robust results under normal lighting conditions, whereas some targeted impersonation attacks failed, indicating sensitivity to environmental noise. Under extremely bright conditions, such as direct sunlight (8F), the attacks consistently failed. Conversely, in dark environments, such as indoors without additional lighting (8E), the laser acted as the

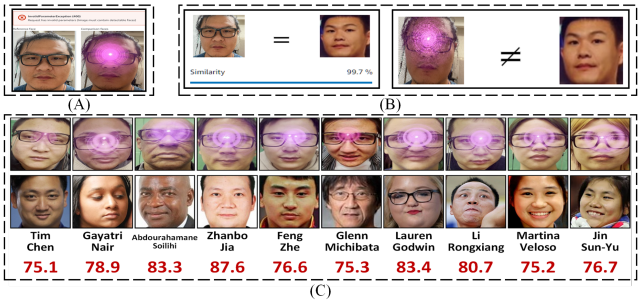

Figure 9: Grey-box and black-box attack against Amazon ReKognition: successful DoS (A), dodging (B), and impersonations with confidence scores (C).

primary light source, causing camera saturation and resulting in dodging and/or denial-of-service (DoS) effects.

6.6 Grey-box and Black-box Attacks

We evaluated the Agile attacks against the *Amazon Rekognition* system [27]. Using its API, we tested DoS and identity dodging attacks using the *Face Comparison* module and the untargeted impersonation attacks using the *Celebrity Recognition* under the black-box setting. Besides, we tested database dodging using the *Searching for a face using an image* function under the grey-box setting, which adopted *LFW* as the target dataset. In this experiment, we have no knowledge about the pre-processing and face recognition models used by Amazon Rekognition.

The benign and attack images of 15 attackers were used to test DoS and dodging attacks. As shown in Figure 9A, the Amazon Rekognition system returned a message 'images must contain detectable faces' for the input attack image, indicating a successful DoS attack. In Figure 9B, for a benign image with a high confidence score of 99.7%, the system failed to recognize the attack image with the laser perturbation, indicating a a successful identity dodging. Similar results were obtained in database dodging attacks.

Next, we tested untargeted impersonation attacks on Amazon Rekognition. If the input face image is "recognized" (with a default threshold of 75.0), the system will return the celebrity's name, the profile page, and a confidence score. We uploaded the original and 2,250 attack images of 15 attackers with 150 laser perturbations generated with $I \in [60mA, 210mA]$ through Amazon Rekognition's API. For 15 attackers, none of their original images was matched with any celebrity in the database. Under the Agile attack, 10 attackers were identified as celebrities, as shown in Figure 9C.

While face recognition models hosted at large companies are expected to be more robust than open-source models and research prototypes, this experiment shows that Agile is quite effective against a commercial face recognition product under black-box and grey-box settings. We had an interesting observation that the impersonated targets shared the same gender and race as the attackers, indicating that these factors may play a role in commercial services.

7 ADDITIONAL ATTACK SURFACES

7.1 The Continual Attack

The real-world face recognition systems often adopt continual learning (a.k.a. incremental learning), in which images from recently

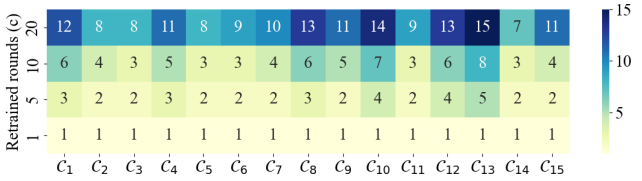

Figure 10: Attack propagation in the continual attack.

recognized subjects are added to the dataset and the model is periodically retrained with updated data. Hence, the number of images for each subject increases, and the recognition accuracy improves over time [26]. However, FR systems are shown to be highly susceptible to backdoor attacks with physical triggers [50]. Continual learning introduces a new attack vector that allows poisoning samples to be injected through a legitimate channel. In such settings, once an Agile attacker successfully impersonates a target in the dataset, the attack image with laser perturbation is added to the training dataset and labeled as the impersonated identity. Once the face recognition model is retrained, the laser perturbation will be "learned" as a feature of the impersonated identity, so that future Agile attacks can more easily succeed using the same attacker with different laser settings or even with varied attackers.

