Adversarial Imaging Pipelines

Buu Phan[1] Fahim Mannan[1] Felix Heide[1,2]
[1]Algolux [2]Princeton University

Abstract

Adversarial attacks play a critical role in understanding deep neural network predictions and improving their robustness. Existing attack methods aim to deceive convolutional neural network (CNN)-based classifiers by manipulating RGB images that are fed directly to the classifiers. However, these approaches typically neglect the influence of the camera optics and image processing pipeline (ISP) that produce the network inputs. ISPs transform RAW measurements to RGB images and traditionally are assumed to preserve adversarial patterns. In fact, these low-level pipelines can destroy, introduce or amplify adversarial patterns that can deceive a downstream detector. As a result, optimized patterns can become adversarial for the classifier after being transformed by a certain camera ISP or optical lens system but not for others. In this work, we examine and develop such an attack that deceives a specific camera ISP while leaving others intact, using the same downstream classifier. We frame this camera-specific attack as a multi-task optimization problem, relying on a differentiable approximation for the ISP itself. We validate the proposed method using recent state-of-the-art automotive hardware ISPs, achieving 92% fooling rate when attacking a specific ISP. We demonstrate physical optics attacks with 90% fooling rate for a specific camera lens.

1. Introduction

Deep neural networks have become a cornerstone method in computer vision [7, 20, 21, 25, 58] with diverse applications across fields, including safety-critical perception for self-driving vehicles, medical diagnosis, video security, medical imaging and assistive robotics. Although a wide range of high-stakes applications base their decision making on the output of deep networks, existing deep models have been shown to be susceptible to adversarial attacks on the image that the network ingests. Specifically, existing adversarial attacks perturb the input image with carefully designed patterns to deceive the model while being imperceptible to a human viewer [34, 41, 44, 48, 37, 52]. As such, understanding and exploring adversarial perturbations offer insights into the failure cases of today's models and it allows researchers to develop defense methods and models that are resilient against proposed attacks [3, 33, 34, 42, 55].

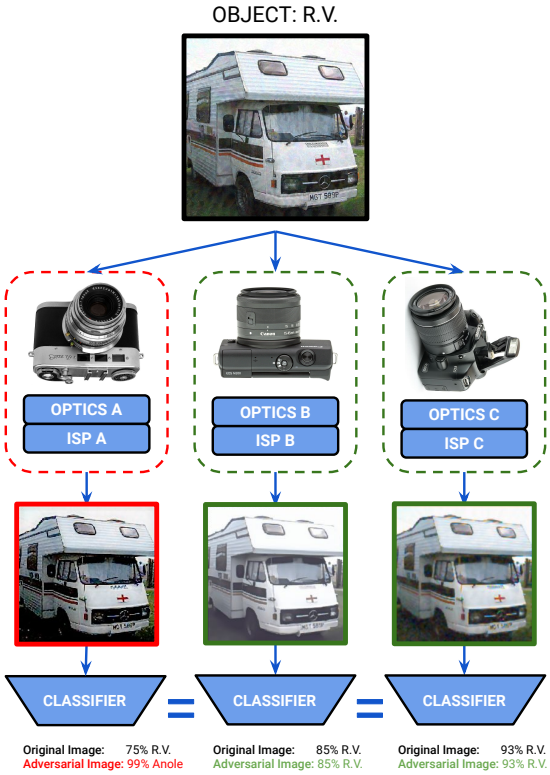


Figure 1: We illustrate and show the camera-specific attack. The image is tampered such that it becomes only adversarial for a specific camera pipeline, even when the three pipelines deploy the same classifier.

Existing adversarial attacks find post-capture adversaries, tampering with the image after capture before it is input to the deep network. Recently, a number of attack methods have been demonstrated in the form of physical objects that are placed in real-world scenes to generate adversarial patterns by capturing images of the physical objects [2, 14, 29]. The most successful methods for computing adversarial perturbations rely on network gradients to form adversarial examples [48, 18, 29, 36, 41, 4] for each input image, that struggle to transfer to other networks or images [48, 32, 39]. Alternative approaches rely

only on the network predictions [24, 38, 47] and use surrogate networks [40] or gradient approximations [1]. All of these methods, both physical and synthetic attacks, *assume that the camera image processing pipeline (ISP) preserves the attack pattern*. Although modern image processing pipelines implement complex algorithms, such as tonemapping, sharpening and denoising [27, 28], which transform RAW measurements to RGB images on embedded camera processors, the influence of this pipeline is ignored by existing attack methods. Some of the processing blocks in camera image processing pipelines have even been suggested as defenses against existing attacks [19, 31].

In this work, we close this gap between scene-based physical attacks and attacks on post-processed images. Specifically, we propose a novel method that allows us to attack cameras with a specific ISP, while leaving the detections of other cameras intact for the *identical* classifier but a different ISP. As such, the attack mechanism proposed in this work is a *camera-specific attack* that not only targets the deep network but conventional hardware ISPs that traditionally have not been considered susceptible to adversarial attacks. As a further camera-specific attack, we also attack the optical system of a camera. The proposed method can incorporate proprietary black-box ISP and complex compound optics, without accurate models, by relying on differentiable approximations as gradient oracles. We validate our method using recent automotive hardware ISP processors and automotive optics, where the novel attack achieves a fooling rate of 92% on RAW images in experimental captures.

Specifically, we make the following contributions

- We introduce the *first method for finding adversarial attacks that deceives a specific camera ISP* and optics while leaving cameras with other ISPs or optics intact although they employ the same classifier network.
- We demonstrate attacks for embedded hardware ISPs that are not differentiable and only available as black-box algorithms. To this end, we learn differentiable approximations of the image processing and sensing pipeline that serves as gradient oracles for our attack.
- We analyze and validate the attack on RAW input measurements for state-of-the-art hardware ISPs.
- We validate physical attacks of the proposed method on recent automotive camera ISPs and automotive optics, achieving more than 90% success rate.

2. Related Work

Our work considers the problem of adversarial attacks on camera pipelines. We review the relevant literature below.

