Spoofing Real-world Face Authentication Systems through Optical Synthesis

Yueli Yan, Zhice Yang
School of Information Science and Technology, ShanghaiTech University, China

***Abstract*—Facial information has been used for authentication purposes. Recent face authentication systems leverage multi-modal cameras to defeat spoofing attacks. Multimodal cameras are able to simultaneously observe the targeting people from multiple physical aspects, such as the visible, infrared, and depth domains. Known spoofing attacks are not effective in evading the detection since they cannot simulate multiple modalities at the same time.**

This paper presents a new class of spoofing attacks on multimodal face authentication systems. Its main idea is to forge each and every modality and then combine them together to present to the camera. The attack is realized with a special display device called Hua-pi display. It costs less than $500 and incorporates dedicated scene generators to optically reproduce multimodal scenes of an authorized user, and then synthesizes the scenes together at the camera's view point through optical combiners to fool face authentication systems. We evaluate the risks of this attack by systematically testing it against the latest commercial face authentication products from major vendors in the field. The results not only demonstrate a successful bypass rate of 80% but also characterize the impacting factors and their feasible regions, revealing a new and realistic threat in the field.

I. Introduction

Face recognition has profoundly changed the world. It drives a growing market of US$ 5 billion, which will reach up to US$ 12 billion in 2028 [1]. With the ever-increasing accuracy and efficiency, face recognition has been applied in security-sensitive areas such as authentication.

Face, however, has raised many debates when using it as the authenticator. Facial features are unique and stable, but unlike other biometric features such as fingerprint and iris print, facial information is much more accessible. As the most common and natural personal identifier, faces are exposed to the public. With the prevalence of social media and short videos, multi-media content containing rich facial information is uploaded to the Internet for sharing and entertainment. These opportunities allow adversaries to gather facial information efficiently, and further abuse for, *e.g.*, impersonation [2]. Without proper defense, showing a piece of paper with a face printed on it is enough to make the face recognition system believe the presence of the printed people. This is because face recognition algorithms alone cannot and are also not designed to judge whether the input content is from spoofing artifacts or not.

Spoofing attacks alike are threats against applying face recognition for authentication. As such, anti-spoofing methods or liveness detectors are developed and employed in face authentication systems. They are designed to verify whether the facial content captured by the authentication system is from a real person appearing in front of the camera or from the artifacts.

There are two types of anti-spoofing methods. Dynamic anti-spoofing methods analyze captured videos or frames to search for spoofing evidence. One common way is similar to CAPTCHA [3]. It challenges the user with certain actions to judge cognition. Static anti-spoofing methods make decisions according to a single shot of the scene. It detects spoofing artifacts according to subtle imaging differences. The key advantage of static methods is user experience, as it takes much less time than dynamic methods and involves no user actions.

The advantages of static anti-spoofing methods are reflected in the market. They have been widely adopted by commercial products. For example, face authentication is a common feature of access control systems [4]. Facial payment terminals are now popular in shopping malls [5], where people show their faces to authorize transactions. Apple's FaceID has evolved for five years and many high-end smartphones have similar features [6]. The success of static anti-spoofing methods is also backed by the fact that, by far, they have not been effectively compromised. It is reported that attackers used high-quality 3D masks and head models to evade the detection [7], but the high fabricating cost makes such attacks not practically feasible at scale.

In this paper, we report a novel class of spoofing attacks on face authentication systems (FAS). The attack is launched with a special display device called *Hua-pi* [1] display. When it is placed in front of the FAS's camera and shows the content of an authorized user. The FAS will be fooled and fail to be aware of the spoofing situation. This allows the adversary to impersonate that user and bypass the protection.

Hua-pi attack is dedicatedly designed to defeat current static anti-spoofing methods. We extensively test and successfully show its effectiveness against ALL (16) latest commercial FAS products from leading smartphones, access control systems, and payment solutions vendors in the worldwide market. Their FASes already cover a wide range of production-level static anti-spoofing technologies and algorithms. *Hua-pi* attack is

Corresponding author: Zhice Yang, yangzhc@shanghaitech.edu.cn

[1]Literal meaning: painted skin, named after the short story of the same name collected in *Strange Tales from a Chinese Studio* [8].

launched in an open and physical environment, and is validated with 20 participants. We have not noticed any public reports revealing such risks of these products at this scale and with similar test conditions.

Hua-pi attack does not make use of vulnerabilities of specific face recognition and anti-spoofing algorithms. Its consistent effectiveness is based on our finding that the security basis of current static anti-spoofing methods is fundamentally flawed. *Hua-pi* display is a heuristically-designed and low-cost device to exploit this finding in practice. It allows instant regeneration and precise presentation of the spoofing content while costing less than US$ 500 (see Table VII). When evaluating real-world products, we further reveal several severe inefficiencies in current FAS designs that further increase the risk, *e.g.*, many of them do not verify the consistency of identities across modalities.

Due to the above results, we view *Hua-pi* attack as a new and true threat to face authentication systems. In the following sections, we first review the literature in Section II and clarify the attack model in Section III. Then, we describe the attack details in Section IV and V.

II. Background and Related Work

Face authentication system verifies the identity of the user through the captured facial content. A typical face authentication system (FAS) consists of a camera and three algorithmic modules: face detection, anti-spoofing, and face recognition. These algorithms take images from the camera as input and work together to make authentication decisions.

The workflow of a typical FAS is shown in Figure 1. The face detection module first extracts the area containing a face from the image captured by the camera. The face recognition module identifies whether the face belongs to the database. When the database contains multiple people, it solves a 1:N matching problem, and is also called face identification [9]. When there is only one authorized person, it solves a 1:1 matching problem and is called face verification [10]. Face recognition alone cannot be used for authentication, since it only judges the similarity of the input face to the database, and cannot verify whether the face image captures the live user or some spoofing artifacts. Therefore, the anti-spoofing module is employed to detect spoofing attempts [11]. It speculates about the truthfulness of the input face, and hunts for evidence from all possible aspects. The decision of the FAS considers the results of both modules. In the following, we review existing attacks and countermeasures targeting face recognition and anti-spoofing, respectively.

A. Attacks on Face Recognition Algorithms

Attacks on the face recognition module mainly focus on exploiting algorithmic vulnerabilities to affect the recognition results. Classical hand-crafted recognition algorithms are sensitive to lighting conditions, face orientations, and image qualities. Imperfect conditions can lead to misclassifications. Recent advances in artificial neural networks have greatly improved face recognition performance, but neural networks

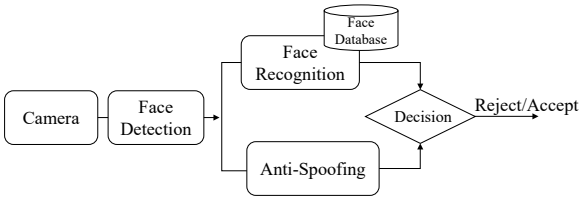


Fig. 1. **Workflow of a Typical Face Authentication System.** The authentication decision is made according to the decisions of the face recognition module and the anti-spoofing module.

are found not stable with respect to small variations of the input. When the input content is slightly disturbed, the inference results may be completely different [12]. This property implies vulnerabilities and has led to extensive discussion. Sharif *et al.* [13] printed dedicated patterns on an eye class to mislead the algorithm for impersonation. Similar approaches were explored in other forms such as fake eyes [14], special hat [15], *etc.* Zhou *et al.* [16] used a projector instead of physical artifacts to cast desired perturbations. Shen *et al.* [17] directly project photos to overwrite the adversary's face.

These attacks mislead face recognition algorithms to draw wrong decisions and usually do not consider FAS's anti-spoofing capabilities, so it is not clear whether the artifact that the adversary makes use of and the perturbation patterns can be detected or not. Moreover, since they exploit algorithm-specific vulnerabilities, white-box models, and/or access to the training set are needed to generate robust adversary patterns for launching attacks in the physical world. Hence, applying them to evade general black-box commercial FAS products is still an open issue. As a comparison, *Hua-pi* attack belongs to spoofing attacks and assumes the recognition algorithms work properly as expected (*e.g.*, the recognition vulnerabilities have been eliminated [18], [19]). It uses a display device to present the unmodified facial content of the authorized user to the FAS and bypasses it by fooling the anti-spoofing module.

B. Spoofing Attacks and Anti-spoofing Methods

There are three known classes of spoofing attacks. They differ in the way of presenting the facial content. 1. Image-based attacks print the face on paper or show it on a display. 2. Video-based attacks play a video recording that contains the face. 3. 3D-model attacks use the face as the reference to fabricate 3D models or face masks. The evolution of spoofing attacks is either to increase presentation quality, *e.g.*, using the latest 3D-mask manufacturing technologies [20], [21], or to reduce attack cost. *e.g.*, synthesizing 3D-model from public 2D photos [22]. To counter them, many anti-spoofing methods are proposed, and they fall into the following two categories.

Dynamic anti-spoofing methods observe the scene in front of the camera for a while and infer dynamics from multiple video frames. Image-based attacks can then be differentiated by abnormal optical flow [23], lens distortion [24], biometric motion [25], *etc.* Improved dynamic methods incorporate challenge and response protocols to thwart video attacks. They ask the user to finish certain actions within a period [26]. A passive way is to use stimulation, *e.g.*, screen flash [27],

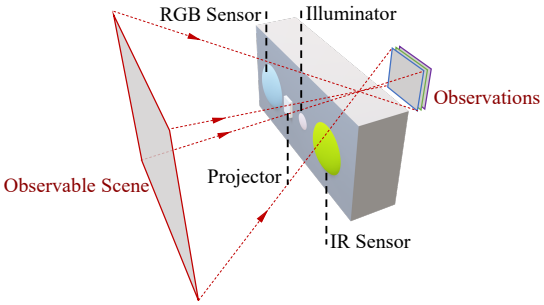


Fig. 2. **Multimodal Camera.** The camera incorporates multiple sensors, *e.g.*, RGB, IR, and depth sensors, to observe the scene in front of it and generate multimodal observations.

to detect whether the face is shown on a plane. Clues other than video frames, *e.g.*, sensor data [28], are used to detect replays. Dynamic methods can be defeated if the adversary is able to mimic the facial dynamics correctly. Recent work by Xu *et al.* [22] proposes to forge a virtual 3D face model in virtual-reality displays in real time to bypass dynamic anti-spoofing detection.

Static anti-spoofing methods are designed with a different philosophy. They examine the scene for a much shorter period and use a single shot to make decisions. In the spatial domain, they search for spoofing evidence in color space [29], texture [30], [31], environmental context [32], *etc.* Recent work makes use of neural networks to extract implicit traits in spoofing attempts [33].

However, for attacks that make use of advanced presentation tools such as high-fidelity 3D models [34], it is still challenging to maintain robust anti-spoofing accuracy with a single shot. This leads to the expansion of the spatial domain. It has been shown that the difference between spoofing artifacts and genuine faces remains large in other physical domains [35]. For example, image and video-based attacks can be detected in the depth domain since display panels are flat and lack facial topography [36]. 3D masks and models can be differentiated since their materials under infrared (IR) illumination is different from that of human skin [37]. The above advantages lead to the trend of adopting multimodal cameras for anti-spoofing. Multimodal cameras can observe an object from multiple physical domains, *i.e.*, modalities, rendering rich information for anti-spoofing.

