VLA: A Practical Visible Light-based Attack on Face Recognition Systems in Physical World

MENG SHEN, ZELIN LIAO, and LIEHUANG ZHU, Beijing Institute of Technology, China
KE XU, Tsinghua University & BNRist, China
XIAOJIANG DU, Temple University, USA

Adversarial example attacks have become a growing menace to neural network-based face recognition systems. Generated by composing facial images with pixel-level perturbations, adversarial examples change key features of inputs and thereby lead to misclassification of neural networks. However, the perturbation loss caused by complex physical environments sometimes prevents existing attack methods from taking effect.

In this paper, we focus on designing new attacks that are effective and inconspicuous in the physical world. Motivated by the differences in image-forming principles between cameras and human eyes, we propose VLA, a novel attack against black-box face recognition systems using visible light. In VLA, visible light-based adversarial perturbations are crafted and projected on human faces, which allows an adversary to conduct targeted or un-targeted attacks. VLA decomposes adversarial perturbations into a perturbation frame and a concealing frame, where the former adds modifications on human facial images while the latter makes these modifications inconspicuous to human eyes. We conduct extensive experiments to demonstrate the effectiveness, inconspicuousness, and robustness of the adversarial examples crafted by VLA in physical scenarios.

CCS Concepts: • **Computing methodologies** → *Neural networks*; • **Security and privacy** → **Spoofing attacks**.

Additional Key Words and Phrases: adversarial example, visible light attack

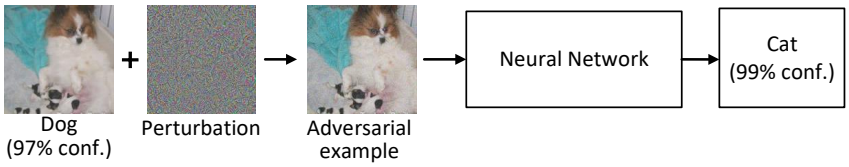
ACM Reference Format:
Meng Shen, Zelin Liao, Liehuang Zhu, Ke Xu, and Xiaojiang Du. 2019. VLA: A Practical Visible Light-based Attack on Face Recognition Systems in Physical World. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 103 (September 2019), 19 pages. https://doi.org/10.1145/3351261

1 INTRODUCTION

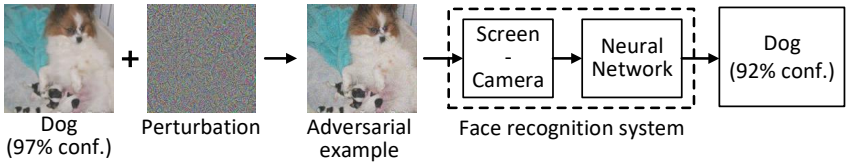
Recent years have witnessed the growing popularity of face recognizers powered by neural networks [6, 37, 40, 41], e.g., FaceNet [31] and SphereFace [24]. An increasing number of neural network-based face recognition systems are being widely deployed in the physical world for access control and decision making [1, 17]. A wide range of studies in ubiquitous computing have granted power from face recognition, e.g., from daily entertainment [5] to e-health [42], mobile user authentication [8] to social assistance [18]. The security concerns for ubiquitous computing applications have also attracted great public attention [12, 13, 20, 33, 34, 43]. Existing studies show that neural networks can be vulnerable to potential attacks [25, 45, 46]. Among all kinds of security threats, adversarial example attacks raise an increasing amount of attention from both academia and industry [16, 36, 38, 46].

Authors' addresses: Meng Shen, shenmeng@bit.edu.cn; Zelin Liao, lzl1918@outlook.com; Liehuang Zhu, liehuangz@bit.edu.cn, Beijing Institute of Technology, No.5, Zhongguancun South Street, Haidian District, Beijing, 100081, China; Ke Xu, xuke@tsinghua.edu.cn, Tsinghua University & BNRist, No.1, Qinghuayuan, Haidian District, Beijing, 100084, China; Xiaojiang Du, dxj@ieee.org, Temple University, 1925 N. 12th Street, Philadelphia, PA, 19122, USA.

2474-9567/2019/9-ART103 $15.00
https://doi.org/10.1145/3351261



(a) Digital scenario



(b) Physical scenario

Fig. 1. An example illustrating perturbation loss of digital adversarial examples in physical conditions.

An adversarial example is an elaborately prepared input to neural networks for face recognition that can lead to an incorrect output, i.e., an output that is basically different from the result of human sense [7, 21, 28, 29]. These attacks can cause unexpected behaviors of neural networks, making the corresponding face recognition systems unreliable. Generally, an adversarial example is derived by composing the original input with *perturbations*, which change the values of some key features and thereby result in a misclassification.

Existing studies on adversarial example attacks against neural networks begin in the digital domain [22, 27, 44]. Digital attacks aim at crafting pixel-level perturbations directly in digital images such that the resulting adversarial examples can mislead neural network based classifiers while being imperceptible to human eyes. In physical-world face recognition systems, however, these attacks may lose efficiency, as the input images to neural networks cannot be manipulated directly by an adversary [26]. Instead, the images for recognition are generally captured by embedded cameras.

The inefficiency of migrating digital attacks in physical scenarios is demonstrated in Fig. 1. Digital attacks generally work as exhibited in Fig. 1(a): given an original image labelled as *dog* with 97% confidence, a digital adversarial example can mislead the neural networks into labeling the sample as *cat* with 99% confidence. To simulate a real face recognition process in the physical world where an adversary cannot manipulate the input image directly, the same adversarial example is first displayed on a screen before being captured by a camera and finally fed to the recognizer. As illustrated in Fig. 1(b), the sample is classified as *dog* with 92% confidence. This is because the modifications on digital objectives crafted by adversarial attack methods become less distinctive in the physical world, which is referred as *perturbation loss* hereafter.

More recent research efforts are dedicated to *physical*, also known as *practical*, adversarial examples [11, 14, 23, 32, 49], where perturbations are physically added to the objectives themselves. Within the domain of face recognition, Sharif et al. [32] showed the feasibility of physical adversarial examples by employing facial decorations, where perturbations are printed on the frames of eyeglasses. Although effective, this attack is observable to nearby people as the eyeglasses colored with generated perturbations are quite conspicuous. Similarly, Zhou et al. [49] proposed adversarial examples based on an infrared mask, which is realized by illuminating objectives with infrared perturbations. The elaborate attack is invisible to human eyes, but it can be simply wiped out by infrared cut-off filters that are commonly equipped in solid state (e.g., CMOS) cameras to block infrared wavelengths while passing visible light.

Considering the limitations of existing methods, it is still a challenging task to craft adversarial examples against face recognition systems in physical scenarios. On one hand, adversarial examples should be applied to a human face in an *imperceptible* and *inconspicuous* way, which means that physical perturbations used in the attack should not be detected by human eyes and unobservable by nearby people. On the other hand, adversarial examples should make the physical perturbations *sensitive* to cameras in face recognition systems such that the perturbed facial images captured by cameras can cheat the face recognition algorithms. It is also desirable that physical perturbations are *robust* in various environmental conditions.

In this paper, we propose a visible light-based attack (VLA) method, which is a brand-new approach to generate adversarial examples against neural networks in physical-world face recognition systems. The basic idea is to project visible light based perturbations on human faces, leading to *targeted* or *un-targeted* misclassification on faces. The adoption of visible light is motivated by the differences in image-forming principles between human eyes and digital cameras: the persistence of vision (POV) which only exists for human eyes can mix the colors of two image frames swapped in a high frequency while cameras cannot, which leads to various appearances from perceptions of human eyes and cameras. POV provides an efficient way to expose specific images to cameras while hiding them from human eyes.

In order to mitigate perturbation loss and make perturbations more sensitive to cameras embedded in face recognition devices, we design region-level, instead of pixel-level, modifications to human faces. Pixel-level perturbations are visually isolating pixel points with various colors, while region-level perturbations are more like comparatively large regions each of which are filled with only one color.

As enlarged perturbations can be more effective against less sensible sensors, the newly generated perturbations can be more noticeable. To ensure the inconspicuousness of region-level perturbations crafted by VLA, they are decomposed into two frames: one is designed to change features of the current user, named the *perturbation frame*, while the other frame is used to hide the perturbation frame from human views, named the *concealing frame*. Both of the two frames are generated with only one environment-related parameter, which can be inferred from the distance between an adversary and the camera.