Camera Image Processing Pipelines. Research on high-level vision tasks has often overlooked the existence of the low-level image signal processing (ISP) pipeline in the camera. In practice, the role of these ISPs is critical in a vision system because their ability to recover high quality images from noisy and distorted RAW measurements directly affects the downstream processing modules [22, 51]. For display applications, domain-specific image processing methods [16, 17, 5, 9, 57, 15, 22] have been successful in tackling low-light, shot-noise and optical aberrations. Unfortunately, these methods are computationally expensive, and, as such, their application is limited to off-line tasks. In contrast, real-time applications, such as robotics and augmented reality demand real-time processing at more than 30 Hz for double-digit megapixel streams. As a result, integrated system-on-chip ISPs are today employed for robotic vision systems, such as autonomous robots, self-driving vehicles, and drones. For example, the ARM Mali-C71 ASIC ISP is capable of processing 12 megapixel streams at up to 100 Hz with less than one Watt power consumption. However, although hardware ISPs are efficient, these processing pipelines are typically highly optimized proprietary compute units that are not differentiable and their behavior is unknown to the user [51]. In this work, we present the first adversarial attack that targets these hardware processing blocks, which, in contrast to deep neural networks, *traditionally have been assumed to be not susceptible to adversarial perturbations* and instead have been suggested as potential defense units [19, 31].

Adversarial Attacks. A large body of work has explored adversarial attacks on deep networks in computer vision. A common formulation describes an attack as an ℓ_p norm-ball constrained perturbation that deceives a specific classifier [34]. Depending on the knowledge of the model (*i.e.*, weights and architecture) that the adversary has, attacks can be grouped into two settings: white-box and black-box attacks. In the white-box setting, the model specification are known and the adversaries leverage it to synthesize the perturbation. By treating the attack as a solution of an optimization problem, techniques ranges from mixed-integer programming [50, 54] to 1$^{\text{st}}$-order gradient method [18, 34, 36, 48] have been proposed. Additionally, by manipulating the optimization objectives and constraints, attacks can reveal interesting properties of the target network, such as sparsity and interpretability [59, 4, 35, 56, 49]. In the black-box setting, adversaries can only query the input-output pairs, and, hence, the target model is more difficult to deceive. Nevertheless, existing approaches have shown that adversaries can successfully approximate the gradients and apply the white-box method. This is achieved by approximating the target network function [40, 1] (transfer methods) or by numerical estimation (score methods) [23, 53, 6, 30, 8]. In this work, we propose a transfer approach that approximates non-differentiable camera pipelines, including the camera

optics and ISP, with differentiable proxy functions.

Going beyond synthetically generated adversarial examples, researchers have shown to be able to recreate them in the wild by placing adversarial patterns on physical objects. Kurakin *et al.* [29] demonstrate such a physical attack by printing the digital adversarial image on paper and capturing it with a camera, assuming that the acquisition and capture pipeline itself is not susceptible to the adversarial pattern. Athalye *et al.* [2] propose an attack which optimizes the perturbation under different image augmentations, a direction further explored by a line of work [14, 26, 46, 13, 10] to achieve higher attack ratios. All of these existing methods have in common that they assume that the scene light transport and acquisition preserve the adversarial patterns, including the optics, sensors and ISP in the camera as non-susceptible image transforms. As a direct result, *existing physical attacks have failed to achieve the high fooling rates of synthetic attacks* [29]. Our work fills this gap and shows that it is possible to achieve high fooling rates when including the acquisition and processing operations in adversarial attacks. Building on this insight, we realize attacks of individual camera types by exploiting slight differences in their acquisition and image processing pipeline.

3. Background

In this section, we review the differentiable proxy framework from Tseng et al. [51] and the projected gradient descent ℓ_p norm-bounded adversarial attack [34], and we introduce relevant notation for the following sections.

3.1. Differentiable Proxy ISPs

A given non-differentiable hardware ISP is approximated by a differentiable proxy function, which implements a mapping from RAW input data to post-ISP images via a convolutional neural network (CNN). We note that this framework can also be extended to include the compound optics in the pipeline, see Supplementary Document.

Proxy ISP Model. We denote $h : \mathbb{R}^d \rightarrow \mathbb{R}^{d\times 3}$ as a black-box ISP function that maps a RAW image $x \in \mathbb{R}^d$ to an RGB image, where d is the RAW image dimension (*e.g.*, 1920×1200). The proxy ISP function $\tilde{h}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d\times 3}$ depends on θ as the learnable parameters (*i.e.*, CNN weights) also map a RAW image to a post-ISP image. As a departure from Tseng et al. [51], we found that bilinear demosaicing as a first layer in this proxy module improves training stability and accuracy. This demosaicing layer is differentiable. The demosaiced RGB image is fed into a U-Net [45], which is trained to approximate the output of the hardware ISP.

Proxy Training. Given a set of RAW captures: $X = \{x_1, x_2, ..., x_N\}$ where each $x_i \in \mathbb{R}^d$, we train the proxy function $\tilde{h}_\theta$ by minimizing the ℓ_1 reconstruction loss.

3.2. Projected Gradient Adversarial Attacks

Let us denote a probabilistic classifier that maps an input $x \in \mathbb{R}^d$ to a categorical distribution vector as $f : \mathbb{R}^d \rightarrow \mathbb{R}^K$, where d is the input dimension and K is the number of classes. We define a decision function $c(x)$, which assigns a label to x as: $c(x) = \underset{k=1,2,...,K}{\arg\max}\, f^k(x)$.

ℓ_p Norm-bounded Attack. For an input x, an additive perturbation $\delta \in \mathcal{B}_d(p;\epsilon)$ is adversarial when $c(x+\delta) = t$, where t is a target label and $\mathcal{B}_d(p;\epsilon) = \{r \in R^d : \|r\|_p < \epsilon\}$ is an ℓ_p norm-ball with radius ϵ[1]. We will use ℓ_∞ throughout this paper. To create such a perturbation, we solve the following constrained optimization problem

$$\underset{\|\delta\|_\infty \leq \epsilon}{\text{minimize}} \quad \mathcal{L}(f(x+\delta), t), \tag{1}$$

where $\mathcal{L}$ is the cross-entropy loss.

