# SPAA: Stealthy Projector-based Adversarial Attacks on Deep Image Classifiers

Bingyao Huang* Haibin Ling†

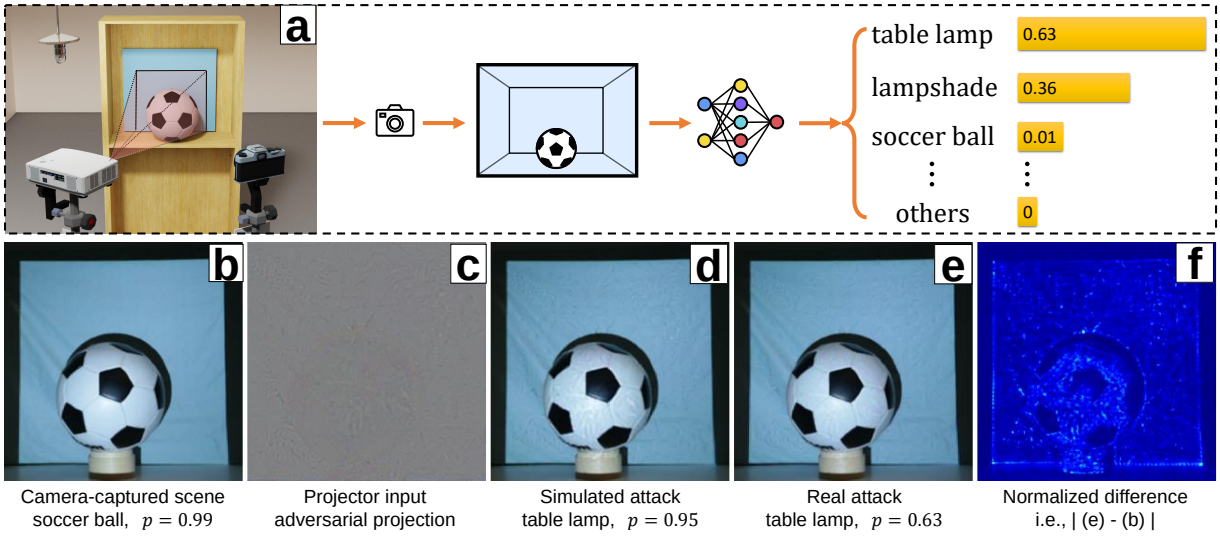


Figure 1: Stealthy projector-based adversarial attack (SPAA): **(a)** System setup: the goal is to project a stealthy adversarial pattern (*e.g.*, (c)), such that the camera-captured scene (*e.g.*, (e)) causes misclassification. **(b)** Camera-captured scene under normal light and the classifier output is **soccer ball** with a probability of $p = 0.99$. **(c)** An adversarial pattern created by our SPAA algorithm. **(d)** Our SPAA *simulated* camera-captured adversarial projection (*i.e.*, (c) virtually projected onto (b)). **(e)** The *actual* camera-captured adversarial projection (*i.e.*, (c) actually projected onto (b)). **(f)** Normalized difference between (b) and (e). It is clear that the camera-captured adversarial projection is stealthy, meanwhile, successfully fools the classifier such that the output is **table lamp** with a probability of $p = 0.63$. More results are provided in § 4 and supplementary.

## Abstract

Light-based adversarial attacks use spatial augmented reality (SAR) techniques to fool image classifiers by altering the physical light condition with a controllable light source, *e.g.*, a projector. Compared with physical attacks that place hand-crafted adversarial objects, projector-based ones obviate modifying the physical entities, and can be performed transiently and dynamically by altering the projection pattern. However, subtle light perturbations are insufficient to fool image classifiers, due to the complex environment and project-and-capture process. Thus, existing approaches focus on projecting clearly perceptible adversarial patterns, while the more interesting yet challenging goal, stealthy projector-based attack, remains open. In this paper, for the first time, we formulate this problem as an end-to-end differentiable process and propose a Stealthy Projector-based Adversarial Attack (SPAA) solution. In SPAA, we approximate the real Project-and-Capture process using a deep neural network named PCNet, then we include PCNet in the optimization of projector-based attacks such that the generated adversarial projection is physically plausible. Finally, to generate both robust and stealthy adversarial projections, we propose an algorithm that uses minimum perturbation and adversarial confidence thresholds to alternate between the adversarial loss and stealthiness loss optimization. Our experimental evaluations show that SPAA clearly outperforms other methods by achieving higher attack success rates and meanwhile being stealthier, for both targeted and untargeted attacks.

**Index Terms:** Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Security and privacy—Privacy protections; Computing methodologies—Object recognition

## 1 Introduction

Adversarial attacks on deep image classifiers aim to generate adversarial perturbation to the input image (*i.e.*, digital attacks) or the physical world (physical or projector-based attacks) such that the perturbed input can fool classifiers. With the rapid advancement of artificial intelligence, adversarial attacks become particularly important as they may be applied to protect user privacy and security from unauthorized visual recognition. It is worth noting that our work is different from existing studies in privacy and security of virtual reality (VR) and augmented reality (AR) [1, 6, 11, 32, 34], because we aim to use spatial augmented reality (SAR) to protect privacy and security rather than studying the privacy and security of VR/AR systems themselves. The most popular type of adversarial attacks are digital attacks [5, 8, 12, 25–27, 33, 39, 41, 47], which directly perturb the input images of a classifier. A common requirement for digital attack is stealthiness, *i.e.*, the perturbation should be relatively small (usually bounded by $L_p$ norm) yet still successfully fools the classi-

*College of Computer and Information Science, Southwest University, Chongqing, China. E-mail: bhuang@swu.edu.cn

†Department of Computer Science, Stony Brook University, Stony Brook, NY 11794, USA. E-mail: hling@cs.stonybrook.edu

fiers. Another type is physical attack [2, 4, 9, 10, 20, 21, 35, 43, 44], which assumes no direct access to the classifier input image. Instead, the perturbation is made on the physical entities, *e.g.*, placing adversarial patches, stickers or 3D printed objects. Usually physical attacks are much harder to achieve stealthiness due to complex physical environment and image capture process [2, 20, 44], and they must be strong enough to fool the classifiers. Another challenge is for targeted attacks, physical ones must manufacture a different adversarial pattern for each target.

Light-based (in the rest of the paper, we use *projector-based* to better describe our setup) attacks, as shown by our example in Figure 1, use SAR techniques to modify the environment light without physically placing adversarial entities to the scene. Thus, the attacks can be transient and dynamic, *e.g.*, by turning on and off the projector or changing the projected patterns. However, similar to physical attacks, projector-based attacks are difficult to fool image classifiers due to the complex environment and the project-and-capture process. Thus, existing methods [22, 29, 30] focus on improving attack success rates using *perceptible* patterns, while *stealthy* projector-based attack remains an open problem.

Note that simply projecting a digital adversarial example to the scene may not produce a successful stealthy projector-based attack, due to the complex geometric and photometric transformations involved in the project-and-capture process. One intuitive solution is to use a two-step pipeline by first performing digital attacks on the camera-captured scene image, then using projector compensation techniques [3, 13, 15] to find the corresponding projector adversarial pattern. However, this two-step method is problematic, because digital attacks may generate physically implausible [44] adversarial examples that cannot be produced by a projector, *e.g.*, perturbations in shadow regions or luminance beyond the projector's dynamic range. As will be shown in our experimental evaluations, such a two-step method has lower attack success rates and stealthiness than our SPAA solution. Another idea is the online one-pixel-based attack [30]. However, this preliminary exploration only allows to perturb one projector pixel and requires at least hundreds of real projections and captures to attack a single $32 \times 32$ low resolution target, making it hardly applicable to higher resolution images in practice, as shown in our experiments.

