# Accessorize in the Dark: A Security Analysis of Near-Infrared Face Recognition

Amit Cohen and Mahmood Sharif*

Tel Aviv University

**Abstract.** Prior work showed that face-recognition systems ingesting RGB images captured via visible-light (VIS) cameras are susceptible to real-world evasion attacks. Face-recognition systems in near-infrared (NIR) are widely deployed for critical tasks (e.g., access control), and are hypothesized to be more secure due to the lower variability and dimensionality of NIR images compared to VIS ones. However, the actual robustness of NIR-based face recognition remains unknown. This work puts the hypothesis to the test by offering attacks well-suited for NIR-based face recognition and adapting them to facilitate physical realizability. The outcome of the attack is an adversarial accessory the adversary can wear to mislead NIR-based face-recognition systems. We tested the attack against six models, both defended and undefended, with varied numbers of subjects in the digital and physical domains. We found that face recognition in NIR is highly susceptible to real-world attacks. For example, $\geq$96.66% of physically realized attack attempts seeking arbitrary misclassification succeeded, including against defended models. Overall, our work highlights the need to defend NIR-based face recognition, especially when deployed in high-stakes domains.

## 1 Introduction

Face-recognition technology has become increasingly popular in recent years, with applications ranging from border security [7] and surveillance [41] to access control [1,2]. Among others, face recognition based on near infrared (NIR) imaging has received wide adoption (e.g., [1,2]) due to its near-invariance to changes in ambient illumination and its ability to capture facial features in dark environments [17]. Because such NIR-based face-recognition systems are deployed to address security-critical problems, it is crucial to analyze their integrity against adversaries seeking to mislead them (e.g., to circumvent surveillance or receive unauthorized access).

Recent work in adversarial machine learning (ML) has demonstrated that ML models in general, and ones for face-recognition in particular, are vulnerable to evasion attacks at deployment time (e.g., [20,32,33,36,37]). Specifically, adversaries generating so-called adversarial examples—minimally but strategically modified variants of benign inputs—can lead ML models to misclassify.

* Corresponding author (e-mail: mahmoods@tauex.tau.ac.il).

These adversarial examples can also be realized in the problem space to mislead systems [29]. For example, adversaries can physically realize and wear accessories such as eyeglasses to impersonate others against face recognition in visible light (VIS) [32,33]. Still, prior work demonstrating evasion attacks against image classification chiefly focused on systems relying on VIS sensors, and the susceptibility of NIR-based face recognition to evasion attacks has yet to be determined. Indeed, because NIR images vary less under changes in imaging conditions [17] and have lower dimensionality (shown to be correlated with susceptibility to attacks [31]) than VIS images, it is plausible that NIR-based face-recognition systems could be less susceptible to evasion than their VIS counterparts.

Our work fills the gap by developing and evaluating attacks against state-of-the-art NIR-based face-recognition models, enabling us to determine whether and to what extent these systems are vulnerable to evasion attacks in the digital and physical domains. We design attacks that enable adversaries to mislead NIR-based face recognition according to different attack objectives (namely, dodging to attain arbitrary misclassifications or impersonation), and further extend them to facilitate realizing adversarial examples in the physical world. For example, among others, we ensure attacks are robust to real-world transformations, such as changes in pose and camera sampling noise. The attacks result in accessories (namely, eyeglasses) adversaries can wear to mislead face recognition.

We extensively tested attacks against six state-of-the-art NIR-based face-recognition models in the digital and physical domains. Our experiments involved varied numbers of subjects, and both undefended and defended [42] models. We found that the models were highly vulnerable to evasion, with a mean of 98.33% of dodging attempts and 77.77% of impersonation attempts succeeding in the physical domain. The defense hindered impersonation attacks to some extent (36.66% mean attack success rate), but was still vulnerable to dodging (96.66% mean attack success rate). Overall, our work highlights that NIR-based face recognition-systems are not inherently more robust than their VIS counterparts, and that defenses to advance their integrity in adversarial settings are crucial.

The paper is structured as follows. Next, we present necessary background (§2) and the threat model (§3). Then, we introduce our methodology (§4), followed by an evaluation of NIR-based face recognition's robustness (§5). Lastly, we close the paper by discussing its limitations (§6) and concluding (§7).

## 2 Background and Related Work

### 2.1 Face Recognition in NIR

NIR is a portion of the electromagnetic spectrum falling between visible light and mid-infrared, with wavelengths ranging between ∼700 and ∼2500 nm. As NIR light can penetrate certain material, such as clothing and wood, it is particularly useful for imaging such objects. Consequently, NIR is commonly used in various applications, ranging from gaze detection in challenging conditions [24] to the analysis of food and agricultural products [25]. In the biometrics field,

NIR has been used for face recognition, including in widely deployed commercial systems (e.g., [1, 2]), due to its ability to capture facial features that may not be otherwise visible. Notably, NIR cameras can capture images in low-light conditions, rendering them useful for settings with limited (VIS) illumination, such as surveillance and biometric authentication in the dark.

Leading NIR-based face-recognition systems rely on deep learning [9, 10, 12, 14, 15, 23, 43, 44, 49]. For example, Lezama et al. presented a deep-learning-based face-recognition system to identify individuals based on their NIR facial images [18]. They attained high recognition performance by leveraging generative models mapping NIR to VIS and an off-the-shelf feature-extraction network as a backbone, and tuning the representations using generated images. Later on, Wu et al. presented a deep convolutional neural network, named LightCNN, designed to be light-weight and effective on multiple tasks [43]. Among others, LightCNN achieves high accuracy on NIR-based face recognition. In a follow-up work, Fu et al. proposed an LightCNN variant, LightCNN-DVG [12], achieving the highest face-recognition accuracy in NIR to date. The primary difference between LightCNN and LightCNN-DVG is that the latter is fine-tuned with NIR-VIS data. During fine-tuning, LightCNN-DVG was trained to map NIR and VIS image pairs of the same person (resp., different people) into feature vectors that are close together (resp., further away) in the feature space. We evaluate our proposed attack against a representative set of such top performing models.