Extensive experiments are conducted to evaluate the performance of VLA. The well-known face dataset LFW [19] along with a custom dataset are utilized to investigate the attack success rate of VLA and a typical adversarial example generation method FGSM [15]. The result shows that both methods could reach high attack success rate in digital scenarios. While in physical scenarios, the effectiveness degradation of VLA is 7.6% vs. 57.3% for FGSM, which is more tolerable. The robustness of VLA which is also revealed as the attack success rate is kept in high level (i.e., less than 10% of success rate degradation) when changing environmental conditions. The experiment shows the high attack success rate of VLA (i.e., 98.6%) when concealing frames are accidentally captured. In addition, the inconspicuousness of generated perturbations for human senses is concluded from an on-spot survey and gets quantified by comparing image differences, i.e., fused images of faces and perturbations perceived by human eyes are similar with the appearance of original human faces.

We summarize the main contributions as follows:

(1) We estimate the feasibility of hiding perturbations using visible light based on differences in image-forming principles for humans and cameras according to the chromatic color mixture model. (Section 3)
(2) We propose VLA to generate and present adversarial examples in physical scenarios. VLA is agnostic to specific neural networks and requires only one input parameter. Each adversarial example can be generated within 3 seconds, achieving high efficiency. (Section 4)
(3) We conduct an on-spot survey and extensive experiments, the results of which reveal the relatively high attack success rate of VLA in physical scenarios. The inconspicuousness and robustness of perturbations generated by VLA are also demonstrated. (Section 5).

Table 1. Summary of typical physical adversarial example attacks

Target	Method	Perturbation presentation method	Imperceptibility or inconspicuousness
Physical objects	Kurakin et al. [23]	Add perturbations onto images	Not focused
Road signs	RP_2 [14]	Tie shapes onto road signs	Minimize the distance of adversarial examples with original images
	DARTS [35]	Print colors onto road signs	Minimize the p-norm of perturbations
Faces	Sharif et al. [32]	Wear eye glasses with printed colors	People wearing glasses is natural
	IMA [49]	Project infrared onto faces	Infrared is not visible for human
	VLA	Project perturbations to human faces	Hide perturbations according to POV

2 RELATED WORK

With the wide adoption of neural networks in various applications, the security and safety of neural networks have attracted increasing research efforts. Adversarial example attacks were first proposed by Szegedy et al. [38], aiming at crafting adversarial perturbations that deceive neural networks to label input objectives incorrectly. Many different approaches have been proposed to generate adversarial examples for specific or general-purpose applications. According to the objectives to which adversarial perturbations are added, existing approaches can be classified into two categories: *digital* adversarial attacks and *physical* adversarial attacks. The summary of several attack methods targeting physical scenarios is shown in Table 1.

Digital adversarial attacks. Methods within this category rely on an implicit assumption that attackers can directly feed adversarial examples to machine learning classifiers. Thus, adversarial perturbations can be applied to digital objectives, such as digital images.

Szegedy et al. [38] proposed an attack against neural networks and demonstrated the feasibility that modifying an input x with a perturbation r can mislead the results given by DNN classifiers. Along the similar line, different methods have been proposed to find r with the minimal modification to the original inputs, leading to visual similarity between the original and perturbed samples. Papernot et al. [30] proposed a method to find r by computing forward derivatives. Their method achieves a relatively high success rate while requiring modifications on only a few features. Xie et al. [44] proposed a gradient descent algorithm to generate adversarial examples, misleading an object detection system into outputting targeted labels. Carlini et al. [10] proposed methods according to three different distance metrics to generate digital adversarial examples. Their methods are significantly effective although being more computational expensive and requiring more time for generation.

Physical adversarial attacks. To better understand the impact adversarial examples have on the performance of neural networks deployed in real-world applications, physical adversarial attacks have been proposed to generate adversarial examples on real objectives, instead of digital ones.

Eykholt et al. [14] proposed a method to mislead the classification result of road signs in consideration of specific challenges in the physical world. By attaching cropped papers with color printed to road signs, the method succeeded in leading the classifier to produce incorrect results. Similarly, Sitawarin et al. [35] proposed a method to deceive autonomous cars by printing adversarial examples onto road signs that are previously generated by photos of road signs. These methods show automatic driving systems are vulnerable to potential attacks. More importantly, they also revealed the feasibility of applying adversarial example attacks against classifiers in physical applications.

In the scenarios of face recognition, the primary goal of physical adversarial examples is to add the minimal perturbations to real faces in an inconspicuous and robust way under varying environmental conditions.

Sharif et al. [32] presented a method to fool DNN-based face recognition systems by printing a pair of eyeglass frames. Their method enables attackers to evade face recognition or impersonate another individual. However, the adversarial examples can be conspicuous as the eyeglasses as well as the colors shown on the frames look unusual. Zhou et al. [49] employed infrared to generate facial perturbations on real faces to deceive face recognizers. The infrared is imperceptible to human eyes but not for cameras, which improves the stealthiness. However, the infrared can be easily filtered out with low-cost lens, which blocks perturbations from being captured, making the adversarial examples inefficient. In addition, the generation of perturbations relies on a white-box assumption, where the attackers can fully access the face recognition algorithm. It also raises health concerns as an attacker may get the eyes hurt if exposed in infrared for a long period of time (around 10 minutes).

Previous discussion reveals that the imperceptibility and robustness are requiring different grades between digital and physical attacks. These two factors lead to conflicts between inconspicuous perturbations against effective modifications of inputs. To reduce the effect of perturbation loss, we enlarge image modifications from pixel level to region level. Perturbations are presented by projection using visible light. To enhance the imperceptibility of adversarial example attacks, with projecting perturbation frames which in fact take a role of changing facial features, we additionally introduce the concealing frames which unify the appearance of perturbations by swapping two frames in a high frequency according to several differences in the imaging theory between human eyes and cameras.

3 MOTIVATION

Research advances in communications have developed visible light wireless communication systems by employing the differences in the image-forming principle between human eyes and cameras [39, 47]. Motivated by these observations, we investigate the feasibility of generating adversarial perturbations using visible light for face recognition, with special considerations on two critical factors: inconspicuousness to human eyes and sensitivity to cameras.

Persistence of Vision (POV). When light changes faster than 25Hz, human brain does not directly process these changes at the exact moment they occur. Instead, the brain will mix the last image with the newly changed image, which is called *persistence of vision* [47]. However, for cameras with fast shutter speeds (i.e., 1/60s), there is no such effect since pixels in an image are derived by reading instantaneous voltages of sensors. POV reveals the feasibility of exposing certain shapes to cameras while making them less perceptible to human eyes. The effect of POV, which is further demonstrated by a concrete example in Appendix A, can be used for generating physical adversarial examples.

The chromatic addition rule. The effect of POV causes a fusion of colors, which is called *color mixture* hereafter in this paper. The mixture of colors R_A and R_B, which produces a color R_C for human eyes, is defined as $R_A \oplus R_B = R_C$. Note that the fusion result of R_A and R_B is independent of the sequence of the two colors, thus we have $R_A \oplus R_B = R_B \oplus R_A$.

For clarity, the main notations used in the rest of this paper are summarized in Table 2. One color R_* can be expressed as a 3-dimension vector $R_* = (x_*, y_*, Y_*)$ in the CIE 1931 color spaces, which were the first defined quantitative links between distributions of wavelengths in the electromagnetic visible spectrum and physiologically perceived colors in human color visions. The mixture result can be expressed using the *chromatic addition rule* [47]:

$$\begin{aligned} x_C &= \frac{Y_A}{Y_A+Y_B} x_A + \frac{Y_B}{Y_A+Y_B} x_B \\ y_C &= \frac{Y_A}{Y_A+Y_B} y_A + \frac{Y_B}{Y_A+Y_B} y_B \\ Y_C &= \frac{Y_A+Y_B}{2} \end{aligned} \tag{1}$$

Table 2. Notations used in this paper

Notation	Description
R_*	A color usually expressed as a 3-dimension vector
P_A	The label of user A
I_A	An image containing the face of the user labelled as P_A
$[R_*]$	An image with all pixels rendered in color R_*
$\mathrm{FR}(I_A) = P_A$	An image I_A is recognized by the face recognition system with a label P_A
r	A perturbation frame generated by VLA
r'	A concealing frame generated by VLA
n_0	A pre-defined threshold to filter color regions containing pixels less than n_0
C_{I_A,I_B}	A set of regions that make up the clustering result of images $I_A - I_B$
$H\left(C_{I_A,I_B}\right)$	A function that recovers an image from a group of regions

The color mixture model has been validated through a series of experiments in the late 1920s. In Section 5.4, our survey also demonstrates its effectiveness.