In this paper, we approach stealthy projector-based attacks from a different perspective by approximating the real Project-and-Capture process using a deep neural network named *PCNet*. Then, we concatenate PCNet with a deep image classifier such that the entire system is end-to-end differentiable. Thus, PCNet adds additional constraints such that the projected adversarial patterns are physically plausible. Finally, to generate robust and stealthy adversarial patterns, we propose an optimization algorithm that uses minimum perturbation and adversarial confidence thresholds to alternate between the minimization of adversarial loss and stealthiness loss.

To validate the effectiveness of the proposed SPAA algorithm, we conduct thorough experimental evaluations on 13 different projector-based attack setups with various objects, for *both targeted and untargeted* attacks. In all the comparisons, SPAA significantly outperforms other baselines by achieving higher success rates and meanwhile being stealthier.

Our contributions can be summarized as follows:

- For the first time, we formulate the stealthy projector-based adversarial attack as an end-to-end differentiable process.
- Based on our novel formulation, we propose a deep neural network named PCNet to approximate the real project-and-capture process.
- By incorporating the novel PCNet in projector-based adversarial attacks, our method generates physically plausible and stealthy adversarial projections.

The source code, dataset and experimental results are made publicly available at `https://github.com/BingyaoHuang/SPAA`.

In the rest of the paper, we introduce the related work in § 2, and describe the problem formulation and the proposed SPAA algorithm in § 3. We show our system configurations and experimental evaluations in § 4, and conclude the paper in § 5.

## 2 Related Work

In this section we review existing adversarial attacks on deep image classifiers in three categories: digital attacks, physical ones and projector-based ones as shown in Figure 2.

**Digital attacks** directly alter a classifier's input digital image such that the classifier's prediction becomes either (a) a specific target (targeted attack) or (b) any target as long as it is not the true label (untargeted attack). The input image perturbation is usually performed by back-propagating the gradient of adversarial loss to the input image, and can be either single-step, *e.g.*, fast gradient sign method (FGSM) [12], or iterative, *e.g.*, L-BFGS based [41], iterative FGSM (I-FGSM) [21], momentum iterative FGSM (MI-FGSM) [8], projected gradient descent (PGD) [25], C&W [5] and decoupling direction and norm (DDN) [33].

The gradient-based methods above require access to the classifier weights and gradients (*i.e.*, white-box attack). To relax such requirements, another type of digital attacks use gradient-free optimization, *e.g.*, one-pixel attack using differential evolution (DE) [39] or black-box optimization [46]. Another advantage of gradient-free attacks is that they can be applied to scenarios where the system gradient is inaccessible or hard to compute (see projector-based attacks below). However, they are usually less efficient than gradient-based methods, and this situation deteriorates when image resolution increases.

**Physical attacks** assume no direct access to the classifier input image, instead they modify the physical entities in the environment by placing manufactured adversarial objects or attaching stickers/graffiti. For example, Brown *et al*. [4] print 2D adversarial patches such that when placed in real scenes, the camera-captured images may be misclassified as certain targets. Sharif *et al*. [35] create a pair of adversarial eyeglass frames such that wearers can evade unauthorized face recognition systems. Similarly, Wu *et al*. [43] create an invisibility cloak to evade object detectors. Li *et al*. [23] alter camera-captured scenes by applying a translucent adversarial sticker to the camera lens. Early approaches often perform attacks in the digital image space first, and then bring the printed versions to the physical world. However, Kurarin *et al*. [20] show that the complex physical environment and the image capture process significantly degrade the attack success rates, because image space perturbations may not be physically meaningful [44] and are sensitive to minor transformations [2].

To fill the gap between the digital and the physical worlds, and to improve transferability, some studies focus on robustness of physical adversarial examples against transformations. For example, Athalye *et al*. [2] propose Expectation Over Transformation (EOT) to generate robust physical adversarial examples over synthetic transformations. Then, Eykholt *et al*. [10] propose Robust Physical Perturbations ($RP_2$) to produce robust adversarial examples under both physical and synthetic transformations. Afterwards, Jan *et al*. [17] present D2P to capture more complex digital-to-physical transformations using an image-to-image translation network.

Despite these efforts, how to make adversarial patterns stealthy remains challenging. Unlike digital attacks where perturbations can be easily made stealthy, subtle physical perturbations are hard to capture using digital cameras and can be easily polluted by sensor noise, lens distortion and camera internal image processing pipeline. Thus, to improve robustness against these factors, most existing physical adversarial examples are designed with strong artificial patterns.

**Projector-based attacks** modify only the environment light condition using a projector instead of changing the physical entities (*e.g.*, placing manufactured adversarial objects in the scene), and very

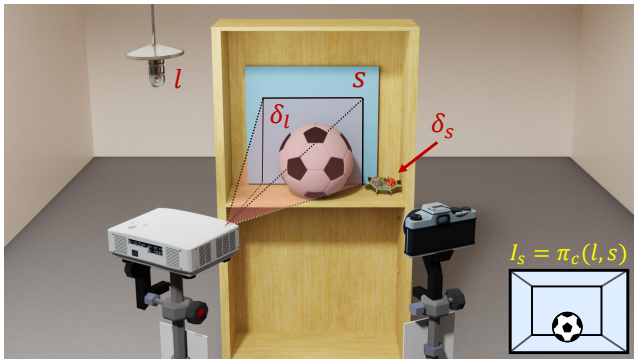


Figure 2: Adversarial attack types. Digital attacks directly perturb the camera-captured image $I_s$. Physical attacks perturb the scene $s$ by adding physical entities, *e.g.*, an adversarial patch $\delta_s$. Projector-based attacks perturb the environment light $l$ by $\delta_l$.

few studies have been dedicated to this direction. A preliminary exploration done by Nichols and Jasper [30] uses a low resolution projector-camera pair (both set to $32 \times 32$) to perturb scene illuminations and capture projections. Because the image resolutions are relatively small, a differential evolution [38] (DE)-based one-pixel attack framework [39] can be applied to solve this problem. In particular, by perturbing only one projector pixel, only five variables need to be optimized, *i.e.*, the pixel's 2D location and its RGB value. Even so, it still requires hundreds of real projections and captures for each targeted attack. Moreover, including the real project-and-capture process in the DE optimization may not only cause efficiency bottlenecks but also makes it hard to run in parallel. Thus, this method is impractical for high resolution cases due to the exponentially increased number of real project-and-capture processes.

Other studies focus on attacking face recognition systems [22, 29, 36, 48]. Special hardware settings are proposed to achieve stealthiness, *e.g.*, Zhou *et al*. [48] use infrared LEDs to project human imperceptible patterns and Shen *et al*. [36] leverage persistence of vision and the chromatic addition rule to control camera shutter speed, such that the camera can capture human imperceptible adversarial patterns.

**Stealthiness** is a common requirement for adversarial attacks, *i.e.*, perturbations should be (nearly) imperceptible to human eyes while still successfully causing misclassification. Usually stealthiness is measured using $L_\text{p}$ norm [5, 12, 20, 27, 41] and used as an additional constraint when optimizing the adversarial attack objective. Recently, Zhao *et al*. [47] show that optimizing perceptual color distance $\Delta E$ (*i.e.*, CIEDE2000 [24]) instead of $L_\text{p}$ norm may lead to more robust attacks yet still being stealthy. Besides pixel-level color losses, neural style similarity constraints can also improve stealthiness, *e.g.*, Duan *et al*. [9] propose an adversarial camouflage algorithm named AdvCam to make physical adversarial patterns look natural. Although it looks less artificial than previous work [4, 10], there is still room for improvement, especially the texture and color.