### 2.2 Attacking ML

Attacks against ML models can be categorized based on attacker objectives and capabilities. Several attack types against ML have been proposed, including, but not limited to, training-time attacks, where adversaries partially control the training data or process to harm model performance (e.g., [5, 16]); privacy attacks, where adversaries aims to extract sensitive information about the training data from access to the model or training process (e.g., [34]); and availability attacks, where attackers seek to craft inputs that increase prediction or training latency (e.g., [35]). By contrast, our work studies evasion attacks in which adversaries have no control over the trained model but can manipulate inputs at inference time to induce misclassifications (e.g., [20, 36]).

Evasion attacks were first popularized by Biggio et al. [4] and Szegedy et al. [36], who demonstrated the vulnerability of ML models to small perturbations of their inputs. Since then, evasion attacks have been studied extensively, with numerous techniques proposed for generating and defending against adversarial examples (e.g., [6, 13, 20, 27]). Formally, evasion attacks seek to find a solution to some variation of the following optimization problem:

$$\arg\max_{\delta} L\big(f(x+\delta), c_x\big)$$

where $f$ is an ML model, $x$ is the input, $\delta$ is an adversarial perturbation, $c_x$ is the input's class (i.e., label), and $L$ is the loss function. Often, the optimization is constrained by requiring that $\delta$'s $\ell_p$-norm is bounded by a constant $\epsilon$,

(i.e., $||\delta||_p = \left(\sum_i |\delta_i|^p\right)^{1/p} \leq \epsilon$, commonly for $p \in \{2, \infty\}$). By solving the optimization, attacks aim to find perturbations increasing the loss, leading $f$ to misclassify. Several first-order (i.e., gradient-based) optimization methods have been proposed to solve the optimization (e.g., [6, 13]), many of which are slight variants of the popular Projected Gradient Descent (PGD) attack [20]. Given a model and a sample input, PGD generates adversarial perturbations in an iterative manner—it calculates the input gradients w.r.t. the loss, and updates the input in the direction maximizing the model's error. More formally, PGD computes the perturbed sample $x^{t+1}$ at iteration $t+1$ by:

$$x^{t+1} = \Pi_S(x^t + \alpha \ sign(\nabla_{x^t} L(f(x^t), c_x)))$$

where $\alpha$ is the step size, and $\Pi_S$ projects samples into a set of allowed perturbations $S$ (e.g., $\epsilon$-ball around $x$), and $x^0$ is set to $x$ or randomly initialized within $S$. The attack we design (§4) is a variant of PGD in which $\delta$'s max-norm is unbounded, but the perturbation can be applied to a specific region in the image covered by an accessory, as defined by a mask.

Attackers performing evasion can vary in their capabilities. In *white-box* settings, attackers have full access to the model parameters and architectures, allowing them to design powerful attacks using their knowledge about the model (e.g. gradients [20]). By contrast, in *black-box* settings, attackers have no access to the model internals, and may only query models [27]. Thus, intuitively, black-box attacks are more challenging than white-box attacks. Attacker goals may also vary. An attacker may aim to produce any misclassification—i.e., conduct an *untargeted* attack—or induce a misclassification to a particular class—i.e., perform a *targeted* attack [28]. Intuitively, targeted attacks impose more constraints and are hence more challenging.

Early evasion attacks primarily explored adversarial perturbations constrained in $\ell_p$-norms. While in those settings the adversarial sample is close to the original benign example, $\ell_p$-norm-bounded attacks are challenging to realize in the problem space (i.e., as an artifact whose corresponding features are misclassified by a model) [29]. By contrast, realizable attacks incorporate domain constraints to produce problem-space artifacts that lead to evasion (e.g., [3, 11, 29, 30, 32].) For example, Sharif et al. showed how to produce eyeglass frames that adversaries can don to evade VIS-based face recognition [32]. The attack was effective under real-world circumstances, allowing adversaries to mislead recognition by wearing eyeglass frames with specific color patterns. Differently than Sharif et al., we develop attacks suited for NIR-based face recognition.

### 2.3 Defending ML

Defending models' integrity against evasion attacks is crucial for ensuring safe and secure deployment. Adversarial training—the process of augmenting the training data with correctly labeled adversarial examples—is one of the most effective techniques for enhancing model robustness (e.g., [13, 20]). Other defenses offer methods to detect adversarial inputs (e.g., [22]), sanitize adversarial perturbations (e.g., [48]), and certify robustness within certain regions (e.g., [8]).

Researchers have also published defenses against patch-based attacks [45–47], however these are either limited to models with small receptive fields or significantly increase inference time. Wu et al. presented a defense method called Defense against Occlusion Attacks (DOA) to defend against physically realizable attacks in the image domain [42]. They suggest adversarially training models with an abstract adversary perturbing a rectangular patch and show this enhances robustness against adversaries using eyeglasses to evade face recognition and ones producing stickers to evade traffic-sign recognition. We evaluate our attack against a model defended via DOA.

## 3 Threat Model

In this paper, we primarily study white-box evasion attacks against NIR-based face-recognition models. Studying white-box attacks is critical, as *(1)* it can help us assess systems' vulnerability when relying on publicly available models (e.g., [43]) or when proprietary models are stolen [38]; *(2)* they help assess the effectiveness of defenses against worst-case adversaries with complete knowledge of the system, and inform means to enhance them; and *(3)* these attacks serve as the basis for black-box attacks using queries to estimate gradients [27] or via transferability [26]. Indeed, we attempt to transfer attacks created against surrogate models to target models, thus simulating black-box attacks, and find that evasion attempts often transfer between NIR-based face-recognition models. We implement both untargeted (dodging) and targeted (impersonation) attacks, and test them both in digital and physical domains against state-of-the-art models.

To maintain stealth and plausible deniability, we consider attacks using everyday accessories (mainly eyeglasses, but also face masks and stickers), in line with prior work [32, 37]. By using accessories, the adversary aims to remain inconspicuous and avoid raising suspicion by observers. Additionally, we aim for the attacks to be (physically) realizable, such that adversaries would be able to mislead the system by slightly changing their own appearance, without altering their surroundings or manipulating the digital representation of their image.

## 4 Methodology

We now present our attacks against NIR-based face recognition, starting with how to evade models before describing how to enable physical realizability.

### 4.1 Evading Recognizers

The face-recognition systems we study classify NIR face images by finding the most similar VIS image from within an image gallery. The process is enabled by neural networks that extract feature vectors of both NIR and VIS images. For classification, the systems compute the cosine similarity $cos(\cdot, \cdot)$ between the NIR features and each of gallery images' features. Eventually, the gallery subject

with the highest similarity is selected as the classification result. After exploring numerous directions (see §5.2), we identified techniques that were most effective at producing dodging and impersonation attacks.