Then, we define the operation by which we extract color R_B from the fused result R_C and the other source R_A. The separation is denoted as $R_B = R_C \ominus R_A$. For ease of calculation, values in all of the three color channels are normalized to $[0, 1]$. It is worth noting that there may not always exist a color R_B such that $R_A \oplus R_B = R_C$, making the resulting channel values out of the range $[0, 1]$. In this case, any value out of the range is approximated to its nearest bound (0 or 1).

Following the definition of color addition and subtraction, given two images I_A and I_B of the same size, we define the image addition $I_A + I_B$ as the color addition of pixels at the same position. Similarly, the image subtraction $I_A - I_B$ is defined as the color subtraction of pixels at the same position.

4 THE PROPOSED METHOD

In this section, we first describe the threat model which focuses on impersonating system users. Then, we present the overview of the proposed visible light-based attack (VLA) and illustrate the methods of generating and presenting adversarial perturbations for human faces.

4.1 Threat Model

Generally, a face recognition system would calculate the feature similarity of input facial image with faces that are previously trained and stored in its model. We denote the face recognition algorithm as $\mathrm{FR}(I_A)$ which returns the identity detected from the input image of user P_A. An ideal face recognition algorithm is always expected to label the input image correctly, i.e., $\mathrm{FR}(I_A) = P_A$.

In this paper, we focus on two types of impersonation attacks based on adversarial perturbations to prevent identities from being correctly recognized.

Targeted impersonation attack. The goal of a targeted attack is to mislead the recognition result of an image of user P_A into the result of a specific user P_B, by adding perturbations r_A, i.e., $\mathrm{FR}(I_A + r_A) = P_B$ $(P_A \neq P_B)$. In real-world applications, such an attack enables an illegal user P_A to pretend to be a legitimate user P_B.

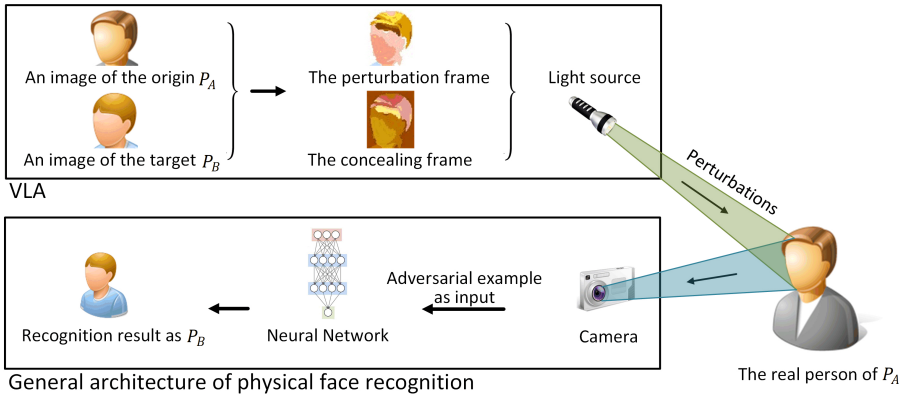


Fig. 2. The architecture of face recognition systems and attacks in the physical world.

Un-targeted impersonation attack. Attacks in this category attempt to make the recognition result of an image of user P_A be any label other than P_A by adding adversarial perturbations r_A, i.e., $\mathrm{FR}(I_A + r_A) \neq P_A$. Such an attack can block a legitimate user P_A from being correctly recognized.

Theoretically, there exists a special case of un-targeted impersonation attacks, which makes the face recognizer fail in detecting any human faces. In this paper, however, we focus on the more dangerous cases of un-targeted impersonation attacks, where an illegal user can spoof face recognition systems by imitating legitimate users.

4.2 Overview of VLA

The key limitation of existing approaches to generate adversarial examples is the unresolved perturbation loss, which becomes the main obstacle to migrating adversarial examples working well in digital attacks to physical scenarios, as shown in Fig. 1. For instance, eye-inconspicuous perturbations can be invisible for cameras, whereas camera-sensible perturbations are noticeable changes for human eyes.

To tackle the challenges of robustness and inconspicuousness in physical scenarios, VLA decomposes adversarial perturbations into separate frames, namely *perturbation frames* and *concealing frames*.

To mitigate the effect of perturbation loss in physical scenarios, pixel-level image modifications are enlarged into region-level. A perturbation frame r is employed for changing and imitating facial features and is generated with exclusive color regions that fill up the whole frame. To hide the perturbation frame from human eyes, a concealing frame r' is introduced according to the effect of POV discussed earlier.

That is, for each adversarial example, VLA generates a perturbation frame and a concealing frame and presents the two frames *alternatively*. The former contains information of how to change facial features of user P_A to the features of a targeted or un-targeted user P_B, while the latter aims at hiding the perturbations in the perturbation frame from being observed by human eyes.

The architecture for presenting adversarial examples to face recognition systems in physical scenarios generally consists of three components: a user, a face recognition system, and a model for generating adversarial perturbations. The workflow of VLA is shown in Fig. 2. A perturbation frame r is generated given the images of a user P_A and another user P_B. When the user P_A presents before the camera by which a face recognition system gains its inputs, the perturbation frame r along with the concealing frame r' would be alternately projected to the face of P_A. The camera captures the composition results of face images and perturbations before feeding them to neural networks for recognition.

Adversarial examples come with the additive effects of faces and perturbations, where the projected frames and human faces are captured together. Due to the existence of the perturbation frame r, the recognition result can finally be P_B for impersonation attacks. According to the color mixture model, by alternately projecting r and r', the perturbed regions in r will take on the appearance of no special pattern or color for human eyes.

Since VLA introduces two separate frames to ensure the robustness and inconspicuousness of adversarial examples, the generation of each adversarial example consists of two phases: 1) generation of perturbation frames, and 2) generation of concealing frames.

4.3 Generation of Perturbation Frames

Given the facial images of an original user P_A and a target user P_B, which are denoted by I_A and I_B respectively, an adversary needs to design a perturbation frame r such that $I_A + r$ would be labelled with a high probability as I_B. We refer to the image addition result of I_A and r as I'_A, i.e., $I'_A = I_A + r$.

To imitate the features of I_B with I'_A, a better adversarial example owns a lower image *distance* (e.g., pixel-level Euclidean distance) between I'_A and I_B. Intuitively, for the ideally best fit where $I'_A = I_B$, we could get the perturbation frame r as $r = I_B - I_A$. In this way, a perturbation frame can create an image I'_A exactly the same with I_B, which owns the highest probability of imitating P_B.

However, such modifications can only be feasible at a pixel level. Since pixel-level modifications may lead to perturbation loss in physical scenarios, we try to divide $I_B - I_A$ into exclusive ranges based on the similarity of containing color values. For simplicity, we denote $I_\omega = I_B - I_A$. Such a division can be described as a clustering task, where nearby similar colors are divided into the same regions.

The first step is to reduce the number of different colors in I_ω by performing a MeanShift clustering over all colors. The algorithm finds several colors which are the center of corresponding clusters divided according to color distances. These colors can be used to make up an image similar to I_ω. Each group of nearby pixels with the same color in the image is regarded as one perturbation region.

Then, the second step is to ensure each region is observable for cameras. As stated earlier, the resolution of the image captured by the camera differs from the resolution of the image projected by the projector and small color regions would get lost in the images captured in physical scenarios. One may come up with a simple solution of resizing (e.g., upscaling or downscaling) the projected images. However, it is unable to address this problem, as resizing cannot superpose the projected perturbation frame to the physical face of P_A. Therefore, in VLA, a clustering method and a region filtering strategy are utilized to ensure that all projected details in a perturbation frame can be successfully captured by the camera.