C. Multimodal Camera

Multimodal cameras have been widely adopted in commercial FASes. They strengthen static anti-spoofing methods without prolonging the authentication process or involving user actions. This section introduces this technology and defines the terminologies used in the context.

As shown in Figure 2, the collection of light signals that can be seen from a point is called a "scene". The observable scene of a camera consists of the light signals from its field of view. The image sensor of the camera is used to measure the light signals of the observable scene, and a shot of the sensor is called an "observation" of the scene. A multimodal camera is equipped with multiple dedicated optical sensors to measure different physical aspects of the observable scene, such as chromatic reflection, infrared emission, physical distance, *etc.* A single shot of a multimodal camera contains observations of multiple modalities from its sensors. Hence, multimodal cameras have a more complete physical figure about the scene than ordinary single modality cameras.

RGB sensor is the most widely used image sensor. It is an array of millions of tiny sensors capable of measuring the intensity of light falling on it. The chromatic information is perceived by applying red, green, and blue (RGB) light filters in front of each sensor. The wavelength of visible light signals falls between 380 to 700 nm, but the light sensor has a wider sensitivity range. To avoid the interference of light of other wavelengths, a bandpass filtering layer is usually coated on the lens in front of the RGB sensor to select light transmitting through.

Infrared (IR) or near-infrared (NIR) sensor, is identical to RGB sensor in the working principle. IR sensor is a single-channel grayscale sensor and does not decompose chromatic channels. 850 nm and 940 nm are the two most common pass wavelengths for IR sensors' passband filters. Since environmental IR light is weak, IR sensors are usually used in conjunction with IR illuminators, *e.g.*, IR LED or lasor diode, to light up the observable scene.

Depth sensor is able to measure its distance to the objects in the observed scene. When its resolution is high, the objects are small and massive. The distances to them outline the 3D topography of the scene. There are mainly two types of depth sensing technologies in FAS cameras - structured light and Time of Flight (ToF). Their working principles are quite different and will be detailed later in Section IV-D1 and Section IV-E1. They all use active light projectors to aid depth measurement. The two technologies are independent of light wavelength, but in practice, they work in IR band to avoid visible interference.

While there are sensors that can measure other modalities such as temperature, most commercial FAS cameras use the above three modalities for cost and forming factor consideration. A typical three-modality camera is shown in Figure 2, but it is also common to see cameras that use two-modality combinations, *e.g.*, RGB and IR. We also note that the components of different modality sensors may be logically separated but physically integrated. For example, IR sensor is usually multiplexed for both IR imaging and depth sensing. The illuminator and projector are combined in some products.

Recent work proposed several static anti-spoofing algorithms based on multimodal cameras [33], [38]–[41], showing superior performance over conventional means. For commercial FAS products, little information about their anti-spoofing algorithms is disclosed, but their adoption of multimodal cameras can be confirmed by the distinct hardware appearance and detective teardown.

III. Attack Model and Goal

Our attack goal is to evade general FASes via spoofing. A successful attack attempt will lead to wrong decisions of

the FAS. Its incorrectness may be caused by false-negative or false-positive decisions of the face detection, anti-spoofing, and/or face recognition modules (see Figure 1). 1. We mainly focus on misleading the FAS by largely increasing the false-negative rate of its *anti-spoofing* module, *i.e.*, let it believe the spoofing content is a genuine face, but keep the remaining functional modules unaffected. Meanwhile, 2. we would like to regenerate the spoofing content to accommodate the facial information of any people in a flexible and low-cost way. Achieving the two points allows the adversary to impersonate any people to the FAS. When the people being impersonated is an authorized user, the adversary can fool the FAS and illegally get the authorities of unlocking screen, issuing payment, opening door, *etc.*

We assume the FAS uses static anti-spoofing methods and makes use of a multimodal camera to enhance its anti-spoofing accuracy. We do not make assumptions on which type of multimodal camera it uses. Most commercial FAS products are within these assumptions.

We assume the adversary is not a network attacker. He/she has to have close proximity but not necessarily physical access to the FAS camera in order to present the spoofing content. For public FASes, such as door access controllers and in-store payment terminals, the adversary has sufficient opportunities for doing so. For personal FASes, such as screen lock, the adversary needs other approaches, *e.g.*, social engineering [42], to get physical proximity.

We assume the adversary has gained sufficient facial information of the authorized user of the FAS prior to the attack. In the context of multimodal camera, facial information has multiple modalities as well. Specifically, we assume the adversary has gained sufficient RGB facial information. Existing studies showed the feasibility of retrieving RGB photos from the public Internet [2] and methods to synthesize them to desired styles [22]. We assume the adversary has gained IR facial information as well. IR facial images are not publicly available, but they might be leaked from compromised IR photo sources, such as surveillance cameras and IR face database of entrance control systems. Interestingly, our result reveals that about half of the tested devices can be fooled without genuine IR photos (Section V-B4), and some can be fooled with fake IR photos forged from RGB photos (Section V-C2 and Appendix A). We do not make assumption on the depth modality information, *i.e.*, 3D face model.

IV. *Hua-pi* Attack

A. Motivation

People's trust in biometric authentication methods is based on the belief that biometric features are unique and hard to replicate. The hardness of replicating certain biometric features is because of either the difficulties of collecting them, *e.g.*, iris texture [43], and/or the complexity of reproducing them. Facial features, however, are exposed to the public, and hence FASes have to mainly rely on the difficulties of reproducing correct facial features.

As a result, multimodal cameras are introduced to strengthen FASes since a live person's face illustrates distinct and unique features in observations of different modalities. To fool such FASes, the spoofing method has to simultaneously present correct facial features in multiple modalities, which is believed hard or cost-ineffective in the past. For example, high-resolution and high-fidelity displays can be tuned to show fake RGB scenes to defeat RGB anti-spoofing methods, but cannot present IR scenes. A carefully-crafted face model that costs thousands of dollars has precise 3D face topography and RGB features [21], [44], but cannot preserve IR features because its materials look different from human skin under IR illumination [37].

In this paper, we challenge the above anti-spoofing basis of multimodal cameras by noticing that the involved modalities are physically independent. For example, RGB and IR observations are of different light wavelengths. Such independence implies the feasibility of decoupling and coupling them in the physical domain. If this is possible, the idea of spoofing is like divide and conquer: as it is not hard to forge a scene to spoof a single modality, the scenes of different modalities can be forged separately and coupled again to result in the desired observations in the multimodal camera.

We next introduce *Hua-pi* display - a multimodal display device designed to realize the above idea. Section IV-B describes its key component to enable modality re/decoupling. Section IV-C, IV-D, and IV-E provide a one-by-one description of its functional modules by showing how it can be adapted to launch attacks on different types of multimodal cameras.

B. Optical Combiner and Scene Synthesis

Optical combiner is a basic optical device. The top view of its optical paths is shown in Figure 3. When light signal incidents on the surface of the optical combiner, a part of the signal is transmitted through the combiner while the remaining is reflected off. When an optical combiner is properly positioned with physical scenes, it is like a "low-quality" mirror that not only reflects the scene in front but also exposes the scene behind. They together have the following properties:

- **P1**: two separate scenes can be synthesized into one scene through an optical combiner.
- **P2**: multiple separate scenes can be synthesized into one scene through multiple optical combiners.
- **P3**: **P1** and **P2** are effective for scenes of different light wavelength regions.

P1 is a rephrase of optical combiners' properties. As shown in Figure 3, Scene 1 and Scene 2 are merged into one scene when viewed through Optical Combiner 1. **P2** indicates that optical combiners can be used to synthesize multiple scenes. **P2** is deduced from **P1** by noticing that any scene can be replaced with a synthesized scene. Figure 3 shows a setup example of synthesizing three scenes with two optical combiners. Further, with appropriate coating materials, optical combiners are effective for a wide range of light wavelength regions, which leads to **P3**.

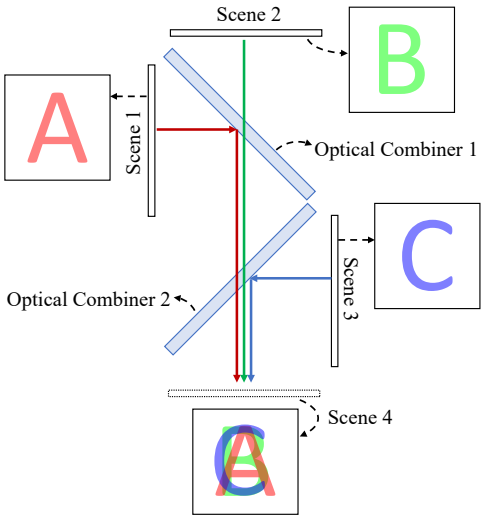


Fig. 3. **Using Optical Combiners to Synthesize Optical Scenes.** Optical combiner is like a semi-transparent mirror. When properly positioned, it synthesizes the reflection and transmission scenes into one scene. Multiple optical combiners can be jointly used to synthesize multiple scenes.

Hua-pi display is based on the above properties. It forges and generates the scenes of multiple modalities separately. Then, it uses optical combiners to synthesize the separated scenes into one consistent scene and present it in front of the FAS camera. We use 14-inch optical combiners. Since the modalities are physically independent of each other, the camera's observations of the synthesized scene are automatically separated by modalities and only capture the corresponding forged scenes.

Hua-pi display is a modular device that generates the scene of a certain modality with the corresponding *scene generator* module. Every scene generator is characterized by the *Scene Display* scheme and the *Scene Content*. This allows it to be flexibly adapted to various multimodal cameras and attack targets.

C. Adaptation to RGB-IR Cameras

RGB-IR cameras use RGB and IR sensors. They are probably the most widely-used multimodal cameras due to their cost-effectiveness. Many door access control systems are using this type of camera. *Hua-pi* display is equipped with RGB and IR scene generators to spoof RGB-IR cameras. The setup is shown in Figure 6 (a).

1) RGB Scene Generator: The goal of this module is to generate appropriate RGB scenes to induce the desired RGB observations in the FAS camera to fool the anti-spoofing algorithms on the RGB aspect.

RGB Scene Display - LCD. Existing RGB anti-spoofing methods seek spoofing evidence in RGB observations through exploiting the texture and chromatic clues caused by the display medium and imaging process. We use a 13.3-inch 4k LCD panel to show the RGB scene. Its pixel density is 332 pixels per inch (ppi). After fine-tuning the screen's color and brightness, the presented scenes can bypass all FASes we tested. Not only is this because the high-pixel density panel eliminates most texture clues[2], but commercial FASes also rely more on chromatic clues[3].