Projected Gradient Descent (PGD). In the case of ℓ_∞, we can solve (1) by first randomly initializing $\delta \in \mathcal{B}_d(\infty;\epsilon)$ and iteratively perform the following PGD update

$$\delta \leftarrow \delta - \alpha \cdot \text{sgn}(\nabla_\delta \mathcal{L}(f(x+\delta), t)). \tag{2}$$

where α is the step size that can depend on the current ordinal iteration number.

4. Camera Pipeline Adversarial Attack

In the following, we consider a camera pipeline consisting of a black-box, non-differentiable ISP followed by a downstream RGB image classifier. A direct RAW attack on such a pipeline involves manipulating the captured RAW image. For a physical camera attack, our pipeline also includes the optical system that captures an adversarial scene. In this section we only explain the direct RAW attack without any loss of generality.

Next, we describe two types of attacks on these pipelines and the method to generate them. The first type of attack, referred to as *untargeted camera attack*, aims to craft an adversarial RAW perturbation to the pipeline, without considering its transferability to the other pipelines. The second type, referred to as *targeted camera attack*, generates a perturbation that deceives a specific pipeline while leaving the other intact, even when the same classifier is deployed. Figure 2 provides an overview of the proposed targeted camera attack and corresponding proxy functions.

We define a black-box ISP function as $h : \mathbb{R}^d \rightarrow \mathbb{R}^{d\times 3}$, a trained proxy function that approximates h as $\tilde{h}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^{d\times 3}$ and an RGB image classifier as $g : \mathbb{R}^{d\times 3} \rightarrow \mathbb{R}^K$. Given a RAW image $x \in \mathbb{R}^d$, we define the camera pipelines using the original ISP and proxy ISP separately as: $f(x) = (g \circ h)(x)$ and $\tilde{f}(x) = (g \circ \tilde{h})(x)$. Similar to Section3.2, $c(x)$ and $\tilde{c}(x)$ are the corresponding decision func-

[1] As an image, $(x+\delta)$ needs to stay within the valid range (*e.g.*, [0,255] for RGB images), which can be achieved by clipping. We implicitly assume this condition throughout this paper without stating it.

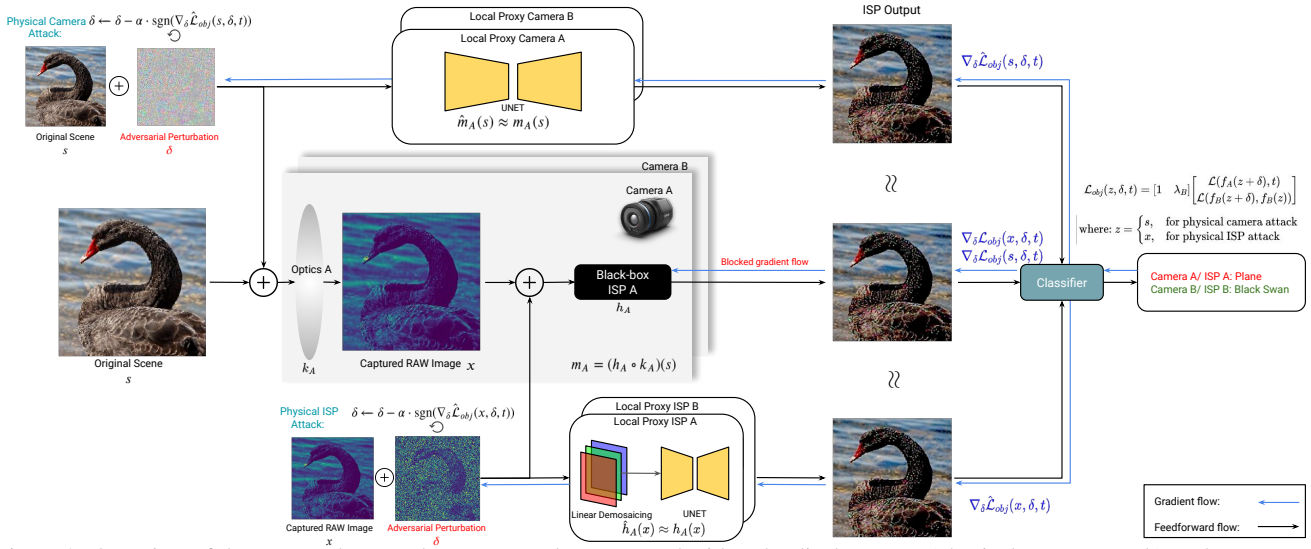


Figure 2: Overview of the proposed targeted camera attack. We perturb either the display scene (physical camera attack) or the captured RAW image (physical ISP attack), whose label is "black swan", such that they are misclassified into "plane" by pipeline A but not by pipeline B. To find such an attack, we solve an optimization problem, using the estimated gradients from the proxies approximation of the black-box, non-differentiable imaging modules. The objective function is a weighted sum of two cross-entropy losses, where the first term encourages the attack to fool pipeline A and the second term prevents it from changing the original prediction probability of pipeline B.

Algorithm 1 Local Proxy Training

Input: h; $\tilde{h}$; g; Number of augmented images M; number of attack iterations n; a list of targeted images S; a predefined bound ϵ; update step size α.

Output: A local proxy function $\hat{h}$

$\hat{S} = S$
$\tilde{f} = (g \circ \tilde{h})$
for all $x_i \in S$ **do**:
  for $m \leftarrow 1...M$ **do**:
    $\hat{\epsilon} \sim \text{uniform}(\alpha, \epsilon + \alpha)$
    $\delta \leftarrow \text{PGD}(x_i, \tilde{f}, n, \hat{\epsilon}; \alpha)$ ▷ perform n-steps PGD update, target random class
    $\hat{S} = \hat{S} \cup \{x_i + \delta, h(x_i + \delta)\}$
  end for
end for
$\hat{h} \leftarrow \text{TRAIN}(\tilde{h}, \hat{S})$ ▷ Train the local proxy $\hat{h}$ from $\hat{S}$ and $\tilde{h}$
return $\hat{h}$

tions. Before describing the two camera attacks, we next introduce the a local proxy function, which is a modification of Tseng et al. [51]'s model that is essential to the success of the proposed attack.

