Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt*[1], Ivan Evtimov*[2], Earlence Fernandes[2], Bo Li[3],
Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], and Dawn Song[3]

[1]University of Michigan, Ann Arbor
[2]University of Washington
[3]University of California, Berkeley
[4]Samsung Research America and Stony Brook University

Abstract

Recent studies show that the state-of-the-art deep neural networks (DNNs) are vulnerable to adversarial examples, resulting from small-magnitude perturbations added to the input. Given that that emerging physical systems are using DNNs in safety-critical situations, adversarial examples could mislead these systems and cause dangerous situations. Therefore, understanding adversarial examples in the physical world is an important step towards developing resilient learning algorithms. We propose a general attack algorithm, Robust Physical Perturbations (RP_2), to generate robust visual *adversarial perturbations under different physical conditions. Using the real-world case of road sign classification, we show that adversarial examples generated using RP_2 achieve high targeted misclassification rates against standard-architecture road sign classifiers in the physical world under various environmental conditions, including viewpoints. Due to the current lack of a standardized testing method, we propose a two-stage evaluation methodology for robust physical adversarial examples consisting of lab and field tests. Using this methodology, we evaluate the efficacy of physical adversarial manipulations on real objects. With a perturbation in the form of only black and white stickers, we attack a real stop sign, causing targeted misclassification in 100% of the images obtained in lab settings, and in 84.8% of the captured video frames obtained on a moving vehicle (field test) for the target classifier.*

1. Introduction

Deep Neural Networks (DNNs) have achieved state-of-the-art, and sometimes human-competitive, performance on many computer vision tasks [11, 14, 36]. Based on these successes, they are increasingly being used as part of control pipelines in physical systems such as cars [8, 17], UAVs [4, 24], and robots [40]. Recent work, however, has demonstrated that DNNs are vulnerable to adversarial perturbations [5, 9, 10, 15, 16, 22, 25, 29, 30, 35]. These carefully crafted modifications to the (visual) input of DNNs can cause the systems they control to misbehave in unexpected and potentially dangerous ways.

This threat has gained recent attention, and work in computer vision has made great progress in understanding the space of adversarial examples, beginning in the digital domain (*e.g.* by modifying images corresponding to a scene) [9, 22, 25, 35], and more recently in the physical domain [1, 2, 13, 32]. Along similar lines, our work contributes to the understanding of adversarial examples when perturbations are physically added to the *objects themselves*. We choose road sign classification as our target domain for several reasons: (1) The relative visual simplicity of road signs make it challenging to hide perturbations. (2) Road signs exist in a noisy unconstrained environment with changing physical conditions such as the distance and angle of the viewing camera, implying that physical adversarial perturbations should be robust against considerable environmental instability. (3) Road signs play an important role in transportation safety. (4) A reasonable threat model for transportation is that an attacker might not have control over a vehicle's systems, but is able to modify the objects in the physical world that a vehicle might depend on to make crucial safety decisions.

The main challenge with generating robust physical perturbations is environmental variability. Cyber-physical systems operate in noisy physical environments that can destroy perturbations created using current digital-only algorithms [19]. For our chosen application area, the most dynamic environmental change is the distance and angle of

*These authors contributed equally.

Figure 1: The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows our a physical perturbation applied to a Stop sign. We design our perturbations to mimic graffiti, and thus "hide in the human psyche."

the viewing camera. Additionally, other practicality challenges exist: (1) Perturbations in the digital world can be so small in magnitude that it is likely that a camera will not be able to perceive them due to sensor imperfections. (2) Current algorithms produce perturbations that occupy the background imagery of an object. It is extremely difficult to create a robust attack with background modifications because a real object can have varying backgrounds depending on the viewpoint. (3) The fabrication process (e.g., printing of perturbations) is imperfect.

Informed by the challenges above, we design *Robust Physical Perturbations (RP_2)*, which can generate perturbations robust to widely changing distances and angles of the viewing camera. RP_2 creates a visible, but inconspicuous perturbation that only perturbs the object (*e.g.* a road sign) and not the object's environment. To create robust perturbations, the algorithm draws samples from a distribution that models physical dynamics (*e.g.* varying distances and angles) using experimental data and synthetic transformations (Figure 2).

