Measurement-driven Security Analysis of Imperceptible Impersonation Attacks

Shasha Li, Karim Khalil, Rameswar Panda, Chengyu Song
Srikanth V. Krishnamurthy , Amit K. Roy-Chowdhury, *Ananthram Swami
University of California Riverside, *US Army Research Laboratory
{sli057,karimk,rpanda002}@ucr.edu, {csong,krish}@cs.ucr.edu,
amitrc@ee.ucr.edu, ananthram.swami.civ@mail.mil

***Abstract*—The emergence of Internet of Things (IoT) brings about new security challenges at the intersection of cyber and physical spaces. One prime example is the vulnerability of Face Recognition (FR) based access control in IoT systems. While previous research has shown that Deep Neural Network (DNN)-based FR systems (FRS) are potentially susceptible to imperceptible impersonation attacks, the potency of such attacks in a wide set of scenarios has not been throughly investigated. In this paper, we present the first systematic, wide-ranging measurement study of the exploitability of DNN-based FR systems using a large scale dataset. We find that arbitrary impersonation attacks, wherein an arbitrary attacker impersonates an arbitrary target, are hard if imperceptibility is an auxiliary goal. Specifically, we show that factors such as skin color, gender, and age, impact the ability to carry out an attack on a specific target victim, to different extents. We also study the feasibility of constructing universal attacks that are robust to different poses or views of the attacker's face. Our results show that finding a universal perturbation is a much harder problem from the attacker's perspective. Finally, we find that the perturbed images do not generalize well across different DNN models. This suggests security countermeasures that can dramatically reduce the exploitability of DNN-based FR systems.**

***Key Words:* face recognition, imperceptible adversarial perturbation, Internet of Things**

I. Introduction

Face-recognition-based biometric authentication has become very popular in Internet of Things (IoT) [25], [32], [45]. In fact, according to the International Biometric Group (IBG), face is the second most widely deployed biometric in terms of market share, right after fingerprints [26]. The most noteworthy applications using face recognition include opening doors [15], activating personalized services by automated identification of users, e.g., smart TV program selector or pervasive software such as Microsoft's Kinect [25], [49].

Face Recognition Systems (FRSs) are typically trained on known faces, and use the trained model to classify test cases (i.e., when a human presents herself to a camera). The deep learning paradigm has seen significant proliferation in FRSs due to its ability to provide high recognition accuracy [30], [33], [41].

Due to the ubiquity of FRSs in security-critical applications, their security and reliability have drawn attention and various attacks have been showcased. Early presentation attacks [1], [10], [44] impersonate a victim's identity by presenting a fake face to FRSs, which could be in the form of photographs, replayed videos, 3D masks etc., as shown in Fig. 1(a). It has been recently shown that Deep Neural Networks (DNNs) are vulnerable to adversarial examples [13], [19], [40]. Adversarial examples are generated in such a manner that humans cannot notice adversarially induced perturbations and correctly classify the images, but the perturbations cause FRSs to misclassify them. Many attack methods [6], [7], [14], [38] have been proposed to generate adversarial examples for impersonation attacks, among which, intensity-based adversarial examples (Fig. 1(c)) can be quickly generated and are effective against a variety of FRSs [13], [19]. Intensity-based impersonation attacks add imperceptible perturbation to the original face images such that the FRSs misclassify the perturbed face images (adversarial examples) to be that of the victim.

While we defer a detailed discussion of related work to § II, we find that none of previous efforts perform an in depth study on the scope and effectiveness of such intensity-based impersonation attacks (referred to as impersonation attacks from hereon). In other words, there seems to be no answer yet to the question "Can an arbitrary attacker impersonate an arbitrary victim easily?" The key term here is *easily*. Specifically, if an attacker were able to add arbitrary amount of perturbations to her own image, she certainly could impersonate any victim. However, this would cause the attacker to be stand out, i.e., her actions could be perceived by observers as strange or even suspicious. Thus, the perturbation has to be imperceptible—the perturbation used must be small and inconspicuous. The question that is of interest therefore becomes "Can the perturbations be kept small in general settings?".

Towards answering this question, we undertake an in depth, systematic measurement study of the exploitability of DNN-based FRSs, using a very large scale dataset of about 2.6 million images. Our measurement study demonstrates that several factors influence the imperceptibility of impersonation attacks. We also find that it is more difficult to fool systems if the attacker has to account for the variability in her pose/orientation and other environmental conditions such as lighting, or use the perturbations generated from one DNN model to attack a different model. Based on the measurements, we suggest security countermeasures that could significantly