4.1. Local Gradient Proxies

In our experiments, we found that, despite $\tilde{h}(x)$ being perceptually similar to $h(x)$, performing the PGD-update based on the estimated gradient from $\tilde{h}$ does not result in high success rate in many cases, especially for the targeted camera attack. To this end, we propose using a local proxy model as an alternative gradient-oracle, which is trained by fine-tuning the existing proxy model $\tilde{h}$ with a set of target images and Jacobian augmentation [40]. We find that such a local proxy model effectively improves the success rate for both untargeted and targeted camera attacks.

Specifically, given an image set S that we wish to attack, we create M different Jacobian-augmented pairs: $\{(x_i + \delta_i), h(x_i + \delta_i)\}$ for each $x_i \in S$, where $\delta_i \in \mathcal{B}_d(p; \hat{\epsilon})$ is the adversarial perturbation on the proxy pipeline $\tilde{f}$ and the bounded radius $\hat{\epsilon}$ is uniformly sampled within $[\alpha, \epsilon + \alpha]$, where α is the PGD update step size. The local proxy model $\hat{h}$ is obtained by finetuning $\tilde{h}$ with the newly augmented training set $\hat{S}$. This method is formalized in Algorithm 1.

4.2. Untargeted Camera Attack

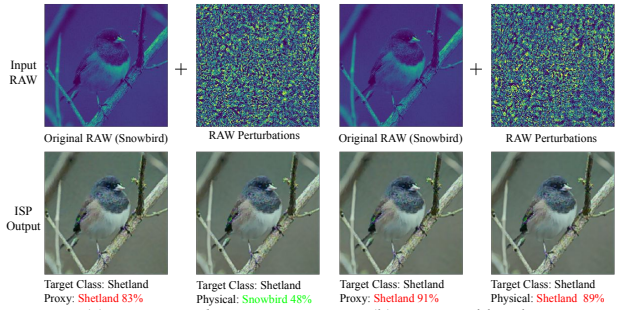
For this attack type, we aim to generate an adversarial perturbation $\delta \in \mathcal{B}_d(\infty; \epsilon)$ to a RAW image x such that: $c(x + \delta) = t$ independent of the camera pipeline. We replace the black-box ISP h with its local proxy function $\hat{h}$ and generate adversarial perturbations δ from the PGD update on $\hat{f} = g \circ \hat{h}$, that is

$$\delta \leftarrow \delta - \alpha \cdot \text{sgn}(\nabla_\delta \mathcal{L}(\hat{f}(x + \delta), t)). \quad (3)$$

We found that, despite both $\tilde{h}(x), \hat{h}(x)$ being perceptually similar to $h(x)$, estimated gradients using $\hat{h}$ consistently yields a higher success rate than $\tilde{h}$, see Supplementary Document for quantitative comparisons. We illustrate this in Figure 3, showing that being trained with different perturbations enables $\hat{h}$ to provide accurate gradients for the attack to transfer well to h.

4.3. Targeted Camera Attack

For this attack type, we find a perturbation that deceives a specific camera pipeline h, while leaving the classifications of other camera pipelines intact, even when all the pipelines deploy the same classifier g. Let h_i, for $i \in \{1, 2, ..., T\}$, be one of the ISPs that we do not want to attack, its associated camera pipeline and decision function are $f_i(x) =$



(a) Tseng *et al.* Proxy (b) Proposed local Proxy

Figure 3: Local Proxies: The proxy from Tseng *et al.* [51] does not approximate the real ISP adequately for an adversarial attack. The proposed local proxy attack successfully causes the physical pipeline to misclassify the image into the target class "Shetland", while Tseng *et al.*'s proxy fails.

$(g \circ h_i)(x)$ and $c_i(x)$. We assume that g is transferable across different ISPs, *i.e.*, accuracy higher than 70% for each ISP in the Imagenet dataset. Ideally, an adversarial perturbation $\delta \in \mathcal{B}_d(\infty; \epsilon)$ to an image x with label y should satisfy: $c(x+\delta) = t$ and $f_i(x+\delta) = f_i(x)$, given that $c(x) = y$ and every $c_i(x) = y$. Such a perturbation can be found as a solution of the following optimization problem

$$\begin{aligned} &\underset{||\delta||_p \leq \epsilon}{\text{minimize}} \quad \mathcal{L}(f(x+\delta), t) \\ &\text{s.t} \qquad\qquad f_i(x+\delta) = f_i(x), \forall i \in 1, ..., T. \end{aligned} \tag{4}$$

The problem from (4) is a challenging nonlinear-equality constrained problem that may only have a feasible solution with large cross-entropy loss.

Soft-Constrained Objective. For known and differentiable h, h_i, we can relax (4) using soft-constraints and apply the PGD update on δ to jointly minimize the objective function and the distance between $f_i(x+\delta)$ and $f_i(x)$

$$\underset{||\delta||_p \leq \epsilon}{\text{minimize}} \quad \mathcal{L}_{obj}(x, \delta, t), \tag{5}$$

where $\mathcal{L}_{obj}(x, \delta, t) = \mathcal{L}(f(x+\delta), t) + \sum_{i=1}^{T} \lambda_i \mathcal{L}(f_i(x+\delta), f_i(x))$. The second term measures the cross-entropy loss between $f_i(x+\delta)$ and $f_i(x)$ which is equivalent to minimizing the KL divergence between the two categorical distributions. We set $\lambda_i = 1, \forall i$ in our experiment.