Taking regions with the area less than n_0 as 'noises', we remove these noises and replace them with their neighbors owning the largest area. n_0 is a pre-defined constant depending on the camera-user distance. Generally, n_0 is related to the scaling ratio of sizes between perturbation frames (with c_p pixels in the diagonal line) and captured facial regions (with c_q pixels in the diagonal line). Regions containing $n_0 = \left(c_p/c_q\right)^2$ pixels in a perturbation frame could be recorded as one pixel by the camera. Considering the perceptive transformation caused by facial depth, n_0 is suggested to be higher than $\left(c_p/c_q\right)^2$. By filtering color details with n_0, a region containing n_0 pixels is reflected as a single pixel in images captured by the camera. As a result, the number of pixels in each region would be no less than n_0.

Denoting a clustering and filtering result of the image $I_B - I_A$ as C_{I_A,I_B}, we define $C_{I_A,I_B} = \{(G_i(p), R_i) \,|\, 0 \leq i \leq m\}$ where $G_i(p)$ indicates whether the color of a pixel p should be set as R_i, and m is the total number of color regions. For each pixel p in the image C_{I_A,I_B}, $G_i(p) = 1$ if p lies within R_i, or $G_i(p) = 0$ otherwise. Finally, we define a generation function $H(\cdot)$, which transforms the clustering result C_{I_A,I_B} into a perturbation frame r, as shown in Eq. (2):

$$r = H(C_{I_A,I_B}) = [R_i \text{ if } G_i(p) = 1] \tag{2}$$

where p is the position of a pixel in the generated perturbation frame r.

Since the generation of perturbation frames requires two images as input, one of which is the facial image of the target person, an adversary can easily derive perturbation frames for targeted impersonation attacks. For un-targeted attacks, it is also feasible as the attacker can randomly select a target user which is different from the original user.

4.4 Generation of Concealing Frames

As stated earlier, perturbations should be *robust* and *inconspicuous* in physical scenarios. The robustness of perturbations is provided by the perturbation frames proposed above. We now describe the generation of concealing frames, which are employed in VLA to make perturbations inconspicuous to human eyes.

The rationale for designing concealing frames is the effect of POV. Remember that two different colors that swap frequently can make human eyes perceive a new color. According to the color mixture model in Eq. (1), we can hide the color R_A in a background color R_C by finding and alternately displaying a color R_B where $R_B = R_C \ominus R_A$. By changing the start time and swapping frequency, we can determine which color (e.g., R_A or R_B) is available to cameras.

Given a base color R_{back}, a straightforward way to hide color R_i from R_{back} is to find a color R'_i that $R'_i = R_{back} \ominus R_i$. By displaying these two colors alternately, it can be difficult for human eyes to feel the color R_i. Notice that the clustering result of image $I_B - I_A$ is denoted by C_{I_A,I_B}, we have

$$C'_{I_A,I_B} = \left\{ (G_i(p), R_{back} \ominus R_i) \,|\, (G_i(p), R_i) \in C_{I_A,I_B} \right\} \tag{3}$$

R_{back} can be simply selected by calculating the gray color converted from the average of colors in a perturbation frame r. The same image generation method can be applied to obtain the concealing frame r' as $r' = H\left(C'_{I_A,I_B}\right)$.

By making the light source swap r and r' in a high frequency, a perturbation frame can be hidden from the background as $r + r' = [R_{back}]$. Thus, the resulting adversarial example can be inconspicuous to human eyes and observable to cameras.

5 PERFORMANCE EVALUATION

The experiments aim at evaluating the effectiveness and efficiency of VLA, in terms of 1) the success rate of VLA in physical scenarios, 2) to what extent VLA can reduce the perturbation loss, 3) the inconspicuousness of VLA, and 4) the robustness of VLA in various environmental conditions.

5.1 Experimental Settings

Testbed. We select FaceNet [31], dlib [3], and SphereFace [24] as the state-of-the-art face recognition systems. They are deployed on a PC with 8GB RAM and an Intel Dual-Core i5-2435M CPU. The camera captures images with a size of 480×640 in pixels. To present adversarial perturbations, an LCD projector NEC ME-300X+ with a lamp NP16LP-ME is used as the source of visible light. The projection resolution is 1366×768 with a frame rate of 60fps.

The default values of the experimental parameters are set as follows: FaceNet as the face recognizer, camera shutter speed of 1/60s, the user-projector distance of 40cm, environmental brightness of 350lux, and a head pose of facing directly forward. The default device we used to present perturbations is an LCD projector. All the experiments are conducted using the default setting unless otherwise noted.

As described previously, a lower camera resolution would require a larger n_0, which loses more details of adversarial examples and thereby makes attacks more likely to fail. In our experiment, we select a relatively lower capturing resolution, i.e., 480×640, which helps to reveal the worst-case performance of VLA.

Methods to compare. We select a widely-used digital adversarial example attack, Fast Gradient Sign Method (FGSM) [15], as the baseline for comparison. Since FGSM focuses on un-targeted impersonation attacks by generating pixel-level perturbations, we obtain two variants of VLA for targeted and un-targeted attacks, which are referred to as VLA_T and VLA_U, respectively.

For VLA_T, an attack is successful only when the recognition result is the same with the targeted label. While for VLA_U, an attack succeeds once the recognition result is different from the label of the original user.

Datasets. An open dataset **LFW** [19] is used as a large-scale dataset, which consists of 13,233 facial images of 5,749 different persons. The trained recognition algorithms in FaceNet, dlib, and SphereFace can achieve an accuracy of 99.7%, 99.4%, and 99.2% over LFW, respectively [4].

Since LFW contains different number of labelled facial images for individuals, the detection accuracy of FaceNet may vary when being applied for different persons. In other words, the detection accuracy of the recognition algorithm in FaceNet may vary over different subsets of LFW. To investigate the performance of different attacks against face recognition systems with various detection accuracies, two subsets LFW_4 and LFW_8 are selected from LFW, which contain individuals that have no less than 4 and 8 labelled facial images.

In addition, to simulate the recognition process on real faces, 9 volunteers from campus are involved in collecting a small-scale dataset named **CusFace**, which contains 10 facial images for each volunteer.

For each dataset of CusFace, LFW_4, and LFW_8, half of the facial images are used for training FaceNet, while the other half for validation. For the full set of LFW, the public available pre-trained FaceNet model [2] is used. The accuracy of the recognition algorithm in FaceNet over different datasets is listed in Table 3.

Table 3. Detail of datasets

Dataset	# Persons (Labels)	# Total facial images	Accuracy of FaceNet
CusFace	9	90	100%
LFW	5,749	13,233	99.7%
LFW_8	217	4,822	92.2%
LFW_4	610	6,733	71.4%

5.2 Evaluation of Impersonation Attacks

The experiments evaluating the success rate of VLA are conducted with the datasets listed in Table 3.

Evaluation with CusFace. For VLA, we generate several groups of *origin-target* attacks as follows: Each volunteer acts as user P_A in Fig. 2 and attempts to imitate one of the rest 8 individuals. Since each targeted individual has 10 facial images in CusFace, 10 adversarial examples are generated for each *origin-target* pair. Thus, there are altogether $9 \times 8 \times 10 = 720$ adversarial examples for evaluation. For each adversarial example, VLA generates a pair of perturbation frame and concealing frame that would be then alternately projected onto the face of user P_A, as shown in Fig. 2.

The success rate of VLA_T can be simply calculated as the percentage of adversarial examples that successfully impersonate the targeted individual. For un-targeted attacks, the success rate of VLA_U is the percentage of adversarial examples that make the face recognizer mislabel an original person as another individual.

As a counterpart for comparison, FGSM focuses on un-targeted attacks. The original implementation of FGSM requires only a single image to generate adversarial perturbations. In our implementation, however, we use a pair of images of the same individual so that the generated adversarial examples would be detected with labels different from both images. Since each label (e.g., an individual) in CusFace has 10 facial images, there are 90 ordered image pairs for each label, among which 80 pairs are randomly picked for evaluation. Thus, we also have

a total of $9 \times 80 = 720$ adversarial examples. The success rate of FGSM is counted as the percentage of adversarial examples that lead to misclassification.