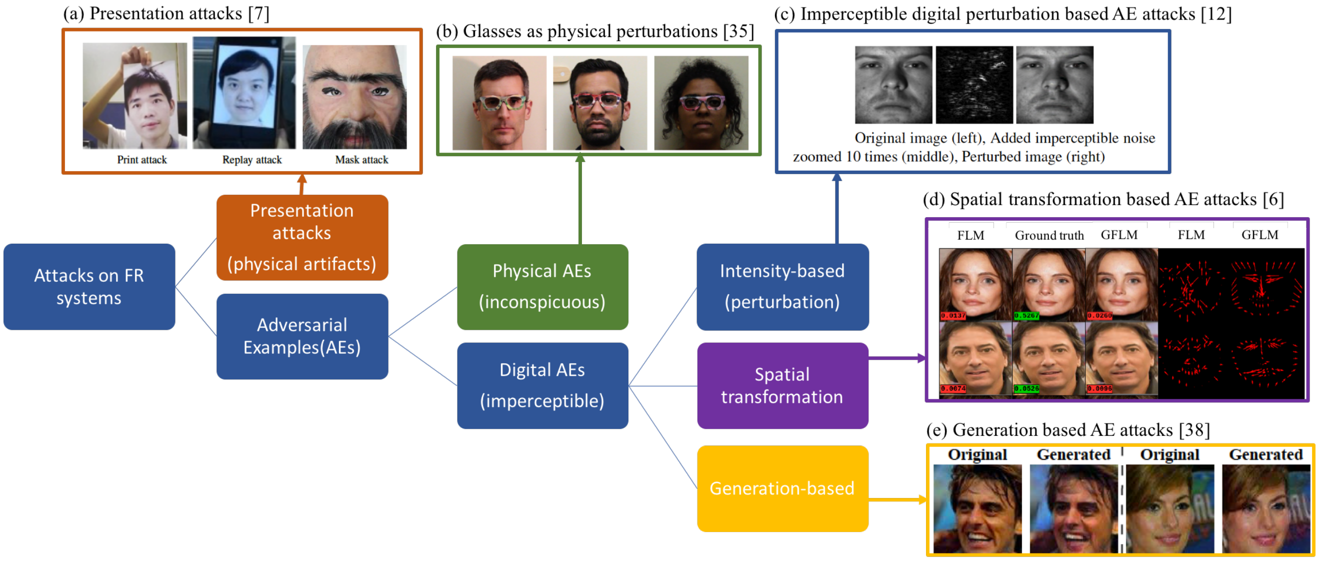

(a) Presentation attacks [7]
Print attack
Replay attack
Mask attack
(b) Glasses as physical perturbations [35]
(c) Imperceptible digital perturbation based AE attacks [12]
Original image (left), Added imperceptible noise zoomed 10 times (middle), Perturbed image (right)
(d) Spatial transformation based AE attacks [6]
FLM
Ground truth
GFLM
FLM
GFLM
(e) Generation based AE attacks [38]
Original
Generated
Original
Generated
Attacks on FR systems
Presentation attacks (physical artifacts)
Adversarial Examples(AEs)
Physical AEs (inconspicuous)
Digital AEs (imperceptible)
Intensity-based (perturbation)
Spatial transformation
Generation-based

Fig. 1: Various attacks on Face Recognition Systems. We focus on intensity-based AE attacks in our analysis since they are the kind of attacks explored the most in the literature. Intensity-based AE attacks are fast to carry out and are proven to have high attack success rates.

enhance the security of FR based IoT access control. In brief, our contributions in this paper are :

- We perform an extensive measurement study which shows that the efficacy/imperceptibility of impersonation attacks depends on several factors such as gender, skin color and age. We quantify the extent to which each of these factors affects the attack.
- We perform an in-depth measurement study to understand the feasibility of constructing universal perturbations that make the attack robust to different poses or facial orientations of the attacker. We find that this is much harder in practice from the attacker's perspective.
- We show that the use of multiple DNNs for performing FR (check faces across DNN models) can render imperceptible impersonation attacks almost infeasible.

II. Related Work

A. DNNs based FRSs

A lot of efforts have targeted the design of highly accurate FRSs. Traditional methods applied hand-crafted features like edges and texture descriptors [8], [20], [21], [29], which have been used for a long time. Due to the convenience of obtaining large training data and the availability of inexpensive computing power and memory, the trend towards replacing the traditional methods by deep learning methods is increasing. Deep Convolutional Neural Networks (DCNNs) can automatically extract high level representative features from large datasets and have been shown to be invariant to illumination variations, brightness variations, age variations and/or facial orientation [2]. Today the state-of-the-art FR algorithms are almost all based on end-to-end DCNNs [24], [30], [33], [39], [41]. We use VGG-Face [30] in our analysis. VGG-Face is a 39-layer DCNN, and is one of the most well-known and highly accurate face recognition systems.

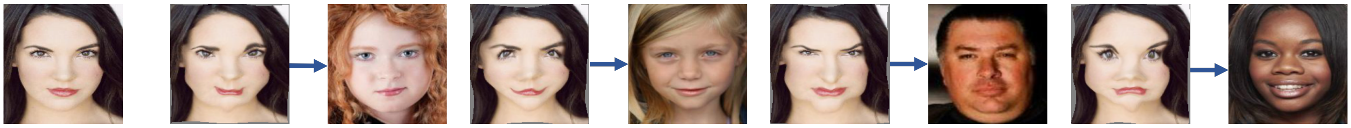
B. Presentation Attacks

It is generally believed that DNN-based FRSs have extremely high recognition accuracy, even better than humans. However, this is based on the implicit assumption that attackers do not actively attempt to fool the system. Recently however, there have been extensive efforts reported in the literature on attacks targeting FRSs [9]–[12], [44].

Many early approaches used by attackers to spoof a FRS, are based on using fake target faces, which is termed *presentation attack*. In general, attackers hold a non-real face of a target person in front of the camera to evade the FRS. The attackers could use photographs [1], [23], replayed videos [4], [47], dummy faces (such as 3D masks) [10], [17], or 3D virtual reality facial models displayed on a screen [44] as shown in Fig. 1(a). While these methods are shown to successfully lead to attacker misclassification as the target identities, such attacks, they however require the attacker to overtly indulge in action that may seem strange or even suspicious to nearby observers.

C. Adversarial Examples for FRSs

More recently, general DNN-based classifiers [22], [40], [48] have been shown to be vulnerable to adversarial example attacks. Adversarial Examples (AEs) refer to perturbed inputs, which are correctly classified by humans, but misclassified by machine learning systems. In [34], [35], the authors demonstrate the potential of using adversarial examples to conduct real face attacks on FRSs, i.e., the attackers use their own faces to mount attacks. By wearing special glasses (physical perturbations), the attacker's face can be misclassified by the DDN as shown in Fig. 1(b).



(a) original image **(b)** impersonation attack **(c)** impersonation attack **(d)** impersonation attack **(e)** impersonation attack

Fig. 2: Impersonation attacks using the Fast Landmark Manipulation (FLM) method proposed in [6]. (a) shows the original image; (b)-(e) show four impersonation attacks, within each the left image is the adversarial example and the target identity is shown in the right image.