**The proposed SPAA** belongs to projector-based attacks, and is most related to the preliminary exploration in [30], with the following main differences: (1) We formulate projector-based adversarial attack as an end-to-end differentiable process, and simulate the real project-and-capture process with a deep neural network. (2) With such a formulation and implementation, our method can perform projector-based attacks using gradient descent, which is more efficient than one-pixel differential evolution [30]. (3) Because the real project-and-capture process is excluded from the gradient descent optimization, our method is more efficient and parallelizable, and multi-classifier and multi-targeted adversarial attacks can be performed simultaneously in batch mode. (4) Our SPAA achieves much higher attack success rates, yet remains stealthy.

## 3 Methods

### 3.1 Problem formulation

Denote $f$ as an image classifier that maps a camera-captured image $I$ to a vector of class probabilities $f(I) \in [0, 1]^N$, for $N$ classes, and denote $f_i(I) \in [0, 1]$ as the probability of the $i$-th class. Typically, *targeted* digital adversarial attacks aim to perturb $I$ by a small disturbance $\delta$ whose magnitude is bounded by a small number $\epsilon > 0$, such that a certain target $t$ (other than the true label $t_\text{true}$) has the highest probability. Similarly, *untargeted* attacks are successful as long as the classifier's output label is not the true class $t_\text{true}$:

$$\begin{aligned} &\underset{i}{\operatorname{argmax}}\, f_i(I+\delta) \begin{cases} = t & \text{targeted} \\ \neq t_\text{true} & \text{untargeted} \end{cases} \\ &\text{subject to} \quad \mathcal{D}(I, I+\delta) < \epsilon, \end{aligned} \tag{1}$$

where $\mathcal{D}$ is a distance metric measuring the similarity between two images, *e.g.*, $L_\text{p}$ norm, which also measures the perturbation stealthiness.

We extend Eqn. 1 to physical world (Figure 2) and denote the camera capture function as $\pi_c$, which maps the physical scene $s$ (*i.e.*, including all geometries and materials in the scene) and lighting $l$ to a camera-captured image $I$ by:

$$I = \pi_c(l, s) \tag{2}$$

Physical adversarial attacks aim to perturb the physical entities $s$ such that the classifier misclassifies the camera-captured image $I$ as a certain target label $t$ (or any label other than $t_\text{true}$ for untargeted attacks). By contrast, projector-based attacks aim to perturb the lighting $l$ by $\delta_l$ such that the camera-captured image causes misclassification, *i.e.*:

$$\begin{aligned} &\underset{i}{\operatorname{argmax}}\, f_i(\pi_c(l+\delta_l, s)) \begin{cases} = t, & \text{targeted} \\ \neq t_\text{true} & \text{untargeted} \end{cases} \\ &\text{subject to} \quad \mathcal{D}\left(\pi_c(l+\delta_l, s), \pi_c(l, s)\right) < \epsilon \end{aligned} \tag{3}$$

In this paper, $\delta_l$ is illumination perturbation from a projector. Denote the projector's projection function and input image as $\pi_p$ and $x$, respectively. Then, the illumination generated by the projector is given by $\delta_l = \pi_p(x)$, and the camera-captured scene under superimposed projection is given by $I_x = \pi_c(l + \pi_p(x), s)$. Denote the composite project-and-capture process above (*i.e.*, $\pi_c$ and $\pi_p$) as $\pi : x \mapsto I_x$, then the camera-captured scene under superimposed projection is:

$$I_x = \pi(x, l, s) \tag{4}$$

Finally, projector-based adversarial attack is to find a projector input adversarial image $x'$ such that:

$$\begin{aligned} &\underset{i}{\operatorname{argmax}}\, f_i\left(I_{x'} = \pi(x', l, s)\right) \begin{cases} = t, & \text{targeted} \\ \neq t_\text{true} & \text{untargeted} \end{cases} \\ &\text{subject to} \quad \mathcal{D}\left(I_{x'}, I_{x_0}\right) < \epsilon, \end{aligned} \tag{5}$$

where $x_0$ is a null projector input image.

This optimization problem involves the real project-and-capture process $\pi$, and it has no analytical gradient. Theoretically, we can compute numerical gradient instead, but it is extremely inefficient, *e.g.*, for a $256 \times 256$ projector resolution, $256 \times 256 \times 3$ real project-and-capture processes are required to compute the Jacobian matrix for a single gradient descent step. To avoid gradient computation and reduce project-and-capture processes, Nichols and Jasper [30] include $\pi$ in a gradient-free optimization (*e.g.*, differential evolution)

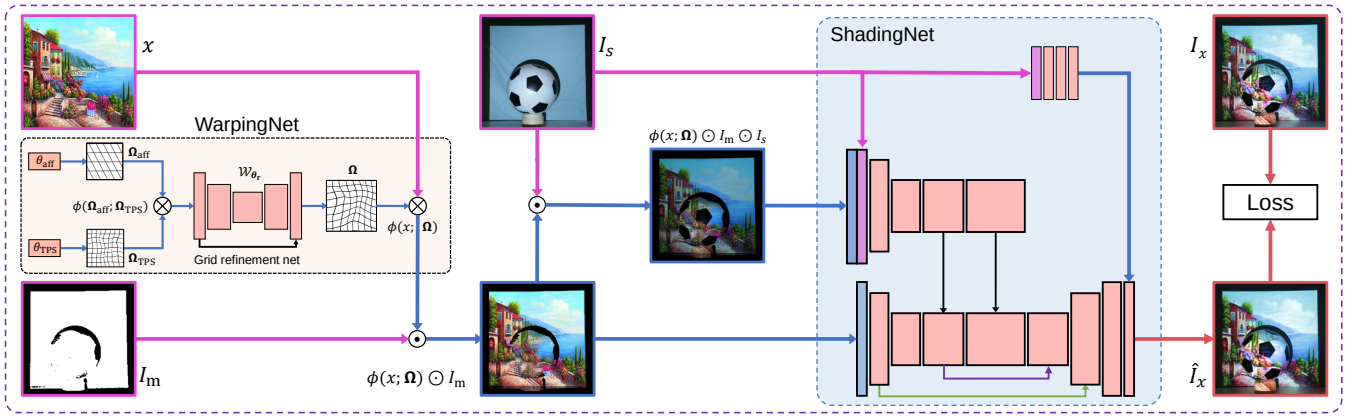


Figure 3: PCNet $\hat{\pi}$ architecture and training. PCNet approximates the real project-and-capture process $\pi$ using a deep neural network (WarpingNet + ShadingNet). The inputs are a projector input image $x$, a camera-capture scene image (under normal light) $I_s$, and a projector direct light mask $I_{\text{m}}$. The output $\hat{I}_x$ is an inferred camera-captured scene (under superimposed projection). **WarpingNet** consists of a learnable affine matrix $\theta_{\text{aff}}$, thin-plate-spline (TPS) parameters $\theta_{\text{TPS}}$ and a grid refinement network $\mathcal{W}_{\theta_r}$. This coarse-to-fine pipeline allows WarpingNet to learn a fine-grained image sampling grid $\Omega$ to warp the projector input image $x$ to the camera's canonical frontal view by $\phi(x, \Omega)$, where $\phi(\cdot;\cdot)$ is a differentiable image interpolator [16] denoted as $\otimes$. Then, we use the input projector direct light mask $I_{\text{m}}$ to exclude occluded pixels by $\phi(x, \Omega) \odot I_{\text{m}}$, where $\odot$ is element-wise multiplication. Afterwards, this warped projector image is further used to compute an intermediate rough shading image $\phi(x, \Omega) \odot I_{\text{m}} \odot I_s$ to enforce the occlusion constraint. **ShadingNet** has a two-branch encoder-decoder structure to capture complex photometric transformations. In particular, it concatenates $I_s$ and $\phi(x, \Omega) \odot I_{\text{m}} \odot I_s$ and feeds them to the middle encoder branch. Similarly, $\phi(x, \Omega) \odot I_{\text{m}}$ is fed to the backbone encoder branch. The skip connections between the two branches model photometric interactions between the three inputs at different levels. In addition, we pass $I_s$ to the output layer through three convolutional layers. Finally, the feature maps are fused into one inferred camera-captured scene (under superimposed projection) $\hat{I}_x$ by the backbone decoder.

and only perturb one projector pixel. However, even for a low resolution image (*e.g.*, $32 \times 32$), hundreds of real project-and-capture processes are required for a single targeted attack, let alone for higher resolutions. Moreover, because only one-pixel perturbation is allowed, this method also suffers from low attack success rates when image resolution increases.