Table 4. Success rate against FaceNet with CusFace

Methods	Physical scenarios	Digital scenarios
FGSM	31.0%	88.3%
VLA_U	84.5%	92.1%
VLA_T	46.2%	90.0%

Table 5. Success rate against FaceNet with LFW

Methods	Dataset	
	LFW_8	LFW_4
FGSM	86.3%	87.9%
VLA_U	90.5%	91.2%
VLA_T	87.3%	89.2%

The experiments are conducted using FaceNet. The success rates of VLA and FGSM with CusFace are summarized in Table 4. It is clearly shown that for the un-targeted attacks in physical scenarios, VLA_U significantly improves the success rate over FGSM. In addition, VLA_T also achieves a moderate success rate for the targeted attacks. The results can be explained that the region-level color areas in perturbation frames generated by VLA are more robust that help to obtain more effective adversarial examples.

Although we focus on physical scenarios in this paper, we still compare the success rate of adversarial examples in *digital* scenarios, as the results would provide more insights into the differences between the two scenarios.

In digital scenarios, attacks rely on a strong and somewhat unrealistic assumption that the generated adversarial examples can be directly fed to the recognizer in FaceNet. From Table 4, we can find that there is a sharp decline in the success rate of FGSM by 57.3%, which demonstrates that pixel-level perturbations experience sever degradation when being applied in physical scenarios. For VLA, the success rate of un-targeted attacks reaches 92.1%, which has only a slight decrease in physical scenarios. For targeted attacks, however, there is a significant drop of success rate. We will further investigate the reasons behind the result in the next subsection.

Evaluation with LFW. Since FGSM requires pairwise facial images in a labelled dataset, it is unfeasible to simulate the real-world face recognition using LFW. Therefore, we are concentrated on evaluating the effectiveness of these methods in digital scenarios, where the results can shed a light on their performance in physical scenarios, with a joint consideration of the results in Table 4.

VLA uses the image pair of two individuals to generate perturbations, which is previously named as *origin-target* pair. We randomly select 20 and 30 labels from LFW_8 and LFW_4 as origins, respectively. For each origin label, we randomly select 8 different labels as targets. Each *origin-target* pair is tested with 5 pairs of images. As a result, there are $20 \times 8 \times 5 = 800$ and $30 \times 8 \times 5 = 1200$ adversarial examples for LFW_8 and LFW_4, respectively.

For FGSM, we randomly select 200 and 300 labels from LFW_8 and LFW_4, respectively. And for each label, we randomly select 4 pairs of images to generate adversarial examples. Thus, we have 800 and 1200 samples with LFW_8 and LFW_4, which have a same scale with those of VLA.

For these attack methods, we ensure that all of the selected images without perturbations are correctly labelled by FaceNet. We summarize the success rates of these methods in Table 5. In general, VLA shows a higher success rate than FGSM. Since a larger dataset (e.g., LFW_4) usually decreases the detection accuracy of FaceNet, impersonation attacks can achieve a higher success rate with a larger dataset.

Time efficiency. The time overhead for VLA to generate adversarial examples is evaluated in physical and digital experiments with LFW. On average, VLA costs less than 3 seconds to generate a frame pair containing a perturbation frame and a concealing frame. Also, since VLA requires only one parameter of n_0 which can be inferred by estimating the human-camera distance, the adversarial examples can be generated in advance.

Table 6. Success rates against various face recognizers with LFW

Methods	Face Recognizers		
	FaceNet	dlib	SphereFace
FGSM	32.6%	26.3%	26.5%
VLA_U	85.6%	86.5%	86.1%
VLA_T	32.2%	32.8%	33.0%

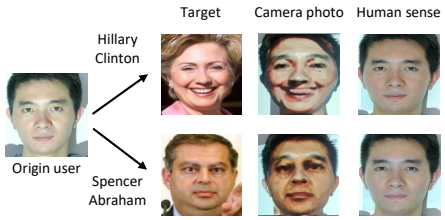


Figure 3. Samples of a user mimicking Hillary Clinton and Spencer Abraham.

5.3 Evaluation with Various Face Recognizers

FGSM is a white-box attack, indicating that the generation of perturbations depends on an acquaintance over the structure of underlying neural networks. In contrast, VLA is a black-box attack model, as the details of the targeted face recognition system is unnecessary for generating adversarial examples.

Targeting FaceNet, we generate the adversarial examples using FGSM and VLA separately. These examples are also used to evaluate with other face recognizers, namely SphereFace and dlib.

Each of the 9 volunteers involved acts as a real user P_A and tries to mimic each of the 60 targets randomly selected from the full set of LFW. Each origin-target pair is tested 3 times, resulting in $9 \times 60 \times 3 = 1620$ cases in total. The attack success rates are summarized in Table 6.

The results of FGSM indicate that the attack success rate against dlib and SphereFace is less than that against FaceNet, as FGSM is a white-box approach and the adversarial examples targeting FaceNet may not fit for other recognizers. VLA is agnostic to face recognizers, exhibiting similar performance against the three recognizers. The results also reveal that FGSM is less effective than VLA_U for un-targeted attacks in real-world scenarios.

5.4 Inconspicuousness Evaluation of Adversarial Examples

The chromatic addition rule describes the mixture result of how human brains sense colors when affected by POV. Concealing frames are generated according to Eq. (3) to weaken the appearance of projected perturbations in human perceptions. Now, we investigate the inconspicuousness of the adversarial examples crafted by VLA.

As described before, the human sense of colors may last for about 1/16s after the colors get changed. Thus, to simulate the human perceived appearances of adversarial examples (i.e., the fusion results of real faces with crafted adversarial perturbations generated by an attack method), we use the corresponding images captured by a camera with a slower shutter speed of 1/14s, which approximates the one in human image-forming principles. Note that the default camera shutter speed is set to 1/60s.

To link the real-world original users with the targeted individuals in LFW, FaceNet is re-trained using the combined dataset of **CusFace** and LFW. Two samples are shown in Fig. 3, where the original user, the target user, and the adversarial examples sensed by the camera and human eyes are accordingly exhibited.

To evaluate the effectiveness of the color mixture model adopted in VLA as well as the methodology in simulating human perceived adversarial examples (e.g., the 4th column in Fig. 3), we conduct the experiment including two parts: 1) an on-spot survey on the effects of adversarial examples crafted by VLA, 2) a quantification of human perceptual differences between adversarial examples and original facial images.

We conducted a survey oriented to 46 individuals without prior knowledge of the attack method. Each participant was presented with 4 attack cases and finished the following questions: 1) whether they saw shining lights on faces, 2) whether they observed any abnormal pattern projected on faces, 3) if choosing *yes* in 2), what

Table 7. Human perceptual average distance between adversarial examples and original facial images

Metric	SelfDis (S_R, S_R)	InterDis (S_R, S_R)	SelfDis (S_R, S_T)	SelfDis (S_T, S_T)	InterDis (S_T, S_T)
# Image Pair	15,930	129,600	32,400	15,930	129,600
Value	0.133 ± 0.03	0.456 ± 0.09	0.192 ± 0.02	0.135 ± 0.05	0.459 ± 0.02

they believe the pattern comes from (options are *uncertain*, *meaningless color regions*, *a picture of something (e.g., a building)*, *someone's face*), and 4) quantifying the similarity between the face perceived by their eyes and the simulated human-sensed image as shown in Fig. 3, ranging from 0 (totally different) to 10 (exactly the same).

The survey shows that all participants can notice additional facial lights. In 10.8% of the total cases, the participants can find something abnormal on faces, among which 40% are uncertain of the pattern source, while the remaining 60% believe the patterns are meaningless color regions. Overall, nobody could tell if these exists face resembles someone else's face. The result demonstrates the effectiveness of the color mixture model used in VLA.

The average similarity score derived from the 4th question in the survey reaches 8.2, which validates the effectiveness of our methodology in simulating human perceptions.

To further quantify the inconspicuousness of adversarial examples, we use a distance metric proposed by Zhang et al. [48] to quantify the similarity between the facial image of a real user (without any perturbation) and the simulated image of the corresponding adversarial example. The metric reduces the effect of perspective differences among images, by which extra image differences caused by changes of user positions and poses are counteracted. We use $\mathrm{D}(I_1, I_2)$ to symbolize the distance between two images I_1 and I_2.