Using the proposed algorithm, we evaluate the effectiveness of perturbations on physical objects, and show that adversaries can physically modify objects using low-cost techniques to reliably cause classification errors in DNN-based classifiers under widely varying distances and angles. For example, our attacks cause a classifier to interpret a subtly-modified physical Stop sign as a Speed Limit 45 sign. Specifically, our final form of perturbation is a set of black and white stickers that an adversary can attach to a physical road sign (Stop sign). We designed our perturbations to resemble graffiti, a relatively common form of vandalism. It is common to see road signs with random graffiti or color alterations in the real world as shown in Figure 1 (the left image is of a real sign in a city). If these random patterns were adversarial perturbations (right side of Figure 1 shows our example perturbation), they could lead to severe consequences for autonomous driving systems, without arousing suspicion in human operators.

Given the lack of a standardized method for evaluating

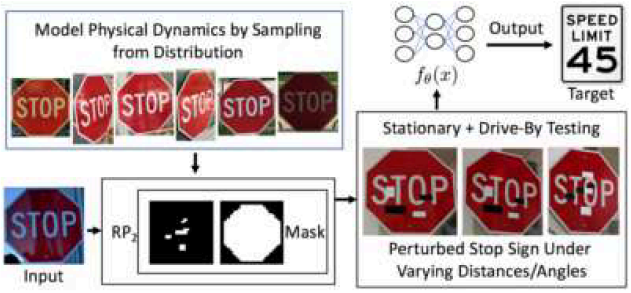


Figure 2: RP_2 pipeline overview. The input is the target Stop sign. RP_2 samples from a distribution that models physical dynamics (in this case, varying distances and angles), and uses a mask to project computed perturbations to a shape that resembles graffiti. The adversary prints out the resulting perturbations and sticks them to the target Stop sign.

physical attacks, we draw on standard techniques from the physical sciences and propose a two-stage experiment design: (1) A lab test where the viewing camera is kept at various distance/angle configurations; and (2) A field test where we drive a car towards an intersection in uncontrolled conditions to simulate an autonomous vehicle. We test our attack algorithm using this evaluation pipeline and find that the perturbations are robust to a variety of distances and angles.

Our Contributions. Figure 2 shows an overview of our pipeline to generate and evaluate robust physical adversarial perturbations.

1. We introduce Robust Physical Perturbations (RP_2) to generate physical perturbations for *physical-world* objects that can consistently cause misclassification in a DNN-based classifier under a range of dynamic physical conditions, including different viewpoint angles and distances (Section 3).

2. Given the lack of a standardized methodology in evaluating physical adversarial perturbations, we propose an evaluation methodology to study the effectiveness of physical perturbations in real world scenarios (Section 4.2).

3. We evaluate our attacks against two standard-architecture classifiers that we built: LISA-CNN with 91% accuracy on the LISA test set and GTSRB-CNN with 95.7% accuracy on the GTSRB test set. Using two types of attacks (object-constrained poster and sticker attacks) that we introduce, we show that RP_2 produces robust perturbations for real road signs. For example, poster attacks are successful in 100% of stationary and drive-by tests against LISA-CNN, and sticker attacks are successful in 80% of stationary testing conditions

and in 87.5% of the extracted video frames against GTSRB-CNN.

4. To show the generality of our approach, we generate the robust physical adversarial example by manipulating general physical objects, such as a microwave. We show that the pre-trained Inception-v3 classifier misclassifies the microwave as "phone" by adding a single sticker.

Our work, thus, contributes to understanding the susceptibility of image classifiers to robust adversarial modifications of *physical objects*. These results provide a case for the potential consequences of adversarial examples on deep learning models that interact with the physical world through vision. Our overarching goal with this work is to inform research in building robust vision models and to raise awareness on the risks that future physical learning systems might face. We include more examples and videos of the drive-by tests on our webpage https://iotsecurity.eecs.umich.edu/#roadsigns

2. Related Work

We survey the related work in generating adversarial examples. Specifically, given a classifier $f_\theta(\cdot)$ with parameters θ and an input x with ground truth label y for x, an adversarial example x' is generated so that it is close to x in terms of certain distance, such as L_p norm distance. x' will also cause the classifier to make an incorrect prediction as $f_\theta(x') \neq y$ (untargeted attacks), or $f_\theta(x') = y^*$ (targeted attacks) for a specific $y^* \neq y$. We also discuss recent efforts at understanding the space of physical adversarial examples.

Digital Adversarial Examples. Different methods have been proposed to generate adversarial examples in the white-box setting, where the adversary has full access to the classifier [3, 5, 9, 13, 23, 29, 35]. We focus on the white-box setting as well for two reasons: (1) In our chosen autonomous vehicle domain, an attacker can obtain a close approximation of the model by reverse engineering the vehicle's systems using model extraction attacks [37]. (2) To develop a foundation for future defenses, we must assess the abilities of powerful adversaries, and this can be done in a white-box setting. Given that recent work has examined the black-box transferability of digital adversarial examples [27], physical black-box attacks may also be possible.