Another intuitive solution is to digitally attack the camera-captured scene image under normal light first, *i.e.*, $I_{x_0} + \delta$ (Eqn. 1), then use a projector compensation method, *e.g.*, CompenNet++ [15], to find its corresponding projector input image by: $x' = \pi^\dagger(I_{x_0} + \delta)$, where $\pi^\dagger : I_x \mapsto x$ (named CompenNet++) is the pseudo-inverse of $\pi$. However, digital attacks are unaware of the physical constraints of the projector-camera system (*e.g.*, dynamic ranges and occlusions), thus the generated digital adversarial image $I_{x_0} + \delta$ may contain physically implausible perturbations. Therefore, even if $\pi^\dagger$ is a perfect approximation of $\pi$'s inverse, the real camera-captured scene under superimposed projection may not match the generated digital version. Moreover, CompenNet++ cannot address occlusions and those regions may become blurry after compensation.

In this paper, we propose a more practical and accurate solution by first approximating the real project-and-capture process $\pi$ with a deep neural network, named PCNet $\hat{\pi}_\theta$ parameterized by $\theta$. Then, we substitute the real project-and-capture process $\pi$ with PCNet $\hat{\pi}$ in Eqn. 5. Finally, fixing the weights of the classifier $f$ and PCNet $\hat{\pi}$, the projector adversarial image $x'$ can be solved by optimizing Eqn. 5 using gradient descent. Our approach brings three advantages: (a) because PCNet $\hat{\pi}$ is differentiable, we can use analytical gradient to improve adversarial attack optimization efficiency; (b) Compared with two-step methods, *e.g.*, digital attack with projector compensation, PCNet can model physical constraints of the projector-camera system, thus it can produce more robust and stealthy adversarial attacks; (c) Because PCNet can be trained offline, it requires only one online project-and-capture process for stealthy projector-based attacks.

### 3.2 PCNet $\hat{\pi}$

**Formulation**. In Eqn. 5, the real project-and-capture process $\pi$ takes three inputs, *i.e.*, a projector input image $x$, the environment light $l$ and the physical scene $s$. For each setup, $l$ and $s$ remain static, and only the projector input image $x$ is varied, thus we can approximate $l$ and $s$ with a camera-captured image $I_s = I_{x_0} = \pi(x_0, l, s)$. In practice, the camera may suffer from large sensor noise under low light, thus we set $x_0$ to a plain gray image to provide some illumination, *i.e.*, $x_0 = [128, 128, 128]^{256 \times 256}$. Another practical issue is occlusion, which may jeopardize PCNet training and adversarial attack if not properly modeled. Thus, we explicitly extract a projector direct light mask $I_{\text{m}}$ using the method in [28]. Then, the camera-captured scene under superimposed projection can be approximated by:

$$\hat{I}_x = \hat{\pi}(x, I_s, I_{\text{m}}) \tag{6}$$

Apparently $\hat{\pi}$ implicitly encodes both geometric and photometric transformations between the projector input and camera-captured images, and may be learned using a general image-to-image translation network. However, previous work (*e.g.*, [15]) shows that explicitly disentangling geometry and photometry significantly improves network convergence, especially for limited training data and time.

**Network design**. As shown in Figure 3, PCNet consists of two subnets: **WarpingNet** (for geometry) and **ShadingNet** (for photometry), and this architecture is inspired by CompenNet++ [15], which uses a CNN for projector compensation by learning the *backward* mapping $\pi^\dagger : I_x \mapsto x$. By contrast, our PCNet learns the *forward* mapping (*i.e.*, $\pi : x \mapsto I_x$) from a projector input image $x$ to the camera-captured scene under superimposed projection. In addition, CompenNet++ is designed for smooth surfaces, and it assumes no occlusions in camera-captured images, thus it may not work well if directly applied to stealthy projector-based attacks where occlusions exist. As shown in our experiments, CompenNet++ produces strong artifacts on our setups (Figure 4), while our PCNet addresses this

**Algorithm 1:** SPAA: Stealthy Projector-based Adversarial Attack.

**Input:**
$x_0$: projector plain gray image
$I_s$: camera-captured scene under $x_0$ projection
$I_{\text{m}}$: projector direct light mask
$t$: target class
$K$: number of iterations
$p_{\text{thr}}$: threshold for adversarial confidence
$d_{\text{thr}}$: threshold for $L_2$ perturbation size
$\beta_1$: step size in minimizing adversarial loss
$\beta_2$: step size in minimizing stealthiness loss
**Output :** $x'$: projector adversarial image

Initialize $x'_0 \leftarrow x_0$
**for** $k \leftarrow 1$ **to** $K$ **do**
  $\hat{I}_{x'} \leftarrow \hat{\pi}(x'_{k-1}, I_s, I_{\text{m}})$
  $d \leftarrow \|\hat{I}_{x'} - I_s\|_2$
  **if** $f_t(\hat{I}_{x'}) < p_{thr}$ **or** $d < d_{thr}$ **then**
    $g_1 \leftarrow \alpha \nabla_{x'} f_t(\hat{I}_{x'})$ // minimize adversarial loss
    $x'_k \leftarrow x'_{k-1} + \beta_1 * \frac{g_1}{\|g_1\|_2}$
  **else**
    $g_2 \leftarrow -\nabla_{x'} d$ // minimize stealthiness loss
    $x'_k \leftarrow x'_{k-1} + \beta_2 * \frac{g_2}{\|g_2\|_2}$
  **end if**
  $x'_k \leftarrow \text{clip}(x'_k, 0, 1)$
**end for**
**return** $x' \leftarrow x'_k$ that is adversarial and has smallest $d$

issue by inputting an additional projector direct light mask $I_{\text{m}}$ to exclude occluded pixels. Moreover, we compute a rough shading image $\phi(x, \Omega) \odot I_{\text{m}} \odot I_s$ as an additional input for ShadingNet, and it brings improved performance compared with CompenNet++'s photometry part (*i.e.*, CompenNet).

Finally, for each scene $s$ under lighting $l$, given a camera-capture scene image $I_s$, a projector direct light mask $I_{\text{m}}$ and projected and captured image pairs $\{(x_i, I_{x_i})\}_{i=1}^{M}$, PCNet parameters $\theta$ (*i.e.*, pink blocks in Figure 3) can be trained using image reconstruction loss $\mathcal{L}$ (*e.g.*, pixel-wise $L_1$+SSIM loss [45]) below:

$$\theta = \underset{\theta'}{\text{argmin}} \sum_i \mathcal{L}\big(\hat{I}_{x_i} = \hat{\pi}_{\theta'}(x_i, I_s, I_{\text{m}}),\ I_{x_i}\big) \tag{7}$$

We implement PCNet using PyTorch [31] and optimize it using Adam optimizer [18] for 2,000 iterations with a batch size of 24, and it takes about 6.5 minutes to finish training on three Nvidia GeForce 1080Ti GPUs.