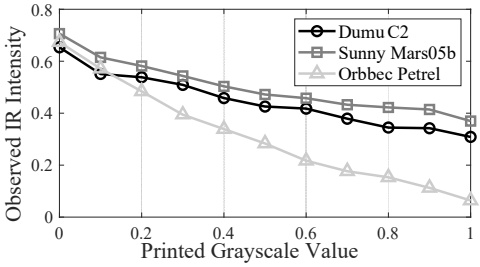


Fig. 4. **Observed IR Intensity vs. Laser-printed Grayscale**. Their values are linearly related, meaning that the laser-printed content can be correctly perceived by IR sensors. Refer to Table I for the camera model.

RGB Scene Content. To launch *Hua-pi* attack, the scene generators have to load appropriate content. We call the people to be impersonated the target user, who is an authorized user of the FAS. The target user's RGB headshot is collected and displayed on the LCD panel. The displayed content is the RGB scene to be shown to the FAS camera.

2) IR Scene Generator: Similar to the RGB scene generator, this module is responsible for the IR aspect. Existing work on anti-spoofing algorithms is usually not dedicated to IR modality [46], but since IR observations also measure light intensity information, we believe that IR anti-spoofing methods are similar to RGB's methods. However, IR observations contain unique facial features due to IR reflectance properties. This is an anti-spoofing feature related to the IR scene content.

IR Scene Display - Laser Print. Similar to the RGB case, we try to reproduce IR scenes with as much fidelity as possible. However, normal displays are not designed to work in the IR band, and IR display devices, such as Digital Light Processing (DLP) projector [47] and IR LCD [48] are either expensive or not mature. To avoid these restrictions, we choose to print the IR scene. We tested different printing technologies and found that laser-printed content can be observed by IR sensors since laser toner also absorbs IR light, but ink printing cannot be used since normal ink is translucent to IR light. Therefore, we use HP LaserJet Pro MFP M427 to print content on XEROX 80g white papers.

When the printed paper is covered by the camera's IR illuminator , the IR reflections show an IR scene. Intuitively, the darker the printed content, the less IR reflection the camera will receive. We print blocks of different grayscales on the paper and observe it with cameras to reveal this relationship. Figure 4 shows that the IR light intensities perceived by the cameras, *i.e.*, via IR observations, are almost linearly related to the printed grayscale values, implying that an IR scene can be precisely reproduced by first mapping the desired

[2]High-resolution cameras with appropriate lenses can still detect, *e.g.*, Moiré patterns [45], but they are not widely used as FAS cameras due to cost and processing overhead.

[3]In many working scenarios of FAS devices, texture clues are hard to observe. For example, a smart lock needs to recognize a face 1.5 m away. The dimension of the face area may be only a few hundred pixels, which provide very limited spatial information.

IR intensities, *i.e.*, the IR scene, to corresponding grayscale values, *i.e.*, a grayscale image, and then printing on a paper.

IR Scene Content. We prepare three types of input content for the scene generator, they differ in acquisition methods and effectiveness.

- IR headshot of the target user. It is similar to RGB mugshot, but there are fewer resources to obtain it at scale. For valued target users, the adversary can choose ad-hoc ways, *e.g.*, taking candid shots, compromising IR surveillance cameras.
- IR headshot of any people. The adversary can use the IR photo of him/herself if necessary. This content can be used to bypass most FASes that also use RGB modality. It sounds ridiculous but is reasonable in some sense. Such IR photos are effective for spoofing because these FASes are designed to use the IR observation for anti-spoofing ONLY (the adversary's IR photo contains all biometric features of a live person except that its identity does not match with the target user), but does not check the content consistency between modalities. We will return to this serious defective design issue later in Section V-B4.
- Forged IR headshot of the target user. Such photos are derived from the user's RGB photos and can be used to spoof FASes that rely on IR modality for both anti-spoofing and face recognition. Our forgery methods make use of the internal similarity between RGB and IR observations. They are described in Appendix A.

3) Launching Hua-pi Attack: Figure 6 (a) shows the setup of *Hua-pi* display when using it to spoof the FASes that are based RGB-IR cameras. The RGB and IR scene generators are installed on the two sides of the optical combiner. Their planes are perpendicular to the horizontal plane. The planes of the two scene generators are perpendicular to each other and both have a 45° bearing angle to the combiner's plane. The FAS's RGB-IR camera is supposed to be at one side of the combiner, where the RGB and IR scenes are synthesized into the observable scene of the camera. The RGB and IR sensors of the camera will separate RGB and IR scenes out with their passband filters, and their observations capture the scenes produced by the corresponding scene generators. All observations contain designated facial content, which will likely fool the FAS to make an "accept" decision, *i.e.*, the success of the attack. *Hua-pi* attacks can be launched similarly with other types of cameras and corresponding setups (see the following two sections IV-D and IV-E).

D. Adaptation to Structured Light Cameras

Depth camera is a multimodal camera that provides depth modality in addition to RGB and/or IR modalities. Depth observation measures the 3D topography of the observed scene, which renders unique information for anti-spoofing. Many recent FAS products use depth cameras in critical applications. Structured light and Time-of-flight (ToF) are the two mainstream depth sensing technologies with very different operation principles. *Hua-pi* display incorporates two dedicated modules to generate depth scenes for them. We discuss structured light camera in the following and ToF later in Section IV-E.

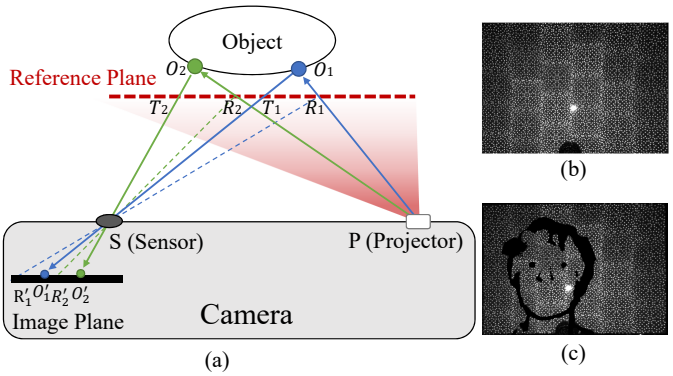


Fig. 5. **Principle of Structured Light Depth Sensing.** (a) The camera projects a static light pattern (b) into the environment as the reference plane. An object will distort the pattern and introduce disparity, according to which the depth information can be calculated. (c) is a forged pattern according to the sensing principle. When putting it in front of the camera, it leads to a 3D face in the depth observation.

1) Principle of Structured Light Depth Sensing: Structured light camera relies on two hardware components for depth sensing: the projector that actively casts structured light patterns into the scene and the image sensor that captures the scene overlaid with the light patterns. The IR sensor of the FAS camera is usually reused as the image sensor. Time-invariant and spatially-random textures are chosen as the light pattern (see Figure 5 (b) and Figure 10).

Figure 5 (a) explains its basic principle. The camera initially captures the light patterns on a plane of known distance, which is recorded as the reference plane. The occurrence of an object in the camera's scene will cause disparities in the observed pattern compared to the reference. The topography of the object can then be computed based on triangulation. Specifically, any points in the scene that reflects the projected light will cause a displacement of the light pattern in the image plane. For instance, the pattern point R_i on the reference plane is projected to the object surface O_i, which leads to the displacement of pattern point from $R_i^{'}$ to $O_i^{'}$ on the image plane[4]. Since the depth of the reference plane is known, $\overline{O_i^{'}R_i^{'}}$ can be used to calculate $\overline{T_iR_i}$ according to the triangle similarity. Further, since the length of $\overline{SP}$ is known, $\overline{T_iR_i}$ can be used to obtain the distance to O_i, *i.e.*, the depth of object O_i.

2) Depth Scene Generator (Structured Light): This module is to present appropriate depth scenes to fool FASes. Its scene display should be able to generate arbitrary 3D shapes in the camera's observation in order to achieve the goal. We explored two display schemes during our research. *TYPE-A* utilizes 3D print technology to physically realize the 3D scene. Its cost is moderate, but it is time-consuming (about 24 hours) to produce a new scene. *TYPE-B* is based on the reverse engineering of the sensing process. It forges the light pattern to generate

[4]The structured light pattern is chosen to ensure that any two regions are sufficiently distinct so that $R_i^{'}$ and $O_i^{'}$ can be paired in the image plane by matching algorithms.

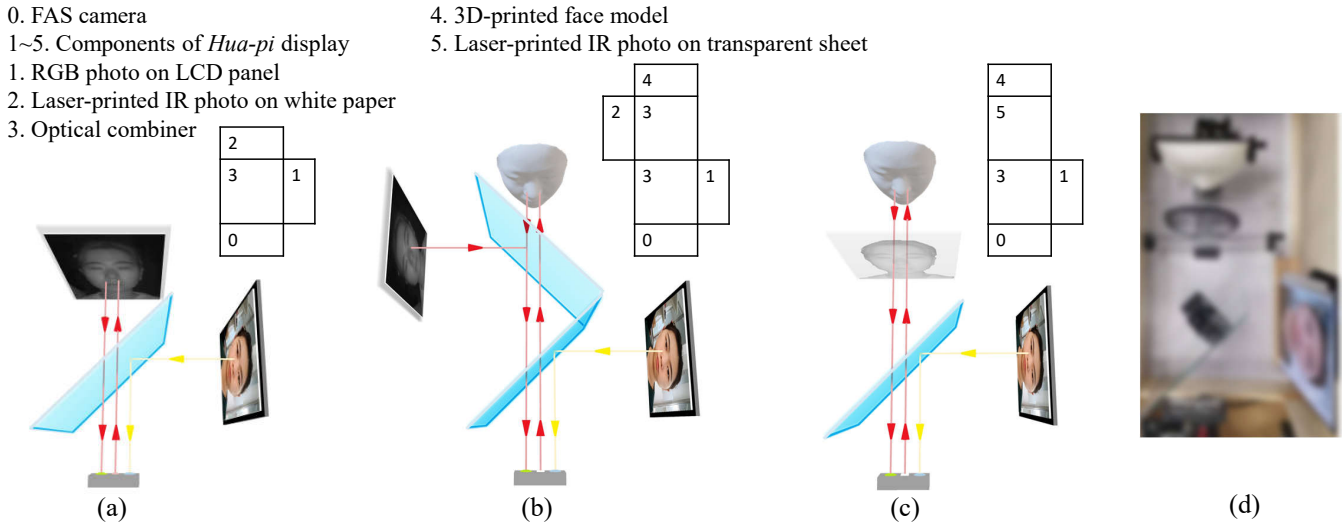


Fig. 6. **Adapting *Hua-pi* Display to Spoof Various Face Authentication Systems**. The windows beside the light path diagrams show the components. (a) is for RGB-IR cameras. (b) is for structured light cameras. (c) is for ToF cameras. (d) is the photo of (c).

the desired depth observation. Its reloading cost is low, but the hardware is more complex. We verified the effectiveness of both types, but due to the same reason we mentioned in the IR scene content in Section IV-C2, it turned out that all tested FAS devices do not examine the consistency of the depth observation. Therefore, there is no need to reproduce the 3D scene to attack current FASes, so a convenient way is to use *TYPE-A* display with the same 3D model to setup the attack. Considering the merit of *TYPE-B*, *e.g.*, in future attacks against improved anti-spoofing methods, we present both types in the following.