In addition to physical AE attacks , various digital AE attack approaches have been proposed, which can be categorized into three kinds as follows.

- *Intensity-based AE attacks.* Imperceptible Perturbations are added to the images to change the intensity of each pixel as shown in Fig. 1(c). [13] hypothesizes that DNNs are vulnerable to AE attacks because of their linear nature and thus proposed the fast gradient sign method (FGSM) for efficiently generating perturbations. [19] extends the FGSM method by applying it multiple times with a small step size. [28] uses a norm minimization based formulation, termed DeepFool, to search for adversarial perturbations by casting it as an optimization problem. [3] introduces new gradient based attack algorithms that are more effective in terms of the adversarial success rates. We use [19] in our analysis since it can generate adversarial perturbations very fast, which is the key requirement for large-scale analysis (needed to generate these perturbations), and at the same time, it achieves very high attack success rates compared to other fast methods.
- *Spatial transformation based AE attacks.* As opposed to manipulating the pixel values, perturbations generated through spatial transformation could result in large L_p distance measures, but are perceptually realistic as shown in Fig. 1(d). [43] estimates the displacement field for all pixel locations in the input images. [6] first detects key landmarks of the faces and the displacement field is only defined for the key landmarks.
- *Generation-based AE attacks.* [38] utilize generative models to generate fake face images as shown in Fig. 1(e), which are visually similar to the original face images, thus hard to cause noticability; at the same time, these have similar feature representations as the target faces, and are thus recognized as the target individuals.

There are two different kinds of attack goals viz.:

- Dodging, where the attacker seeks to have one face misidentified as any other different face.
- Impersonation, where the attacker seeks to have one face classified as a specific target victim's face, which is harder than the dodging attacks.

While dodging attacks are of interest in evading surveillance, impersonation attacks, which are much more targeted, are of more relevance to IoT security. Attackers can leverage this method to gain unauthorized entry, for instance, by bypassing a smart locking mechanism. Our work thus focuses on impersonation attacks. The spatial transformation based attacks, that is, Fast Landmark Manipulation Method (FLM) and Grouped Fast Landmark Manipulation Methods (GFLM), are proposed for realizing dodging attacks. We extend these two methods to the impersonation attack. We observe that FLM gives largely deformed facial images as shown in Fig. 2, which is not imperceptible at all. GFLM, which aims to generate more natural adversarial examples, fails in all the four impersonation attacks. Therefore, it is evident that these types of attacks are not appropriate for impersonation and thus, we do not perform additional measurements on such spatial transformation based attack methods.

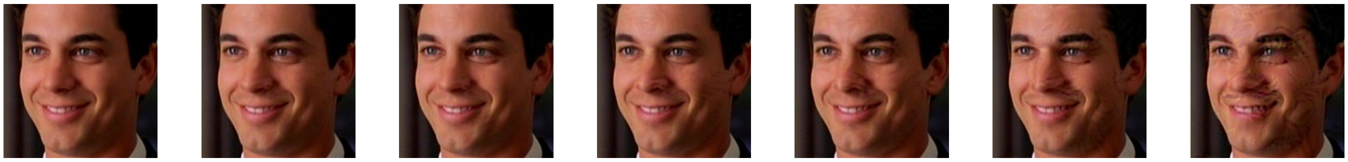
We focus on intensity-based AE attacks in our analysis since they are the kind of attacks explored the most in the literature. Intensity-based AE attacks are fast to carry out and have been proven to have extremely high attack success rates. Unlike prior works which simply showcase the possibility of such attacks, we do extensive measurements to provide a detailed view of the potency of such attacks in various scenarios and unearth various factors that affect this potency.

III. Imperceptible Impersonation Attack

To ensure that an impersonation attack is imperceptible (i.e., does not raise suspicion for human observers), the attackers should modify the faces such that visibility of the modifications is minimal. In this section, we describe the attack model and how the magnitude of the perturbation are meatured. The lower the magnitude of the perturbation, the higher the imperceptibility [34], [40].

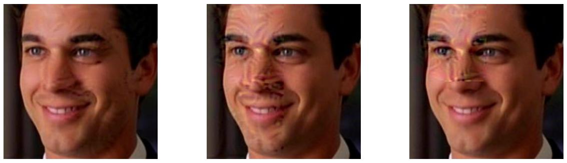
A. Attack Model

We assume that the attacker mounts the impersonation attack after the system has been trained. This implies that the adversary cannot "poison" the FRS by altering training data or by injecting mislabeled data. Rather, the adversary can only alter the composition of input images based on the knowledge of the underlying DNN model. Our attack model is consistent with IoT access control attack scenarios where the attacker cannot tamper with the manufacturing of the commercial smart devices. In this paper, we mainly focus on a white-box model in which the attacker knows the DNN architecture and the parameters of the FRSs being attacked. This is supported by the fact that it is possible to train local models that can infer the functionality of the target FRSs [37] and carry transfer attacks to the target FRSs. However, in Section IV-E, we also examine a black-box model by evaluating how well the perturbed images generated for one model can be successful in the impersonation attack on another model.



(a) $\sigma = 0$ **(b)** $\sigma = 0.8277$ **(c)** $\sigma = 1.0423$ **(d)** $\sigma = 1.5407$ **(e)** $\sigma = 2.0735$ **(f)** $\sigma = 2.7160$ **(g)** $\sigma = 5.00$

Fig. 3: Perturbed images for different levels of perturbation σ. In (g) with large value of $\sigma = 2.7160$, patterns are visible on the forehead, left cheek and nose. (The patterns are more visible in color version.)



(a) **(b)** **(c)**

Fig. 4: Perturbed images with restrictions on the location of pixels to be perturbed. In (a), all pixels are to be perturbed, $\sigma = 2.7160$. In (b), only left half of the image is allowed to be perturbed to achieve the same goal as (a). In (c), only top left quarter is allowed to be perturbed to achieve the same goal as (a).