***Dodging*** In dodging, the adversary's goal is to produce an arbitrary misclassification to any class other than the true class. We find that evading classification by increasing the similarity w.r.t. the closest incorrect class (in a given attack iteration) and decreasing it w.r.t. the true class is most effective (§5.2). Given an input $x$ pertaining to class (i.e., gallery subject) $c_x$, we denote the feature array of the gallery images by $G \in \mathbb{R}^{k \times d}$, where $k$ is the number of classes, and $d$ is the dimensionality of the features extracted by the model $f$, and by $max_{c \neq c_x}\big(cos(f(x), G[c])\big)$ the closest class to $x$ which is not $c_x$. To produce a misclassification, we find a perturbation $\delta$ that maximizes the dodging loss:

$$L_{dodge}(x, c_x) = -\alpha cos\big(f(x+\delta), G[c_x]\big) + \beta max_{c \neq c_x}\big(cos(f(x+\delta), G[c])\big)$$

where $\alpha$ and $\beta$ are two non-negative constants, aiming to balance the first objective (decreasing the distance from $c_x$) and the second objective (increasing similarity with the most similar class $c \neq c_x$), respectively. After running a grid search, we found that setting both $\alpha$ and $\beta$ to one led to the highest success.

***Impersonation*** In impersonation, the adversary selects a target class (i.e., subject) $c_t$ to impersonate. To achieve this objective, besides increasing similarity with $c_t$ and decreasing similarity with $c_x$, we found that it is crucial to decrease similarity with all gallery subjects that are more similar to the input than the target, or are less similar to the input than the target but only slightly so. Said differently, our attack aims to ensure that the similarity with $c_t$ is higher than all other classes by a significant margin, increasing the confidence that the (adversarial) input pertains to $c_t$. In doing so, we could increase the likelihood that attacks would succeed when realized, even when similarity with $c_t$ is decreased after realization (e.g., due to imperfect fabrication of the accessory). To this end, we define the high-margin ($hm$) loss:

$$L_{hm} = \frac{1}{k}\sum_{c} ReLU\big(cos(f(x+\delta), G[c]) - cos(f(x+\delta), G[c_t]) + \tau\big)$$

where $\tau$ is a small non-negative constant set to ensure that the perturbation decreases the similarity w.r.t. classes sufficiently similar to $x$ (i.e., with similarity higher or up to a small margin of $c_t$). We empirically found that $\tau$=0.2 leads to successful attacks (see §5). Accordingly, the impersonation attacks aim to maximize the impersonation loss, defined by:

$$L_{imp}(x, c_x) = \alpha cos\big(f(x+\delta), G[c_t]\big) - \beta cos\big(f(x+\delta), G[c_x]\big) - \gamma * L_{hm}$$

where $\alpha$, $\beta$, and $\gamma$ are non-negative constants to balance between the attack goals, of increasing similarity with $c_t$, decreasing similarity with $c_x$, and decreasing $L_{hm}$. We set $\alpha$ and $\beta$ to 6, and $\gamma$ to 15, as we found these to work best after performing a grid search.

(a) (b)

Fig. 1: Attacks generated with (a) and without (b) minimizing TV.

### 4.2 Realizing Attacks

To implement attacks in the physical world, measures to aid in fabricating the adversarial artifacts and improve their robustness to varying imaging conditions (e.g., scale and pose) are necessary [32]. We address this by adding constraints to the attack to encourage the creation of objects that resemble their digital counterpart when printed and photographed with an NIR camera, and are robust to transformations encountered in the real world, as elaborated below.

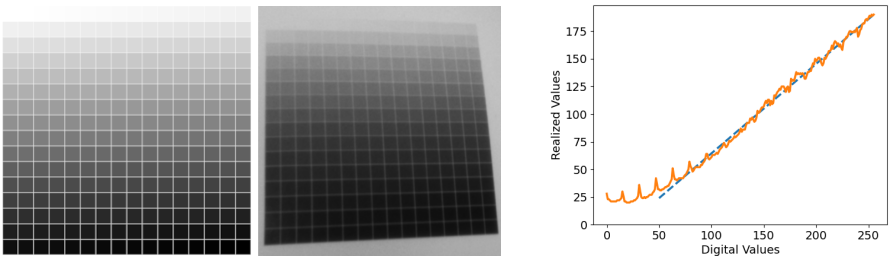
***Total Variation (TV)*** When not restricted, the attack may produce sharp, unnatural transitions between neighboring pixels. Such transitions would be challenging to realize, as they would require high-resolution printers and cameras to produce and capture them [21, 32]. Thus, to facilitate realizability and promote inconspicuousness, we use TV as part of the loss, similarly to Sharif et al.'s work [32]. Given an input $x \in \mathbb{R}^{d \times d}$, TV measures the distance between neighboring pixels via the following formula:

$$TV(x) = \sum_{i,j} [(x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2]^\beta$$

where $\beta$ is a configurable parameter that we set to 1, in line with prior work [32]. Fig. 1 shows artifacts produced with and without TV—by minimizing TV, attacks produce artifacts with smooth textures more amenable for realization.

***Printability*** To produce adversarial artifacts containing colors that can be physically realized via printing, we define and use a Non-Printability Score (NPS) metric tailored for the NIR domain. To define NPS, we first identify the color ranges our printer can produce and model how they are captured by cameras. Empirically, this works by printing a grayscale palette covering the entire [0, 255] range and photographing it (see Fig. 2a).[1] Doing so showed that the range of printable colors is a consecutive sub-range $[v_{lb}, v_{ub}]$=[40,180] of the the full [0,255] range (see Fig. 2b). Moreover, we observe a roughly linear relationship between printed colors an their captured counterparts, enabling us to pre-process

[1] We use grayscale as NIR contains a single channel and we found grayscale covers the value range more comprehensively than RGB.