Depth Scene Display (TYPE-A) - Physical Model. The depth scene is physically generated by a 3D model, which can be arbitrarily fabricated with 3D printing. When presented in the projection of the camera, the model will distort the pattern. Then, the sensor's depth observation reflects the shape of the printed 3D model.

The setup example is shown in Figure 6 (b). The RGB, IR scene generators, and the 3D head model generate three separate scenes. Two optical combiners are used to synthesize them together in front of the FAS camera. The light patterns from the camera are projected on the model after transmitting through the two combiners, and then scattered back to the camera along the reverse path. The IR scene from the IR scene generator encounters and merges with the depth scene at the IR sensor. We note that the two scenes are separable since they multiplex the IR sensor in the time domain. The projection is not always-on, and the IR observation is captured only when the projection is off. The depth observation is also valid, because the pattern matching algorithms will ignore content other than the structured patterns. As a result, the FAS camera will generate three observations of the corresponding scenes for authentication.

Depth Scene Display (TYPE-B) - Forged Pattern. This method presents the depth scene of an object by showing the corresponding light patterns. It blocks the camera's projector to force it to see a forged IR scene, whose content is a counterfeit light pattern identical to that when the object really presents. To obtain this pattern, we reverse engineer the projection process. We first put a white and blank plane in front of the camera at a known distance and use another IR camera to capture the pattern on the blank plane as the "actualized" reference plane. When the object presents, as depicted in Figure 5 (a), a point R_i on the reference plane is projected to O_i and then reflected to point T_i. From the camera's perspective, the light coming from point O_i is like coming from point T_i. So, we can generate the light pattern of R_i at point T_i to simulate the reflection from point O_i. Therefore, the content of point T_i on the counterfeit pattern is obtained by copying from the texture of point R_i on the reference plane. This process is repeated until the object's surface is traversed. Figure 5 (c) is an example of the counterfeit pattern generated from Figure 5 (b) and a 3D face model. The pattern is shown to the camera through another IR scene generator, which is installed by replacing the 3D model in Figure 6 (b). More details about the *TYPE-B* design are presented in Appendix B.

Depth Scene Content. We prepare three types of 3D model content for the depth scene generator.

- Forged 3D face of the target user. Existing work shows 3D face models can be inferred through RGB photos [22]. We apply Deep3DFaceRecon [49] to the target user's RGB photo to generate the 3D face model.
- Face model of any people. For the same reason as the IR scene content, the mean female face model [50] is chosen to fool anti-spoofing algorithms.
- Ball model of head size. This is related to another serious defective design issue, which will be detailed in Section V-C3. Most of FASes examine the depth scene in an extremely coarse-grained way. We use this model to reveal the issue.

E. Adaptation to ToF Cameras

ToF camera is another type of depth camera. Our experience is that its effective range is longer, depth resolution, precision, and cost are higher than that of structured light cameras. Some high-end smart devices use ToF cameras. Spoofing FASes that use ToF cameras is challenging since their IR and depth observations are tightly coupled.

1) Principle of ToF Depth Sensing: Time of Flight (ToF) is a distance measurement method. ToF cameras have two key components. ToF projector is a laser diode like the ones used in optical fiber communication. Unlike structured light projectors that produce static spatial patterns, it generates dense and short-period light pulses and scatters them evenly into the environment. The reflected pulses from surrounding objects are received by a special image sensor, the ToF sensor. In addition to light intensity, each pixel (a photodiode) of the ToF sensor can measure the elapsed time since the received pulse was emitted, *i.e.*, the ToF of the pulse, according to which, the depth of the objects reflecting the pulse can be calculated.

ToF sensors have several variants. Figure 7 (a) shows the operation principle of the common type used in FASes. A pulse that lasts for T_w is projected out, and then its reflected copy is received by a pixel of the ToF sensor. The time of flight of the pulse is $T_2 - T_1$. To measure this value, the pixel uses three counters to precisely record the received energy in three periods that all last for T_w. The end of the first period is the start of the pulse. The second period is aligned with the pulse. The third period follows the end of the pulse. The received light energy in the corresponding counting periods is V_1, V_2, and V_3. The projected pulses are reflected by environmental objects. Due to propagation delay, the bright period of the reflected pulse will intersect with the second and third periods. The farther the object is, the longer the delay is and the larger the intersection area with the third period is. The delay is just the time of flight. As shown in Figure 7 (a)(b), by noticing that the intersection areas of the two periods are proportional to the intensity of the received energy V_3 and V_2, the time of flight of the pulse and the depth of the object reflecting it can be estimated by the proportional relationship.

While performing ToF sensing, each pixel of the ToF sensor also measures the intensity of the received light, *i.e.*, imaging the scene. But unlike normal image sensors, the output of ToF sensor usually omits the light emitted by the environment. As shown in Figure 7 (c), this is because it only uses the intensity of the reflected pulses for imaging.

2) Depth Scene Generator (ToF): Since a ToF sensor only images the reflection of its pulses, the IR scene generated by the previous IR generator is not visible in ToF sensor's IR observation. According to Figure 7 (b)(c), ToF sensor's intensity and depth are jointly determined. We denote the counter reading of pixel i as V_1^i, V_2^i, and V_3^i. When the ambient lighting condition is stable, V_1^i can be ignored. The depth is determined by V_3^i/V_2^i and the intensity is determined by $V_3^i + V_2^i$. So, even if the IR scene can be precisely turned on to affect the intensity, *i.e.*, $V_3^i + V_2^i + \Delta^i$, it is still challenging to spatially let every pulse maintain a specific V_3^i/V_2^i. As such,

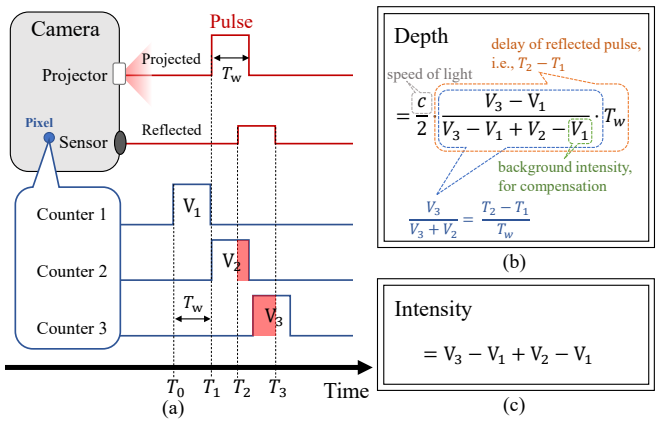


Fig. 7. **Principle of Time of Flight (ToF) Depth Sensing.** (a) ToF sensor records the intensity of pulse reflection in different time instances as V_1, V_2, and V_3. The depth is calculated according to the relation in (b). The intensity of the pulse reflection is calculated by (c).

we propose the following display scheme to jointly generate the two scenes.

IR & Depth Joint Scene Display - Optical Filter & Physical Model. Similar to the scheme for structured light cameras, we use a physical 3D model to produce the depth scene. The remaining problem is how to generate the correct IR scene. Note that if we can apply a per-pixel attenuation factor α_i to pixel i, *i.e.*, change V_2^i and V_3^i to $(1-\alpha^i)V_2^i$ and $(1-\alpha^i)V_3^i$, then the measured intencity is modified to $(1-\alpha^i)(V_3^i+V_2^i)$, and the measured depth remains unchanged. This is because $((1-\alpha^i)V_3^i)/((1-\alpha^i)V_2^i) = V_3^i/V_2^i$. Our scheme is motivated by photographic filters [51], which are widely used to manipulate scene intencity. The above per-pixel attenuation can be achieved through inserting an attenuation filter between the depth scene and the camera. The filter is spacially heterogeneous and aligned with the ToF sensor to induce per-pixel attenuation by absorbing the energy of the light that passes through towards pixel i by a factor of α^i.

The filter has to be changeable, since the generator has to present arbitrary scene content. In the light of the IR scene generator (Section IV-C2), we find a viable way to fabricate such filters cost-effectively. We use the laser printer to paste toner on a transparency film to absorb IR light. The spatial location and amount of toner can be precisely controlled by the printer. Figure 8 shows the optical properties of the printed content. We print grayscale blocks on the film, and place the film in front of a solid white background plane. The ToF camera is placed 350 mm in front of the background to observe the scene. The grayscale values are within $[0, 1]$. Values larger than 1 are generated by printing two times. When the printed grayscale is within some range, *e.g.*, 0.1 to 0.4 for Sunny and 0.4 to 0.6 for device #15, the depth is barely affected while the attenuation factor is linearly changed, indicating feasible regions for generating IR and depth scenes simultaneously. The intensity of the original IR scene content is scaled into these regions and then printed to fabricate the attenuation filter. Figure 6 (c) shows the setup of this module. The filter is inserted between the 3D model and the combiner. Figure 6 (d)

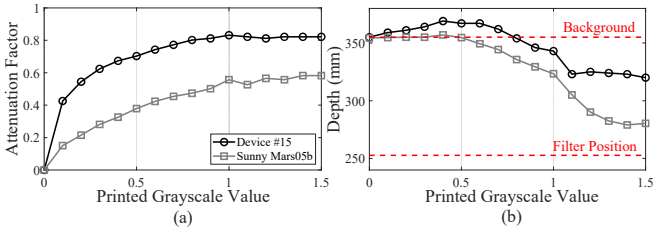


Fig. 8. **Optical Properties of Printed Attenuation Filter**. (a) Darker printed content lead to greater attenuation. (b) When the printed content is light and within some range, it will not affect the depth measurement. Refer to Table I for the camera model.

is the photo of the prototype system.

Depth & IR Joint Scene Content is identical to the previous IR (Section IV-C2) and depth content (Section IV-D2).

V. Evaluation

We extensively test *Hua-pi* attack against leading FAS products in physical environments and real-world settings. We also explore factors affecting attack efficiency.

A. Methodology

1) Test Devices: Table I summarizes the devices used for the experiments. Table IX in the Appendix provides more information about the devices and test settings. They are classified according to their camera technology. RGB-IR cameras use RGB and IR modalities only, while the structured light and ToF categories use the depth modality. Each category includes several devices, and they are further divided into subcategories according to their roles in the supply chain of FAS products.

Camera devices are bare metal cameras. They output multimodal observations, and we direct their outputs to commercial FAS algorithms for testing[5]. Module devices are minimum and functional FASes. They are built upon camera devices and process FAS algorithms locally. They output confidence levels or decisions of spoofing detection and face recognition. Some of them also provide interfaces to log their camera observations. Module devices are rarely used by end users and are usually integrated into product devices. Product devices, including consumer devices such as smartphones and enterprise devices such as door access controllers, make use of FAS modules, along with other functional components, to render concrete services to users. Product devices block most output from FAS modules and only feedback binary decisions - accept or reject.

We collected FAS-related devices available on the market at our best and validated the risk of *Hua-pi* attack in all of them. The 16 devices in Table I are selected for extensive testing. They are either flagship products in the field or come from major vendors of the supply chain. In this sense, their results reflect the performance of current production-level FAS technologies.

For the security consideration, we anonymized some devices throughout the paper. Detailed information may be disclosed through other channels once the vendors have effective countermeasures.

