Robust Physical-World Attacks on Deep Learning Models

Visit https://iotsecurity.eecs.umich.edu/#roadsigns for an FAQ

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

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

***Abstract*—Although deep neural networks (DNNs) perform well in a variety of applications, they are vulnerable to adversarial examples resulting from small-magnitude perturbations added to the input data. Inputs modified in this way can be mislabeled as a target class in targeted attacks or as a random class different from the ground truth in untargeted attacks. However, recent studies have demonstrated that such adversarial examples have limited effectiveness in the physical world due to changing physical conditions—they either completely fail to cause misclassification or only work in restricted cases where a relatively complex image is perturbed and printed on paper. In this paper, we propose a general attack algorithm—Robust Physical Perturbations (RP_2)—that takes into account the numerous physical conditions and produces robust adversarial perturbations. Using a real-world example of road sign recognition, we show that adversarial examples generated using RP_2 achieve high attack success rates in the physical world under a variety of conditions, including different viewpoints. Furthermore, to the best of our knowledge, there is currently no standardized way to evaluate physical adversarial perturbations. Therefore, we propose a two-stage evaluation methodology and tailor it to the road sign recognition use case. Our methodology captures a range of diverse physical conditions, including those encountered when images are captured from moving vehicles. We evaluate our physical attacks using this methodology and effectively fool two road sign classifiers. Using a perturbation in the shape of black and white stickers, we attack a real Stop sign, causing targeted misclassification in 100% of the images obtained in controlled lab settings and above 84% of the captured video frames obtained on a moving vehicle for one of the classifiers we attack.**

I. Introduction

Deep Neural Networks (DNNs) are being applied with great success to a variety of areas ranging from speech processing [6] to medical diagnostics [4] and are increasingly being used to control physical objects such as cars [14], UAVs [2], and robots [30]. These DNNs used for control of physical objects rely on sensory input to inform decisions. Recent work in computer vision, however, has demonstrated that DNNs are vulnerable to adversarial perturbations [3], [5], [8], [12], [13], [22], [23], [28]. Such maliciously crafted modifications to the sensory input of DNNs can cause the systems they control to misbehave in unexpected and potentially dangerous ways.

Although there is significant progress in creating digital adversarial perturbations, *e.g.*, by modifying an image representing a real-world scene that a cyber-physical system might perceive [7], [9], a fundamental open question, which we answer in this paper, is whether it is possible to create *robust physical adversarial perturbations*—small modifications to real-world objects themselves that can trigger misclassifications in a DNN under widely varying physical conditions.

We identify several challenges that must be overcome in order for an effective physical adversarial perturbation to be created: (1) A perturbation should be constrained to the targeted object and cannot be added to the object's background because that can vary. Many digital adversarial example generation algorithms do not consider this constraint (*i.e.*, they add perturbations to the entire area of a digital image, which includes both the targeted object and its background). (2) A perturbation should be robust against various dynamic physical conditions that can potentially decrease its effectiveness. For instance in recognition systems, the generated physical adversarial example should be robust against different viewing conditions. (3) A perturbation in the digital world can be so low in magnitude that humans cannot perceive them.[1] But, such small magnitude perturbations may not be captured by real world sensors due to sensor imperfections, and more generally, physical limitations of the sensor technology. (4) A perturbation should account for imperfections in the fabrication process. Printers, for example, cannot generate the entire color spectrum [26]. Therefore, it might not be feasible to fabricate a digital perturbation in a way that preservers all of its desired properties.

In addition to these general challenges, there are further barriers to realizability in our chosen application area of road-sign classification: (1) A physical road sign is a relatively simple object that makes concealing adversarial perturbations in it harder. This is in contrast to previous work that generates adversarial examples for relatively complex scenes. (2) Road signs exist in an uncontrolled noisy physical environment. There are multiple sources of error in such an environment that can reduce the effectiveness of an attack: (a) The distance between the camera in the vehicle and the road sign constantly varies depending on the road's shape, the specific vehicle, the number of lanes, and the position of the vehicle with respect to the

*All authors are in alphabetical order.

[1]This is especially relevant for object recognition tasks.

Fig. 1: Graffiti on real Stop signs.

sign. (b) The angle of the camera varies with respect to the sign. (c) Lighting varies depending on the time of day and the weather. (d) Debris either on the road sign or on the vehicle may occlude the view of the camera. Therefore, a robust and effective attack on DNNs used in vision for vehicles must be able to tolerate multiple sources of error.

Motivated by our analysis above, and by recent work, we design and evaluate *Robust Physical Perturbations (RP_2)*, the first attack algorithm that is robust to changing distances and angles of the camera and to varying resolutions of the image. RP_2 creates a visible, but inconspicuous perturbation that only perturbs the object, a road sign, and not the object's environment. We add a masking matrix to our objective function to realize such spatial constraints for perturbations.