### 3.3 Stealthy projector-based adversarial attack

Once PCNet $\hat{\pi}$ is trained, we replace the real project-and-capture process $\pi$ in Eqn. 5 by $\hat{\pi}$ using Eqn. 6, then stealthy projector-based adversarial attacks are to find an image $x'$ such that

$$\begin{aligned} &\underset{i}{\text{argmax}}\, f_i\left(I_{x'} = \hat{\pi}(x', I_s, I_{\text{m}})\right) \begin{cases} = t, & \text{targeted} \\ \neq t_{\text{true}} & \text{untargeted} \end{cases} \\ &\text{subject to} \quad \mathcal{D}\left(I_{x'},\ I_s\right) < \epsilon \end{aligned} \tag{8}$$

Here, we choose $L_2$ norm as our image distance/stealthiness metric $\mathcal{D}$, results on other image distance metrics such as $\Delta E$ and $\Delta E + L_2$ can be found in the supplementary. Then, we propose to solve Eqn. 8 by minimizing the following loss function with gradient descent:

$$x' = \underset{x'}{\text{argmin}}\ \underbrace{\alpha f_t(I_{x'})}_{\text{adversarial loss}} + \underbrace{\|I_{x'} - I_s\|_2}_{\text{stealthiness loss}} \tag{9}$$

where $\alpha = -1$ for targeted attacks and $\alpha = 1$ for untargeted attacks.

To get higher attack success rates while remaining stealthy, we develop an optimization algorithm (Algorithm 1) that alternates between the adversarial loss and stealthiness loss in Eqn. 9. Note that our method is inspired by digital attack algorithms PerC-AL [47] and DDN [33] with the following differences: (a) PerC-AL and DDN are digital attacks while our algorithm is designed for projector-based attacks by including a deep neural network approximated project-and-capture process $\hat{\pi}$; (b) We add two hyperparameters, perturbation size threshold $d_{\text{thr}}$ and adversarial confidence threshold $p_{\text{thr}}$ to improve transferability from $\hat{\pi}$ to $\pi$. It is worth noting that we have tried simply optimizing the weighted sum of adversarial and stealthiness losses, and it led to an inferior performance compared with the alternating algorithm.

For Algorithm 1, we initialize $x'$ with a projector plain gray image $x_0$ and run optimization for $K = 50$ iterations. After experiments on different settings, we set the step sizes to $\beta_1 = 2, \beta_2 = 1$. The adversarial confidence threshold is set to $p_{\text{thr}} = 0.9$ and the perturbation size threshold $d_{\text{thr}}$ is varied from 5 to 11 (§ 4.3). Note that Algorithm 1 is highly parallelizable and multi-classifier and multi-targeted attacks can simultaneously run in batch mode.

## 4 Experimental Evaluations

### 4.1 System configurations

Our setup consists of a Canon EOS 6D camera and a ViewSonic PA503S DLP projector, as shown in Figure 1. Their resolutions are set to $320 \times 240$ and $800 \times 600$, respectively. The projector input image resolution is set to $256 \times 256$. The distance between the projector-camera pair and the target object is around 1.5 meters.

Note that PCNet is trained/tested individually for each setup. We capture 13 different setups with various objects (see supplementary). For each setup, we first capture a scene image $I_s$ and two shifted checkerboard patterns to extract the scene direct illumination component using the method in [28], and obtain the projector direct light mask $I_{\text{m}}$ by thresholding the direct illumination component. Then, we capture $M = 500$ sampling image pairs $\{(x_i, I_{x_i})\}_{i=1}^{M}$ (took 3 minutes) for training PCNet $\hat{\pi}$. Afterwards, for each setup we apply Algorithm 1 to ten projector-based targeted attacks and one untargeted attack on three classifiers *i.e.*, ResNet-18 [14], VGG-16 [37] and Inception v3 [40]. In total, it takes 34 seconds to generate the adversarial projection patterns and another 17 seconds to project and capture all of them.

### 4.2 Evaluation benchmark

We evaluate stealthy projector-based attack methods by targeted and untargeted attack success rates and stealthiness measured by similarities between the camera-capture scene $I_s$ and the camera-captured scene under adversarial projection $I_{x'}$ using $L_2$ norm, $L_\infty$ norm, perceptual color distance $\Delta E$ [24] and SSIM [42].

We first compare with the gradient-free differential evolution (DE)-based baseline [30], named *One-pixel DE*, which only alters one projector pixel. Originally, it was designed for attacking classifiers trained on $32 \times 32$ CIFAR-10 [19] images, with both the projector and camera resolutions set to $32 \times 32$ as well. However, as shown in the last three rows of Table 1, the top-1 targeted attack success rates are 0, meaning that in our higher resolution setups, this method failed to fool the three classifiers (ResNet-18 [14], VGG-16 [37] and Inception v3 [40]) trained on ImageNet [7]. To increase its attack success rates, we increase the original perturbed projector pixel size from $1 \times 1$ to $41 \times 41$, and then we see a few successful untargeted attacks. In terms of efficiency, we use the same DE parameters as [30], and it takes one minute to attack a single image and 33 minutes to attack three classifiers in total, while our method only takes 10 minutes including PCNet training, adversarial attack and real project-and-capture. Note that our method can simultaneously attack multiple classifiers and targets while *One-pixel DE* involves a

Table 1: Quantitative comparison of projector-based adversarial attacks on Inception v3 [40], ResNet-18 [14] and VGG-16 [37]. Results are averaged on 13 setups. The top section shows our SPAA results with different thresholds for $L_2$ perturbation size $d_{\text{thr}}$ as mentioned in Algorithm 1. The bottom section shows two baselines *i.e.*, PerC-AL+CompenNet++ [15, 47] and One-pixel DE [30]. The 4th to 6th columns are targeted (T) and untargeted (U) attack success rates, and the last four columns are stealthiness metrics. Please see **supplementary** for more results.

| | | **Classifier** | **T. top-1 (%)** | **T. top-5 (%)** | **U. top-1 (%)** | $L_2 \downarrow$ | $L_\infty \downarrow$ | $\Delta E \downarrow$ | **SSIM**$\uparrow$ |
|---|---|---|---|---|---|---|---|---|---|
| Our method | $d_{\text{thr}} = 5$ | Inception v3 | 41.54 | 67.69 | 84.62 | 6.273 | 5.101 | 2.588 | 0.937 |
| | | ResNet-18 | 73.08 | 90.00 | 100.00 | 6.304 | 5.158 | 2.701 | 0.940 |
| | | VGG-16 | 69.23 | 83.85 | 100.00 | 6.629 | 5.428 | 2.824 | 0.934 |
| | $d_{\text{thr}} = 7$ | Inception v3 | 67.69 | 84.62 | 100.00 | 7.603 | 6.199 | 3.135 | 0.904 |
| | | ResNet-18 | 92.31 | 94.62 | 100.00 | 7.786 | 6.396 | 3.349 | 0.907 |
| | | VGG-16 | 83.08 | 97.69 | 100.00 | 8.117 | 6.668 | 3.435 | 0.899 |
| | $d_{\text{thr}} = 9$ | Inception v3 | 76.15 | 90.00 | 100.00 | 9.336 | 7.620 | 3.766 | 0.872 |
| | | ResNet-18 | 95.38 | 98.46 | 100.00 | 9.640 | 7.923 | 4.066 | 0.874 |
| | | VGG-16 | 90.00 | 99.23 | 100.00 | 9.978 | 8.211 | 4.156 | 0.864 |
| | $d_{\text{thr}} = 11$ | Inception v3 | 76.92 | 92.31 | 100.00 | 11.190 | 9.156 | 4.386 | 0.843 |
| | | ResNet-18 | 97.69 | 100.00 | 100.00 | 11.605 | 9.545 | 4.785 | 0.846 |
| | | VGG-16 | 94.62 | 99.23 | 100.00 | 11.750 | 9.671 | 4.784 | 0.835 |
| Baselines | PerC-AL+CompenNet++ [15, 47] | Inception v3 | 20.00 | 42.31 | 84.62 | 7.430 | 6.006 | 2.690 | 0.949 |
| | | ResNet-18 | 40.77 | 52.31 | 100.00 | 7.713 | 6.249 | 2.823 | 0.943 |
| | | VGG-16 | 33.85 | 49.23 | 100.00 | 7.526 | 6.099 | 2.753 | 0.946 |
| | One-pixel DE [30] | Inception v3 | 0.00 | 1.54 | 15.38 | 8.388 | 6.550 | 2.460 | 0.973 |
| | | ResNet-18 | 0.00 | 0.00 | 7.69 | 8.034 | 6.276 | 2.401 | 0.976 |
| | | VGG-16 | 0.00 | 1.54 | 23.08 | 8.233 | 6.410 | 2.473 | 0.975 |