B. Perturbation Vector

Suppose the finite set of people's identities (i.e., labels) to be detected by the FRS is $\mathcal{C}$, with $|\mathcal{C}| = N$. Further, suppose that each input image is given as an RGB vector $\mathbf{x}$ and the ground truth label of $\mathbf{x}$ is given by $c_x \in \{1, 2, \cdots, N\}$.

A DNN-based FRS implements a high-dimensional nonlinear function which maps an input $\mathbf{x}$ to an output probability vector $f(\mathbf{x})$ of length N, where each element in the output vector represents the probability that $\mathbf{x}$ matches the corresponding label. In addition, the label that corresponds to the largest entry in $f(\mathbf{x})$ is output as the recognition result. Consequently, a correct recognition result is realized when c_xth entry of $f(\mathbf{x})$ is the maximum entry. Thus, the ideal output $f(\cdot)$ is a one-hot vector, i.e., only the c_xth entry has value 1 and all the other entries are zero.

To impersonate a target c_t, the attacker with an input image vector $\mathbf{x}_a$ thus finds a perturbation vector $\mathbf{r}$ such that c_tth entry of $f(\mathbf{x}_a + \mathbf{r})$ is the maximum one. To measure the error in the output of the FRS with the adversarial input $\mathbf{x}_a + \mathbf{r}$, we adopt the *softmaxloss* score [30]. For an input vector $\mathbf{x}_a$ and a given label c_t, the *softmaxloss* function is defined as:

$$softmaxloss(f(\mathbf{x}_a), c_t) = -\log\left(\frac{e^{<h_{c_t}, f(\mathbf{x}_a)>}}{\sum_{c=1}^{N} e^{<h_c, f(\mathbf{x}_a)>}}\right), \quad (1)$$

where $< \cdot, \cdot >$ denotes inner product between two vectors and h_c is the one-hot vector corresponding to label c. Note that the value of *softmaxloss* score is low when the DNN outputs the label as c_t and high otherwise. The attacker's goal is to achieve a $softmaxloss(f(\mathbf{x}_a + \mathbf{r}), c_t)$ that is low enough such that c_xth entry of $f(\mathbf{x})$ is the maximum entry, while minimizing $||\mathbf{r}||$. In other words, the attacker solves the following optimization problem.

$$\mathbf{r}^* = \arg\min_{\mathbf{r}} softmaxloss(f(\mathbf{x}_a + \mathbf{r}), c_t) + \alpha||\mathbf{r}||. \quad (2)$$

In (2), α is weight factor used to balance impersonation error and imperceptibility. As discussed in § II, BIM algorithm [19] as shown in Algorithm 1 is used to solve this optimization problem.

Algorithm 1 Computing perturbation vector.

```
Input: image $\mathbf{x}_a$, target identity $c_t$
Output: impersonation perturbation $\mathbf{r}$
Initialize $\mathbf{r} \leftarrow \mathbf{0}$
while $\mathbf{x}_a + \mathbf{r}$ is not recognized as $c_t$ do
    $\Delta\mathbf{r}$ = argmin $softmaxloss(f(\mathbf{x}_a + \mathbf{r} + \Delta\mathbf{r}), c_t)$
    Quantize the additional perturbation: $\Delta\mathbf{r}' \leftarrow \Delta\mathbf{r}$
    Update the perturbation: $\mathbf{r} \leftarrow \mathbf{r} + \Delta\mathbf{r}'$
end while
```

C. Measuring Imperceptibility

Using (2), the attacker can always find perturbation vectors that allow desired misclassification of input vectors. However, the produced attack image, i.e., $\mathbf{x}_a + \mathbf{r}^*$ is not guaranteed to be "imperceptible" to humans. In other words, the perturbation vector could be too large. This would cause the produced perturbed image to be quite distinguishable from the original attacker image. To quantify the effectiveness of the attack in various settings, we measure per pixel per color channel magnitude of perturbation using the root mean square error (RMSE) between the original and perturbed images. In particular, suppose that the images are of width W, height H and number of color channels D. Let the total number of dimensions in an image vector be $M = W \times H \times D$. Given an input and perturbed image vectors $\mathbf{x}, \mathbf{x}' \in \{0, 1, \cdots, 255\}^M$, the RMSE (we also use the term "noise level") is given by the following.

$$\sigma(\mathbf{x}, \mathbf{x}') = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(x(i) - x'(i))^2}, \quad (3)$$

where $x(i)$ is the ith component of $\mathbf{x}$, and σ is in pixel-value units, where $\sigma \in [0, 255]$.

To get a sense of what values of σ renders a perturbed attack image easy to identify, we show images of an attacker with varying levels of perturbation in Fig. 3. We note that for

$\sigma > 2$, it is easy to identify the noisy pixels in the perturbed images.

D. Physical Imperceptibility

If the attackers want to realize this perturbation physically (via using various paraphernalia such as dummy faces, or 3D-printed glasses), the amount of perturbation will need to be limited in terms of either (a) the maximum number of pixels to which the noise is added, or (b) the locations of those pixels [34], or (c) both. In Fig. 4, we study the effects of such limitations. We fix the attacker image and a target label, and then find the adversarial images when the entire image can be perturbed, as well as when only the left half and top left quarters of the image pixels are to be perturbed. As shown in the figure, the noise level increases significantly and the pattern is perceptible. Thus, one can expect the attack to be much harder in these cases. In the rest of the paper, we only consider scenarios in which the full attacker image is subject to perturbation. This reflects a worst case scenario analysis from the defender's perspective. Even in this scenario, we show that it can be hard for an attacker to launch the attack in all possible scenarios.

IV. Experiments

In this section, we detail the results of our measurement study towards getting an in depth understanding of the practicality of imperceptible impersonation attacks on DNN-based FRS and the factors that influence such attacks.