Goodfellow *et al*. proposed the fast gradient method that applies a first-order approximation of the loss function to construct adversarial samples [9]. Optimization based methods have also been proposed to create adversarial perturbations for targeted attacks [5, 18]. These methods contribute to understanding digital adversarial examples. By contrast, our work examines physical perturbations on real objects under varying environmental conditions.

Physical Adversarial Examples. Kurakin *et al*. showed that printed adversarial examples can be misclassified when viewed through a smartphone camera [13]. Athalye and Sutskever improved upon the work of Kurakin *et al*. and presented an attack algorithm that produces adversarial examples robust to a set of two-dimensional synthetic transformations [1]. These works do not modify physical objects—an adversary prints out a digitally-perturbed image on paper. However, there is value in studying the effectiveness of such attacks when subject to environmental variability. Our object-constrained poster printing attack is a reproduced version of this type of attack, with the additional physical-world constraint of confining perturbations to the surface area of the sign. Additionally, our work goes further and examines how to effectively create adversarial examples where the object itself is physically perturbed by placing stickers on it.

Concurrent to our work,[1] Athalye *et al*. improved upon their original attack, and created 3D-printed replicas of perturbed objects [2]. The main intellectual differences include: (1) Athalye *et al*. *only* use a set of synthetic transformations during optimization, which can miss subtle physical effects, while our work samples from a distribution modeling both physical *and* synthetic transformations. (2) Our work modifies *existing* true-sized objects. Athalye *et al*. 3D-print small-scale replicas. (3) Our work simulates realistic testing conditions appropriate to the use-case at hand.

Sharif *et al*. attacked face recognition systems by printing adversarial perturbations on the frames of eyeglasses [32]. Their work demonstrated successful physical attacks in relatively stable physical conditions with little variation in pose, distance/angle from the camera, and lighting. This contributes an interesting understanding of physical examples in stable environments. However, environmental conditions can vary widely in general and can contribute to reducing the effectiveness of perturbations. Therefore, we choose the inherently unconstrained environment of road-sign classification. In our work, we explicitly design our perturbations to be effective in the presence of diverse physical-world conditions (specifically, large distances/angles and resolution changes).

Finally, Lu *et al*. performed experiments with physical adversarial examples of road sign images against *detectors* and show current detectors cannot be attacked [19]. In this work, we focus on *classifiers* to demonstrate the physical attack effectiveness and to highlight their security vulnerability in the real world. Attacking detectors are out of the scope of this paper, though recent work has generated digital adversarial examples against detection/segmentation algorithms [6, 20, 38], and our recent work has extended RP_2 to attack the YOLO detector [7].

[1] This work appeared at arXiv on 30 Oct 2017.

3. Adversarial Examples for Physical Objects

Our goal is to examine whether it is possible to create robust physical perturbations for real-world objects that mislead classifiers to make incorrect predictions even when images are taken in a range of varying physical conditions. We first present an analysis of environmental conditions that physical learning systems might encounter, and then present our algorithm to generate physical adversarial perturbations taking these challenges into account.

3.1. Physical World Challenges

Physical attacks on an object must be able to survive changing conditions and remain effective at fooling the classifier. We structure our discussion of these conditions around the chosen example of road sign classification, which could be potentially applied in autonomous vehicles and other safety sensitive domains. A subset of these conditions can also be applied to other types of physical learning systems such as drones, and robots.

Environmental Conditions. The distance and angle of a camera in an autonomous vehicle with respect to a road sign varies continuously. The resulting images that are fed into a classifier are taken at different distances and angles. Therefore, any perturbation that an attacker physically adds to a road sign must be able to survive these transformations of the image. Other environmental factors include changes in lighting/weather conditions, and the presence of debris on the camera or on the road sign.

Spatial Constraints. Current algorithms focusing on digital images add adversarial perturbations to all parts of the image, including background imagery. However, for a physical road sign, the attacker cannot manipulate background imagery. Furthermore, the attacker cannot count on there being a fixed background imagery as it will change depending on the distance and angle of the viewing camera.

Physical Limits on Imperceptibility. An attractive feature of current adversarial deep learning algorithms is that their perturbations to a digital image are often so small in magnitude that they are almost imperceptible to the casual observer. However, when transferring such minute perturbations to the real world, we must ensure that a camera is able to perceive the perturbations. Therefore, there are physical limits on how imperceptible perturbations can be, and is dependent on the sensing hardware.