(a) (b)

Fig. 2: (a) Digital grey-scale palette (left) compared to a printed and photographed palette (right). (b) Comparison of digital colors and their realized counterparts (after being printed and recaptured via an NIR camera). A dotted line is added to emphasize the roughly linear relationship.

the accessories' colors prior to printing to preserve similarity between the printed and captured colors (§5.3). Accordingly, we defined the NPS formula as follows:

$$NPS(x) = \sum_{i,j}[ReLU(v_{lb} - x_{i,j}) + ReLU(x_{i,j} - v_{ub})].$$

Intuitively, the NPS accumulates a penalty for each pixel that is lower than the lower bound or higher than the upper bound color we could realize. Therefore, by minimizing NPS, our attack pushes colors on the adversarial artifacts to become printable, thus aiding in realizability.

***Expectation Over Transformation (EOT)*** is a measure aiming to enhance robustness against changes likely to be encountered in the physical world [3]. For instance, when an attacker wears accessories (e.g., eyeglasses), we cannot assume they will be located exactly as intended on the face, that the attacker pose will be completely frontal, or that they will stand at a fixed distance w.r.t. the camera. To ensure that attacks succeed across input variations, we adapt EOT to face recognition such that we maximize the expected impersonation and dodging losses over potential variations. Formally, given an image $x$ and a perturbation $\delta$, instead of maximizing $L_{\{dodge|imp\}}$ over $x+\delta$, we maximize it over $t_1\big(x+t_2(\delta)\big)$ for $t_1 \sim T_1$ and $t_2 \sim T_2$, where $T_1$ are transformations applied to the face and accessory combined (e.g., changes in pose or distance), and $T_2$ are transformations applied to accessory alone (e.g., slight translation due to dislocation and noise due to sampling errors). Specifically, for $T_2$, we use slight rotations ($\in$ [-2,2] degrees), scaling ($\times$[0.98,1.02]), and translations along the x- and y-axes ($\in \{-2,\ldots,2\}$), to account for potential variations that might occur when attackers wear the accessory. Furthermore, we add small amount of zero-centered Gaussian noise ($\sigma$=0.04) to $\delta$ to account for slight color noise during sampling. To simulate transformations of the face and accessory (i.e., $T_1$), we

take multiple images per attacker with slight variations in pose, distance, and lighting and attach the accessories to them via perspective transformation.

***Overall Objective*** To physically realize attacks we find $\delta$ that maximizing

$$\arg\max_{\delta} E_{t_1,t_2\sim T_1,T_2}\left[L_{\{dodge|imp\}}\left(f\left(t_1\left(x+t_2(\delta)\right)\right),c_{\{x|t\}}\right)\right]-\omega_1 TV(\delta)-\omega_2 NPS(\delta).$$

The optimization process searches for a perturbation $\delta$ maximizing $L_{dodge}$ or $L_{imp}$ (depending on the attack objective) over expected input transformations, while minimizing the TV and NPS of $\delta$. $\omega_1$ and $\omega_2$ are non-negative constants for balancing the objectives tuned to maximize the success of realized attacks.

***Implementation Details*** We solve the optimization using PGD, after initializing the accessory colors to a uniform grayscale value of 76/255, allowing the accessory's values to range $\in$ [0,1] while not perturbing values not covered by the accessory. We run PGD for 400 iterations and set its step size to 1/255. To enable a more direct comparison with prior work, we use Sharif et al.'s eyeglasses covering 8% of the image [32] as the adversarial accessory. We tested other accessories (e.g., face masks and stickers) in the digital domain and found they led to relatively lower success than eyeglasses (§6). We implemented attacks using `PyTorch` and published our code to aid in reproducibility.[2]

## 5 Evaluation

Our experiments examined the vulnerability of several NIR-based face-recognition systems to dodging and impersonation attacks in the digital and physical domains. Next, we describe our experimental setup before reporting the results of attacks in the digital (§5.2) and physical (§5.3) domains.

### 5.1 Experimental Setup

***Data*** For our evaluation, we relied on the CASIA NIR-VIS 2.0 dataset [19]. This dataset consists of frontal face images of 725 subjects collected using NIR and VIS sensors. The VIS images were collected in visual light, while NIR images were collected in complete darkness, using an NIR camera surrounded by 850 nm NIR light-emitting diodes (LEDs). Fig. 3 presents samples from the dataset. The number of VIS images per subject varies between one and 22 while that of NIR images varies between five and 50. For testing, the dataset contains a gallery of 358 VIS images, one per subject, and a probe set consisting of 6,000 NIR images for the same 358 subjects. The objective is to map the NIR images from the probe set to the correct identity from the gallery. The dimensionality of the images is 480×640, and we aligned them to a fixed pose and cropped them to 224×224 centered around the face, per standard practice [32].

[2] Code available at `https://github.com/AmitCohen3/Accessorize-in-the-dark`

Fig. 3: NIR (top) and VIS (bottom) images of three subjects (columns) from the CASIA NIR-VIS 2.0 dataset.

To conduct experiments in the physical domains, we further augmented the dataset by enrolling three additional subjects—two males and a female 28–31 years of age. We refer to them by $S_1$–$S_3$. For each subject enrolled, we captured a VIS image and 20 NIR images, all using an Intel RealSense D415 camera. Similarly to CASIA NIR-VIS 2.0 dataset [19], both VIS and NIR images were taken with the subject's face positioned in the middle of the frame with a frontal pose. When capturing images in NIR, the subjects were wearing eyeglasses frames and were asked to slightly move their faces in a circular motion. All images were taken in a dark room, with closed window blinds to prevent external light, while turning on NIR LEDs positioned around the camera to faithfully simulate CASIA NIR-VIS 2.0's conditions. For printing, we used a Xerox B230 printer.

***Models*** We evaluated attacks against state-of-the-art architectures for NIR-VIS face-recognition: LightCNN, LightCNN-DVG, LightCNN-Rob, and ResNeSt. Wu et al. proposed LightCNN and trained it using multiple VIS datasets after converting inputs to one-dimensional (i.e., grayscale) images [43]. They showed that, by training model on noisy labels, LightCNN can attain high performance on the NIR-VIS face-recognition task. We acquired the LightCNN weights published by the authors. LightCNN-DVG was proposed in a follow-up work by the same group, in which they fine-tuned LightCNN using generated pairs of NIR-VIS face images to further improve the model's accuracy [12]. We trained a LightCNN-DVG model on our dataset using the official code. To enhance model robustness against attacks, we also followed Wu et al.'s protocol to adversarially train a model [42]. In particular, we fine-tuned LightCNN-DVG using the DOA method, running 10 epochs of adversarial training.[3] Finally, because some NIR-based face recognition systems leverage typical VIS models receiving three channels as input (e.g., [18]), we complemented the LightCNN variants with a Residual Neural Network with Split Attention (ResNeSt) model [50]. We acquired pre-trained ResNeSt weights through Wang et al.'s project [39], and found it was markedly more accurate than other 11 models ingesting three channels

[3] We used the Adam optimizer with a $1e$-5 learning rate for best performance.