There are 9 volunteers (i.e., real users) involved in the experiment. For each volunteer, 60 facial images without and with perturbations are respectively collected into image sets S_R and S_T. As a result, both of the two sets contain $60 \times 9 = 540$ facial images from 9 volunteers. Accordingly, two types of image pair sets are generated: 1) Ω_{S_1,S_2} contains image pairs belonging to the same volunteer selected from image sets S_1 and S_2, 2) Δ_{S_1,S_2} is made up of image pairs of different volunteers selected from S_1 and S_2, respectively.

$$\begin{aligned} \text{SelfDis}\,(S_1, S_2) &= \text{Average}\left(\left\{\mathrm{D}\,(I_1, I_2) \,|\, (I_1, I_2) \in \Omega_{S_1,S_2}\right\}\right) \\ \text{InterDis}\,(S_1, S_2) &= \text{Average}\left(\left\{\mathrm{D}\,(I_1, I_2) \,|\, (I_1, I_2) \in \Delta_{S_1,S_2}\right\}\right) \end{aligned} \tag{4}$$

As shown in Eq. (4), we employ two kinds of distance metrics to quantify the average image distances among various image sets. $\text{SelfDis}\,(S_1, S_2)$ is the average image distance of the same volunteer in the sets S_1 and S_2, and $\text{InterDis}\,(S_1, S_2)$ is the average image distance of two different volunteers in S_1 and S_2, respectively.

Human perceptual average distance between adversarial examples and original facial images is summarized in Table 7. The image distance of a user with perturbations is closer to the image distance of the same user without perturbations rather than the image distance of different users (c.f. $\text{SelfDis}\,(S_R, S_T)$-$\text{SelfDis}\,(S_R, S_R)$ vs. $\text{SelfDis}\,(S_R, S_T)$-$\text{InterDis}\,(S_R, S_R)$-$\text{InterDis}\,(S_T, S_T)$), which means that the appearances of simulated images of a user are similar with the original images of the same user.

Thus, by displaying a perturbation and a concealing frames alternately, the effect of POV could successfully prevent perturbations from being sensed by humans, while these perturbations can be captured by cameras.

5.5 Effects of Perturbation and Concealing Frames

For human eyes, the effect of POV joins both the perturbation frame r and the concealing frame r' on forming images of presented faces, from which features contained in r cannot be easily sensed as $r + r'$ is designed to be unified in color, i.e., $r + r' = [R_{back}]$.

For cameras, as the perturbation frame r and the concealing frame r' are alternately displayed, only one frame (either r or r') can be captured in a single image. If the perturbation frame r is captured, the perturbed facial features in r can be successfully applied to the original face via projection.

Since there is a possibility that the camera captures the fusion result of the real face and the concealing frame, we investigate 880 cases with FaceNet where only concealing frames are captured by the camera. The experimental settings are the same as those in Fig. 3.

In 90% of the cases, the recognition system cannot detect any facial structure, as a concealing frame diminishes the structural features of a real face. About 1.4% of the cases are labelled correctly as the original user, and the rest 8.6% are mislabeled as other persons.

Affected by the concealing frame, the facial structure of a real user is damaged, making FaceNet fail in detecting any faces. As a result, in 90% of the cases when facial structure cannot be detected with a concealing frame only, the camera in face recognition systems will attempt to take another image again, which potentially provides more chances for the attacker to adjust the projection sequence or add several pauses so that perturbation frames can be captured by the camera.

5.6 Evaluation of VLA with Varying Influencing Factors

To investigate the performance of VLA in different conditions, we exploit the success rate of VLA by varying several influencing factors, including the camera shutter speed, user-projector distance, head pose, and environmental brightness level. We employ FaceNet as the face recognizer and evaluate 180 cases in each specific parameter setting: 9 volunteers as original users, 10 individuals randomly selected from LFW as target users, and 2 adversarial examples for each origin-target pair.

Table 8. Success rates of VLA with various camera shutter speeds and projector types

Shutter Speed (s)		1/8	1/14	1/20	1/30	1/40	1/60	1/80	1/125	1/180	1/250	1/500	1/1000
FaceNet acc. (%)		83.6	90.4	96.3	97.3	100	100	100	100	100	100	100	100
LCD	VLA_U (%)	36.3	38.8	42.4	69.7	67.5	85.6	85.5	85.7	85.5	85.3	85.6	81.3
	VLA_T (%)	2.3	2.2	3.5	9.2	16.7	32.2	32.2	32.4	32.1	32.3	32.3	21.1
DLP	VLA_U (%)	35.5	35.6	42.9	67.9	68.9	85.8	84.7	85.2	85.1	70.3	71.1	69.6
	VLA_T (%)	2.2	2.3	5.9	10.7	15.6	32.1	32.7	32.8	26.8	20.3	16.9	17.4

Camera Shutter Speeds. Camera shutter speed is a critical factor impacting the quality of captured images and thereby affects VLA's performance. We evaluate the success rate of VLA against FaceNet, by increasing the shutter speed from 1/8s to 1/1000s.

To investigate the impact of different projection techniques on VLA's performance, we leverage a DLP projector (i.e., NEC NP-V302X+) in addition to the default LCD projector (i.e., NEC ME-300X+) in our experiments. The results are summarized in Table 8. To better understand the impact of shutter speed on face recognition systems, we also present the corresponding accuracy of the original FaceNet over facial images without perturbations.

Based on the results with the *LCD* projector, we make several observations corresponding to *four* types of camera shutter speeds: 1) Slow shutter speeds (i.e., 1/8s and 1/14s), which allow camera sensors to detect environmental lights with a longer duration, are employed to simulate human perceived images under the effect of POV. The results demonstrate that a slow shutter speed can prevent the perturbations from being observed and thus the success rate of VLA significantly decreases. Meanwhile, a slow shutter speed can result in blurred facial images, leading to performance degradation of FaceNet. 2) Medium shutter speeds, ranging from 1/20s to 1/40s, can gradually remove image blurring and improve the accuracy of FaceNet. Low-quality perturbations in these

situations can be captured, making the success rate of VLA increase with the shutter speed. 3) Fast shutter speeds (i.e., from 1/60s to 1/500s) can eliminate face blurring completely and make the perturbations firmly captured. Thus, FaceNet maintains 100% accuracy and the VLA variants achieve their highest success rate. And 4) very fast shutter speeds (i.e., 1/1000s) catch artifacts from the projector in the attack scenarios, which means that partial of the perturbation frame is replaced by the content of the concealing frame. The resulting perturbed facial images can maintain most of the expected perturbations, but they cannot completely imitate the facial features of a targeted user. As a result, the success rate of VLA_U decreases slightly, whereas that of VLA_T drops significantly.

According to the results with the *DLP* projector, we can also categorize the shutter speeds into 4 types. The main difference lies in that the range of fast shutter speeds shrinks with the DLP projector, staring from 1/60s to 1/180s. Cameras with a shutter speed of 1/250s begin to capture artifacts from the projector. This is basically because that the color wheel of a DLP projector displays a single color segment in a short time interval and thereby a whole frame cannot be captured entirely by the camera with a fast shutter speed. For example, assuming that a 6-segment color wheel rotates at the speed of 120Hz (i.e., each color segment would be displayed 1/720s), it requires 1/120s to display the whole frame. As a result, cameras with a shutter speed faster than 1/120s cannot capture the expected perturbation frames. A color wheel with faster rotation speeds can enlarge the range of fast shutter speeds.

Note that an attacker can carefully select a light source with more stable performance, e.g., using an LCD projector instead of a DLP projector. In addition, we envision that increasing the refresh rate of a projector may help mitigate the effect of artifacts. Due to time and device limitation, we leave these attempts as the future work.

Distances. In physical scenarios, an attacker can adjust the distance to a camera and make his face fit exactly to the recognition area. Thus, we evaluate the success rate of VLA with typical user-projector distances of 40cm (by default), 1m, 2m, 3m, and 4m. The results are shown in Table 9.

Table 9. Success rate with varying user-projector distances

Distance	40 cm	100 cm	200 cm	300 cm	400 cm
VLA_U	85.6%	83.3%	82.2%	80.6%	80.6%
VLA_T	32.2%	28.3%	28.1%	24.4%	23.2%

The distance is also an indicator of projector intensity, where a shorter distance allows stronger light projected on human faces and thus leads to a higher success rate. When the distance increases, not only the strength of projected light is weaken, but also a larger n_0 has to be set, which reduces the textural detail in perturbations and decreases the attack success rate.