We evaluated the effectiveness of the Agile attacks in a continual learning setting using FaceNet with an SVM classifier for face recognition. 143 identities in the *LFW* dataset, each with more than 10 images, were selected to train the initial classifier, denoted as C_0. Next, we identified 15 pairs of attackers and targets where each attacker can successfully impersonate the target under laser perturbations. Finally, we inserted the images of each attacker into the training dataset to retrain 15 new classifiers, one for each attacker, for 20 rounds. In each round, we inserted one image of the attacker into the corresponding training set.

Let (A_i, T_i) denote a pair of attacker and target and C_i denote the classifier polluted by A_i, where $1 \leq i \leq 15$. In each round, we tested each C_i with the images of all 15 attackers and recorded the number of attackers who were mis-identified as the target T_i. Figure 10 reports the results at rounds 1, 5, 10, and 20. The attack images inserted into the training sets after the first round performed a data poisoning attack against the classifiers. Consequently, in the following rounds, besides the intended attacker, an increasing number of other attackers with laser perturbations were misidentified by each classifier. At round 20, most of the attackers can impersonate the initial targets, indicating that the laser perturbations become a discriminative feature learned by the classifiers.

7.2 Predictable Untargeted Impersonation

Untargeted impersonation attacks are considered straightforward but are rarely implemented in real-world scenarios due to the uncertainty of which identities could be impersonated. This information is crucial in systems requiring additional verification, such as secondary authentication. Conversely, due to the limitations of physical-domain attacks, targeted impersonation typically has a lower success rate compared to untargeted impersonation.

To fully leverage the capability of Agile attacks, we propose a variation of untargeted impersonation, called *predictable untargeted impersonation*. This approach identifies a small set of candidate targets who are highly likely to be impersonated by the selected

attacker(s). We consider this method more practical for impersonation attacks in real-world scenarios compared to both targeted and traditional untargeted approaches.

Approach. In untargeted impersonation, we utilized two filters and synthetic attacks to assess an attacker's capability of impersonation. Here, we reverse the process to identify the most robust target for each attacker and the corresponding optimal current settings. In particular, for each attacker, we select the targets where the attacker-target pairs have optimal ranges exceeding 10 mA. Among these pairs, we choose the one with the smallest distance and output the target as the "predictable target" for that attacker, with the laser settings that perform best established as the optimal range.

Experiment Results. We conducted predictable untargeted impersonation experiments using the I-1K dataset from LFW and YouTube Faces DB (YTF) [51] against three SOTA models. We employed the proposed approach to identify the candidate target and the optimal range for each attacker. This information guides real-world attacks. Table 10 shows the predictable targets for 15 attackers in three FR models and two identity datasets with 1,000 individuals. We observed that highly robust FR models (e.g., SFace) yield results consistent with real-world scenarios, where the attackers always impersonated the predictable targets. In models of medium robustness (e.g., FaceNet), for a small fraction of attackers, the predictable targets may not be found. In less robust models (e.g., DeepFace), the same predictable target may be identified for different attackers.

Table 10: Result of predictable untargeted impersonation.

	Dataset	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
SFace	YTF	$T_{1,1}$	$T_{1,2}$	$T_{1,3}$	$T_{1,4}$	$T_{1,5}$	$T_{1,6}$	$T_{1,7}$	$T_{1,8}$	$T_{1,9}$	$T_{1,10}$	$T_{1,11}$	$T_{1,12}$	F	$T_{1,14}$	$T_{1,15}$
	LFW	$T_{2,1}$	$T_{2,2}$	$T_{2,3}$	$T_{2,4}$	$T_{2,5}$	N_i	$T_{2,7}$	$T_{2,8}$	F	$T_{2,10}$	F	F	F	$T_{2,14}$	$T_{2,15}$
FaceNet	YTF	$T_{3,1}$	$T_{3,2}$	$T_{3,3}$	$T_{3,4}$	$T_{3,5}$	$T_{3,6}$	$T_{3,7}$	N	$T_{3,9}$	$T_{3,10}$	N	N	$T_{3,13}$	$T_{3,14}$	$T_{3,15}$
	LFW	$T_{4,2}$	$T_{4,3}$	$T_{4,4}$	$T_{4,5}$	$T_{4,6}$	$T_{4,7}$	$T_{4,7}$	N	$T_{4,9}$	$T_{4,10}$	N	N	$T_{4,13}$	$T_{4,14}$	$T_{4,15}$
DeepFace	YTF	$T_{5,1}$	N	$T_{5,3}$	$T_{5,4}$	$T_{5,5}$	$T_{5,1}$	$T_{5,7}$	$T_{5,8}$	$T_{5,9}$	$T_{5,10}$	$T_{5,11}$	$T_{5,12}$	W	$T_{5,1}$	$T_{5,15}$
	LFW	$T_{6,1}$	$T_{6,2}$	$T_{6,3}$	$T_{6,4}$	$T_{6,5}$	$T_{6,6}$	$T_{6,7}$	$T_{6,8}$	$T_{6,9}$	$T_{6,10}$	$T_{6,11}$	$T_{6,12}$	$T_{6,13}$	W	$T_{6,15}$