Table 1: The models' benign accuracy with the enrolled subjects included. The standard deviation is negligible (<1$e$-4), thus excluded.

| **Model** | **Benign accuracy** |
|---|---|
| LightCNN | 98.27% |
| LightCNN-DVG | 99.80% |
| LightCNN-DVG-100 | 99.84% |
| LightCNN-DVG-10 | 99.85% |
| LightCNN-Rob | 99.56% |
| ResNeSt | 91.10% |

Wang et al. offer (including ResNet and VGG models). Lastly, to assess how the number of subjects affects attack success, we evaluated variants of LightCNN-DVG, LightCNN-DVG-10 and LightCNN-DVG-100, on a subset of ten and 100 subjects, respectively, both of which include the three subjects we enrolled.

To measure benign accuracy, we followed CASIA's protocol [19]: we divided the data into ten folds while adding the enrolled subjects to each of the folds and measured the mean accuracy over the folds. For models with ten or 100 subjects, we randomly chose the subjects from the dataset to compute benign accuracy, and calculated the mean over ten repetitions. Table 1 reports the benign accuracy of all models. All models were highly accurate, and, as expected, the most advanced model, LightCNN-DVG, was most accurate, with an increasing accuracy as the number of subjects decreased.

***Metrics*** We measured attack performance by their success rate (SR) and margin. SR estimates how often the attack achieves its objective—i.e., the percentage of time the attacker is misclassified as someone else (resp. target class) in dodging (resp. impersonation) attacks. The margin is a proxy for the confidence in the (mis)classification result. We measured it by the difference between similarity with the top prediction (resp. target class) and the true class in dodging (resp. impersonation) attacks.

### 5.2 Digital Attacks

We tested attacks in the digital domain to find loss functions that maximize attack success and assess the security of NIR-based face recognition in ideal settings, where adversaries can precisely produce adversarial accessories. In these attacks we ignored the TV, printability, and EOT objectives, and mainly focused on misclassifications using a single adversary image. We evaluated both dodging and impersonation attacks, selecting the impersonation targets at random. We ran each attack type ten times, each time with different 1,024 NIR images (or all images available for the subjects, if less than 1,024), and measured the average and standard deviation (std) of the SR and margin over these repetitions. Lastly, we tested the transferability of attacks—i.e., how often attacks created against one model succeed against other models—to simulate black-box attacks.

Table 2: Comparison between dodging losses against LightCNN-DVG.

| Loss | SR | Margin (std) |
|---|---|---|
| $L^1_{dodge} = -\alpha cos\big(f(x+\delta), G[c_x]\big)$ | 100.00% | 0.38 (0.13) |
| $L^2_{dodge} = -\alpha cos\big(f(x+\delta), G[c_x]\big) + \beta \max_{c \neq c_x} \big(cos(f(x), G[c])\big)$ | 100.00% | 0.30 (0.11) |
| $L^3_{dodge} = -\alpha cos\big(f(x+\delta), G[c_x]\big) + \beta \max_{c \neq c_x} \big(cos(f(x+\delta), G[c])\big)$ | **100.00%** | **0.50 (0.13)** |

Table 3: Comparison between impersonation losses against LightCNN-DVG.

| Loss | SR | Margin (std) |
|---|---|---|
| $L^1_{imp} = cos\big(f(x+\delta), G[c_t]\big)$ | 84.37% | 0.20 (0.15) |
| $L^2_{imp} = \alpha cos\big(f(x+\delta), G[c_t]\big) - \beta cos\big(f(x+\delta), G[c_x]\big)$ | 72.07% | 0.40 (0.16) |
| $L^3_{imp} = \alpha cos\big(f(x+\delta), G[c_t]\big) - \beta \max_c (cos(f(x+\delta), G[c]))$ | 87.07% | 0.02 (0.02) |
| $L^4_{imp} = \alpha cos\big(f(x+\delta), G[c_t]\big) - \beta cos\big(f(x+\delta), G[c_x]\big) - \gamma \max_c (cos(f(x+\delta), G[c]))$ | 80.23% | 0.40 (0.16) |
| $L^5_{imp} = \alpha cos\big(f(x+\delta), G[c_t]\big) - \beta cos\big(f(x+\delta), G[c_x]\big) - \gamma \max_{c \neq c_x} (cos(f(x+\delta), G[c]))$ | 85.89% | 0.34 (0.15) |
| $L^6_{imp} = \alpha cos\big(f(x+\delta), G[c_t]\big) - \beta cos\big(f(x+\delta), G[c_x]\big) - \gamma * L_{hm}$ | **91.01%** | **0.34 (0.17)** |

Table 4: SRs and margins for digital-domain dodging and impersonation attacks.

| Model | Dodging | | Impersonation | |
|---|---|---|---|---|
| | SR | Margin (std) | SR | Margin (std) |
| LightCNN | 100.00% | 0.48 (0.15) | 90.92% | 0.30 (0.17) |
| LightCNN-DVG | 100.00% | 0.50 (0.13) | 91.01% | 0.34 (0.17) |
| LightCNN-DVG-100 | 100.00% | 0.50 (0.12) | 94.95% | 0.39 (0.17) |
| LightCNN-DVG-10 | 100.00% | 0.47 (0.12) | 98.78% | 0.35 (0.17) |
| LightCNN-Rob | 100.00% | 0.36 (0.13) | 52.66% | 0.09 (0.19) |
| ResNeSt | 100.00% | 0.35 (0.10) | 89.05% | 0.20 (0.11) |