Objective Function with Local Proxy ISP. Since h and h_i can be non-differentiable, we optimize δ on the new objective function, which replace h and h_i with their corresponding local proxy $\hat{h}$ and $\hat{h}_i$, that is

$$\hat{\mathcal{L}}_{obj}(x, \delta, t) = \mathcal{L}(\hat{f}(x+\delta), t) + \sum_{i=1}^{T} \lambda_i \mathcal{L}(\hat{f}_i(x+\delta), f_i(x)). \tag{6}$$

Similar to previous work on multi-task optimization [11, 59], we found that alternately updating δ on $\mathcal{L}(\hat{f}(x+\delta), t)$ and $\sum_{i=1}^{T} \lambda_i \mathcal{L}(\hat{f}_i(x+\delta), f_i(x))$ yields a better result, that is

$$\begin{cases} \delta \leftarrow \delta - \alpha \cdot \text{sgn}(\nabla_\delta \mathcal{L}(\hat{f}(x+\delta), t) \\ \delta \leftarrow \delta - \alpha \cdot \text{sgn}(\nabla_\delta \sum_{i=1}^{T} \lambda_i \mathcal{L}(\hat{f}_i(x+\delta), f_i(x))) \end{cases} \tag{7}$$

Algorithm 2 Targeted Camera Adversarial Perturbation

Input: Targeted ISP h; Untargeted ISPs $\{h_1, h_2, ..., h_T\}$; Pre-trained local proxies: $\hat{h}, \{\hat{h}_1, \hat{h}_2, ..., \hat{h}_T\}$; RGB classifier g; number of attack iterations n; targeted image x; targeted class t; perturbation bound ϵ.

Output: adversarial image $x' \in \mathbb{R}^d$

// Construct the proxy pipelines :
$\hat{f} = (g \circ \hat{h})$; $\hat{f}_i = (g \circ \hat{h}_i)$
// Attack the targeted image:
$\delta \sim \text{uniform}(-\epsilon, \epsilon)$
for $k \leftarrow 1...n$ **do**:
  $\delta \leftarrow \text{clip}(x+\delta) - x$ ▷ Clip δ to the valid range
  $\delta \leftarrow \delta - \alpha \cdot \text{sgn}(\nabla_\delta \mathcal{L}(\hat{f}(x+\delta), t))$
  $\delta \leftarrow \delta - \alpha \cdot \text{sgn}(\nabla_\delta \sum_{i=1}^{T} \lambda_i \mathcal{L}(\hat{f}_i(x+\delta), f_i(x)))$
  // Alternately: $\delta \leftarrow \delta - \alpha \cdot \text{sgn}(\nabla_\delta \hat{\mathcal{L}}_{obj}(x, \delta, t))$
end for
$x' = \text{clip}(x+\delta)$ ▷ Clip $x+\delta$ to the valid range
return x'

We note that replacing h, h_i with $\tilde{h}, \tilde{h}_i$ does not give a high success rate, even for large ϵ, as the approximated gradients from $\tilde{h}, \tilde{h}_i$ are not accurate enough for multiple constraints. Finally, while the proposed objective (6) only minimizes the KL divergence between $\hat{f}_i(x+\delta)$ and $f_i(x)$, the local proxies $\hat{h}, \hat{h}_i$ were trained to indirectly minimized the distance between $\hat{f}_i(x+\delta)$ and $f_i(x+\delta)$ around the perturbation radius ϵ. We formalize this method in Algorithm 2.

5. Assessment

We validate our methods using hardware ISPs and optical assemblies for direct RAW and physical camera attacks.

5.1. Validation Experiments

Dataset. For all the experiments, we use a subset of 1,000 ImageNet validation images [12].

Image Processing Pipelines. We evaluate our method for the black-box/non-differentiable hardware ARM Mali C71 and Movidius Myriad 2 ISPs. In addition to the two hardware ISPs, we also jointly evaluate with two differentiable ISPs. The first one only performs bilinear demosaicing, and will be referred to as the Demosaicing ISP. The second one performs bilinear demosaicing operation followed by bilateral filtering [43], and referred to as Bilateral Filter ISP, see details in the Supplementary Document.

Optics. We use a Fujinon CF12.5HA-1 lens with 54° field of view as the default lens for our experiments. As this compound optics is a proprietary design, we evaluate the proposed attacks on a Cooke triplet optimized for image quality using Zemax Hammer optimization and fabricated using PMMA (more details in the Supplementary Document).

Classifier. We use a large Resnet-101 [12] classifier, which achieves 76.4% Top-1 accuracy. Since each ISP has a different set of parameters (such as white-balance coefficients, color-correction matrix, etc.), we prevent the domain-shift problem by finetuning the pretrained Resnet-101 model on

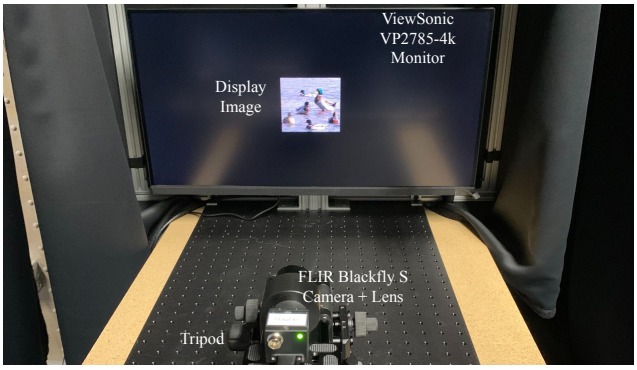


Figure 4: Setup for Evaluation of Camera-Specific Attacks. We employ a monitor placed in front of the target camera system, which is attacked by the proposed method. The proposed setup allows us to evaluate attacks on specific cameras, including their ISPs and camera optics using physical captures.

a set of ISP output images.

Evaluation Metrics. We evaluate success rate, transfer rate and targeted success rate as metrics in our evaluation. Success rate measures whether an attack for a given camera pipeline is able to change that pipeline's prediction to the target class. Transfer rate measures whether an adversarial RAW is misclassified by other pipelines. Targeted success rate measures if an attack pattern changes the targeted pipeline's prediction to the target class while leaving other camera pipelines unaffected (class prediction does not change and the confidence difference between the original and adversarial RAW is below 0.15).

5.2. Physical Setup

To validate the proposed method in a physical setup, we display the attacked images on the ViewSonic VP2785-4k monitor, as shown in Figure 4. This allows us to collect large-scale evaluation statistics in a physical setup, departing from sparse validation examples presented in existing works with RGB printouts [29, 2]. We capture images using a FLIR Blackfly S camera employing a Sony IMX249 sensor. The camera is positioned on a tripod and mounted such that the optical axis aligns with the center of the monitor. The camera and monitor are connected to a computer, which is used to jointly display and capture thousands of validation images. Each lens assembly is focused at infinity with the screen beyond the hyperfocal distance. The captured RAW image acquired by the sensor in this setup is fed to the ISPs and then resized to the resolution of 224×224 before going through the Resnet-101 classifier.