Fabrication Error. To fabricate the computed perturbation, all perturbation values must be valid colors that can be reproduced in the real world. Furthermore, even if a fabrication device, such as a printer, can produce certain colors, there will be some reproduction error [32].

In order to successfully physically attack deep learning classifiers, an attacker should account for the above categories of physical world variations that can reduce the effectiveness of perturbations.

3.2. Robust Physical Perturbation

We derive our algorithm starting with the optimization method that generates a perturbation for a single image x, without considering other physical conditions; then, we describe how to update the algorithm taking the physical challenges above into account. This single-image optimization problem searches for perturbation δ to be added to the input x, such that the perturbed instance $x' = x+\delta$ is misclassified by the target classifier $f_\theta(\cdot)$:

$$\min \quad H(x+\delta, x), \quad \text{s.t.} \quad f_\theta(x+\delta) = y^*$$

where H is a chosen distance function, and y^* is the target class.[2] To solve the above constrained optimization problem efficiently, we reformulate it in the Lagrangian-relaxed form similar to prior work [5, 18].

$$\underset{\delta}{\operatorname{argmin}} \; \lambda||\delta||_p + J(f_\theta(x+\delta), y^*) \tag{1}$$

Here $J(\cdot\,,\,\cdot)$ is the loss function, which measures the difference between the model's prediction and the target label y^*. λ is a hyper-parameter that controls the regularization of the distortion. We specify the distance function H as $||\delta||_p$, denoting the ℓ_p norm of δ.

Next, we will discuss how the objective function can be modified to account for the *environmental conditions*. We model the distribution of images containing object o under both physical and digital transformations X^V. We sample different instances x_i drawn from X^V. A physical perturbation can only be added to a specific object o within x_i. In the example of road sign classification, o is the stop sign that we target to manipulate. Given images taken in the physical world, we need to make sure that a single perturbation δ, which is added to o, can fool the classifier under different physical conditions. Concurrent work [2] only applies a set of transformation functions to synthetically sample such a distribution. However, modeling physical phenomena is complex and such synthetic transformations may miss physical effects. Therefore, to better capture the effects of changing physical conditions, we sample instance x_i from X^V by both generating experimental data that contains actual physical condition variability as well as synthetic transformations. For road sign physical conditions, this involves taking images of road signs under various conditions, such as changing distances, angles, and lightning. This approach aims to approximate physical world dynamics more closely. For synthetic variations, we randomly crop the object within the image, change the brightness, and add spatial transformations to simulate other possible conditions.

To ensure that the perturbations are only applied to the surface area of the target object o (considering the *spatial*

[2] For untargeted attacks, we can modify the objective function to maximize the distance between the model prediction and the true class. We focus on targeted attacks in the rest of the paper.

constraints and *physical limits on imperceptibility*), we introduce a mask. This mask serves to project the computed perturbations to a physical region on the surface of the object (*i.e*. road sign). In addition to providing spatial locality, the mask also helps generate perturbations that are visible but inconspicuous to human observers. To do this, an attacker can shape the mask to look like graffiti—commonplace vandalism on the street that most humans expect and ignore, therefore hiding the perturbations "in the human psyche." Formally, the perturbation mask is a matrix M_x whose dimensions are the same as the size of input to the road sign classifier. M_x contains zeroes in regions where no perturbation is added, and ones in regions where the perturbation is added during optimization.

In the course of our experiments, we empirically observed that the position of the mask has an impact on the effectiveness of an attack. We therefore hypothesize that objects have strong and weak physical features from a classification perspective, and we position masks to attack the weak areas. Specifically, we use the following pipeline to discover mask positions: (1) Compute perturbations using the L_1 regularization and with a mask that occupies the entire surface area of the sign. L_1 makes the optimizer favor a sparse perturbation vector, therefore concentrating the perturbations on regions that are most vulnerable. Visualizing the resulting perturbation provides guidance on mask placement. (2) Recompute perturbations using L_2 with a mask positioned on the vulnerable regions identified from the earlier step.

To account for *fabrication error*, we add an additional term to our objective function that models printer color reproduction errors. This term is based upon the Non-Printability Score (NPS) by Sharif *et al*. [32]. See the supplemental materials for a formal definition of NPS.

Based on the above discussion, our final robust spatially-constrained perturbation is thus optimized as:

$$\begin{aligned} &\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p + NPS \\ &+ \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*) \end{aligned} \tag{2}$$

Here we use function $T_i(\cdot)$ to denote the alignment function that maps transformations on the object to transformations on the perturbation (*e.g*. if the object is rotated, the perturbation is rotated as well).