***Loss-Function Selection*** We evaluated various loss function for dodging and impersonation to identify the ones maximizing attack success. In these experiments, we ran attacks against LightCNN-DVG, as it was the most robust amongst the undefended models. Table 2 presents the three dodging losses considered and their corresponding SRs and margins. $L^1_{dodge}$ aims to decrease the similarity with the true class; $L^2_{dodge}$ extends $L^1_{dodge}$ by increasing similarity with the closest subject (excluding $c_x$) prior to running the attack; and $L^3_{dodge}$ ($L_{dodge}$ in §4.1) extends $L^1_{dodge}$ by increasing similarity with the closest subject to $x+\delta$ in the current iteration. $L^3_{dodge}$ led to markedly higher margins, hence we used it in subsequent attacks. Table 3 lists the six impersonation losses we tested and their respective SRs and margins. $L^1_{imp}$ seeks to increase similarity with $c_t$; $L^2_{imp}$ also aims to decrease similarity with $c_x$; $L^3_{imp}$ extends $L^1_{imp}$ by decreasing similarity with the current top prediction; $L^4_{imp}$ combines $L^2_{imp}$ and $L^3_{imp}$; $L^5_{imp}$ refines $L^4_{imp}$ by excluding $c_x$ when decreasing similarity with the top prediction; and $L^6_{imp}$ is equivalent to $L_{imp}$ described in §4.1. It can be immediately seen that $L^6_{imp}$ reached remarkably higher SR than other losses. Thus, we used $L^6_{imp}$ to perform impersonations in the following experiments.

Table 5: Transferability of digital dodging (left) and impersonations (right).

| Surrogate \ Target | LightCNN-DVG | LightCNN | LightCNN-Rob | ResNeSt |
|---|---|---|---|---|
| LightCNN-DVG | 100.00% | 98.82% | 76.95% | 38.67% |
| LightCNN | 100.00% | 100.00% | 68.93% | 41.28% |
| LightCNN-Rob | 96.24% | 89.52% | 100.00% | 40.31% |
| ResNeSt | 1.32% | 11.45% | 20.04% | 100.00% |

| Surrogate \ Target | LightCNN-DVG | LightCNN | LightCNN-Rob | ResNeSt |
|---|---|---|---|---|
| LightCNN-DVG | 91.02% | 66.99% | 25.00% | 0.00% |
| LightCNN | 65.42% | 90.9% | 37.10% | 0.00% |
| LightCNN-Rob | 41.30% | 26.48% | 52.56% | 0.39% |
| ResNeSt | 0.16% | 0.17% | 0.22% | 89.04% |

***Attack Evaluation*** Table 4 reports the performance of digital-domain dodging and impersonation attacks against all models, using $L_{dodge}$ and $L_{imp}$, respectively. It can be observed that all dodging attempts against all models succeeded. Impersonation attacks' SRs, on the other hand, ranged between 52.66% and 98.78%. The defended model, LightCNN-Rob, was the most challenging to mislead, with 52.66% impersonation SR and a 0.09 margin, compared to $\geq$89.05% and $\geq$0.20 margins for the undefended models. Moreover, perhaps intuitively, impersonation attacks against models with fewer subjects (i.e., LightCNN-DVG-10 and LightCNN-DVG-100) were relatively more successful than the model with all subjects (i.e., LightCNN-DVG).

Although attacks were not optimized for transferability, we found that they often transfer successfully, especially between the LightCNN variants (Table 5). Between different LightCNN models, the mean SR of transferred attacks ranged between 68.93%–100.00% for dodging and 25.00%–65.42% for impersonation. Impersonation attacks transferred from and to ResNeSt had low SRs ($\leq$0.39%), but dodging attacks against LightCNN variants often misled ResNeSt (38.67%–41.28% mean SR). We expect higher SRs would be achievable by integrating techniques to promote transferability (e.g., [40]).

### 5.3 Physical Attacks

We tested physical-domain attacks against all models. In these experiments, the three subjects introduced to the dataset simulated attackers. For each subject, we ran dodging and impersonation attacks against each model, for a total of $3\times2\times6=36$ attack attempts. As in the digital-domain, we randomly chose the target in each impersonation attack. For each attack, we solved the corresponding optimization with all objectives (§4.2) to generate eyeglass textures, which we then printed, cut, and affixed to 3D frames. To this end, we used all NIR images available for the subject to estimate the EOT of the loss and solve the optimization. Prior to printing, we increased the accessory's pixels' brightness by 40/255 to ensure the printed value of each pixel corresponds to its digital counterpart (per Fig. 2b). We then collected ten images of the person simulating the attacker while wearing the adversarial eyeglasses to measure attack SR. Besides white-box attacks, we again evaluated the transferability of attacks between models to simulate black-box settings. Next, we report how we weighted each term in the overall attack objective, followed by the attack performance.

***Setting TV's and NPS' Weights*** In our preliminary experiments, we found that adversarial eyeglasses with a TV value of $\sim$200 and an NPS value of $\sim$250

Table 6: Digital-domain impersonation SR against LightCNN-DVG for varied TV and NPS weights.

| TV w. \ NPS w. | 0 | 1e-4 | 1e-3 | 1e-2 |
|---|---|---|---|---|
| 0 | 92.18% | 91.60% | 91.01% | 85.54% |
| 2e-4 | 91.99% | 91.99% | **91.99%** | 90.82% |
| 2e-3 | 85.15% | 90.62% | 90.62% | 90.42% |
| 2e-2 | 85.74% | 87.69% | 87.5% | 87.69% |

Table 7: Mean NPS (left) and TV (right) values for varied TV and NPS weights.

| TV w. \ NPS w. | 0 | 1e-4 | 1e-3 | 1e-2 |
|---|---|---|---|---|
| 0 | 319.24 | 273.99 | 165.10 | 32.4481 |
| 2e-4 | 263.72 | 263.72 | **244.79** | 157.34 |
| 2e-3 | 32.24 | 182.62 | 175.07 | 128.75 |
| 2e-2 | 30.03 | 116.12 | 112.13 | 90.14 |

| TV w. \ NPS w. | 0 | 1e-4 | 1e-3 | 1e-2 |
|---|---|---|---|---|
| 0 | 347.33 | 295.46 | 201.50 | 112.38 |
| 2e-4 | 229.74 | 229.74 | **217.67** | 167.73 |
| 2e-3 | 100.15 | 95.20 | 93.62 | 83.04 |
| 2e-2 | 61.34 | 23.04 | 22.96 | 22.27 |

Fig. 4: $S_1$ physically dodging (left) and impersonating (middle) target ID 00041 (right) against LightCNN-DVG.

preserve the digital-domain SR best when realized. To this end, to appropriately tune the TV and NPS weights and attain values in the desirable range while maximizing attack SRs, we performed a grid search, evaluating attack SRs in the digital domain with different weights assigned to the TV and NPS objectives. Specifically, we conducted digital-domain impersonation attacks against LightCNN-DVG, using a single adversary image at a time. We repeated the experiment 512 times, each time with different attacker image and a target selected at random. Tables 6–7 report the mean attack SRs, and the mean TV and NPS scores. Per these results, we set the TV and NPS weights to 2e-4 and 1e-3, respectively, as they resulted in the highest attack SRs while attaining TV and NPS conducive for faithful realization.