$T_{i,j}$ denote an anonymized user in two datasets; F means the attacker failed in the physical-domain impersonation; N means no predictable target was found; W means the attacker impersonated a target other than the predictable target.

8 EXISTING DEFENSES AND LIMITATIONS

Existing defenses against physical laser attacks operate on three layers: optical, digital image, and FR model. We will discuss the limitations of these defenses against our novel attack.

Optical Filters. The simplest defense against infrared-based attacks is to mount an *IR cut-off filter* on the camera lens. However, some cameras are designed to capture infrared light to enhance image quality in low light conditions, and an IR filter would interfere with their normal operations. Attackers can bypass this defense by using a wavelength near the red region, partially sacrificing stealth. For instance, common cut-off wavelengths are around 700 nm, but using a laser diode like the HL6750MG with a 685 nm wavelength can bypass the filter. Despite this adjustment, our agile attack remains effective even when an IR cut-off filter is used.

Adversarial Image Detection. An adversarial image detection mechanism can be integrated into the image processing pipeline. Agile images can be detected using image features or a fine-tuned binary deep-learning model for distinguishing between adversarial and benign face images. However, the infrared laser patterns resemble normal red circles (e.g., red ornaments or stickers), which may increase the true negative rate. This resemblance can also be exploited as a new attack surface for DoS attacks.

Model Robustness. Adversarial training involves "teaching" DL models with known adversarial examples and can be used to defend against Agile or other physical attacks. However, this approach is costly, as it requires generating adversarial images and retraining the large face recognition model. Our attack creates adversarial samples that closely resemble normal users. This complicates the distinction between legitimate users and attackers. While careful design and refinement of the training dataset may reduce the increase in true negative cases, attackers can counteract these measures by adjusting the strength, wavelength, and shape of the laser perturbation.

Other Monitoring Cameras. Figure 5 shows that the laser's power is concentrated within a narrow cone. At a distance of $2m$, the cone's radius is $27cm$, but the power density drops to $0.0001w/cm^2$. Therefore, it is impractical for another camera to capture the laser unless it is positioned next to the target camera.

Attack Limitations. Like other physical attacks, Agile cannot work over long distances, under extreme lighting conditions, or with large laser-to-camera angles. FR systems typically automate identity recognition services without human oversight. Traditionally, security personnel patrol areas to identify suspicious individuals, and Agile appears less suspicious than other physical attacks (Figure 1). However, if a human observer monitors the output of the FR system, it presents a challenge similar to other physical attacks.

9 CONCLUSION

In this paper, we present Agile, a novel, highly stealthy, and very effective physical attack against deep face recognition systems. It generates adjustable, invisible laser perturbations and emits them into the camera CMOS to launch DoS, dodging, and impersonation attacks. We develop a theoretical model to generate synthesized attack signals and utilize them to guide the implementation of the physical attacks. Our method's efficacy and feasibility are validated through diverse attacks on both SOTA open-source FR models and sophisticated commercial models. Given Agile's effectiveness in exploiting practical limitations and providing more attack surfaces, coupled with the inadequacy of existing protective measures, there is a highlighted need for enhanced defensive strategies.

ACKNOWLEDGMENTS

Ye Wang, Zeyan Liu, Bo Luo, and Fengjun Li were supported in part by NSF IIS-2014552, DGE-1565570, and the Ripple University Blockchain Research Initiative. The authors would like to thank the anonymous reviewers and the shepherd for their valuable comments and suggestions, and the volunteers in the user study.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *NeurIPS* 31 (2018).
[2] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6, 2 (2016), 20.
[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing robust adversarial examples. In *ICML*. PMLR, 284–293.