A. Experimental Setup

The FRS used in our experiments is VGG-Face [30], one of the most well-known and highly accurate face recognition systems as discussed in § II. The analysis is based on the VGG-Face dataset [30], which contains $N = 2622$ identities of celebrities, and approximately 1000 facial images per identity; this translates to a total of about 2.6 million images.

B. Case studies

To begin with, we use the face image of Micheal Crichton as the attacking image,(i.e., the input) and study whether some targeted individuals are harder than others to impersonate with the attacking image. Fig. 5 shows the minimum perturbations needed for the attacking image to impersonate three different individuals. We observe that it is rather easy for Micheal Crichton to impersonate A.J. Buckley. However, when it comes to impersonating Boris Kodjoe, the perturbation gets larger and is noticeable by human.

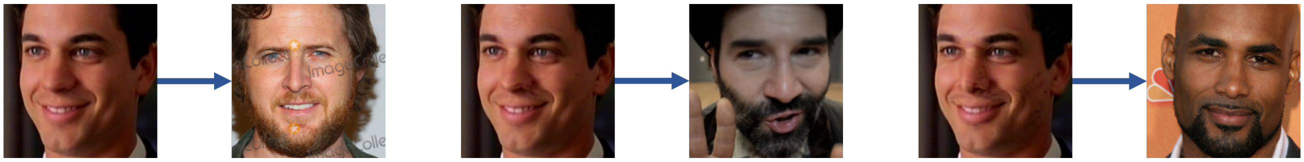
For a more general case study, Fig. 6 shows the noise level σ needed to achieve a successful attack by each attacker depicted on the column, to impersonate each target depicted on the row. It is clear that, different attackers need different values of σ to successfully impersonate different targets. Interestingly, the patterns of large perturbations (marked in red) seen in Fig. 6 suggest that it is easier for the considered attackers (e.g., who are all male with pale skin color) to impersonate targets who are also male with pale skin color, as compared to impersonating other targets. In addition, the noise levels needed to impersonate target 1 are all large, which is possibly due to a difference in gender. Furthermore, we see that impersonating targets 6-10 seems to require larger noise levels. This can be attributed to differences in skin color, age, or a combination of both. This motivates our study to further examine the impact of these factors in Section § IV-C.

Having performed the above preliminary studies, we next look at the statistical distribution of the ability of an attacker to impersonate different targets, subject to a constraint on the noise level $\sigma \leq \bar{\sigma}$. We define the attack success rate $\eta(\bar{\sigma})$ as the percentage of target labels which an attacker can impersonate for a given $\bar{\sigma}$. In Fig. 7, we show the success rates η for three different attackers impersonating all other remaining labels in the VGG-Face dataset. One can see that Abbie Cornish (female, white, young) can more successfully impersonate others, on average, compared to A.R. Rahman (male, Indian, young) and Aaron Yoo (male, Asian, young). For example, with the threshold $\bar{\sigma} = 2$, Abbie Cornish can successfully impersonate 58% of all the labels while A.R. Rahman achieves a success rate of only 6.5% and Aaron Yoo achieves a success rate of 17.7%. This could be attributed to the fact that the VGG-Face dataset contains more white people than people of other races. We observe that the gender distribution is almost balanced in the dataset.

To get aggregate results, we randomly sample the VGG-Face dataset to get a 100-identity subset $\mathcal{S}$. We fix each identity in $\mathcal{S}$ as a specific attacker, and then find the perturbation vector with each of the remaining labels in $\mathcal{S}$ as targets, and we compute $\eta(\bar{\sigma})$ for each attacker for a range of values of $\bar{\sigma}$. We then repeat this experiment 10 times and compute the average attack success rate across attackers in all the samples. The results show that, on average, it is not easy for an attacker to impersonate *any* target identity. In particular, with $\bar{\sigma} = 1.5$, the success rate is only 10.4%. We take a deeper look into how the success rate breaks down within different groups of people in the following section.

C. Factors that influence the attack

Next, we take a closer look at the extent to which various factors, discussed in § IV-B, influence an attacker's ability to carry out an imperceptible impersonation attack. Specifically, we consider different groups of identities based on gender, skin color, and age attributes. We manually label the dataset to produce four groups: (a) white young male (100 identities), (b) white young female (100 identities), (c) black young male (69 identities), and (d) white old male (100 identities). We do not consider other attribute combinations, such as black young female, or white old female, because the majority of the images in the VGG-Face dataset are for white skin color, and young people. For group (c), we only have 69 identities due to limitedness of data points matching such attributes. To reduce errors in labeling, each of the authors of the paper manually labeled the dataset independently and we considered only the images with unanimously common labels in our group dataset.



(a) Impersonating A.J. Buckley, $\sigma = 0.88$ **(b)** Impersonating Adam Buxton, $\sigma = 1.65$ **(c)** Impersonating Boris Kodjoe, $\sigma = 2.26$

Fig. 5: Different noise levels needed for Micheal Crichton to impersonate three different identities. It is rather easy for Micheal Crichton to impersonate A.J. Buckley; and hard to impersonate Boris Kodjoe

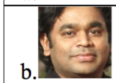
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
a.	2.55	2.28	1.26	1.23	1.89	1.95	3.20	3.15	3.17	1.57
b. 	2.12	1.87	2.00	2.45	1.70	2.04	2.38	1.98	2.39	2.65
c.	2.14	1.30	1.56	1.39	1.63	2.61	2.27	5.96	3.03	1.35
d.	2.75	1.99	1.78	1.97	1.54	1.81	2.09	1.78	1.66	1.98

Fig. 6: Noise level σ required for an attacker (a-d) to impersonate a target (1-10). It is easier for the considered attackers (who are all male with pale skin color) to impersonate targets who are also male with pale skin color, as compared to impersonating other targets. The noise levels needed to impersonate target 1 are all large. Impersonating targets 6-10 seems to require larger noise levels.