Using the proposed algorithm, we show that attackers can physically modify objects using low-cost techniques to reliably cause classification errors in DNN-based vision systems under widely varying distances, angles, and resolutions. For example, our attacks cause a DNN to interpret a subtly-modified physical Stop sign as a Speed Limit 45 sign. Specifically, we introduce two classes of physical attacks: (1) *poster-printing attacks*, where an attacker prints an actual-sized road sign on paper and overlays it onto an existing sign, and (2) *sticker attacks*, where an attacker fabricates only the perturbations and then sticks them to an existing road sign. Our poster-printed attacks appear as faded signs and our sticker attacks resemble common forms of vandalism. We believe that both attacks are inconspicuous to human observers. It is common to see road signs with random graffiti or color alterations in the real world as shown in Figure 1 (these photos are of real-world road signs). If these random patterns are actually adversarial perturbations, they could lead to severe consequences for autonomous driving systems.

For poster-printing attacks, we demonstrate subtle perturbations—small-magnitude modifications that are applied to the entire area of the target object (but *not* its background). These remain inconspicuous because the resulting perturbation appears as a faded version of the original. For sticker attacks, we demonstrate camouflage attacks with graffiti and abstract art—large-magnitude perturbations that are focused in a smaller region of the target object. These remain inconspicuous as they mimic modifications that are commonly seen on road signs. Both of our attack classes do not require special resources and can be mounted using a color printer and a camera.

Given the lack of a standardized method for evaluating physical attacks, we propose a two-stage evaluation methodology consisting of a lab test in which the target object is perceived under controlled conditions and a field test in which the target object is perceived in a scenario appropriate for the application at hand. We evaluate RP_2 against two convolutional neural network road sign classifiers: the LISA-CNN network that is trained on the LISA dataset of U.S. road signs [18] and has an accuracy of 91%, and the GTSRB*-CNN network that is trained on a modified GTSRB (German Traffic Sign Recognition Benchmark) dataset [27] and has an accuracy of 95% (Section III-A).

Figure 2 shows our overall attack and evaluation pipeline. We use images from the domain of road signs recognition to demonstrate our evaluation methodology. We first use real road sign images to train DNNs with high recognition accuracy. Then, we use RP_2 to generate different types of adversarial examples for the trained classifiers. We evaluate the success rate of these attacks using our proposed evaluation methodology. In the lab (stationary) test, we take images of an adversarially perturbed physical sign from different viewpoints. In the field (drive-by) test, we record videos of the same sign from a vehicle at varying speed and extract frames from the video. We test these experimental images for mis-labeling using the high-accuracy classifiers. Our goal with these tests is to demonstrate that the perturbations are robust to multiple sources of variations (changing viewpoints, etc.) in the physical world.

Our experimental results based on varying distances, angles, and resolutions indicate that it is possible to create robust physical attacks: In a sticker attack on a real Stop sign with camouflage abstract art perturbations, we show that the Stop sign is reliably misclassified as a Speed Limit 45 sign in all of our stationary testing conditions. Our work, thus, answers a key, fundamental open question regarding the susceptibility of image classifiers to adversarial modifications of real objects in the physical world.

Our Contributions.

1) We design Robust Physical Perturbations (RP_2), the first algorithm that generates physical adversarial examples. To the best of our knowledge, it is the first time to show that it is possible to build physical attacks robust against different physical conditions, such as various viewing conditions in recognition systems.
2) We apply RP_2 to generate physical adversarial examples for road sign recognition systems. We achieve high attack success rate against varying real-world control variables, including distances, angles, and resolutions, in both lab and drive-by settings.
3) Focusing on the DNN application domain of road sign recognition, we introduce two attack classes on different physical road signs:
   Poster-Printing, where an attacker prints an actual-sized road sign with adversarial perturbations and then overlays it over an existing sign.
   Sticker Perturbation, where an attacker prints perturbations on paper, and then sticks them to an existing sign. For these attacks, we physically realize two types of perturbations: (1) *subtle perturbations* that occupy the entire region of the sign, and (2) *camouflage perturbations* that take the form of graffiti and abstract art. These attacks do not require special resources—only access to a color printer.