non-parallelizable real project-and-capture process, and this advantage may become more significant when the numbers of adversarial targets and classifiers increase.

We then compare with a two-step baseline that first performs digital attacks on the camera-captured image by $\hat{I}_x = I_s + \delta$. For this step, we adapt the state-of-the-art PerC-AL [47] to our projector-based attack problem. The original PerC-AL assumes a just sufficient adversarial effect, *i.e.*, the generated digital adversarial examples just successfully fool the classifiers without pursuing a higher adversarial confidence. However, in our task, these examples failed to fool the classifiers after real project-and-capture processes, due to the complex physical environment and the image capture process of projector-based attacks. Thus, similar to our SPAA, we add an adversarial confidence threshold $p_{\text{thr}}$ to PerC-AL's optimization to allow this algorithm to pursue a more robust adversarial attack, *i.e.*, a digital adversarial example is adversarial only when its probability is greater than $p_{\text{ptr}}$. Then we use CompenNet++ [15] to find the corresponding projector adversarial image $x' = \pi^\dagger(\hat{I}_x, I_s)$. In practice, CompenNet++ is trained using the same sampling image pairs as PCNet, but with the network input and output swapped. Moreover, unlike PCNet, CompenNet++ does not use occlusion mask $I_m$ or compute a rough shading image. We name this method *PerC-AL + CompenNet++*. Note that we do not compare with [36, 48] because they are specifically designed for faces only.

**Quantitative comparisons**. As shown in Table 1, the proposed SPAA significantly outperforms *One-pixel DE* [30] and the two-step *PerC-AL + CompenNet++* [15, 47] by having higher attack success rates (the 4th to 6th columns of Table 1) and stealthiness. Note that *One-pixel DE* has very low targeted attack success rates, because it only perturbs a $41 \times 41$ projector image block, and such camera-captured images have strong square patterns (see the 3rd row of Figure 4) that are clearly far from the adversarial target image distributions, they are also less stealthy. In our experiments, we find this method can reduce the confidence of the true label (untargeted attacks) but can rarely increase the probability of a specific adversarial target. Moreover, digital targeted attacks on classifiers trained on ImageNet ($224 \times 224$, 1,000 classes) are already much harder than those trained on CIFAR-10 ($32 \times 32$, 10 classes), due to higher image resolutions and 100 times more classes, let alone applying it to the more challenging stealthy projector-based attacks. By contrast, our SPAA and *PerC-AL + CompenNet++* have higher success rates and stealthiness than *One-pixel DE*. These results are also shown in qualitative comparisons below.

**Qualitative comparisons**. Exemplar projector-based *targeted* and *untargeted* adversarial attack results are shown in Figure 4 and Figure 5, respectively. In Figure 4, clearly our method can achieve successful attacks while remaining stealthy. *PerC-AL + CompenNet++* failed this targeted attack, and we see two particular problems: **(1)** it produces a blurry bucket-like projection pattern (2nd row, 2nd column), because CompenNet++ cannot learn compensation well under occlusions. Thus, when the adversarial pattern is projected to the scene, we see large dark artifacts on the bucket (2nd row, 4th-5th columns). By contrast, our SPAA addresses occlusions by computing a projector direct light mask, then explicitly generates a rough shading image to enforce the occlusion constraint. Clearly, our generated adversarial projections (1st row, 2nd column) show much weaker artifacts. **(2)** We also see strong adversarial patterns in the bucket shadow (2nd row, 3rd column), however, the projector is unable to project to this occluded region. This is caused by the first step that performs a digital attack by $\hat{I}_x = I_s + \delta$. Without any prior knowledge about the real project-and-capture process, this step may generate physically implausible adversarial patterns like this. By contrast, our SPAA uses an end-to-end differentiable formulation, with which we include a neural network approximated project-and-capture process, *i.e.*, PCNet in the projector-based attack optimization. Then, physical constraints are explicitly applied, such that the generated adversarial pattern is physically plausible. Thus, we do not see undesired adversarial patterns in the bucket shadow of the 1st row, 3rd column.

For untargeted attacks, as shown in the 4th column of Figure 5, all three methods successfully fooled Inception v3 [40], as the classifier predicted labels are NOT **lotion**. In addition, compared with