5.3. Physical ISP Attack

In this setting, we acquire the RAW images from the screen-projected images, using the Sony IMX249 camera. These RAW images are then fed to different ISPs and the adversarial perturbation are added directly to these RAW images. For each RAW image, we target a random class, choose ϵ=2000 for reliable success rates and use a total of 30 iterations, with α=50.

Targeted ISP \ Deployed ISP	Movidius Myriad 2	ARM Mali C71	Bilateral Filter ISP	Demosaicing ISP
Movidius Myriad 2	**93.5%**	28.4% ‖ 50.6%	31.3 % ‖ 53.6%	39.7% ‖ 59.4%
ARM Mali C71	34.3%‖46.1%	**94.4%**	14.5%‖30.3%	18.7%‖44.3%
Bilateral Filter ISP	29.8%‖44.8%	18.9%‖32.4%	**97.3%**	94.3%‖97.3%
Demosaicing ISP	25.2%‖45.3%	23.3%‖35.1%	40.4%‖66.6%	**98.2%**

(a) Untargeted Physical ISP Attack

Targeted ISP \ Deployed ISP	Movidius Myriad 2	ARM Mali C71	Bilateral Filter ISP	Demosaicing ISP
Movidius Myriad 2	**92.2%**	4.3% ‖ 5.4%	0.0 % ‖ 0.0%	0.0 % ‖ 0.0%
ARM Mali C71	4.8%‖7.1%	**93.8%**	0.0 % ‖ 0.0%	0.0 % ‖ 0.0%
Bilateral Filter ISP	4.7%‖6.9%	4.5%‖5.6%	**97.3%**	0.0 % ‖ 0.0%
Demosaicing ISP	4.8%‖7.0%	4.2%‖5.1%	0.0 % ‖ 0.0%	**98.2%**

(b) Targeted Physical ISP Attack

Table 1: Success and transfer rate for the proposed (a) untargeted and (b) targeted physical ISP attack. Each row shows the attack success rate on the targeted ISP (diagonal cells) and transfer rate to other ISPs (non-diagonal cells[2]). The proposed targeted method significantly reduces the transfer rate across different ISPs.

Targeted Optics \ Deployed Optics	Fujinon CF12.5HA-1	Cooke Triplet
Fujinon CF12.5HA-1	**90.7%**	4.5%‖7.9%
Cooke Triplet	5.2%‖8.1%	**91.5%**

Table 2: Success and transfer rate for the targeted physical optics attack. Refer to Table 1 for table notation.

Untargeted Camera Attack. We measure the transferability of the untargeted camera attack described in Section 4.2 in Table 1a. We observe that the attacks on one ISP are more transferable to certain ISPs, *e.g.*, attacks on the ARM Mali C71 ISP are more transferable to the Movidius Myriad 2 than the two differentiable ones. Also, attacks on the Bilateral Filter ISP and Demosaicing ISP are likely to be transferable to each other, but not to the hardware ISPs.

Targeted Camera Attack. Next, we use the proposed targeted attack Algorithm 2 as an attack that only comes into effect when fed into a specific pipeline. We show the result in Table 1b, where our method significantly reduces the transfer rate across different ISPs. For each targeted black-box ISP attack, it reduces the transfer rate of the Bilateral Filter ISP and Demosaicing ISP to 0.0%, and the transfer rate to the other black-box ISP is reduced to below 8.0%.

Figure 5 shows the adversarial RAW images, perturbations (targeting different ISPs) and their associated ISP outputs. Interestingly, despite having the same adversarial RAW image as input, each ISP produces distinct RGB perturbations. For example, in the attack on Movidius Myriad 2, unlike other ISPs, the ARM Mali C71 suppresses the perturbation around the top left black regions. Also, while the output RGB perturbations seem to contain similar macro structures, only the one from the targeted ISP becomes adversarial to the classifier, while others pose no threat at all. Since the untargeted RGB perturbations do not change the prediction, it means that they are considered as noise by some hidden projections in the classifier. As such, the perturbations are specifically tailored to a specific ISP. In general, for each targeted ISP, our method is able to deceive the

[2]The first entry in the non-diagonal cell is the transfer rate. The second entry measures the percentage of images whose confidence for the adversarial image significantly differs from the adversarial-free image (if their confidence difference is greater than 0.15).