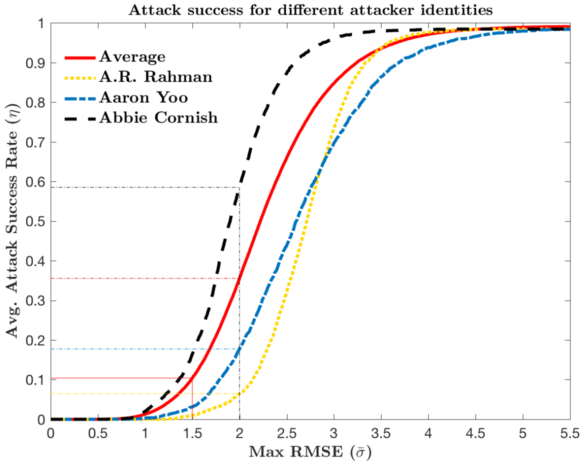


Fig. 7: Impersonation attack performance. Abbie Cornish (female, white, young) can more successfully impersonate others, on average, compared to A.R. Rahman (male, Indian, young) and Aaron Yoo (male, Asian, young).

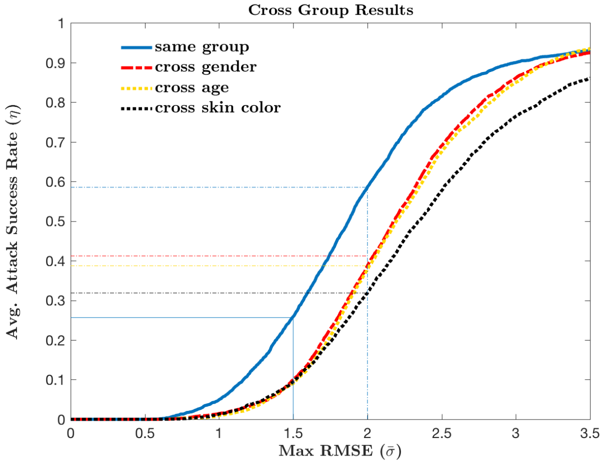


Fig. 8: Cross group impersonation attack performance. It is easier for an attacker to impersonate a target identity having the same attributes (gender, skin color, age). Impersonation across different skin color is the most hardest.

In addition, when labeling, we discard an identity whenever its attributes are hard to label manually.

To investigate the impact of the aforementioned attributes on the imperceptible impersonation attack, we conduct four experiments based on the four groups:

- Take people in group (a) as attackers trying to impersonate the other people in group (a); this case reflects *same group* impersonation measurements;
- take people in group (a) as attackers trying to impersonate people in group (b); this case represents *cross gender* impersonation measurements;
- take people in group (a) as attackers trying to impersonate people in group (c); this case reflects *cross skin color*

impersonation measurements;

- take people in group (a) as attackers trying to impersonate people in group (d); this case counts for *cross age* impersonation measurements.

In Fig. 8, we plot the average attack success rate versus different perturbation constraint $\bar{\sigma}$, for each of the four aforementioned experiments. We note that it is easier for an attacker to impersonate a target identity having the same attributes (gender, skin color, age). For the same group experiment, with the threshold $\tilde{\sigma} = 1.5$, the success rate is 25.65%. Recall that the aggregate success rate in § IV-B is only 10.4%. Moreover, as shown in the figure, it is relatively easier for an attacker to impersonate a target with a different gender or age than to impersonate a target with different skin color. For example, with the threshold $\tilde{\sigma} = 2$, the success rate for cross skin color is only 31.85% while the success rate for cross age and gender are around 40%. These results seem consistent with (and can be explained by) observations that have been previously reported in computer vision literature [5], [16], [36], [42]. Specifically, these papers show that in several scenarios, shape and texture cues suffer from degradation (affecting age or gender) and the color feature becomes dominant [46]. Thus, we conclude that the VGG-Face model relies less on features such as shape and texture as compared to color.

D. Universal Perturbation Results

In a realistic setting, an attacker may want one universal perturbation to impersonate the target identity for all the face images captured in different settings such as pose, camera angle, and lighting conditions. In order to launch the impersonation attack in the presence of these variations, an attacker will need to find a single perturbation vector $\tilde{\mathbf{r}}$ that allows misclassification of a set of his/her own images $\mathcal{X}_a$ of size K, to the target victim label, thus accounting for as many conditions as possible. In other words, the attacker needs to construct a vector $\tilde{\mathbf{r}}$ such that $f(\mathbf{x}_a + \tilde{\mathbf{r}}) = c_t$, $\forall \mathbf{x}_a \in \mathcal{X}_a$ for some given target label c_t.

The approach for calculating $\tilde{\mathbf{r}}$ is similar to the one described in § III-B. The only difference is that now the objective function changes to the following.

$$\tilde{\mathbf{r}} = \arg\min_{\mathbf{r}} \sum_{\mathbf{x}_a \in \mathcal{X}_a} softmaxloss(f(\mathbf{x}_a + \mathbf{r}), c_t) + \alpha ||\mathbf{r}||. \tag{4}$$

The condition to stop the iterations now requires that all K images to be misclassified as the target label, upon adding on the same perturbation vector.

In Fig. 9, we show an example of an attacker with three images, $\mathcal{X}_a = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$. In Fig. 10, we show the output perturbed image $\mathbf{x}_1 + \tilde{\mathbf{r}}$ when $\tilde{\mathbf{r}}$ is computed using only image $\{\mathbf{x}_1\}$, images $\{\mathbf{x}_1, \mathbf{x}_2\}$, and all the images $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, respectively. It is evident that the attacker image is more perceptible as more attack images are considered in computing the universal perturbation vector $\tilde{\mathbf{r}}$.