[4] Q. Cao, L.i Shen, Weidi Xie, O. M. Parkhi, and A. Zisserman. 2017. VGGFace2: A dataset for recognising faces across pose and age. (2017). arXiv:1710.08092

[5] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*. 39–57.

[6] CBP. [online; accessed 19-January-2023]. Say hello to the new face of speed, security and safety. https://biometrics.cbp.gov/.

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE WACV*.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF CVPR*.

[9] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. *CoRR* abs/1905.00641 (2019). arXiv:1905.00641 http://arxiv.org/abs/1905.00641

[10] Fred M. Dickey and Scott C. Holswade. 1996. Gaussian laser beam profile shaping. *Optical Engineering* 35, 11 (1996), 3285 – 3295. https://doi.org/10.1117/1.601069

[11] H. Du, H. Shi, D. Zeng, X. Zhang, and T. Mei. 2022. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM CSUR* 54, 10s (2022).

[12] Ranjie Duan, Xiaofeng Mao, Alex K. Qin, Yun Yang, Yuefeng Chen, Shaokai Ye, and Yuan He. 2021. Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink. *2021 IEEE/CVF CVPR* (2021).

[13] Erik Learned-Miller et. al. [online; accessed 19-January-2023]. Labeled Faces in the Wild Home. http://vis-www.cs.umass.edu/lfw/index.html.

[14] Elaine Glusac. 2021. Your Face Is, or Will Be, Your Boarding Pass. The New York Times. https://www.nytimes.com/2021/12/07/travel/biometrics-airports-security.html [Accessed 20-Dec-2022].

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[16] Nitzan Guetta, Asaf Shabtai, Inderjeet Singh, Satoru Momiyama, and Yuval Elovici. 2021. Dodging attack using carefully crafted natural makeup. *arXiv preprint arXiv:2109.06467* (2021).

[17] Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales, and Javier Galbally. 2019. Introduction to face presentation attack detection. In *Handbook of Biometric Anti-Spoofing*. Springer, 187–206.

[18] Gary B Huang and Erik Learned-Miller. 2014. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep* 14, 003 (2014).

[19] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.

[20] Stepan Komkov and Aleksandr Petiushko. 2019. AdvHat: Real-world adversarial attack on ArcFace Face ID system. (2019). arXiv:1908.08705

[21] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*.

[22] Dinh-Luan Nguyen, Sunpreet S Arora, Yuhang Wu, and Hao Yang. 2020. Adversarial light projection attacks on face recognition systems: A feasibility study. In *Proceedings of the IEEE/CVF CVPR Workshops*.

[23] OpenCV. 2015. Open Source Computer Vision Library.

[24] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *Euro S&P*.

[25] Mikhail Pautov, Grigorii Melnikov, Edgar Kaziakhmedov, Klim Kireev, and Aleksandr Petiushko. 2019. On adversarial patches: real-world attack on arcface-100 face recognition system. In *2019 SIBIRCON*. IEEE.

[26] S. Prasad, A. Sawant, R. Shettigar, and S. Sinha. 2011. Real-Time Face Recognition System with Dynamic Training and Enhanced Multi-Algorithm Face Recognition. In *ICWET*. https://doi.org/10.1145/1980022.1980047

[27] Amazon AWS Rekognition. 2018. Amazon Rekognition. https://aws.amazon.com/rekognition/

[28] Research and Markets. 2022. Biometrics Global Market Report 2022. https://www.researchandmarkets.com/reports/5652782/biometrics-global-market-report-2022.

[29] Gwonsang Ryu, Hosung Park, and Daeseon Choi. 2021. Adversarial attacks by attaching noise markers on the face against deep face recognition. *Journal of Information Security and Applications* 60 (2021), 102874.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE CVPR*.

[31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE ICCV*.

[32] Sefik Serengil. 2023. deepface. https://github.com/serengil/deepface

[33] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 ASYU*. IEEE.

[34] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac CCS*.

[35] Meng Shen, Zelin Liao, Liehuang Zhu, Ke Xu, and Xiaojiang Du. 2019. Vla: A practical visible light-based attack on face recognition systems in physical world. *Proceedings of the ACM on IMWUT* 3, 3 (2019).