2) Facial Information Collection: The scene content of *Hua-pi* display relies on RGB and IR photos of the target user. We collect them from 20 participants of different gender (6 female and 14 male), ethnicity, and age groups (from 18 to 85). We did not obtain their photos from online media, but directly took them with our devices in Table II in various daily environments with relaxed poses, diverse backgrounds, and lighting conditions. Photos of different modalities are not strictly aligned and are not taken at the same time. All photos are portrait shots, which are further cropped and scaled to headshots to make sure the test device has a clear view of the facial content. Additionally, all participants are asked to follow the official instructions to register with the test devices. For face identification devices, *e.g.*, smartphones, that only allow one or two registered users, registering and testing take turns for different participants. Due to the sensitivity of facial information and the risks we reveal, all information collection and experimental protocols have been approved by the Ethical Review Board of the authors' institute. Without the participant's approval, the collected information is kept permanently offline and will not be used for experiments other than this study.

3) Adaptation to FAS Cameras: Hua-pi display's modules are selected and equipped according to the modalities of the FAS to be tested. While the camera technology is sometimes mentioned in the device manual and promotional content, we confirm this information manually through tearing down, functional testing, and optical probing. We note that some devices have the hardware of certain modalities but their FASes do not make use of them. We will return to this issue later in Section V-B3. The confirmed modalities of the test devices are shown in Table I with the abbreviation `RGB`, `IR`, and `D` (depth). The prototype of *Hua-pi* display is installed and fixed into a container of 57cm×33cm×37cm. Figure 6 (d) depicts the situation when launching the attack on ToF cameras. The targeting FAS device is seized by a robotic arm with 0.05 mm positioning accuracy and heads towards the synthesized scene from the *Hua-pi* display.

4) Determine Default Parameters: Due to various camera parameters, the position of a clear view of the synthesized scene is unknown and different for different devices. Parameters of the *Hua-pi* display, such RGB brightness, saturation, *etc.*, also need tuning to achieve good presentation quality. These parameters are mainly device-specific and not user-specific, and can be determined prior to the attack. We measure them with our own headshots and record those leading to successful attack attempts as the feasible parameters. As we will show later in Section V-C, the ranges of feasible parameters are not narrow for most devices and we use the mean values as the default parameters. For example, to determine the

[5]Algorithm vendors usually provide reference designs to facilitate the adoption. These designs are officially used with a specific multimodal camera model and have no further requirement on the remaining system. Device #8 and #11 belong to this case. Device #1, #2, and #14 are assembled strictly according to the reference designs. We did some basic tests on them to ensure that they function properly.

Test Device							Pass Rate of Participant (%)																		
Tech.	Type	Vendor	Model	Modality	Algorithm	Index	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	Avg.
RGB+IR	Camera	Dumu	C2	RGB+IR	SDK1	1	68	86	77	78	70	81	98	91	94	99	98	86	89	99	90	99	77	100	86	89	88
		Dumu	C2	RGB+IR	SDK2	2	98	92	73	70	90	11	43	59	35	80	78	0.0	90	94	90	98	57	72	96	83	71
	Module	Dumu	C2	RGB+IR	built-in	3	74	50	96	100	86	45	47	90	66	86	63	0.0	74	96	91	100	98	91	94	100	77
		□□□	□□□	RGB+IR	built-in	4	38	51	100	76	34	0.0	88	64	29	92	70	0.0	38	89	86	100	94	100	79	100	67
		NXP	SLN-VIZNAS-IOT	RGB+IR	built-in	5	100	100	100	100	100	100	100	100	100	100	100	100	100	×	100	100	100	100	100	100	100
		Intel	RealSense F455	RGB+IR	built-in	6	82	96	90	74	65	75	54	67	69	77	83	72	89	94	100	59	64	64	33	70	74
	Product	□□□	(door access)	RGB+IR	built-in	7	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Structured	Camera	Orbbec	Petrel	RGB +D^	SDK1	8	100	100	100	85	62	91	100	100	100	100	82	100	83	50	100	100	100	100	100	76	92
Light	Module	□□□	□□□	IR+D	built-in	9	~~0.0~~	100	100	100	100	100	100	100	100	~~0.0~~	100	20	100	×	100	100	100	100	~~0.0~~	100	80
		NXP	SLN-VIZN3D-IOT	IR+D	built-in	10	75	100	90	100	94	100	100	80	90	90	100	98	86	80	100	89	100	100	0.0	100	89
	Product	□□□	□□□	RGB+IR+D	SDK3	11	67	67	96	96	67	100	70	78	100	86	56	62	26	×	61	66	74	34	60	65	70
		□□□	(smartphone)	IR+D	built-in	12	82	43	100	80	20	0.0	94	61	100	40	87	75	82	0.0	99	61	95	96	98	83	70
		□□□	(smartphone)	IR+D	built-in	13	86	100	29	100	86	100	100	100	100	82	100	100	100	100	100	100	100	100	100	100	94
ToF	Camera	Sunny#	Mars05b	RGB+IR+D	SDK1	14	70	82	84	74	65	71	70	66	84	64	77	0.0	83	75	83	83	92	82	90	47	72
				RGB+IR			82	81	86	81	80	71	74	72	88	67	83	0.0	86	79	86	85	96	83	93	54	76
				RGB +D			94	98	96	88	89	92	98	84	96	86	89	0.0	91	83	94	91	98	98	99	54	86
				IR			84	84	90	90	86	86	78	86	91	90	91	88	90	90	92	91	95	88	93	90	89
				RGB			97	98	97	92	97	92	99	90	97	90	93	0.0	96	85	96	96	100	100	100	64	89
	Module	□□□##	□□□	IR+D	built-in	15	58	~~0.0~~	100	100	100	100	80	100	100	~~0.0~~	×	~~0.0~~	0.0	×	36	59	0.0	60	22	~~0.0~~	51
	Product	□□□	(smartlock)*	IR+D	built-in	16	○	○	●	●	●	○	●	●	○	●	○	○	●	●	●	●	●	●	○	●	/
						Avg.	73	78	89	89	76	72	83	84	84	73	85	54	76	80	89	88	83	87	71	81	

xx highlights the pass rates lower than 50%.
~~0.0~~ denotes the cases that can bypass anti-spoofing but cannot be recognized.
× denotes registration failure.
A major supplier of smartphone optical modules. Its ToF modules are based on Sony ToF sensor.
Its ToF sensor is from OPNOUS, a Chinese ToF sensor company.
* Due to the inconvenience of moving it, we only record pass (●) or no pass (○).
^ Its IR modality is not used for face recognition since the IR sensor captures the structured patterns, which affect anti-spoofing and face recognition algorithms.

TABLE I
OVERALL RESULTS (PART 1). TABLE IX IN THE APPENDIX PROVIDES MORE INFORMATION ABOUT THE TESTS.

Camera	Modality	IR Wavelength	Resolution
iPhone 13 or iPhone 13 mini	RGB	\	2016×1512
Dumu C2	IR	850 nm	1280×720
Orbbec Petrel	IR	940 nm	640×400
Sunny Mars05b	IR	940 nm ToF	640×480

TABLE II
DEVICES USED TO CAPTURE SCENE CONTENT.

default observing position, the robotic arm is programmed to traverse a 10cm×10cm×10cm cube with a 0.5 cm step size in front of the container. The average of the positions bringing in successful spoofing attempts is chosen as the default observing position of that device.

5) Metric: Since *Hua-pi* attacks are launched in a physical environment, the observations of the same scene may be slightly different due to environmental noise, which randomly disturbs FAS decisions. Thus, the result of a single observation cannot reliably reflect the spoofing performance. Meanwhile, many devices allow at least 3 consecutive authentication failures before reporting errors or issuing alerts. Due to the above considerations, we define a "test" of *Hua-pi* attack as: at a fixed camera position, the test device is triggered to perform 3 consecutive face authentication attempts. The "pass" of the test is: at least one of the three attempts results in the accept decision of the FAS, *i.e.*, the spoofing artifact is not detected and the target user is correctly recognized by the FAS. Some devices constantly perform face authentication without notification intervals, *e.g.*, Device #3. Its pass is defined as at least one FAS accept output during 3 seconds.

As we mentioned previously in Section V-A4, the relative position between the *Hua-pi* display and the target device affects the FAS result. In practice, the attacker needs to aim the *Hua-pi* display to the default observing position. The required aiming precision is an important factor of this attack. As such, we define the "pass rate" as the number of passed tests over the total number of performed tests in an area. The area is a 2cm×2cm×2cm cube with the default observing position as the cube center. The device is moved by the robot arm to traverse the cube with a 0.5 cm step size. In total, the device is moved to 125 different positions to test the attack. We use pass rate in our experiments as the metric to statistically quantify the probability or difficulty of forging a certain user against a certain FAS device. Device #16 is a smartlock, which is not convenient to move. Its default observing position is sought by manually adjusting the lock's position. We then fix it in that position to perform tests and only record pass or no pass based on the first three attempts triggered by touching its doorknob.

FAS's authentication outcome is determined by both face recognition and anti-spoofing. Since the scene content for recognition is from the participant, recognition algorithms tend to accept the test. In this sense, the pass rate mainly reflects the performance of anti-spoofing methods.

B. Overall Results

The overall results are shown in Table I. We maintain a consistent test protocol across devices and participants. We have the following findings, and we highlight important ones to provide insights for improving future FAS designs.

1) Participant Heterogeneity: The mean pass rates of most participants are around 80%, implying the attack tends to

be effective for most people[6]. There are several outliers to mention first:

- A,J,S-9, and B,J,L,T-15 in Table I are from two devices and have a 0% pass rate. These cases all passed anti-spoofing but were rejected by face recognition. This situation is rare in other devices and probably because the two devices' face recognition modules are relatively weak.
- We could not register participant #N on several devices, so we could not launch attacks on these devices to impersonate him. We noticed that this participant had a thick beard, and that device #11's log issued mouth-blocked messages. As such, the failure to register is likely due to the fact that the algorithms of these devices are not strong enough to cover different facial features.
- The attack cannot impersonate participant #L in many devices using RGB modality. The reasons are two folds. First, his skin tone is dark. Due to the reflection loss of the optical combiner, some cameras cannot seize certain details. However, his cousin (participant #M), who has an identical skin tone, has a typical performance, so we think the inadequacy of these devices' FAS algorithms contributes another factor, *e.g.*, they may not be equally effective for different ethnic groups.

Except for the outliers, only #F and #S have a pass rate about 10% lower than the others. Further, when we add subtle noise to the non-attack authentication process by letting FAS cameras observe participants' faces through a transparent glass or a mirror, the only failures are also from participant #F. These experiments suggest that the effectiveness of *Hua-pi* attack has some dependence on the target user. For a specific device, some people are naturally at lower risks than others. This is rooted in FAS algorithms. Both face recognition and anti-spoofing rely on features learned from training examples. If a participant's features are close to the margin of the classifier, he/she is more likely to be rejected when encountering noise, *e.g.*, attacks, and vice versa.