Finally, an attacker will print out the optimization result on paper, cut out the perturbation (M_x), and put it onto the target object o. As our experiments demonstrate in the next section, this kind of perturbation fools the classifier in a variety of viewpoints.[3]

[3]For our attacks, we use the ADAM optimizer with the following parameters: $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \eta \in [10^{-4}, 10^0]$

4. Experiments

In this section, we will empirically evaluate the proposed RP_2. We first evaluate a safety sensitive example, Stop sign recognition, to demonstrate the robustness of the proposed physical perturbation. To demonstrate the generality of our approach, we then attack Inception-v3 to misclassify a microwave as a phone.

4.1. Dataset and Classifiers

We built two classifiers based on a standard crop-resize-then-classify pipeline for road sign classification as described in [28, 31]. Our LISA-CNN uses LISA, a U.S. traffic sign dataset containing 47 different road signs [21]. However, the dataset is not well-balanced, resulting is large disparities in representation for different signs. To alleviate this problem, we chose the 17 most common signs based on the number of training examples. LISA-CNN's architecture is defined in the Cleverhans library [26] and consists of three convolutional layers and an FC layer. It has an accuracy of 91% on the test set.

Our second classifier is GTSRB-CNN, that is trained on the German Traffic Sign Recognition Benchmark (GTSRB) [33]. We use a publicly available implementation [39] of a multi-scale CNN architecture that has been known to perform well on road sign recognition [31]. Because we did not have access to German Stop signs for our physical experiments, we replaced the German Stop signs in the training, validation, and test sets of GTSRB with the U.S. Stop sign images in LISA. GTSRB-CNN achieves 95.7% accuracy on the test set. When evaluating GTSRB-CNN on our own 181 stop sign images, it achieves 99.4% accuracy.

4.2. Experimental Design

To the best of our knowledge, there is currently no standardized methodology of evaluating physical adversarial perturbations. Based on our discussion from Section 3.1, we focus on angles and distances because they are the most rapidly changing elements for our use case. A camera in a vehicle approaching a sign will take a series of images at regular intervals. These images will be taken at different angles and distances, therefore changing the amount of detail present in any given image. Any successful physical perturbation must cause targeted misclassification in a range of distances and angles because a vehicle will likely perform voting on a set of frames (images) from a video before issuing a controller action. Our current experiments do not explicitly control ambient light, and as is evident from experimental data (Section 4), lighting varied from indoor lighting to outdoor lighting.

Drawing on standard practice in the physical sciences, our experimental design encapsulates the above physical factors into a two-stage evaluation consisting of controlled lab tests and field tests.

Stationary (Lab) Tests. This involves classifying images of objects from stationary, fixed positions.

1. Obtain a set of clean images C and a set of adversarially perturbed images ($\{\mathcal{A}(c)\}, \forall c \in C$) at varying distances $d \in D$, and varying angles $g \in G$. We use $c^{d,g}$ here to denote the image taken from distance d and angle g. The camera's vertical elevation should be kept approximately constant. Changes in the camera angle relative the the sign will normally occur when the car is turning, changing lanes, or following a curved road.

2. Compute the attack success rate of the physical perturbation using the following formula:

$$\frac{\sum_{c \in C} \mathbb{1}_{\{f_\theta(\mathcal{A}(c^{d,g}))=y^* \ \wedge f_\theta(c^{d,g})=y\}}}{\sum_{c \in C} \mathbb{1}_{\{f_\theta(c^{d,g})=y\}}} \quad (3)$$

   where d and g denote the camera distance and angle for the image, y is the ground truth, and y^* is the targeted attacking class.[4]

Note that an image $\mathcal{A}(c)$ that causes misclassification is considered as a successful attack only if the original image c with the same camera distance and angle is correctly classified, which ensures that the misclassification is caused by the added perturbation instead of other factors.

Drive-By (Field) Tests. We place a camera on a moving platform, and obtain data at realistic driving speeds. For our experiments, we use a smartphone camera mounted on a car.

1. Begin recording video at approximately 250 ft away from the sign. Our driving track was straight without curves. Drive toward the sign at normal driving speeds and stop recording once the vehicle passes the sign. In our experiments, our speed varied between 0 mph and 20 mph. This simulates a human driver approaching a sign in a large city.

2. Perform video recording as above for a "clean" sign and for a sign with perturbations applied, and then apply similar formula as Eq. 3 to calculate the attack success rate, where C here represents the sampled frames.