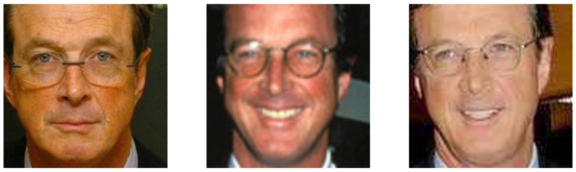


Fig. 9: A set of face images of Micheal Crichton. $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$

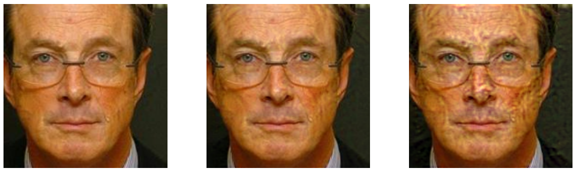


Fig. 10: Universal perturbations visualization. Three perturbations are universal to different number of attacking images. Left: universal for $\{\mathbf{x}_1\}$, $\sigma = 1.7509$; Middle: universal for $\{\mathbf{x}_1, \mathbf{x}_2\}$, $\sigma = 3.6027$; Right: universal for $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $\sigma = 7.8877$. The perturbations are more perceptible as more attacking images are considered in computing the universal perturbation vector.

In Fig.11, we plot the average success rate for an attacker employing universal perturbation. Here, we randomly sample 100 identities from the VGG-Face dataset and let them impersonate each other. We conduct this experiment 10 times and average the results. The results show that the success rate is strictly decreasing when a universal perturbation vector is required to perturb multiple attacker images. More importantly, the attackers' ability to impersonate a given target is significantly reduced with even slight increases in K. For example, the success rate with threshold $\tilde{\sigma} = 2$ is 39.9% for $K = 1$ (the case considered in § IV-B and § IV-C). However, when we increase K to 2, the success rate drops dramatically to 2.28% and the success rate when $K = 3$ is only 0.6%, which suggests that the impersonations can almost fail all the time, if the attacker seeks to be imperceptible.

E. Cross Model Measurements

Recently, it has been shown that adversary examples that are successfully misclassified by one trained DNN model can also cause misclassifications in other (different) DNN models that have different hyperparameters [27], [40]. However, it is unclear whether different models could misclassify the perturbed images to the same target classes, which is the key characteristic for determining if white-box impersonation attacks can easily extend to black-box attacks. To check whether our perturbed images targeting impersonation generalize across different DNN models, we fine-tune the AlexNet DNN [18] on the VGG-Face dataset, and test the impersonation attack success ratio on the AlexNet model using the perturbed images generated using VGG-Face model.

We test $10,000$ perturbed images generated by VGG-Face, and find that *none* are classified as the victims by the AlexNet, but most of them are misclassified by the AlexNet. This significant result indicates that impersonation attacks do not easily transfer across different DNN models. It will be extremely hard for the attacker to use the perturbation vectors to fool

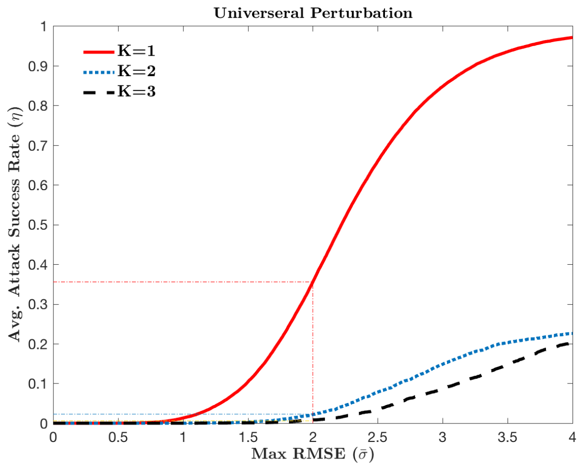


Fig. 11: Universal perturbation impersonation attack performance. K is the number of attacking images to generate the universal perturbations. The attackers' ability to impersonate given targets is significantly reduced when the perturbations are required to be universal to multiple attacking images.

a DNN model different from the one used to generate them. Thus cross model validation could significantly enhance the robustness of face recognition based access control in IoT systems.

F. Detecting and removing perturbations

Finally, we test whether de-noising [31] (which could be used by an IoT access control system) affects the potency of the attack. Three standard de-noising filters are considered in our experiments: average filter, median filter, and Wiener filter. We test 100 different perturbed images, and find that all of them are still misclassified as the targets. This suggests that de-noising does not hurt the attack. This is because de-noising filters assume a certain pattern of noise, which is unlikely to be what is used by the attacker for generating the perturbations.

We conclude that traditional noise detection and de-noising algorithms are not helpful in countering the imperceptible impersonation attack since the perturbation generated is structured.

G. Summary of results

Below is a summary of our take-aways based on the results in § IV-B to § IV-F. **(a)** DNNs are vulnerable to adversary examples. However, in contrast to recent work in the literature, we find that the average success rates of the imperceptible impersonation attack are low. **(b)** Attackers can achieve better success rates by choosing targets with similar attributes; in particular choosing targets with same skin color helps. **(c)** When variations, such as pose, camera angle and lighting conditions are considered, the attack is significantly less successful. **(d)** Perturbed images do not generalize well across different DNN models. **(e)** Current noise estimation and de-noising methods do not adversely impact the imperceptible impersonation attack.

V. Conclusion

The security of face recognition is an important toptic as face recognition is more and more used in IoT access control. In this paper, we perform an in-depth measurement study of the generality and efficacy of imperceptible impersonation attacks that have recently gained popularity. Our study is done using a very large dataset. We find that it is hard for a given adversary to impersonate an arbitrary target victim without making perceptible changes to her face. Further, we show that several factors such as age, race and gender of the attacker and victim influence the efficacy of the attack and we quantify the impact of each. We also show that, in a realistic scenario where the attacker seeks to be robust to different poses or variations in environmental conditions, the attack becomes more difficult or even impossible. Based on this, we suggest the use of cross-model verifications as well as multi-views, which can potentially counter such attacks very effectively.

Acknowledgments

This research was partially sponsored by the U.S. Army Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Anjos, A., and Marcel, S. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *2011 IEEE international joint conference on Biometrics (IJCB)* (2011), pp. 1–7.