2) Device Heterogeneity: The pass rates of the 15 devices are mostly around 80%, indicating close FAS performance across vendors. There is no one that significantly outperforms the other. This suggests that the current understanding of spoofing methods has been restricted into a region, and has not been aware of *Hua-pi* attacks. The only exception is device #15. It has a large number of registration failure cases, and has very biased pass rate values. Its high pass rates are concentrated in East Asian participants, so we speculate that its result may not be due to stronger anti-spoofing algorithms, but implementation insufficiency.

The lowest average pass rate is from device #4. The fourth lowest pass rate is attributed to device #2, where we apply the same vendor's algorithms to a different camera. This suggests this vendor's RGB anti-spoofing is stronger than the others against *Hua-pi* display. The second lowest pass rate is from device #12. For this leading industry product, being more secure than others is reasonable, but its current advantage is not fundamental.

3) Effectiveness of Modalities: Device #14 is a ToF camera of three modalities. We direct its observations to SDK1 to emulate a fully functional FAS device. The sdk has three anti-spoofing interfaces, RGB-only, IR-only, and RGB+Depth. Each of them returns a confidence level and a binary decision. We use logical "AND" to combine the three decisions of the three interfaces to understand the impact of different modalities. The unindexed rows under device #14 show all feasible combinations. The average pass rates of the rows imply that using more modalities brings more accurate anti-spoofing decisions. This is reasonable since different modalities contain complementary information. Numerically, according to the incremental changes of the pass rates, at least for SDK1's algorithms, the depth modality contains the least information. We will return to this point shortly in Section V-C3.

Some devices' cameras have more modalities than the listed ones. For example, smartphones #12, #13, and smartlock #16 have RGB front cameras, but they only use IR and depth for face authentication. This is probably a design choice because their targeting situations include low-light conditions, where RGB observations are not clear. But for some devices, it is more like design flaws. For example, when testing with device #6, Intel RealSense F455, we thought it has a structured camera inside, like other RealSense cameras. It indeed contains a projector and the pattern is recorded in Figure 10. The module has three anti-spoofing levels, and the projector is only enabled at the highest level, but it turns out it does not matter if the projector is enabled or not. When the projector is physically blocked, registered users can still pass face authentication. As such, *Hua-pi* display only use the RGB and IR display modules to test it. The above experience suggests the following insufficiency of some FASes:

➥ **Defective Design 1 (D1): Available modalities are not (properly) in use.**

4) Physical Consistency: Hua-pi display separately generates scenes. This not only allows for multimodal scene generation but also reveals other vulnerabilities hidden in FAS designs. As highlighted by the "IR ID" and "3D Model" columns in Table IX in the Appendix, all FAS devices can be fooled without the participant's 3D face model, and all FASes using RGB modality can be fooled without the participant's IR photo. The results imply that most multimodal FAS devices only use one modality, *e.g.*, RGB, for face recognition and do not verify the face identity in multimodal observations.

The first half of the above design choice is reasonable. RGB observations contain richer facial information than IR and depth, and depth is less suitable than IR due to the high computational cost and insufficient imaging quality [52], [53]. Therefore, FASes only need to choose the best available method for face recognition. The second half becomes problematic when considering *Hua-pi* display. Current FASes are

[6]A common performance metric in anti-spoofing techniques is False Acceptance Rate (FAR), which is the ratio of spoofing attempts that are incorrectly recognized as genuine attempts over the total number of spoofing attempts. A typical FAR in commercial products is $\leq$0.5% (see Table IX). Considering the relation of pass rate and FAR, we can roughly estimate that the FAR is increased by 2 orders of magnitude.

Parameter	Adjustment				
	-40%	-20%	0%	20%	40%
Brightness	0.0%	72.0%	96.0%	96.8%	0.0%
Contrast	38.4%	94.4%	96.0%	95.2%	98.4%
Observed Size	0.0%	61.6%	96.0%	98.4%	94.4%
Compression Level	5	10	lossless	\	\
	0.0%	95.2%	96.0%		
Resolution	126×95	252×189	2016×1512	\	\
	51.2%	97.6%	96.0%		

Test Device: □□□ Index 3

TABLE III
IMPACT OF RGB SCENE GENERATOR. PASS RATE IN %.

not aware of the possibility of scene decoupling and ignore the intrinsic fact that multimodal observations measure the same object at a certain moment and their content should be temporarily and spatially consistent, *i.e.*, physical consistency. Physical consistency is manifested in many aspects. For example, the facial components of multimodal observations should be aligned. The facial features should reflect the same identity, expression, age, makeup, *etc.* The lack of consistency check allows the adversary to prepare the scene content in a much easier way, and makes FAS devices more vulnerable to *Hua-pi* attacks. Therefore, we have:

➥ **Defective Design 2 (D2): The physical consistency of multimodal observations is not jointly verified.**

C. Impacting Factors

In this section, we repeat the tests of participant #C, and adjust the default test settings in Table I to reveal the impacting factors. We explore individual scene generators first, and then the alignment issue when composing them up. Different devices have different sensitivity to these factors, and the following test devices are chosen for their representative performance.

1) RGB Scene Generator: We alter the default display parameters for device #3 and record the results in Table III. The brightness and contrast of the participant's photo are modified with image processing tools [54]. The results show the FAS's decisions are affected by both factors and are more sensitive to brightness changes. The RGB anti-spoofing algorithm takes chromatic features into consideration, but both factors have a relatively wide feasible range.

The observed size quantifies the area of the face as seen from the camera. It is magnified and shrunk by zooming in and out of the displayed photo. The results show the face size is not relevant to the FAS decision unless the face area is too small to be detected. We also note that device #4's default face size is larger than other devices. This is probably due to its camera, since the same algorithm applied to device #2 does not have this issue.

The compression level quantifies the loss in image quality due to compression. The original RGB photo is compressed with different quality levels in [0,100] by the standard JPEG encoder. The results show that this FAS is not sensitive to compression loss until the photo becomes blocky. This is probably because the FAS algorithm is designed to be robust in harsh conditions or with low-end cameras.

Test Device		Index	Device IR Wavelength	IR Scene Content			
				850 nm	940 nm ToF	940 nm	Fake
□□□	□□□	4	850 nm	100.0%	0.0%	0.0%	0.0%
Intel	RealSense F455	6	850 nm	89.6%	50.4%	76.0%	0.0%
NXP	SLN-VIZN3D-IOT	10	940 nm	59.2%	44.0%	97.6%	90.4%

TABLE IV
IMPACT OF IR SCENE GENERATOR. PASS RATE IN %.

The resolution is the actual pixel dimension of the displayed photo. High-resolution photos are not necessarily clear due to compression and focus, but low-resolution photos are blurry. We downsample the original photo to reduce its resolution and scale it up to fit the default observed size. The results show the FAS is also not sensitive to the photo quality until it becomes completely blurry.

2) IR Scene Generator: The IR scenes are generated by printing IR photos. Unlike RGB sensors, IR sensors work in two different wavelengths, *i.e.*, 850 and 940 nm. Observations of the two wavelengths are slightly different due to the different reflective properties of biological tissues. 850 nm observations are brighter and the highlight regions are more even. The face outline is clearer but the facial contours are less contrasted. The eye regions, *i.e.*, sclera (white), iris (gray), and pupil (black) are more distinctive in reflection intensity. Further, in the same 940 nm band, the ToF sensor and normal IR sensor are slightly different. ToF observations are less contrasted, the pupil usually shows bright rather than black.

Due to the above difference, some FASes are selective to IR photos, since they may have been trained with IR photos of certain wavelengths. So, we use 1 RGB camera and 3 different IR cameras to collect facial data. To study the impact of this factor, we repeat the tests of participant #C with the IR photos captured by different IR cameras. Device #4, #6, and #10 are selected since they use built-in algorithms, which are likely trained with and specific to the device's native IR wavelength. Table IV shows the results. The three devices are selective, while to different extents, to IR photos, implying that choosing correct IR photos is essential to defeating FASes.

The IR photo of the last column of Table IV is forged from the RGB photo with the method described in Appendix A. The main idea is to use a 3D face model overlaid with the facial features from the RGB photo to simulate the reflective properties of IR observations. The preliminary method only works for device #10, but the room for improvement remains large. Forged IR photos eliminate the need of collecting the target user's IR photos, which increases the risk of *Hua-pi* attack in many situations.

3) Depth Scene Generator: As we mentioned in Section V-B4, due to the ignorance of physical consistency check, a mean face model is able to fool all FAS devices. But this reason cannot explain the incremental benefits of the depth modality. Device #9, #11, and #14 in Table I make use of depth modality. However, compared with RGB+IR devices, the pass rate shows no evidence that depth modality brings obvious security advantages. We dig more into it.

In Table V, we measure the pass rate by replacing the default

Test Device		Index	3D Model					
			Real	Male	Female	Stacked	Foam	Ball
□□□	□□□	9	100.0%	100.0%	100.0%	100.0%	100.0%	58.4%
□□□	□□□	11	94.7%	92.0%	96.0%	90.4%	100.0%	88.8%
Sunny	Mars05b	14	84.5%	84.0%	84.0%	0.0%	94.4%	41.6%

TABLE V
IMPACT OF DEPTH SCENE GENERATOR. PASS RATE IN %.

Test Device		Index	Shift Scene	Shift Distance				
				up 1cm	up 0.5cm	aligned	right 0.5cm	right 1cm
Dumu	C2	3	RGB	91.2%	90.4%	96.0%	86.4%	96.0%
□□□	□□□	11	RGB	96.8%	96.8%	96.0%	90.4%	78.4%
			IR	34.4%	48.0%	96.0%	92.8%	7.2%
Sunny	Mars05b	14	RGB	20.8%	61.6%	84.0%	32.0%	10.4%
			IR	2.4%	54.4%	84.0%	84.0%	45.6%

TABLE VI
IMPACT OF SCENE ALIGNMENT. PASS RATE IN %.

Device Cost				
Item	Description	Unit Price	Quantity	Amount
LCD 13"	RGB Display	$100	1	$100
Optical Combiner		$75	2	$150
Glass Plate 15"	Holding IR Sheet	$10	1	$10
Camera Clamp	Holding Items	$25	4	$100
3D Head Model		$15	1	$15
IKEA Wooden Box	Device Container	$35	1	$35
+Woodworking	Customization	$40	1	$40
			Total	$450
Per-attack Cost				
Laser Printing	White Paper	$0.05	1	$0.05
Laser Printing	Transparent Sheet	$0.25	1	$0.25
3D Printing	(If Applicable)	$15	1	$15

TABLE VII
COST OF *Hua-pi* DISPLAY AND ATTACK CONSUMABLES.

mean female 3D face model in Figure 6 with different objects. The real model (generated from the participant's ToF depth image) and the mean male model [50] are printed with the same material. The Stacked model is derived from the female model by deliberately reducing the vertical printing resolution to 2 mm. The foam model's [55] material is different from the 3D-printed ones. It is smoother and more reflective. The ball is a 3D-printed ellipsoid of human-head size.