An autonomous vehicle will likely not run classification on every frame due to performance constraints, but rather, would classify every j-th frame, and then perform simple majority voting. Hence, an open question is to determine whether the choice of frame (j) affects attack accuracy. In our experiments, we use $j = 10$. We also tried $j = 15$ and did not observe any significant change in the attack success rates. If both types of tests produce high success rates, the attack is likely to be successful in commonly experienced physical conditions for cars.

[4]For untargeted adversarial perturbations, change $f_\theta(e^{d,g}) = y^*$ to $f_\theta(e^{d,g}) \neq y$.

4.3. Results for LISA-CNN

We evaluate the effectiveness of our algorithm by generating three types of adversarial examples on LISA-CNN (91% accuracy on test-set). For all types, we observe high attack success rates with high confidence. Table 1 summarizes a sampling of stationary attack images. In all testing conditions, our baseline of unperturbed road signs achieves a 100% classification rate into the true class.

Object-Constrained Poster-Printing Attacks. This involves reproducing the attack of Kurakin *et al*. [13]. The crucial difference is that in our attack, the perturbations are confined to the surface area of the sign excluding the background, and are robust against large angle and distance variations. The Stop sign is misclassified into the attack's target class of Speed Limit 45 in 100% of the images taken according to our evaluation methodology. The average confidence of predicting the manipulated sign as the target class is 80.51% (second column of Table 2).

For the Right Turn warning sign, we choose a mask that covers only the arrow since we intend to generate subtle perturbations. In order to achieve this goal, we increase the regularization parameter λ in equation (2) to demonstrate small magnitude perturbations. We achieve a 73.33% targeted-attack success rate (Table 1). Out of 15 distance/angle configurations, four instances were not classified into the target. However, they were still misclassified into other classes that were not the true label (Yield, Added Lane). Three of these four instances were an Added Lane sign—a different type of warning. We hypothesize that given the similar appearance of warning signs, small perturbations are sufficient to confuse the classifier.

Sticker Attacks. Next, we demonstrate how effective it is to generate physical perturbations in the form of stickers, by constraining the modifications to a region resembling graffiti or art. The fourth and fifth columns of Table 1 show a sample of images, and Table 2 (columns 4 and 6) shows detailed success rates with confidences. In the stationary setting, we achieve a 66.67% targeted-attack success rate for the graffiti sticker attack and a 100% targeted-attack success rate for the sticker camouflage art attack. Some region mismatches may lead to the lower performance of the LOVE-HATE graffiti.

Drive-By Testing. Per our evaluation methodology, we conduct drive-by testing for the perturbation of a Stop sign. In our baseline test we record two consecutive videos of a clean Stop sign from a moving vehicle, perform frame grabs at $k = 10$, and crop the sign. We observe that the Stop sign is correctly classified in all frames. We similarly test subtle and abstract art perturbations for LISA-CNN using $k = 10$. Our attack achieves a targeted-attack success rate of 100% for the subtle poster attack, and a targeted-attack success rate of 84.8% for the camouflage abstract art attack. See the supplemental materials for sample frames from the drive-by video.

Table 1: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN.

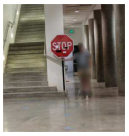
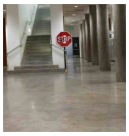
Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Table 2: Targeted physical perturbation experiment results on LISA-CNN using a poster-printed Stop sign (subtle attacks) and a real Stop sign (camouflage graffiti attacks, camouflage art attacks). For each image, the top two labels and their associated confidence values are shown. The misclassification target was Speed Limit 45. See Table 1 for example images of each attack. Legend: SL45 = Speed Limit 45, STP = Stop, YLD = Yield, ADL = Added Lane, SA = Signal Ahead, LE = Lane Ends.