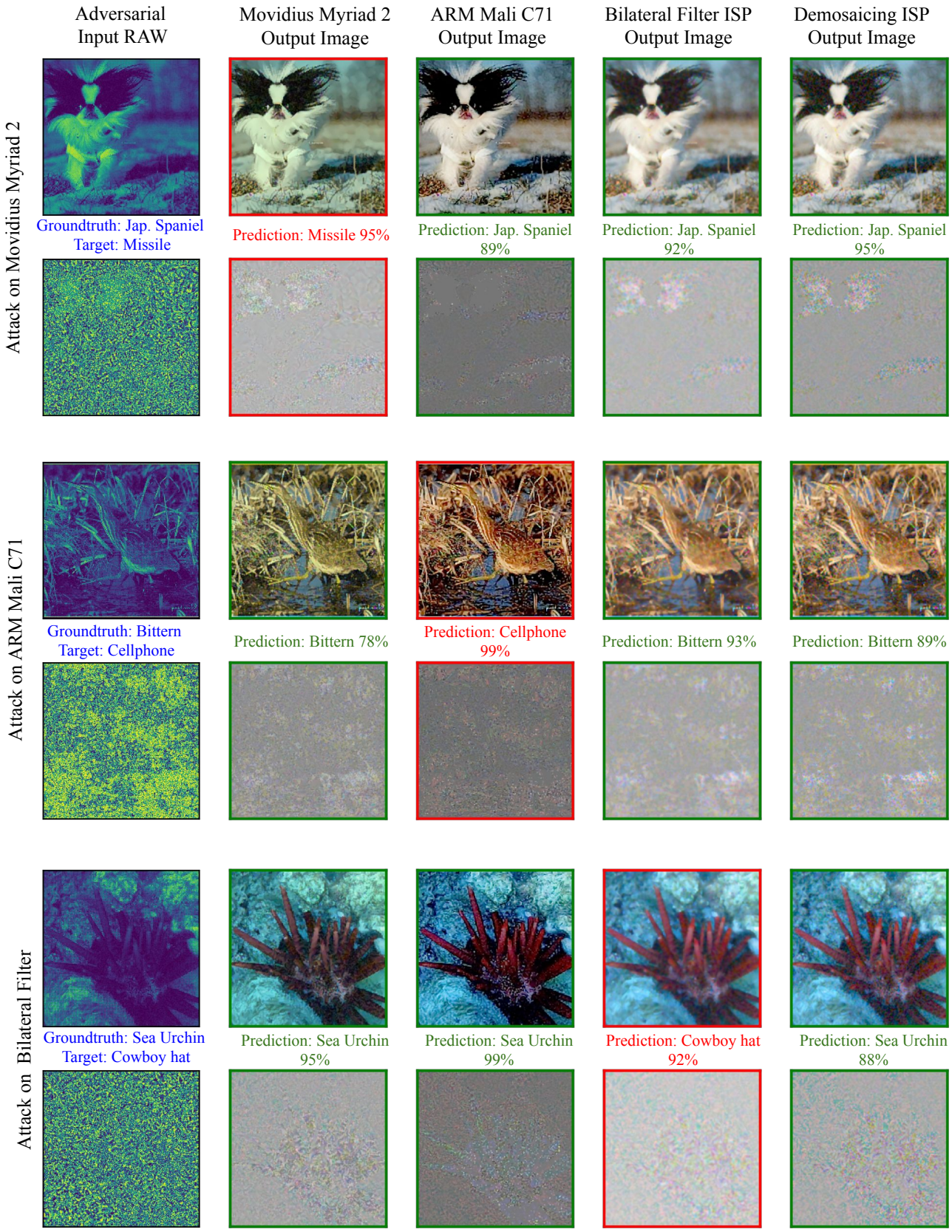


Figure 5: Visualization of the adversarial images and perturbations for the targeted ISP attack. Each pair of rows shows the attack on the Movidius Myriad 2, ARM Mali C71 and Bilateral Filter ISP, respectively. For each targeted ISP attack, we show in the first column the adversarial RAW (top) and perturbations (bottom). The next four columns show the corresponding RGB images and perturbations from the ISPs. The RGB perturbation is visualized by subtracting the ISP output of adversarial RAW from the unattacked output.

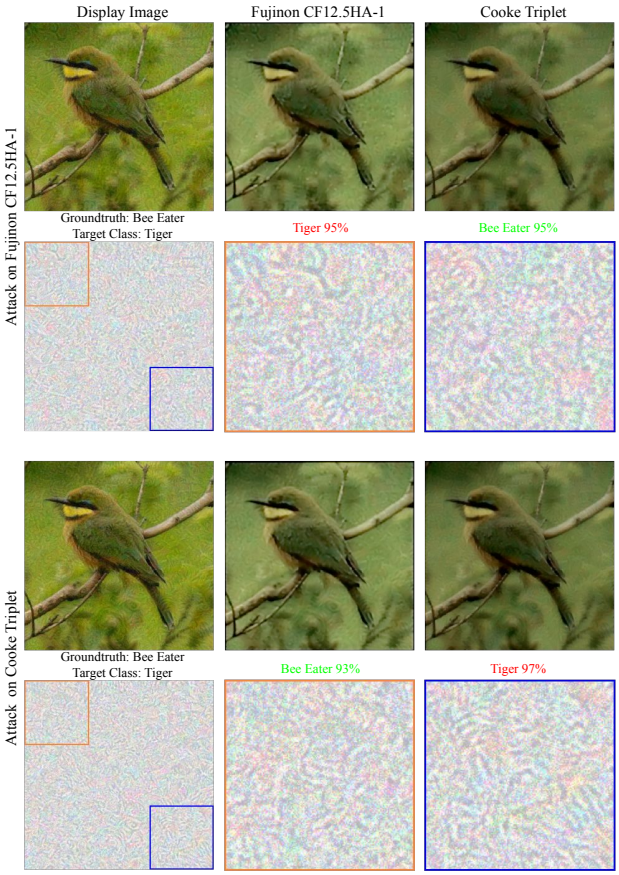


Figure 6: Visualization of the targeted optics attack on the Fujinon CF12.5HA-1 and Cooke Triplet optics. For each attack, we show the displayed adversarial and post-processed images (top row). In the bottom row, we visualize (from left to right) the additive perturbations on the display image and its zoomed in 150×150 top-left and bottom right region.

target pipeline with more than 87% success rate[3].

5.4. Attacking Camera Optics

Next, we use the proposed method to target a compound optical module instead of a hardware ISP. The proxy function now models the entire transformation from the displayed image to optics, sensor, and ISP processing that results in the final RGB image that is fed to the image classifier. In this experiment, all the pipelines deploy an *identical* ARM Mali C71 ISP, which allows us to assess adversarial pattern that targets only one optical system but not another. For each attacked image, we use ϵ=20/255, target a random class and use a total of 30 iterations, with the step size α=0.005. Note that ϵ is larger than the standard value of 8/255 since we need to compensate for the attenuation loss during the acquisition process. We use Algorithm 2 for the targeted optics attack and report its success and transfer rate[3] in Table 2. The proposed method is able to achieve a high success rate of 90% while keeping the transfer rate below 10%. We visualize the attacks in Figure 6.

[3]See Supplementary Document for results per ISP and untargeted optics attack.

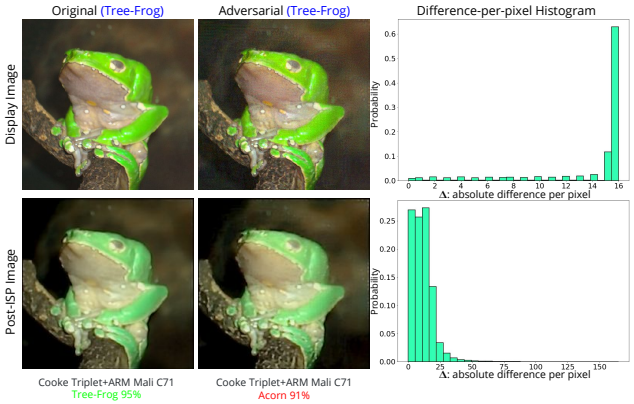


Figure 7: Successful attack on a robust classifier where the ISP amplifies perturbations. The original, adversarial, and pixel difference distribution shown for display and post-ISP images.