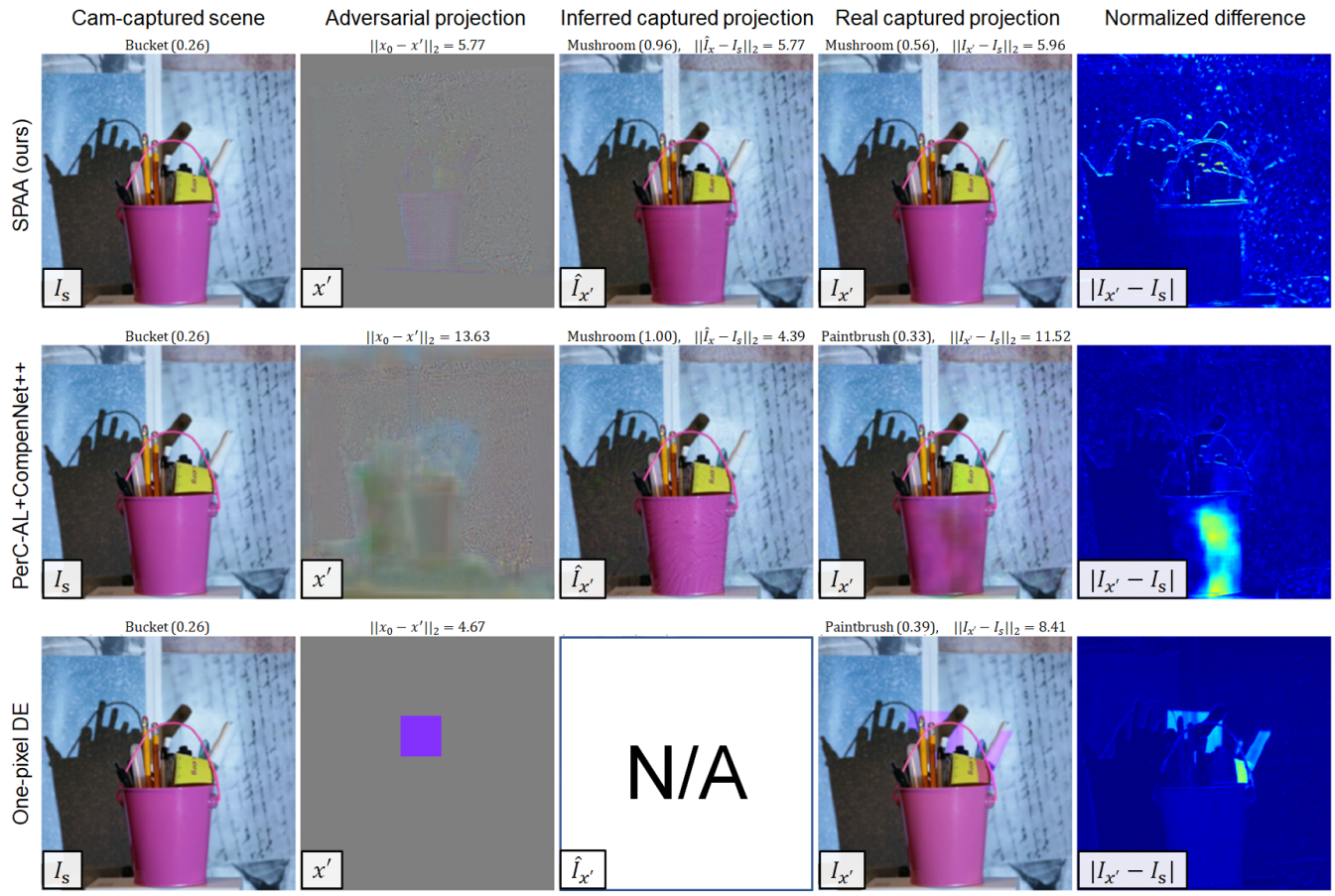


Figure 4: **Targeted projector-based adversarial attack on VGG-16**. The goal is to use adversarial projections to cause VGG-16 to misclassify the camera-captured scene as **mushroom**. The 1st to the 3rd rows are our SPAA, PerC-AL + CompenNet++ [15, 47] and One-pixel DE [30], respectively. The 1st column shows the camera-capture scene under plain gray illumination. The 2nd column shows inferred projector input adversarial patterns. The 3rd column plots model inferred camera-captured images. The 4th column presents real captured scene under adversarial projection *i.e.*, the 2nd column projected onto the 1st column. The last column provides normalized differences between the 4th and 1st columns. On the top of each camera-captured image, we show the classifier's predicted labels and probabilities. For the 2nd to 4th columns, we also show the $L_2$ norm of perturbations. Note that for One-pixel DE, the 3rd column is blank because it is an online method and no inference is available. Note that both baselines fail in this *targeted* attack. Please see **supplementary** for more results.

the two baselines, our method has the smallest perturbation size ($L_2$ norm is 4.33), and the projected adversarial image (the 2nd column) and camera-captured adversarial projection (the 4th column) are also stealthier. More untargeted attack results can be found in **supplementary** Figures 14-26, where *One-pixel DE* [30] shows successful untargeted attacks in Figures 14 and 16. For other scenes, although *One-pixel DE* [30] failed untargeted attacks, it decreases the classifiers' confidence of the true labels.

### 4.3 Ablation study

In this section, we study the proposed SPAA's success rates with different perturbation size thresholds ($d_{\text{thr}}$) and the effectiveness of PCNet's direct light mask and rough shading image.

**Perturbation size threshold** $d_{\text{thr}}$ is the minimum perturbations of the PCNet $\hat{\pi}$ inferred camera-captured scene under adversarial projection. As shown in Algorithm 1, a higher $d_{\text{thr}}$ can lead to a stronger adversary and higher projector-based attack success rates. In Table 1, we show different $d_{\text{thr}}$ ranging from 5 to 11. Clearly, attack success rates and real camera-captured perturbation sizes (*i.e.*, $L_2$, $L_\infty$, $\Delta E$ and SSIM) increase as $d_{\text{thr}}$ increases. Thus, it controls the trade-off between projector-based attack success rates and stealthiness.

**PCNet direct light mask and rough shading image.** For each setup, we project and capture 200 colorful and textured images $x$, then we compare the similarities between the real camera-captured

Table 2: Quantitative comparisons between **PCNet** and PCNet without the direct light mask and rough shading image (**PCNet w/o mask and rough**). The image similarity metrics below are calculated between the real camera-captured scene under adversarial projection $I_x$ (GT) and the model inferred camera-captured scene under adversarial projection $\hat{I}_x$. Results are averaged on 13 setups.

| **Model name** | $L_2 \downarrow$ | $L_\infty \downarrow$ | $\Delta E \downarrow$ | **SSIM**$\uparrow$ |
|---|---|---|---|---|
| PCNet | 10.461 | 8.408 | 3.066 | 0.947 |
| PCNet w/o mask and rough | 11.952 | 9.567 | 3.385 | 0.932 |