Intuitively, the way we judge whether a 3D surface shows a human face is to see if it has facial components, *e.g.*, slightly raised nose and sunken eyes, but through the tests, we find that some FASes are not quite sensitive to these features. A smooth and convex surface is enough to fool them. Take device #9 for example. It is a widely-used smart lock module manufactured by the major supplier of structured light cameras. Its depth modality is more like a tool to roughly determine whether there is an uneven surface in front, and does not examine details. The above indicates:

➥ **Defective Design 3 (D3): The use of depth information is superficial.**

4) Scene Alignment: While many FASes ignore physical consistency checking (Section V-B4), it does not mean the scenes could be arbitrarily presented. A reason is from face detection. Like face recognition, usually only one modality is used for face detection. The detected face area is used as the reference to crop the corresponding face areas of other observations. Then, all cropped face areas are forwarded to anti-spoofing and face recognition algorithms. Hence, if the scenes are not aligned, the cropped observations may not contain correct facial information.

By default, we align the scenes by adjusting the size and location of the scene content to get their eyes and noses aligned. Table VI shows the tolerance of scene misalignment. We deliberately shift the physical position of a scene off the aligned position. When the device uses RGB and IR, we shift the RGB scene only. When the device uses RGB, IR, and depth, we shift the position of RGB or IR scenes and keep the remaining two aligned. A shift distance of 0.5 cm is roughly about 4% of the face width displayed in the scene generators. The results show that some devices are sensitive to misalignment. Recall that the camera needs to be positioned at default observing locations (Section V-A5). They are two side of the same issue. Good alignment leads to a wide range of feasible observing positions. We determine the default alignment positions with help of a similar camera or our eyes' observations. Table VI also shows that compared with RGB, the results are slightly more sensitive to shifting the IR scene. This is because IR and depth observations are measured by the same sensor and have more correlations.

D. Cost Analysis

As shown in Table VII, the prototype of *Hua-pi* display costs about $500, which likely can be halved with cheaper components. The display consumes office printing supplies, which cost less than $1 for one target. When the target user's 3D head models are needed (not necessary for current FASes, see Section V-B4), the per-attack cost is about $10. The literature explored using carefully-crafted 3D model/masks to evade FASes. Regardless of the effectiveness, we estimate the price for comparison. A high-quality custom face mask costs more than $3000 [21], [44], and it is the per-attack cost, since the model/masks have to be tuned/manufactured for different target users.

VI. COUNTERMEASURES AND DISCUSSION

This study covers several representative devices, but it affects numerous similar products. Hence, it is not possible to contact vendors one by one. In the first place, we have reported our results to the authorities that certify or standardize FAS products, *e.g.*, BCTC (Bank Card Test Center) [56] and major companies.

A timely upgrade of FAS algorithms can largely thwart the attack. The aforementioned design issues are constructive. **D1** and **D3** suggest that using more available modalities and using them correctly can increase anti-spoofing performance. **D2** suggests enforcing the physical consistency check. An immediate patch is to reuse the face recognition module to check whether the faces in different modalities are similar and aligned. Further, the device can treat FAS as a secondary authentication option that must be used with other methods,

such as PINs and fingerprint. Also, disabling face authentication after several failures can effectively limit illegal spoofing attempts.

The above methods cannot completely eliminate *Hua-pi* attacks and some of them involve user overhead, which diminishes the advantages of static anti-spoofing. Permanent solutions can focus on impairing scene decoupling. Since the scene combiner used in *Hua-pi* display is only effective for light signals, a potential method is to leverage non-optical modalities, such as radio [57] and sonic imaging [58], to defeat optical spoofing. Another way is to complicate optical decoupling. For example, the FAS can use multiple cameras, *e.g.*, binocular camera[7], to observe the scene from different positions. Their observations have spatial disparity. Replicating the disparity requires multiple precisely-aligned scene generators.

On the other hand, *Hua-pi* attack can be further improved. First, if *Hua-pi* display becomes portable, then, it can be freely moved to aim at stationary targeting devices, *e.g.*, payment terminals, door access controllers, *etc*. The components of *Hua-pi* display are light in weight but require sufficient space to generate scenes to a suitable size. Sophisticated optical lens systems could help aggregate and compact the light paths. Further, many FASes heavily rely on the IR modality, whose content contains distinct features but is mostly similar to the RGB modality. It is interesting to explore the feasibility of high-precision translation of the two modalities. We believe this will not only facilitate the attack but may also reveal other security and privacy issues.

VII. Conclusion

Using facial information for authentication is like a paradox that the information that people expose daily is used as the identity proof. Latest face authentication systems leverage multimodal cameras to thwart spoofing attempts, but are based on an unvalidated assumption that the modalities can hardly be forged simultaneously. We challenge this assumption by showing the feasibility of decoupling the modalities with a low-cost optical display device. We then apply it to evade various commercial face authentication systems. The risks that the attack causes echo people's skeptical attitude towards face recognition applications.

Acknowledgements

We thank anonymous reviewers for their valuable comments. We are grateful to Dr. Hao Ren and Zhihao Li from ShanghaiTech for their assistance in laser optics.

References

[1] "Facial Recognition Market Size," https://www.prnewswire.com/news-releases/facial-recognition-market-size-worth-12-67bn-globally-by-2028-at-14-2-cagr---exclusive-report-by-the-insight-partners-301489784.html, 2023.

[2] Y. Li, K. Xu, Q. Yan, Y. Li, and R. H. Deng, "Understanding osn-based facial disclosure against face authentication systems," in *Proceedings of the 9th ACM symposium on Information, computer and communications security*, 2014, pp. 413–424.

[3] L. v. Ahn, M. Blum, N. J. Hopper, and J. Langford, "Captcha: Using hard ai problems for security," in *International conference on the theory and applications of cryptographic techniques*. Springer, 2003, pp. 294–311.

[4] "Facial Recognition: A Safer Access Control Solution," https://americansecuritytoday.com/facial-recognition-a-safer-access-control-solution/, 2023.

[5] "WeChat Pay," https://www.cnbc.com/2019/11/19/tencents-wechat-china-may-soon-use-facial-recognition-for-payments.html, 2023.

[6] A. Bud, "Facing the future: The impact of apple faceid," *Biometric technology today*, vol. 2018, no. 1, pp. 5–7, 2018.

[7] "Apple Face ID 'Fooled Again'. This Time By $200 Evil Twin Mask," https://www.forbes.com/sites/thomasbrewster/2017/11/27/apple-face-id-artificial-intelligence-twin-mask-attacks-iphone-x/?sh=1c1a3ad42775, 2023.

[8] "The painted skin," https://en.wikipedia.org/wiki/The_Painted_Skin, 2023.

[9] P. Grother, M. Ngan, and K. Hanaoka, "Face recognition vendor test (frvt) part 2: identification," https://pages.nist.gov/frvt/html/frvt1N.html, 2023.

[10] P. Grother, M. Ngan, K. Hanaoka, and A. Hom, "Face recognition vendor test (frvt) part 1: verification," https://pages.nist.gov/frvt/html/frvt11.html, 2023.

[11] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

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

[14] L. Qin, F. Peng, M. Long, R. Ramachandra, and C. Busch, "Vulnerabilities of unattended face verification systems to facial components-based presentation attacks: An empirical study," *ACM Transactions on Privacy and Security*, vol. 25, no. 1, pp. 1–28, 2021.

[15] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 819–826.

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

[17] M. Shen, Z. Liao, L. Zhu, K. Xu, and X. Du, "Vla: A practical visible light-based attack on face recognition systems in physical world," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–19, 2019.

[18] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 135–147.

[19] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[20] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3d masks," *IEEE transactions on information forensics and security*, vol. 9, no. 7, pp. 1084–1097, 2014.

[21] R. Ramachandra, S. Venkatesh, K. B. Raja, S. Bhattacharjee, P. Wasnik, S. Marcel, and C. Busch, "Custom silicone face masks: Vulnerability of commercial face recognition systems & presentation attack detection," in *2019 7th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2019, pp. 1–6.

[22] Y. Xu, T. Price, J.-M. Frahm, and F. Monrose, "Virtual u: Defeating face liveness detection by building virtual models from your public photos," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 497–512.

[23] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *2009 International Conference on Image Analysis and Signal Processing*. IEEE, 2009, pp. 233–236.

[7] Binocular is a basic depth measurement method [59]. Its principle is similar to structured light, but lacks active patterns to enhance disparity. Although not a common choice, it is used by some FAS products [60].

[24] "Facetec," https://dev.facetec.com/, 2023.

[25] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[26] E. Uzun, S. P. H. Chung, I. Essa, and W. Lee, "rtcaptcha: A real-time captcha based liveness detection system." in *NDSS*, 2018.

[27] D. Tang, Z. Zhou, Y. Zhang, and K. Zhang, "Face flashing: a secure liveness detection protocol based on light reflections," *arXiv preprint arXiv:1801.01949*, 2018.

[28] Y. Li, Y. Li, Q. Yan, H. Kong, and R. H. Deng, "Seeing your face is not enough: An inertial sensor-based liveness detection for face authentication," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1558–1569.

[29] V. Costa, A. Sousa, and A. Reis, "Image-based object spoofing detection," in *Combinatorial Image Analysis*, R. P. Barneva, V. E. Brimkov, and J. M. R. Tavares, Eds. Cham: Springer International Publishing, 2018, pp. 189–201.

[30] A. Agarwal, R. Singh, and M. Vatsa, "Face anti-spoofing using haralick features," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2016, pp. 1–6.

[31] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*. IEEE, 2012, pp. 1–7.

[32] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–8.

[33] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5042–5051.

[34] "We Broke Into A Bunch Of Android Phones With A 3D-Printed Head," https://www.forbes.com/sites/thomasbrewster/2018/12/13/we-broke-into-a-bunch-of-android-phones-with-a-3d-printed-head/?sh=7505f0781330, 2023.

[35] G. Heusch, A. George, D. Geissbühler, Z. Mostaani, and S. Marcel, "Deep models and shortwave infrared information to detect face presentation attacks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 4, pp. 399–409, 2020.

[36] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 42–55, 2019.

[37] H. Steiner, S. Sporrer, A. Kolb, and N. Jung, "Design of an active multispectral swir camera system for skin detection and face verification," *Journal of Sensors*, vol. 2016, 2016.

[38] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1179–1187.

[39] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 919–928.

[40] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5295–5305.

[41] Z. Yu, C. Zhao, K. H. Cheng, X. Cheng, and G. Zhao, "Flexible-modal face anti-spoofing: A benchmark," *arXiv preprint arXiv:2202.08192*, 2022.

[42] I. S. Winkler and B. Dealy, "Information security technology? don't rely on it. a case study in social engineering." in *USENIX Security Symposium*, vol. 5, 1995, pp. 1–1.

[43] S. Sanderson and J. Erbetta, "Authentication for secure environments based on iris scanning technology," 2000.

[44] "REAL-f CO.,LTD," https://real-f.jp/en_news.html, 2023.

[45] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE transactions on information forensics and security*, vol. 11, no. 10, pp. 2268–2283, 2016.

[46] X. Sun, L. Huang, and C. Liu, "Context based face spoofing detection using active near-infrared images," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 4262–4267.