***Attack Evaluation*** Table 8 reports the attack SRs against all models. Dodging attacks were highly successful, with $\geq 9/10$ attempts leading to misclassification in all cases, and all attempts being misclassified for most subject and model pairs. The models were also relatively susceptible to impersonation at-

Table 8: SRs of physical attacks. For each model, we report the dodging and impersonation attack SRs per subject simulating the attacker out of ten attempts, as well as the mean SR across attackers. In the interest of reproducibility, we also report the randomly selected impersonation targets.

| Model | Attacker | Dodging SR | Dodging Mean(SR) | Impersonation Target | Impersonation SR | Impersonation Mean(SR) |
|---|---|---|---|---|---|---|
| LightCNN | $S_1$ | 10/10 | | 10047 | 10/10 | |
| | $S_2$ | 9/10 | 93.33% | 20476 | 10/10 | 100.00% |
| | $S_3$ | 9/10 | | 20361 | 10/10 | |
| LightCNN-DVG | $S_1$ | 10/10 | | 20370 | 9/10 | |
| | $S_2$ | 10/10 | 100.00% | 20389 | 10/10 | 96.66% |
| | $S_3$ | 10/10 | | 00050 | 10/10 | |
| LightCNN-DVG-100 | $S_1$ | 10/10 | | 10123 | 0/10 | |
| | $S_2$ | 10/10 | 100.00% | 20472 | 10/10 | 66.66% |
| | $S_3$ | 10/10 | | 00140 | 10/10 | |
| LightCNN-DVG-10 | $S_1$ | 10/10 | | 20387 | 10/10 | |
| | $S_2$ | 10/10 | 100.00% | 20364 | 10/10 | 76.66% |
| | $S_3$ | 10/10 | | 30565 | 3/10 | |
| LightCNN-Rob | $S_1$ | 9/10 | | 00041 | 10/10 | |
| | $S_2$ | 10/10 | 96.66% | 30778 | 0/10 | 36.66% |
| | $S_3$ | 10/10 | | 10210 | 1/10 | |
| ResNeSt | $S_1$ | 10/10 | | 20349 | 10/10 | |
| | $S_2$ | 10/10 | 100.00% | 00202 | 7/10 | 90.00% |
| | $S_3$ | 10/10 | | 00122 | 10/10 | |

Table 9: Transferability of physical dodging (left) and impersonations (right).

| Surrogate \ Target | LightCNN-DVG | LightCNN | LightCNN-Rob | ResNeSt |
|---|---|---|---|---|
| LightCNN-DVG | 100.00% | 63.33% | 63.33% | 0.00% |
| LightCNN | 43.33% | 100.00% | 40.00% | 13.33% |
| LightCNN-Rob | 30.00% | 23.33% | 96.66% | 0.00% |
| ResNeSt | 0.00% | 3.33% | 33.33% | 100.00% |

| Surrogate \ Target | LightCNN-DVG | LightCNN | LightCNN-Rob | ResNeSt |
|---|---|---|---|---|
| LightCNN-DVG | 100.00% | 36.66% | 30.00% | 0.00% |
| LightCNN | 33.33% | 96.66% | 26.66% | 0.00% |
| LightCNN-Rob | 23.33% | 0.00% | 36.66% | 0.00% |
| ResNeSt | 0.00% | 0.00% | 0.00% | 90.00% |

tacks, with $\geq 1/10$ attempts succeeding in 16 of the 18 impersonation attacks, and 36.66%–100.00% mean SR across the six models. Naturally, the adversarially trained model, LightCNN-Rob, was the most challenging to mislead, however, even against it, two of the three impersonation attacks succeeded in at least 1/10 attempts, with one attack succeeding in all attempts. An example of a physical attack is depicted in Fig. 4.

Similarly to the digital domain, attacks exhibited strong transferability between LightCNN variants (Table 9)—mean SRs ranged between 30.00%–63.33% for transferred physical-world dodging attempts and reached up to 36.66% for impersonation. However, despite 33.33% mean SR for dodging attempts transferred from ResNeSt to LightCNN-Rob, physical-world attacks transferred be-

tween ResNeSt and other models had limited SRs (0.00%–13.33% in all other cases). Overall, the SRs of transferred attacks were non-negligible, but we expect they could be further improved via techniques geared to enhance transferability.

## 6 Limitations

Our findings should be interpreted in light of certain limitations. We evaluated physical attacks in relatively controlled settings, in a single room, with three subjects acting as adversaries. Hence, the generalizability of the results to more settings with other attackers remains to be determined. Still, we expect our results to inform us about the susceptibility of NIR-based face recognition systems in real-world deployments, where imaging variations may resemble those in our experiments (e.g., internal deployment in airports). We also primarily studied evasion attacks using eyeglasses. However, when testing other accessories, such as face masks and stickers [37], we found that attack SRs in the digital environment were significantly lower than with eyeglasses (e.g., 58.43% impersonation SR with stickers against LightCNN-DVG) or that attacks were conspicuous (e.g., attacks with face masks added odd facial features to masks).

## 7 Conclusion

Prior work has shown VIS-based face-recognition systems to be vulnerable to evasion attacks (e.g., [32, 33, 37]). To the best of our knowledge, we are the first to demonstrate realizable evasion attacks against NIR-based face recognition. As those systems are widely employed in security-critical settings (e.g., [1, 2]), our work highlights the need for enhancing their robustness, especially as existing defenses [42] remain vulnerable to attacks (§5). Relatively expensive defense techniques, such as human supervision to ascertain the absence of facial accessories, can be implemented immediately. However, further research is needed to establish technical means to enhance NIR-based face recognition's adversarial robustness. We hope that the attacks presented in this work can help inform the design of such defenses and aid in evaluating them.