We find that in both attacks, the perturbations show distinctive frequency-dependent patterns. We interpret this attack as one that efficiently exploits the frequency bands specific to the optical transfer functions of the employed optics.

Effectiveness Against Defense Methods. Optics and ISPs tuned for image quality can *inadvertently amplify the adversarial perturbation in the scene*. Thus, our attack poses a realistic threat to current defense methods. We illustrate this in Figure 7, by attacking a pipeline deployed with a state-of-the-art robust ResNet152 with denoising modules by Xie et al.[4][55]. Our attacks achieve up to 58.2% success rate, which is much higher than the success rate of 26.6% reported in Xie et al. [55]. Note that this does not disqualify Xie et al. [55] since the network input perturbation exceeds the bound of 16/255, due to the effects of the ISP. Hence, training ISP-aware robust classifiers and optimal trade-off between robustness and image quality are essential for a robust vision system.

6. Conclusion

This work introduces the *first method for finding adversarial attacks that deceives a specific camera ISP and optics* while leaving cameras with other ISPs or optics intact although they employ the *same classifier network*. Departing from existing adversarial attacks, that assume camera pipelines to preserve adversarial perturbations, we propose an optimization method that employs a local proxy network, making it possible to attack embedded hardware ISPs that are not differentiable and only available as black-box algorithms. We validate the method experimentally on recent automotive camera ISPs and optics, achieving more than 90% targeted success rate for both ISP and optics attacks.

Building on the proposed methods, we envision not only research on defense mechanism to improve future image processing and camera optics but the method also suggests end-to-end multi-modal sensor design as a potential avenue to design systems resilient against adversarial attacks.

[4]Pretrained model from: https://github.com/facebookresearch/ImageNet-Adversarial-Training. We set ϵ=16/255 for display images, α=1/255 and 100 iterations following Xie et al. [55].

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the International Conference on Machine Learning, (ICML)*, 2018. 2

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 1, 3, 6

[3] Tejas Borkar, Felix Heide, and Lina Karam. Defending against universal attacks through selective feature regeneration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 709–719, 2020. 1

[4] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 1, 2

[5] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to See in the Dark. *ArXiv e-prints*, May 2018. 2

[6] Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A frank-wolfe framework for efficient and effective adversarial attacks. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. 2

[9] Q. Chen, J. Xu, and V. Koltun. Fast image processing with fully-convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2516–2525, Oct 2017. 2

[10] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 3

[11] Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *arXiv preprint arXiv:2010.06808*, 2020. 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[13] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A Kai Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1000–1008, 2020. 3

[14] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 1, 3

[15] Qingnan Fan, Jiaolong Yang, David Wipf, Baoquan Chen, and Xin Tong. Image smoothing via unsupervised learning. *ACM Transactions on Graphics (Proceedings of SIGGRAPH ASIA 2018)*, 37(6), 2018. 2

[16] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics (TOG)*, 35(6):191, 2016. 2

[17] M. Gharbi, J. Chen, J. Barron, S. Hasinoff, and F. Durand. Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph. (SIGGRAPH)*, 2017. 2

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[19] Puneet Gupta and Esa Rahtu. Ciidefence: Defeating adversarial attacks by fusing class-specific image inpainting and image denoising. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6708–6717, 2019. 2

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[22] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajak, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian, J. Kautz, and K. Pulli. FlexISP: A flexible camera image processing framework. *ACM Trans. Graph. (SIGGRAPH Asia)*, 33(6), 2014. 2

[23] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. *arXiv preprint arXiv:1911.07140*, 2019. 2

[24] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *CoRR*, abs/1804.08598, 2018. 2

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[26] Steve TK Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 962–969, 2019. 3

[27] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision*, pages 429–444. Springer, 2016. 2

[28] Seon Joo Kim, Hai Ting Lin, Zheng Lu, Sabine Süsstrunk, Stephen Lin, and Michael S Brown. A new in-camera imaging model for color computer vision and its application.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(12):2289–2302, 2012. 2

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 1, 3, 6

[30] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. *arXiv preprint arXiv:1905.00441*, 2019. 2

[31] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 2

[32] Y Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial samples and black-box attacks, 2016. 1

[33] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 446–454, 2017. 1

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3

[35] Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9087–9096, 2019. 2

[36] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 1, 2

[37] Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019. 1

[38] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *CoRR*, abs/1612.06299, 2016. 2

[39] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1607.02533, 2016. 1

[40] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 2, 4

[41] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 1

[42] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016. 1

[43] Sylvain Paris, Pierre Kornprobst, Jack Tumblin, and Frédo Durand. *Bilateral filtering: Theory and applications*. Now Publishers Inc, 2009. 5

[44] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 1

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[46] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018. 3

[47] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *CoRR*, abs/1710.08864, 2017. 2

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[49] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *Advances in Neural Information Processing Systems*, pages 7717–7728, 2018. 2

[50] Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2018. 2

[51] Ethan Tseng, Felix Yu, Yuting Yang, Fahim Mannan, Karl ST Arnaud, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. Hyperparameter optimization in black-box image processing using differentiable proxies. *ACM Trans. Graph.*, 38(4):27–1, 2019. 2, 3, 4, 5

[52] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, number 2019, 2019. 1

[53] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019. 2

[54] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018. 2

[55] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019. 1, 8

[56] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin.

Structured adversarial attack: Towards general implementation and better interpretability. In *International Conference on Learning Representations*, 2018. 2

[57] Li Xu, Jimmy Ren, Qiong Yan, Renjie Liao, and Jiaya Jia. Deep edge-aware filters. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1669–1678, Lille, France, 07–09 Jul 2015. PMLR. 2

[58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1

[59] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2020. 2, 5