[47] "NIR Projector Module," https://www.aeonimaging.com/nir-projector-module/, 2023.

[48] "Nir tn liquid crystal rotator," https://boldervision.com/product/nir-tn-liquid-crystal-rotator/, 2023.

[49] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[50] "Mean faces - using mean depth map of registered model sets," http://www.clementcreusot.com/phd/, 2023.

[51] "Photographic filter," https://en.wikipedia.org/wiki/Photographic_filter, 2023.

[52] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang, "Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5773–5782.

[53] Q. Li, X. Dong, W. Wang, and C. Shan, "Cas-air-3d face: A low-quality, multi-modal and multi-pose 3d face database," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[54] "Image manipulation toolbox," http://mimtdocs.rf.gd/manual/index.html?i=1, 2023.

[55] "Male head form," https://www.amazon.com/Craft-Foam-Wig-Head-Polystyrene/dp/B07C95WQ2W/, 2023.

[56] "Bank card test center," https://en.bctest.com/, 2023.

[57] Y. Zhu, Y. Zhu, B. Y. Zhao, and H. Zheng, "Reusing 60ghz radios for mobile radar imaging," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 103–116.

[58] A. O'Donovan, R. Duraiswami, and J. Neumann, "Microphone arrays as generalized cameras for integrated audio visual processing," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[59] "Depth Map from Stereo Images," https://docs.opencv.org/4.x/dd/d53/tutorial_py_depthmap.html, 2023.

[60] "SenseTime Launches Facial Verification Smart Lock for Smart-Home Industry," https://www.sensetime.com/cn/product-detail?categoryId=32498&gioNav=1, 2023.

[61] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[62] H. Wang, H. Zhang, L. Yu, L. Wang, and X. Yang, "Facial feature embedded cyclegan for vis-nir translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1903–1907.

(a) (b) (c) (d)

Fig. 9. **Forging IR Photos from RGB Photos**. (a) RGB Photo. (b) True IR Photo. (c) Forged IR photo by neural network style transfer. (d) Forged IR Photo by heuristic image processing.

Test Device		Index	Forgery Method	Participant			
				C	F	G	Q
□□□	□□□	7	GAN	92.8%	86.4%	93.6%	57.6%
NXP	SLN-VIZN3D-IOT	10	Heuristic	84.8%	52.0%	~~0.0%~~	50.4%
Sunny	Mars05b	14 RGB+IR	Heuristic	62.4%	92.8%	100.0%	57.6%
		14 RGB	Heuristic	84.8%	99.2%	100.0%	64.8%
		14 IR	Heuristic	86.4%	94.4%	100.0%	95.2%

TABLE VIII
PERFORMANCE OF ATTACKING WITH ONLINE RGB PHOTOS AND FORGED IR PHOTOS. PASS RATE IN %.

APPENDIX

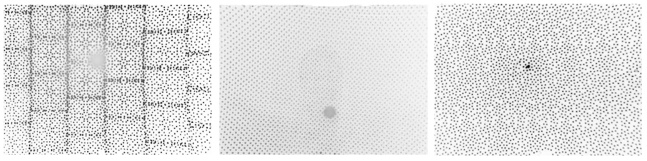
A. IR Photo Forgery

To ease the collection of IR scene content, we explored two methods to transform RGB photos into IR ones. The first is style transfer, which is based on the generative adversarial network (GAN) [61]. The network model treats RGB and IR as different styles of the same image and learns the style relation from RGB and IR photos. We trained an FFE-CycleGAN model [62] with the Oulu-CASIA NIR-VIS dataset. Then, the model is applied to covert RGB face photos to forge corresponding IR photos. For example, Figure 9 (a)(b) are original RGB and IR photos. Figure 9 (c) is generated by the model from (a).

The second is a heuristic method. We summarize the features of IR photos and use image processing techniques to forge them. IR photos are taken under IR illumination, and hence have distinct reflectivity and highlight areas. To emulate these features, we first construct the 3D face model from the RGB photo [49] and apply a point light source to the model to generate the lighting features. Then, the red channel of the RGB photo is overlaid to the 2D front view of the 3D model to sketch the facial component. Figure 9 (d) is the example of the synthesized IR photo.

With the above two methods, it is possible to use the target user's online RGB photos to feed the *Hua-pi* display to launch the attack. Participant #C, #F, #G, and #Q each provided two RGB photos sourced from their vlog screenshot, account avatar, personal homepage, and social networks. We select one photo from each as the RGB scene content. Then, we forge the IR photos from the selected RGB photos with the heuristic and GAN methods, and use them as the IR scene content. Except for the RGB and IR photos, the rest of the settings are identical to tests in Table I.

The results in Table VIII show the feasibility of using online photos as the scene content to attack some devices. The pass rate values are lower than that of Table I. This is because some RGB online photos are of lower-resolution and the forged IR photos are less realistic. The tests of device #14 reveal the performance losses in each modality.



(a) NXP SLN-VIZN3D-IOT (b) Intel RealSense F455 (c) □□□

Fig. 10. **Structured Light Patterns**.

Besides, the attack cannot impersonate participant #G on device #10 because of recognition failures. We guess the main reason is that the RGB photo she provided was taken a few years ago. This is another noteworthy issue when attacking FAS: up-to-date facial information is beneficial for bypassing the face recognition module.

B. Extended Discussion of Depth Scene Display

The *TYPE-B* design in Section IV-D2 is to forge light patterns to generate depth scenes of arbitrary objects for structured cameras. In practice, the patterns of different cameras are different, so the reference plane needs to be recorded prior to calculating the counterfeit pattern. This is a one-time procedure since the pattern is static and (likely) consistent among devices of the same model. Figure 10 depicts several examples we recorded.

Another issue is synchronization. In Section IV-D2, we mentioned that depth cameras multiplex IR sensors in the time domain to observe the IR scene and depth scene. Most depth cameras keep their projector off unless they are capturing light patterns. Otherwise, the IR scene will be interfered with by the pattern. It means the IR generator presenting the counterfeit pattern should also react likewise, *i.e.*, turning off/on according to the projector's state. Our measurement suggests that the on period of the patterns usually last for several milliseconds. We did not find appropriate photoelectric switches and made a circuit for this special case. We use a photodiode to track the light emission of the projector to drive the switch of the IR illumination source.

Due to the above reasons, when using the *TYPE-B* display to launch *Hua-pi* attacks, physical access to the FAS camera is needed to block the projector and attach the detector. This is the main drawback compared to the *TYPE-A* design. Another limitation of *TYPE-B* is that it is based on Figure 5 (a), but some structured light cameras, such as device #8 in Table I and Intel RealSense D400 Series, leverage dual sensors to observe the light pattern to improve the depth measurement. The counterfeit patterns for the two sensors are different since the sensor locations are different. Modifications of the current *TYPE-B* design are needed to handle these cases.

Test Device									Test Settings				
Vendor	Model	Index	IR Wave.	Firmware Version	Algorithm Version	Performance	RGB Rec.	IR Rec.	RGB ID	IR ID	IR Photo	3D Model	Options
Dumu	C2	1	850 nm	/	□□□	FAR<0.5%, FRR<1%	●	○	match	#C	850 nm	/	default
Dumu	C2	2	850 nm	/	□□□	/	●	○	match	#C	850 nm	/	default
Dumu	C2	3	850 nm	v0.7.11-040fd36	Atthis v0.6.8-a4a0eee	/	●	○	match	#C	850 nm	/	default
□□□	□□□	4	850 nm	□□□	□□□	/	●	○	match	#C	850 nm	/	default
NXP	SLN-VIZNAS-IOT	5	850 nm	v2.0.32	OASIS LITE v4.7.5	/	●	○	match	#C	850 nm	/	door access (heavy)
Intel	RealSense F455	6	850 nm	v4.3.0.8200	/	FAR<0.1%	●	○	match	match	850 nm	/	high
□□□	(door access)	7	850 nm	□□□	□□□	FAR<0.5%	●	○	match	#C	850 nm	/	default
Orbbec	Petrel	8	940 nm	/	□□□	FAR<0.5%, FFR<1%	●	○	match	#C	/	female	default
□□□	□□□	9	940 nm	□□□	□□□	FAR<0.1%, FRR<0.1%	○	●	/	match	940 nm	female	default
NXP	SLN-VIZN3D-IOT	10	940 nm	v1.1.4	OASIS v1.27.0	FAR<0.5%	○	●	/	match	fake	female	default
□□□	□□□	11	940 nm	/	□□□	FAR<0.1%	●	○	match	#C	940 nm	female	default
□□□	(smartphone)	12	940 nm	□□□	/	/	○	●	/	match	940 nm	female	no face mask
□□□	(smartphone)	13	940 nm	□□□	/	/	○	●	/	match	940 nm	female	default
Sunny	Mars05b	14	940 nm ToF	/	□□□	FAR<0.5%, FFR<1%	●	○	match	#C	940 nm ToF	female	default
□□□	□□□	15	940 nm	/	/	FAR<0.1%, FRR<1%	○	●	/	match	940 nm	female	default
□□□	(smartlock)	16	940 nm	□□□	/	/	○	●	/	match	940 nm	female	default

IR Wave.	Wavelength of the IR camera and illuminator.	RGB ID	Use the RGB photo of the participant (match) for the test.
RGB Rec.	Use (●) or do not use (○) RGB observation for face recognition.	IR ID	Use the IR photo of participant #C or the participant (match) for the test.
IR Rec.	Use (●) or do not use (○) IR observation for face recognition.	IR Photo	Photos are taken by IR cameras in Table 2, or forged from the RGB photos (fake).

Performance Anti-spoofing performance in FAR and FRR from the product datasheet. For face recognition performance, typical values are: true positive rate > 99% when false positive rate = 10^{-6}.

FAR False acceptance rate (false positive pate). Ratio of spoofing attempts that are incorrectly recognized as genuine attempts (real people).

FRR False rejection rate (false negative rate). Ratio of genuine attempts (real people) that are incorrectly recognized as spoofing attempts.

TABLE IX

OVERALL RESULTS (PART 2). EXTENDED INFORMATION FOR TABLE I.

	Participant																			
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
Age	20	20	20	20	40	20	20	30	20	20	30	20	20	40	<20	<20	>80	70	30	30
Gender	F	F	M	M	F	M	F	M	M	F	M	M	M	M	M	M	M	F	M	M
Ethnicity	EA	EA	EA	EA	EA	EA	EA	EA	EA	EA	SA	A	A	E	EA	EA	EA	EA	E	E
Background	L	L	L	L	L	L	L	L	L	L	O	R	C	O	L	L	H	S	L	O

Acronyms:

Age **<20**, [**20**, 30), [**30**, 40), [**40**, 50), [**50**, 60), [**60**, 70), [**70**, 80), **>80**

Gender **M**ale, **F**emale

Ethnicity **E**urope, **A**frica, **E**ast **A**sia, **S**outh **A**sia

Background **L**aboratory, **O**ffice, **C**afe, **S**tore, **R**estaurant, **H**ospital

TABLE X

PARTICIPANT INFORMATION.