Head Poses. As discussed above, the shape of facial surface would cause perspective transformation of perturbations. We use several typical kinds of head poses to explore the effect: normal (NL, the default head pose), turning the head to the right for 20deg (HR) or to the left for 20deg (HL), raising the head for 20deg (HT) or lowering the head for 20deg (HB). The success rates with these head poses are shown in Table 10.

Table 10. Comparison of success rates in various head poses

Pose	NL	HL	HR	HT	HB
VLA_U	85.6%	81.3%	82.0%	77.3%	75.3%
VLA_T	32.2%	26.7%	26.0%	25.3%	24.7%

Table 11. Comparison of success rates in various environmental brightness levels

Brightness	VLA_U	VLA_T
25lux	85.7%	32.5%
150lux	86.0%	32.7%
350lux	85.6%	32.2%

Generally, head poses would not significantly reduce the attack success rate. Vertical head movements may lead to facial distortion in camera-captured adversarial images, and thus having a greater impact on the performance of VLA than horizontal rotations.

Brightness Levels. We evaluate the success rate of VLA with a fixed user-projector distance of 40cm with varying environmental brightness, i.e., 25lux, 150lux, and 350lux (by default). The results are shown in Table 11.

The environmental illumination has a trivial impact on the performance of VLA, because the projector would provide enough light for the camera to successfully capture facial details. Although a dark environment may make a light source (e.g., a projector) noticeable, the perturbation frame generated in VLA remains inconspicuous to human eyes, which is guaranteed by POV.

6 DISCUSSION

The results of our study show that VLA is an effective tool for generating physical-world adversarial examples against face recognition systems. Based on our observations during experimental evaluations, we now discuss several directions in both improving the performance of VLA and exploring designs for powerful countermeasures.

Pose adjustment. As shown in Table 10, an irregular head pose of an adversary may degrade the success rate of VLA, as it brings difficulties in aligning perturbations with facial components. As a result, image transformations [9], e.g., rotating and skewing, can be applied to generate pose-aware perturbation frames.

Perturbation imperceptibility. We adopt a black-box threat model in this paper, where an adversary is assumed to have no knowledge about the face recognition algorithms except their output. In certain situations, however, a powerful adversary may get access to the implementation detail of a face recognizer. Thus, VLA can be further optimized by generating perturbation frames with less crafted perturbations, which helps to improve the imperceptibility of attacks.

Countermeasures. As shown in Table 8, a straightforward way to defend against VLA without involving additional hardware is to reduce the shutter speed of cameras embedded in face recognition systems. However, a slower shutter speed is a double-edged sword, which also reduces the recognition accuracy of legitimate users. An alternative way is to employ multiple shutter speeds to detect the existence of perturbations by comparing captured images. It requires that the face recognizer be capable of distinguishing between *blurred* faces and *polluted* faces. Capturing and analyzing multiple images would also prolong the recognition duration, which has a negative impact on user experience. Addressing these challenges would be interesting for future work.

7 CONCLUSION

In this paper, we proposed a method named VLA to automatically generate adversarial examples, misleading the result of black-box face recognition systems in the physical world. We employed region-level perturbations to deal with the perturbation loss in physical scenarios and introduced concealing frames to make crafted perturbations imperceptible to human eyes. Extensive experimental results demonstrated that VLA can achieve high effectiveness and robustness while keeping inconspicuous to human eyes in physical scenarios. In future work, we will further investigate the methods to improve the effectiveness of VLA and explore powerful defenses.

ACKNOWLEDGMENTS

This work is partially supported by the National Key Research and Development Program of China under Grant 2018YFB0803405, the National Natural Science Foundation of China under Grants 61602039 and 61872041, the Beijing Natural Science Foundation under Grant 4192050, the China National Funds for Distinguished Young Scientists under Grant No. 61825204, the Beijing Outstanding Young Scientist Project, and CCF-Tencent Open Fund WeBank Special Funding.

REFERENCES

[1] 2018. Arrivals SmartGate. https://www.homeaffairs.gov.au/trav/ente/goin/arrival/smartgateor-epassport. (Accessed on 07/05/2018).

[2] 2018. davidsandberg/facenet: Face recognition using Tensorflow. https://github.com/davidsandberg/facenet. (Accessed on 11/05/2018).

[3] 2018. dlib C++ Library. http://dlib.net/. (Accessed on 11/05/2018).

[4] 2018. LFW : Results. http://vis-www.cs.umass.edu/lfw/results.html. (Accessed on 11/05/2018).

[5] 2019. Facebook's New Facial Recognition Photo Tagging. https://vtldesign.com/digital-marketing/social-media/nh-facebook-marketing/how-to-disable-facebook-facial-recognition-photo-tagging-nhmarketing/. (Accessed on 04/22/2019).

[6] Abdulbasit Alazzawi, Osman N. Ucan, and Oguz Bayat. 2018. Robust Face Recognition Algorithm Based on Linear Operators Discrete Wavelet Transformation and Simple Linear Regression. In *Proceedings of the First International Conference on Data Science, E-learning and Information Systems (DATA '18)*. ACM, New York, NY, USA, Article 1, 7 pages. https://doi.org/10.1145/3279996.3279997

[7] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. 2018. On the Robustness of Semantic Segmentation Models to Adversarial Attacks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 888–897. https://doi.org/10.1109/CVPR.2018.00099

[8] Mozhgan Azimpourkivi, Umut Topkara, and Bogdan Carbunar. 2017. Camera Based Two Factor Authentication Through Mobile and Wearable Devices. *IMWUT* 1, 3 (2017), 35:1–35:37. https://doi.org/10.1145/3131904

[9] Wenming Cao and Shoujue Wang. 2005. An Algorithm For Face Pose Adjustment Based On Gray-scale Static Image. In *Adaptive and Natural Computing Algorithms*, Bernardete Ribeiro, Rudolf F. Albrecht, Andrej Dobnikar, David W. Pearson, and Nigel C. Steele (Eds.). Springer Vienna, Vienna, 474–477.

[10] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. 39–57. https://doi.org/10.1109/SP.2017.49

[11] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. 2018. Robust Physical Adversarial Attack on Faster R-CNN Object Detector. *CoRR* abs/1804.05810 (2018). arXiv:1804.05810 http://arxiv.org/abs/1804.05810

[12] Xiaojiang Du, Mohsen Guizani, Yang Xiao, and Hsiao-Hwa Chen. 2009. A Routing-driven Elliptic Curve Cryptography Based Key Management Scheme for Heterogeneous Sensor Networks. *Trans. Wireless. Comm.* 8, 3 (March 2009), 1223–1229. https://doi.org/10.1109/TWC.2009.060598

[13] X. Du, Y. Xiao, S. Ci, M. Guizani, and H. . Chen. 2007. A Routing-Driven Key Management Scheme for Heterogeneous Sensor Networks. In *2007 IEEE International Conference on Communications*. 3407–3412. https://doi.org/10.1109/ICC.2007.564

[14] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 1625–1634. https://doi.org/10.1109/CVPR.2018.00175

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *ArXiv e-prints* (Dec. 2014). arXiv:stat.ML/1412.6572

[16] A. Graese, A. Rozsa, and T. E. Boult. 2016. Assessing Threat of Adversarial Examples on Deep Neural Networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 69–74. https://doi.org/10.1109/ICMLA.2016.0020

[17] R. Gusain, H. Jain, and S. Pratap. 2018. Enhancing bank security system using Face Recognition, Iris Scanner and Palm Vein Technology. In *2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU)*. 1–5. https://doi.org/10.1109/IoT-SIU.2018.8519850

[18] Till Hellmund, Andreas Seitz, Juan Haladjian, and Bernd Bruegge. 2018. IPRA: Real-Time Face Recognition on Smart Glasses with Fog Computing. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. ACM, New York, NY, USA, 988–993. https://doi.org/10.1145/3267305.3274122

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

[20] Jen-Yan Huang, I-En Liao, and Hao-Wen Tang. 2011. A Forward Authentication Key Management Scheme for Heterogeneous Sensor Networks. *EURASIP J. Wirel. Commun. Netw.* 2011, Article 6 (Jan. 2011), 10 pages. https://doi.org/10.1155/2011/296704