Distance & Angle	Poster-Printing		Sticker			
	Subtle		Camouflage–Graffiti		Camouflage–Art	
5′ 0°	SL45 (0.86)	ADL (0.03)	STP (0.40)	SL45 (0.27)	SL45 (0.64)	LE (0.11)
5′ 15°	SL45 (0.86)	ADL (0.02)	STP (0.40)	YLD (0.26)	SL45 (0.39)	STP (0.30)
5′ 30°	SL45 (0.57)	STP (0.18)	SL45 (0.25)	SA (0.18)	SL45 (0.43)	STP (0.29)
5′ 45°	SL45 (0.80)	STP (0.09)	YLD (0.21)	STP (0.20)	SL45 (0.37)	STP (0.31)
5′ 60°	SL45 (0.61)	STP (0.19)	STP (0.39)	YLD (0.19)	SL45 (0.53)	STP (0.16)
10′ 0°	SL45 (0.86)	ADL (0.02)	SL45 (0.48)	STP (0.23)	SL45 (0.77)	LE (0.04)
10′ 15°	SL45 (0.90)	STP (0.02)	SL45 (0.58)	STP (0.21)	SL45 (0.71)	STP (0.08)
10′ 30°	SL45 (0.93)	STP (0.01)	STP (0.34)	SL45 (0.26)	SL45 (0.47)	STP (0.30)
15′ 0°	SL45 (0.81)	LE (0.05)	SL45 (0.54)	STP (0.22)	SL45 (0.79)	STP (0.05)
15′ 15°	SL45 (0.92)	ADL (0.01)	SL45 (0.67)	STP (0.15)	SL45 (0.79)	STP (0.06)
20′ 0°	SL45 (0.83)	ADL (0.03)	SL45 (0.62)	STP (0.18)	SL45 (0.68)	STP (0.12)
20′ 15°	SL45 (0.88)	STP (0.02)	SL45 (0.70)	STP (0.08)	SL45 (0.67)	STP (0.11)
25′ 0°	SL45 (0.76)	STP (0.04)	SL45 (0.58)	STP (0.17)	SL45 (0.67)	STP (0.08)
30′ 0°	SL45 (0.71)	STP (0.07)	SL45 (0.60)	STP (0.19)	SL45 (0.76)	STP (0.10)
40′ 0°	SL45 (0.78)	LE (0.04)	SL45 (0.54)	STP (0.21)	SL45 (0.68)	STP (0.14)

Table 3: A camouflage art attack on GTSRB-CNN. See example images in Table 1. The targeted-attack success rate is 80% (true class label: Stop, target: Speed Limit 80).

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
5′ 0°	Speed Limit 80 (0.88)	Speed Limit 70 (0.07)
5′ 15°	Speed Limit 80 (0.94)	Stop (0.03)
5′ 30°	Speed Limit 80 (0.86)	Keep Right (0.03)
5′ 45°	Keep Right (0.82)	Speed Limit 80 (0.12)
5′ 60°	Speed Limit 80 (0.55)	Stop (0.31)
10′ 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.006)
10′ 15°	Stop (0.75)	Speed Limit 80 (0.20)
10′ 30°	Speed Limit 80 (0.77)	Speed Limit 100 (0.11)
15′ 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.01)
15′ 15°	Stop (0.90)	Speed Limit 80 (0.06)
20′ 0°	Speed Limit 80 (0.95)	Speed Limit 100 (0.03)
20′ 15°	Speed Limit 80 (0.97)	Speed Limit 100 (0.01)
25′ 0°	Speed Limit 80 (0.99)	Speed Limit 70 (0.0008)
30′ 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)
40′ 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)

4.4. Results for GTSRB-CNN

To show the versatility of our attack algorithms, we create and test attacks for GTSRB-CNN (95.7% accuracy on test-set). Based on our high success rates with the camouflage-art attacks, we create similar abstract art sticker perturbations. The last column of Table 1 shows a subset of experimental images. Table 3 summarizes our attack results—our attack fools the classifier into believing that a Stop sign is a Speed Limit 80 sign in 80% of the stationary testing conditions. Per our evaluation methodology, we also conduct a drive-by test ($k = 10$, two consecutive video recordings). The attack fools the classifier 87.5% of the time.

4.5. Results for Inception-v3

To demonstrate generality of RP_2, we computed physical perturbations for the standard Inception-v3 classifier [12, 34] using two different objects, a microwave and a coffee mug. For the microwave, our adversarial sticker causes the classifier to misclassify it as our target class, "phone," in 90% of the tests (Figure 3). For the coffee mug, our adversarial sticker causes the classifier to misclassify it as our target class, "cash machine", in 71.4% of the tests. Figure 3 shows an example of the adversarial sticker for microwave. See the supplemental materials for more detailed results and adversarial images.

5. Discussion

Black-Box Attacks. Given access to the target classifier's network architecture and model weights, RP_2 can generate a variety of robust physical perturbations that fool the classifier. Through studying a white-box attack like RP_2, we can analyze the requirements for a successful attack using the strongest attacker model and better inform future defenses. Evaluating RP_2 in a black-box setting is an open question.

Figure 3: Physical adversarial example against the Inception-v3 classifier. The left shows the original cropped image identified as microwave (85.2%) while the right shows the cropped physical adversarial example identified as phone (77.8%).