[36] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[37] Ove Steinvall. 2021. The potential role of laser in combating UAVs: Part 2. In *Technologies for Optical Countermeasures XVIII and High-Power Lasers: Technology and Systems, Platforms, Effects V*, Vol. 11867. SPIE, 14–30.

[38] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. *NeurIPS* 27 (2014).

[39] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. 2015. Deepid3: Face recognition with very deep neural networks. *arXiv:1502.00873* (2015).

[40] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *AAAI*.

[41] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*.

[42] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. 2021. Demon in the Variant: Statistical Analysis of {DNNs} for Robust Backdoor Contamination Detection. In *30th USENIX Security Symposium (USENIX Security 21)*. 1541–1558.

[43] Thorlabs. 2023. Thorlabs-HL6750MG 685 nm. https://www.thorlabs.com/thorproduct.cfm?partnumber=HL6750MG. [Accessed 04-Jun-2023].

[44] Thorlabs. January 16, 2019. *785 nm Laser Diode*, L785H1. https://www.thorlabs.com/thorproduct.cfm?partnumber=L785H1

[45] Tim Timesler. 2023. Facenet-Pytorch. https://github.com/timesler/facenet-pytorch

[46] Mairs Turnstile. 2024. Turnstile Gate with Face Recognition. https://mairsturnstile.com/turnstile-gate-with-face-recognition.html.

[47] F. Vakhshiteh, A. Nickabadi, and R. Ramachandra. 2021. Adversarial attacks against face recognition: A comprehensive study. *IEEE Access* 9 (2021).

[48] Zhenting Wang, Juan Zhai, and Shiqing Ma. 2022. BppAttack: Stealthy and Efficient Trojan Attacks against Deep Neural Networks via Image Quantization and Contrastive Adversarial Learning. In *Proceedings of the IEEE/CVF CVPR*.

[49] Xingxing Wei, Ying Guo, and Jie Yu. 2023. Adversarial Sticker: A Stealthy Attack Method in the Physical World. *IEEE TPAMI* 45, 3 (2023), 2711–2725.

[50] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. 2021. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF CVPR*. 6206–6215.

[51] Lior Wolf, Tal Hassner, and Itay Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*. 529–534.

[52] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. 2021. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF CVPR*.

[53] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. 2020. Adversarial t-shirt! evading person detectors in a physical world. In *ECCV*. Springer.

[54] Takayuki Yamada, Seiichi Gohshi, and Isao Echizen. 2013. Privacy visor: Method for preventing face image detection by using differences in human and device sensitivity. In *IFIP CMS*. Springer.

[55] C. Yan, Z. Xu, Z. Yin, X. Ji, and W. Xu. 2022. Rolling Colors: Adversarial Laser Exploits against Traffic Light Recognition. In *USENIX Security*.

[56] Xiao Yang, Yinpeng Dong, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. 2022. Controllable Evaluation and Generation of Physical Adversarial Patch on Face Recognition. *arXiv e-prints* (2022), arXiv–2203.

[57] Man Yanmao, Li Ming, and Gerdes Ryan. 2020. *GhostImage*: Remote Perception Attacks against Camera-based Image Classification Systems. In *RAID 2020*.

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

[59] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[60] Yaoyao Zhong and Weihong Deng. 2020. Towards transferable adversarial attack against deep face recognition. *IEEE TIFS* 16 (2020).

[61] Yaoyao Zhong, Weihong Deng, Jiani Hu, Dongyue Zhao, Xian Li, and Dongchao Wen. 2021. SFace: Sigmoid-Constrained Hypersphere Loss for Robust Face Recognition. *IEEE Transactions on Image Processing* (2021).

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *IEEE CVPR*.

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

[64] Sijie Zhu, Taojiannan Yang, and Chen Chen. 2021. Visual Explanation for Deep Metric Learning. *IEEE Transactions on Image Processing* 30 (2021).

[65] Zheng-An Zhu, Yun-Zhong Lu, and Chen-Kuo Chiang. 2019. Generating adversarial examples by makeup attacks on face recognition. In *2019 IEEE ICIP*.

[66] Alon Zolfi, Shai Avidan, Yuval Elovici, and Asaf Shabtai. 2021. Adversarial Mask: Real-World Universal Adversarial Attack on Face Recognition Models. *arXiv preprint arXiv:2111.10759* (2021).