[2] Arsenovic, M., Sladojevic, S., Anderla, A., and Stefanovic, D. Facetime–deep learning based face recognition attendance system. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)* (2017), IEEE, pp. 000053–000058.

[3] Carlini, N., and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)* (2017), IEEE, pp. 39–57.

[4] Chingovska, I., Anjos, A., and Marcel, S. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)* (2012), IEEE, pp. 1–7.

[5] Choi, J. Y., Ro, Y. M., and Plataniotis, K. N. Color face recognition for degraded face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39*, 5 (2009), 1217–1230.

[6] Dabouei, A., Soleymani, S., Dawson, J., and Nasrabadi, N. Fast geometrically-perturbed adversarial faces. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), IEEE, pp. 1979–1988.

[7] Deb, D., Zhang, J., and Jain, A. K. Advfaces: Adversarial face synthesis. *arXiv preprint arXiv:1908.05008* (2019).

[8] Ding, C., and Tao, D. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST) 7*, 3 (2016), 1–42.

[9] Duc, N. M., and Minh, B. Q. Your face is not your password face authentication bypassing lenovo–asus–toshiba. *Black Hat Briefings* (2009).

[10] Erdogmus, N., and Marcel, S. Spoofing face recognition with 3d masks. *IEEE Transactions on Information Forensics and Security 9*, 7 (2014), 1084–1097.

[11] Findling, R. D., and Mayrhofer, R. Towards face unlock: on the difficulty of reliably detecting faces on mobile phones. In *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia* (2012), ACM, pp. 275–280.

[12] Goel, A., Singh, A., Agarwal, A., Vatsa, M., and Singh, R. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (2018), IEEE, pp. 1–7.

[13] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[14] Goswami, G., Ratha, N., Agarwal, A., Singh, R., and Vatsa, M. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[15] Ibrahim, R., and Zin, Z. M. Study of automated face recognition system for office door access control application. In *2011 IEEE International Conference on Communication Software and Networks (ICCSN)* (2011), IEEE, pp. 132–136.

[16] Karimi, B. *Comparative analysis of face recognition algorithms and investigation on the significance of color*. PhD thesis, Concordia University, 2006.

[17] Kose, N., and Dugelay, J.-L. On the vulnerability of face recognition systems to spoofing mask attacks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), IEEE, pp. 2357–2361.

[18] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.

[19] Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).

[20] Li, J., Li, S., Hu, J., and Deng, W. Adaptive lpq: An efficient descriptor for blurred face recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2015), vol. 1, IEEE, pp. 1–6.

[21] Li, S., and Deng, W. Face recognition based on random feature. In *2015 Visual Communications and Image Processing (VCIP)* (2015), IEEE, pp. 1–4.

[22] Li, S., Neupane, A., Paul, S., Song, C., Krishnamurthy, S. V., Roy-Chowdhury, A. K., and Swami, A. Stealthy adversarial perturbations against real-time video classification systems. In *NDSS* (2019).

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

[24] Liu, J., Deng, Y., Bai, T., Wei, Z., and Huang, C. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310* (2015).

[25] Manjunatha, R., and Nagaraja, R. Home security system and door access control based on face recognition. *International Research Journal of Engineering and Technology (IRJET)* (2017).

[26] Market, B. Industry report 2009-2014. *International Biometric Group* (2008).

[27] Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. *arXiv preprint arXiv:1610.08401* (2016).

[28] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2016), pp. 2574–2582.

[29] Park, U., Tong, Y., and Jain, A. K. Age-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 5 (2010), 947–954.

[30] Parkhi, O. M., Vedaldi, A., and Zisserman, A. Deep face recognition. In *BMVC* (2015), vol. 1, p. 6.

[31] Patidar, P., Gupta, M., Srivastava, S., and Nagawat, A. K. Image de-noising by various filters for different noise. *International Journal of Computer Applications 9*, 4 (2010).

[32] Pentland, A., and Choudhury, T. Face recognition for smart environments. *Computer 33*, 2 (2000), 50–55.

[33] Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 815–823.

[34] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. 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), ACM, pp. 1528–1540.

[35] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS) 22*, 3 (2019), 1–30.

[36] Shih, P., and Liu, C. Comparative assessment of content-based face image retrieval in different color spaces. *International Journal of Pattern Recognition and Artificial Intelligence 19*, 07 (2005), 873–893.

[37] Shokri, R., Stronati, M., and Shmatikov, V. Membership inference attacks against machine learning models. *arXiv preprint arXiv:1610.05820* (2016).

[38] Song, Q., Wu, Y., and Yang, L. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. *arXiv preprint arXiv:1811.12026* (2018).

[39] Sun, Y., Wang, X., and Tang, X. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 2892–2900.

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

[41] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1701–1708.

[42] Torres, L., Reutter, J.-Y., and Lorente, L. The importance of the color information in face recognition. In *In 1999 International Conference on Image Processing (ICIP)* (1999), vol. 3, IEEE, pp. 627–631.

[43] Xiao, C., Zhu, J. Y., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018* (2018).

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

[45] Yang, J.-C., Lai, C.-L., Sheu, H.-T., and Chen, J.-J. An intelligent automated door control system based on a smart camera. *Sensors 13*, 5 (2013), 5923–5936.

[46] Yip, A., and Sinha, P. Role of color in face recognition. *Journal of Vision 2*, 7 (2002), 596–596.

[47] Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., and Li, S. Z. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)* (2012), IEEE, pp. 26–31.

[48] Zhu, S., Wang, Z., Chen, X., Li, S., Iqbal, U., Qian, Z., Chan, K. S., Krishnamurthy, S. V., and Shafiq, Z. A4: Evading learning-based adblockers. *arXiv preprint arXiv:2001.10999* (2020).

[49] Zuo, F., and de With, P. Real-time embedded face recognition for smart home. *IEEE Transactions on Consumer Electronics 51*, 1 (2005), 183–190.