[21] Uyeong Jang, Xi Wu, and Somesh Jha. 2017. Objective Metrics and Gradient Descent Algorithms for Adversarial Examples in Machine Learning. In *Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017)*. ACM, New York, NY, USA, 262–277. https://doi.org/10.1145/3134600.3134635

[22] J. Kos, I. Fischer, and D. Song. 2018. Adversarial Examples for Generative Models. In *2018 IEEE Security and Privacy Workshops (SPW)*. 36–42. https://doi.org/10.1109/SPW.2018.00014

[23] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *CoRR* abs/1607.02533 (2016). arXiv:1607.02533 http://arxiv.org/abs/1607.02533

[24] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE

Computer Society, 6738–6746. https://doi.org/10.1109/CVPR.2017.713
[25] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. 2018. Trojaning Attack on Neural Networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-5_Liu_paper.pdf
[26] Jiajun Lu, Hussein Sibai, and Evan Fabry. 2017. Adversarial Examples that Fool Detectors. *CoRR* abs/1712.02494 (2017). arXiv:1712.02494 http://arxiv.org/abs/1712.02494
[27] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. 2018. Towards Imperceptible and Robust Adversarial Example Attacks Against Neural Networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 1652–1659. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16217
[28] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. 2017. Universal Adversarial Perturbations Against Semantic Image Segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2774–2783. https://doi.org/10.1109/ICCV.2017.300
[29] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. 2016. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*. 582–597. https://doi.org/10.1109/SP.2016.41
[30] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. The Limitations of Deep Learning in Adversarial Settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (2016), 372–387.
[31] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823. https://doi.org/10.1109/CVPR.2015.7298682
[32] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2016. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*. ACM, New York, NY, USA, 1528–1540. https://doi.org/10.1145/2976749.2978392
[33] M. Shen, B. Ma, L. Zhu, R. Mijumbi, X. Du, and J. Hu. 2018. Cloud-Based Approximate Constrained Shortest Distance Queries Over Encrypted Graphs With Privacy Protection. *IEEE Transactions on Information Forensics and Security* 13, 4 (April 2018), 940–953. https://doi.org/10.1109/TIFS.2017.2774451
[34] M. Shen, X. Tang, L. Zhu, X. Du, and M. Guizani. 2019. Privacy-Preserving Support Vector Machine Training over Blockchain-Based Encrypted IoT Data in Smart Cities. *IEEE Internet of Things Journal* (2019), 1–1. https://doi.org/10.1109/JIOT.2019.2901840
[35] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. 2018. DARTS: Deceiving Autonomous Cars with Toxic Signs. *CoRR* abs/1802.06430 (2018). arXiv:1802.06430 http://arxiv.org/abs/1802.06430
[36] Lu Sun, Mingtian Tan, and Zhe Zhou. 2018. A survey of practical adversarial example attacks. *Cybersecurity* 1, 1 (06 Sep 2018), 9. https://doi.org/10.1186/s42400-018-0012-9
[37] M. Sushama and E. Rajinikanth. 2018. Face Recognition Using DRLBP and SIFT Feature Extraction. In *2018 International Conference on Communication and Signal Processing (ICCSP)*. 994–999. https://doi.org/10.1109/ICCSP.2018.8524427
[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *CoRR* abs/1312.6199 (2013). arXiv:1312.6199 http://arxiv.org/abs/1312.6199
[39] Anran Wang, Chunyi Peng, Ouyang Zhang, Guobin Shen, and Bing Zeng. 2014. InFrame: Multiflexing Full-Frame Visible Communication Channel for Humans and Devices. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks (HotNets-XIII)*. ACM, New York, NY, USA, Article 23, 7 pages. https://doi.org/10.1145/2670518.2673867
[40] Jie Wang and Zihao Li. 2018. Research on Face Recognition Based on CNN. *IOP Conference Series: Earth and Environmental Science* 170, 3 (2018), 032110. http://stacks.iop.org/1755-1315/170/i=3/a=032110
[41] Mei Wang and Weihong Deng. 2018. Deep Face Recognition: A Survey. *CoRR* abs/1804.06655 (2018). arXiv:1804.06655 http://arxiv.org/abs/1804.06655
[42] Rui Wang, Andrew T. Campbell, and Xia Zhou. 2015. Using Opportunistic Face Logging from Smartphone to Infer Mental Health: Challenges and Future Directions. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 683–692. https://doi.org/10.1145/2800835.2804391
[43] Yang Xiao, Venkata Krishna Rayi, Bo Sun, Xiaojiang Du, Fei Hu, and Michael Galloway. 2007. A Survey of Key Management Schemes in Wireless Sensor Networks. *Comput. Commun.* 30, 11-12 (Sept. 2007), 2314–2341. https://doi.org/10.1016/j.comcom.2007.04.009
[44] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. 2017. Adversarial Examples for Semantic Segmentation and Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1378–1387. https://doi.org/10.1109/ICCV.2017.153
[45] Chaofei Yang, Qing Wu, Hai Li, and Yiran Chen. 2017. Generative Poisoning Attack Method Against Neural Networks. *CoRR* abs/1703.01340 (2017). arXiv:1703.01340 http://arxiv.org/abs/1703.01340
[46] P. Zawistowski. 2018. Adversarial examples: A survey. In *2018 Baltic URSI Symposium (URSI)*. 295–298. https://doi.org/10.23919/URSI.2018.8406730

[47] Lan Zhang, Cheng Bo, Jiahui Hou, Xiang-Yang Li, Yu Wang, Kebin Liu, and Yunhao Liu. 2015. Kaleido: You Can Watch It But Cannot Record It. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom '15)*. ACM, New York, NY, USA, 372–385. https://doi.org/10.1145/2789168.2790106

[48] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 586–595. https://doi.org/10.1109/CVPR.2018.00068

[49] Zhe Zhou, Di Tang, Xiaofeng Wang, Weili Han, Xiangyu Liu, and Kehuan Zhang. 2018. Invisible Mask: Practical Attacks on Face Recognition with Infrared. *CoRR* abs/1803.04683 (2018). arXiv:1803.04683 http://arxiv.org/abs/1803.04683

A THE DEMONSTRATION OF POV

(a) Simulated human perception (camera shutter speed is 1/14s)

(b) Appearance for cameras (camera shutter speed is 1/60s)

Fig. 4. An example illustrating the differences in image-forming principles between human eyes and cameras. Red and green are alternately displayed at a frequency of 1/60 second. For human eyes, the appearance is brown, as revealed by the image on the left. While the camera can only capture one color in a single image, as shown by the image on the right.

Existing work reveals that the human brain may have a color perception different from the appearance in digital camera images when environmental lights get changed in a high frequency. Due to the effect of POV, when color changes, the human brain would keep the perception of the last color for about 1/16s while accepting the new environmental color [47]. For instance, one will not perceive complete darkness when he/she blinks, even though there exists a moment when no light passes into his/her eyes. However, for cameras, since pixels in an image are derived by reading instantaneous voltages of sensors, the camera may capture completely black images when no lights passing into the camera lens.

In order to demonstrate the impact of POV on the appearance of objects, we conducted a simple experiment. We first prepared two frames: a pink frame with a green rectangle area in the center, and a green frame with a pink rectangle area in the center. Then, these two frames were displayed alternatively with a duration of 1/60 second on a computer screen. The effects perceived by human eyes and cameras are illustrated in Figs. 4 (a) and (b), respectively. To help readers find the rectangle, the corresponding area is highlighted with a white dashed box.

For human eyes, the whole frame is perceived as brown, which results from mixing green and pink. Fig. 4(a) shows a simulated human perceptual effect, which is taken by the camera by setting the shutter speed similar to that in human image-forming principles (i.e., 1/14s). While for the camera with a shutter speed of 1/60s, the sensors can only capture one color at a time, and thus the rectangle can be clearly observed.

Also, the above simulation, as well as the experimental result revealed in Table 8 show the feasibility of cameras getting similar perceptions to the current environment considering POV. The camera can extend its exposal duration with a slower shutter speed (e.g., 1/14s) to simulate the effect of POV. However, a low shutter speed can easily cause blur to human faces and lead to a degradation to the recognition accuracy for face recognizers.