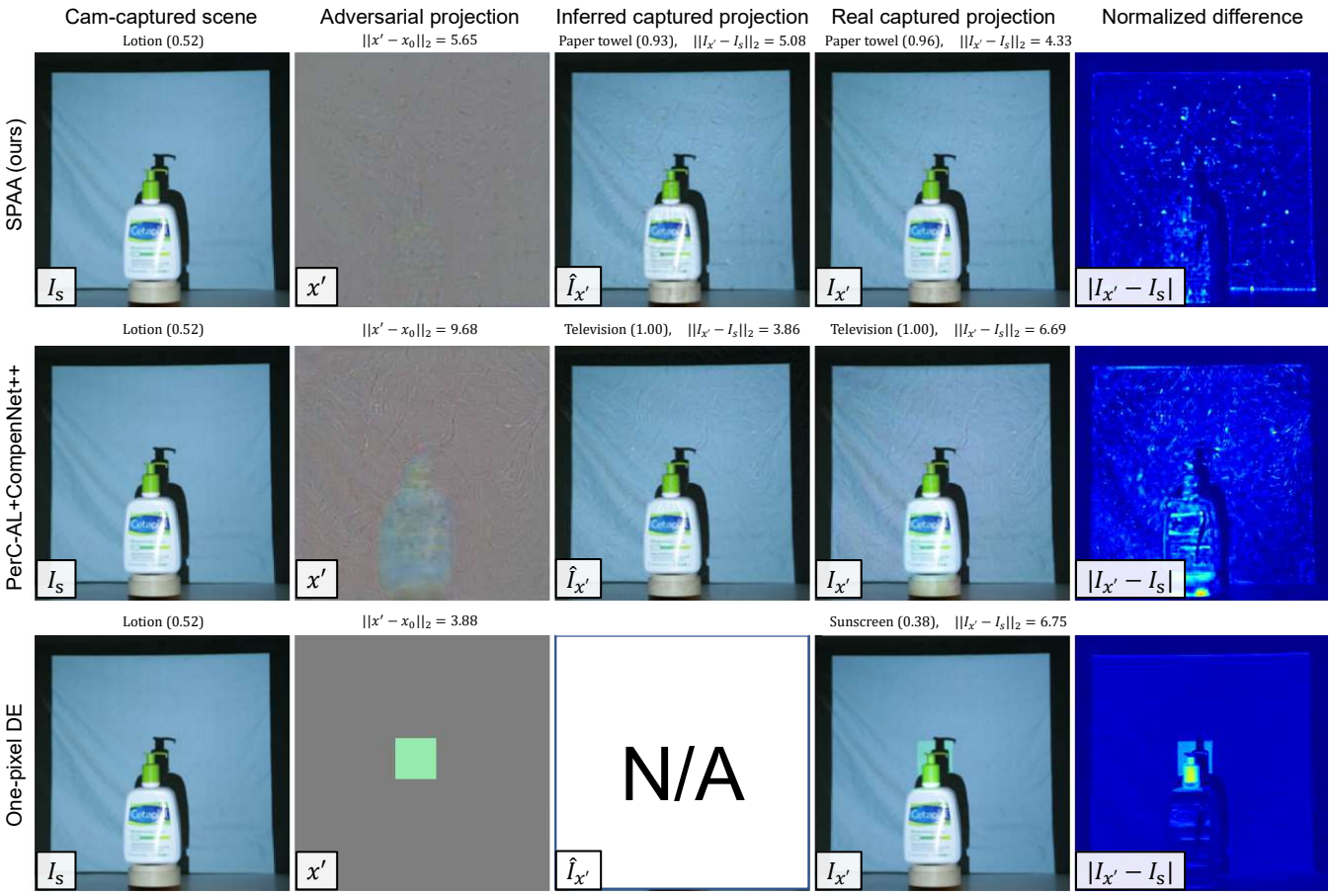


Figure 5: **Untargeted projector-based adversarial attack on Inception v3**. The goal is to use adversarial projections to cause Inception v3 to misclassify the camera-captured scene as any label other than **lotion**. The 1st to the 3rd rows are our SPAA, PerC-AL + CompenNet++ [15, 47] and One-pixel DE [30]. On the top of each camera-captured image, we show the classifier's predicted labels and probabilities. For the 2nd to 4th columns, we also show the $L_2$ norm of perturbations. Note that for One-pixel DE, the 3rd column is blank because it is an online method and no inference is available. See **supplementary** for more results.

scene under adversarial projection $I_x$ and PCNet inferred camera-captured scene under adversarial projection $\hat{I}_x$ using $L_2$ norm, $L_\infty$ norm, $\Delta E$ and SSIM. The results are shown in Table 2 and PCNet outperforms the degraded version that is without direct light mask and rough shading image, demonstrating that we need to model the essential factors, *i.e.*, direct light mask and rough shading image for better project-and-capture approximation. Ablation study on different stealthiness loss functions can be found in supplementary.

## 5 Conclusion

In this paper, for the first time, we formulate stealthy projector-based adversarial attack as an end-to-end differentiable process, and propose a solution named SPAA (Stealthy Projector-based Adversarial Attack). In SPAA, we approximate the real project-and-capture process using a deep neural network named PCNet (Project-And-Capture Network) that provides additional constraints for adversarial attack optimization, such that the generated adversarial projection is physically plausible. In addition, we propose an algorithm to alternate between the adversarial loss and stealthiness loss using minimum perturbation and adversarial confidence thresholds. In our thorough experiments, SPAA significantly outperforms other methods by significantly higher attack success rates and stealthiness, for both targeted and untargeted attacks.

**Limitations and future work.** Although our PCNet can better model the project-and-capture process than CompenNet++ [15], it is not perfect, and we can see some discrepancies between the simulated and the real attacks in Figure 1 (d) and (e). In future work, we can improve PCNet by incorporating physically based rendering domain knowledge in network design. Another limitation of our SPAA is its sensitivity to environment light, and improving its robustness under different light conditions is also an interesting direction to explore in the future.

**Acknowledgements.** This work was supported in part by Fundamental Research Funds for the Central Universities (SWU122001).

## REFERENCES


[1] A. Al Arafat, Z. Guo, and A. Awad. Vr-spy: A side-channel attack on virtual key-logging in vr headsets. In *IEEE VR*, pp. 564–572. IEEE, 2021.

[2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *ICML*, pp. 284–293. PMLR, 2018.

[3] O. Bimber, D. Iwai, G. Wetzstein, and A. Grundhöfer. The visual computing of projector-camera systems. In *Computer Graphics Forum*, p. 84, 2008.

[4] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. *NeurIPS*, 2017.

[5] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *S&P*, pp. 39–57. IEEE, 2017.

[6] B. David-John, D. Hosfelt, K. Butler, and E. Jain. A privacy-preserving approach to streaming eye-tracking data. *TVCG*, 27(5):2555–2565, 2021.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

[8] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *CVPR*, pp. 9185–9193, 2018.

[9] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *CVPR*, pp. 1000–1008, 2020.

[10] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, pp. 1625–1634, 2018.

[11] C. George, M. Khamis, D. Buschek, and H. Hussmann. Investigating the third dimension for authentication in immersive virtual reality and in the real world. In *IEEE VR*, pp. 277–285. IEEE, 2019.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[13] A. Grundhöfer and D. Iwai. Robust, error-tolerant photometric projector compensation. *IEEE TIP*, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

[15] B. Huang and H. Ling. Compennet++: End-to-end full projector compensation. In *ICCV*, 2019.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015.

[17] S. T. Jan, J. Messou, Y.-C. Lin, J.-B. Huang, and G. Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *AAAI*, vol. 33, pp. 962–969, 2019.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[19] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[20] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ICLR-W*, 2017.

[21] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ICLR*, 2017.

[22] H. Li, Y. Wang, X. Xie, Y. Liu, S. Wang, R. Wan, L.-P. Chau, and A. C. Kot. Light can hack your face! black-box backdoor attack on face recognition systems. *arXiv preprint arXiv:2009.06996*, 2020.

[23] J. Li, F. R. Schmidt, and J. Z. Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. *ICML*, 2019.

[24] M. R. Luo, G. Cui, and B. Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350, 2001. doi: 10.1002/col.1049

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

[26] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, pp. 1765–1773, 2017.

[27] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pp. 2574–2582, 2016.

[28] S. K. Nayar, G. Krishnan, M. D. Grossberg, and R. Raskar. Fast separation of direct and global components of a scene using high frequency illumination. In *ACM Trans. Graph.*, vol. 25, pp. 935–944. ACM, 2006.

[29] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang. Adversarial light projection attacks on face recognition systems: A feasibility study. In *CVPR-W*, pp. 814–815, 2020.

[30] N. Nichols and R. Jasper. Projecting trouble: Light based adversarial attacks on deep learning classifiers. In *AAAI Fall Symposium: ALEC*, 2018.

[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NeurIPS-W*, 2017.

[32] F. Roesner, T. Kohno, and D. Molnar. Security and privacy for augmented reality systems. *Communications of the ACM*, 57(4):88–96, 2014.

[33] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *CVPR*, pp. 4322–4330, 2019.

[34] Y. A. Sekhavat. Privacy preserving cloth try-on using mobile augmented reality. *IEEE Transactions on Multimedia*, 19(5):1041–1049, 2016.

[35] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *ACM SIGSAC*, pp. 1528–1540, 2016.

[36] M. Shen, Z. Liao, L. Zhu, K. Xu, and X. Du. Vla: A practical visible light-based attack on face recognition systems in physical world. *ACM IMWUT*, 3(3):1–19, 2019.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[38] R. Storn and K. Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[39] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pp. 2818–2826, 2016.

[41] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2014.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

[43] Z. Wu, S.-N. Lim, L. Davis, and T. Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. *ECCV*, 2020.

[44] X. Zeng, C. Liu, Y.-S. Wang, W. Qiu, L. Xie, Y.-W. Tai, C.-K. Tang, and A. L. Yuille. Adversarial attacks beyond the image space. In *CVPR*, pp. 4302–4311, 2019.

[45] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE TCI*, 2017.

[46] P. Zhao, S. Liu, P.-Y. Chen, N. Hoang, K. Xu, B. Kailkhura, and X. Lin. On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method. In *ICCV*, pp. 121–130, 2019.

[47] Z. Zhao, Z. Liu, and M. Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *CVPR*, pp. 1039–1048, 2020.

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