***Acknowledgments*** We thank Amir Barda and Amit Bermano for their help printing the 3D frames, and the PLUS research group's members for helpful feedback. This work was supported in part by Len Blavatnik and the Blavatnik Family foundation; by a Maof prize for outstanding young scientists; by a scholarship from the the Shlomo Shmeltzer Institute for Smart Transportation at Tel-Aviv University; and by the Neubauer Family foundation.

## References

1. Face ID security. `https://help.apple.com/pdf/security/en_US/apple-platform-security-guide.pdf`

2. Windows Hello. https://docs.microsoft.com/en-us/windows/security/identity-protection/hello-for-business/hello-overview
3. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proc. ICML (2018)
4. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. In: Proc. ECML PKDD (2013)
5. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition **84**, 317–331 (2018)
6. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: Proc. IEEE S&P (2017)
7. Carlos-Roca, L.R., Torres, I.H., Tena, C.F.: Facial recognition application for border control. In: Proc. IJCNN (2018)
8. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: Proc. ICML (2019)
9. Deng, Z., Peng, X., Li, Z., Qiao, Y.: Mutual component convolutional neural networks for heterogeneous face recognition. IEEE Transactions on Image Processing **28**(6), 3102–3114 (2019)
10. Duan, B., Fu, C., Li, Y., Song, X., He, R.: Cross-spectral face hallucination via disentangling independent factors. In: Proc. CVPR (2020)
11. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: Proc. CVPR (2018)
12. Fu, C., Wu, X., Hu, Y., Huang, H., He, R.: DVG-face: Dual variational generation for heterogeneous face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2021)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proc. ICLR (2015)
14. Hu, W., Hu, H.: Orthogonal modality disentanglement and representation alignment network for NIR-VIS face recognition. IEEE Transactions on Circuits and Systems for Video Technology **32**(6), 3630–3643 (2021)
15. Hu, W., Yan, W., Hu, H.: Dual face alignment learning network for nir-vis face recognition. IEEE Transactions on Circuits and Systems for Video Technology **32**(4), 2411–2424 (2021)
16. Huang, H., Mu, J., Gong, N.Z., Li, Q., Liu, B., Xu, M.: Data poisoning attacks to deep learning based recommender systems. In: Proc. NDSS (2021)
17. Kong, S.G., Heo, J., Abidi, B.R., Paik, J., Abidi, M.A.: Recent advances in visual and infrared face recognition—a review. Computer vision and image understanding **97**(1), 103–135 (2005)
18. Lezama, J., Qiu, Q., Sapiro, G.: Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding. In: Proc. CVPR (2017)
19. Li, S., Yi, D., Lei, Z., Liao, S.: The CASIA NIR-VIS 2.0 face database. In: Proc. CVPRW (2013)
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proc. ICLR (2018)
21. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proc. CVPR (2015)
22. Metzen, J.H., Genewein, T., Fischer, V., Bischoff, B.: On detecting adversarial perturbations. In: Proc. ICLR (2017)
23. Miao, Y., Lattas, A., Deng, J., Han, J., Zafeiriou, S.: Physically-based face rendering for NIR-VIS face recognition. In: Proc. NeurIPS (2022)

24. Naqvi, R.A., Arsalan, M., Batchuluun, G., Yoon, H.S., Park, K.R.: Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. Sensors **18**(2), 456 (2018)
25. Osborne, B.G.: Near-infrared spectroscopy in food analysis. Encyclopedia of analytical chemistry: Applications, theory and instrumentation (2006)
26. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint (2016)
27. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proc. AsiaCCS (2017)
28. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: Proc. IEEE EuroS&P (2016)
29. Pierazzi, F., Pendlebury, F., Cortellazzi, J., Cavallaro, L.: Intriguing properties of adversarial ml attacks in the problem space. In: Proc. S&P (2020)
30. Schönherr, L., Kohls, K., Zeiler, S., Holz, T., Kolossa, D.: Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding (2019)
31. Shamir, A., Safran, I., Ronen, E., Dunkelman, O.: A simple explanation for the existence of adversarial examples with small hamming distance. arXiv preprint (2019)
32. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proc. CCS (2016)
33. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security (TOPS) **22**(3), 16:1–16:30 (2019)
34. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Proc. IEEE S&P (2017)
35. Shumailov, I., Zhao, Y., Bates, D., Papernot, N., Mullins, R., Anderson, R.: Sponge examples: Energy-latency attacks on neural networks. In: Proc. IEEE EuroS&P (2021)
36. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proc. ICLR (2014)
37. Tong, L., Chen, Z., Ni, J., Cheng, W., Song, D., Chen, H., Vorobeychik, Y.: FaceSec: A fine-grained robustness evaluation framework for face recognition systems. In: Proc. CVPR (2021)
38. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: Proc. USENIX Security (2016)
39. Wang, J., Liu, Y., Hu, Y., Shi, H., Mei, T.: FaceX-Zoo: A PyTorch toolbox for face recognition. In: Proc. MM (2021)
40. Wang, X., He, X., Wang, J., He, K.: Admix: Enhancing the transferability of adversarial attacks. In: Proc. ICCV (2021)
41. Wang, Y., Bao, T., Ding, C., Zhu, M.: Face recognition in real-world surveillance videos with deep learning method. In: Porc. ICIVC (2017)
42. Wu, T., Tong, L., Vorobeychik, Y.: Defending against physically realizable attacks on image classification. In: Proc. ICLR (2020)
43. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018)
44. Wu, X., Huang, H., Patel, V.M., He, R., Sun, Z.: Disentangled variational representation for heterogeneous face recognition. In: Proc. AAAI (2019)

45. Xiang, C., Bhagoji, A.N., Sehwag, V., Mittal, P.: PatchGuard: A provably robust defense against adversarial patches via small receptive fields and masking. In: Proc. USENIX Security (2021)
46. Xiang, C., Mahloujifar, S., Mittal, P.: PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In: Proc. USENIX Security (2022)
47. Xiang, C., Mittal, P.: PatchGuard++: Efficient provable attack detection against adversarial patches. In: Proc. ICLRW (2021)
48. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: Proc. NDSS (2018)
49. Yu, A., Wu, H., Huang, H., Lei, Z., He, R.: LAMP-HQ: A large-scale multi-pose high-quality database and benchmark for NIR-VIS face recognition. International Journal of Computer Vision (IJCV) **129**(5), 1467–1483 (2021)
50. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: ResNeSt: Split-attention networks. In: Proc. CVPR (2022)