Image Cropping and Attacking Detectors. When evaluating RP_2, we manually controlled the cropping of each image every time before classification. This was done so the adversarial images would match the clean sign images provided to RP_2. Later, we evaluated the camouflage art attack using a pseudo-random crop with the guarantee that at least most of the sign was in the image. Against LISA-CNN, we observed an average targeted attack rate of 70% and untargeted attack rate of 90%. Against GTSRB-CNN, we observed an average targeted attack rate of 60% and untargeted attack rate of 100%. We include the untargeted attack success rates because causing the classifier to not output the correct traffic sign label is still a safety risk. Although image cropping has some effect on the targeted attack success rate, our recent work shows that an improved version of RP_2 can successfully attack object detectors, where cropping is not needed [7].

6. Conclusion

We introduced an algorithm (RP_2) that generates robust, physically realizable adversarial perturbations. Using RP_2, and a two-stage experimental design consisting of lab and drive-by tests, we contribute to understanding the space of physical adversarial examples when the *objects themselves* are physically perturbed. We target road-sign classification because of its importance in safety, and the naturally noisy environment of road signs. Our work shows that it is possible to generate physical adversarial examples robust to widely varying distances/angles. This implies that future defenses should not rely on physical sources of noise as protection against physical adversarial examples.

Acknowledgements. We thank the reviewers for their insightful feedback. This work was supported in part by NSF grants 1422211, 1616575, 1646392, 1740897, 1565252, Berkeley Deep Drive, the Center for Long-Term Cybersecurity, FORCES (which receives support from the NSF), the Hewlett Foundation, the MacArthur Foundation, a UM-SJTU grant, and the UW Tech Policy Lab.

References

[1] A. Athalye. Robust adversarial examples. https://blog.openai.com/robust-adversarial-inputs/, 2017.

[2] A. Athalye and I. Sutskever. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[3] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.

[4] H. Bou-Ammar, H. Voos, and W. Ertel. Controller design for quadrotor uavs using reinforcement learning. In *Control Applications (CCA), 2010 IEEE International Conference on*, pages 2130–2135. IEEE, 2010.

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

[6] M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

[7] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, D. Song, T. Kohno, A. Rahmati, A. Prakash, and F. Tramer. Note on Attacking Object Detectors with Adversarial Stickers. Dec. 2017.

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012.

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

[10] J. Kos, I. Fischer, and D. Song. Adversarial examples for generative models. *arXiv preprint arXiv:1702.06832*, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

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

[14] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3361–3368, Washington, DC, USA, 2011. IEEE Computer Society.

[15] B. Li and Y. Vorobeychik. Feature cross-substitution in adversarial classification. In *Advances in Neural Information Processing Systems*, pages 2087–2095, 2014.

[16] B. Li and Y. Vorobeychik. Scalable optimization of randomized operational decisions in adversarial classification settings. In *AISTATS*, 2015.

[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[18] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[19] J. Lu, H. Sibai, E. Fabry, and D. Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017.

[20] J. H. Metzen, M. C. Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. *arXiv preprint arXiv:1704.05712*, 2017.

[21] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *Trans. Intell. Transport. Sys.*, 13(4):1484–1497, Dec. 2012.

[22] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *CoRR*, abs/1610.08401, 2016.

[23] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *arXiv preprint arXiv:1511.04599*, 2015.

[24] C. Mostegel, M. Rumpler, F. Fraundorfer, and H. Bischof. Uav-based autonomous image acquisition with multi-view stereo quality assurance by confidence prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2016.

[25] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[26] N. Papernot, I. Goodfellow, R. Sheatsley, R. Feinman, and P. McDaniel. cleverhans v1.0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 2016.

[27] N. Papernot, P. McDaniel, and I. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[28] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. 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. ACM, 2017.

[29] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[30] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet. Adversarial manipulation of deep representations. *arXiv preprint arXiv:1511.05122*, 2015.

[31] P. Sermanet and Y. LeCun. Traffic sign recognition with multi-scale convolutional networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2809–2813. IEEE, 2011.

[32] 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 *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540. ACM, 2016.

[33] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 2012.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, 2015.

[35] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

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

[37] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction apis. In *USENIX Security*, 2016.

[38] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. *arXiv preprint arXiv:1703.08603*, 2017.

[39] V. Yadav. p2-trafficsigns. `https://github.com/vxy10/p2-TrafficSigns`, 2016.

[40] F. Zhang, J. Leitner, M. Milford, B. Upcroft, and P. Corke. Towards vision-based deep reinforcement learning for robotic motion control. *arXiv preprint arXiv:1511.03791*, 2015.