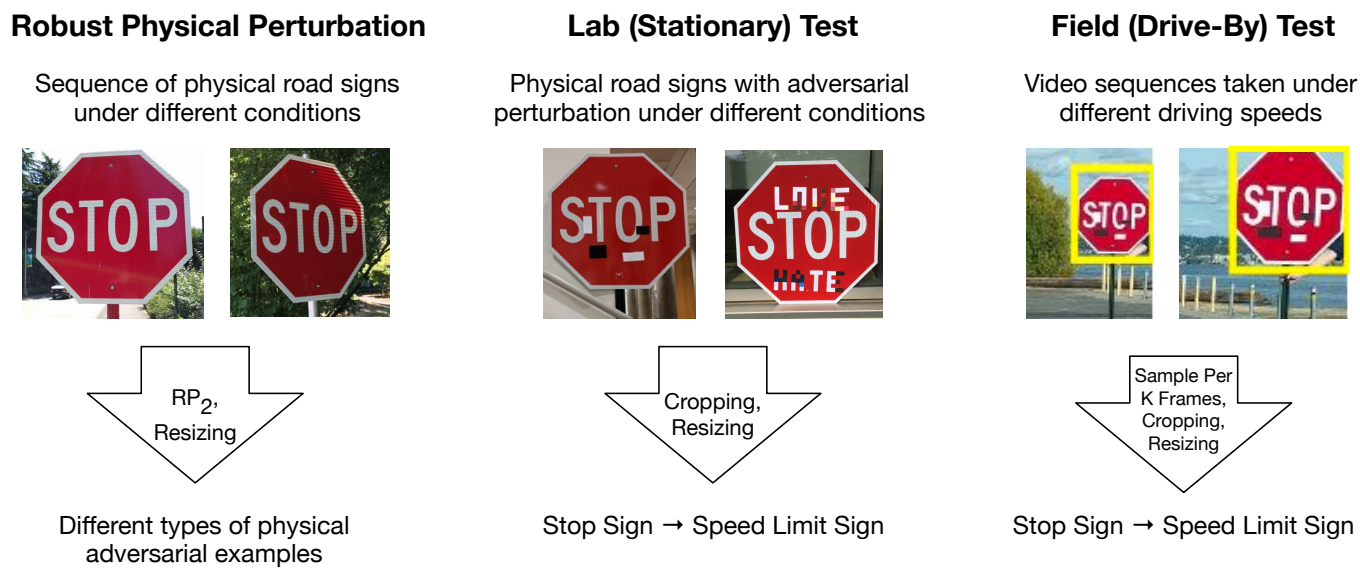


Fig. 2: Pipeline for generating and evaluating physical adversarial perturbations in real world.

4) Given the lack of a standardized methodology in evaluating physical adversarial perturbation, we propose an evaluation methodology to study the effectiveness of physical perturbations in real world scenarios. For the road sign recognition system, the methodology consists of two stages: a stationary test, and a drive-by test using a vehicle. The tests aim to capture dynamic real-world conditions which an autonomous vehicle might experience (Section V).
5) We provide a thorough evaluation of our physical adversarial examples against LISA-CNN and GTSRB*-CNN using the proposed methodology. We find that:

   LISA-CNN In a stationary test: a subtle poster attack causes a Stop sign to be misclassified as a Speed Limit 45 sign in 100% of the testing conditions (15 out of 15 images); a camouflage graffiti attack and an abstract art attack cause a Stop sign to be misclassified as a Speed Limit 45 sign in 66.67% (10 out of 15) and 100% (15 out of 15) of the test cases, respectively.
   LISA-CNN In a drive-by test: a subtle poster attack causes a Stop sign to be misclassified as a Speed Limit 45 sign in 100% of the extracted video frames (37 out of 37 frames); a camouflage abstract art attack has the same misclassification effect in 84.8% of the extracted video frames (28 out of 33 frames; we sampled every 10 frames in both cases).
   GTSRB*-CNN In a stationary test: a camouflage abstract art attack causes a Stop sign to be misclassified as a Speed Limit 80 sign in 80% of all test cases (12 out of 15 images).
   GTSRB*-CNN In a drive-by test: a camouflage abstract art attack causes a Stop sign to be misclassified as a Speed Limit 80 sign in 87.5% of the extracted video frames (28 out of 32 frames; we sampled every 10 frames).

These results provide a strong case for the potential consequences of adversarial attacks on deep learning models that interact with the physical world. We believe this work can serve to inform future defense research and raise awareness on risks that physical learning systems might face. Please visit https://iotsecurity.eecs.umich.edu/#roadsigns for sample images, videos, and other resources..

II. Related Work

We survey the related work in generating adversarial examples and note that most of these works assume digital access to the input vectors. Specifically, given a classifier $f_\theta(\cdot)$ with parameters θ and an input x with ground truth label y for x, it is possible to generate an adversarial example x' that is close to x but causes the classifier to output $f_\theta(x') \neq y$ (in untargeted attacks), or for a specific y', $f_\theta(x') = y'$ (in targeted attacks). We also provide a discussion of recent efforts at making adversarial examples work in the physical world.

A. Adversarial Examples

Adversarial examples are an active area of research. Goodfellow *et al.* proposed the fast gradient sign (FGS) method to add small magnitude perturbations that fool classifiers by calculating the gradient once, leading to untargeted attacks [5]. The generated adversarial instance x' is computed as follows:

$$x' = P(x + \epsilon\ \textbf{sign}(\bigtriangledown_\theta J(x, y))), \tag{1}$$

where **sign** is a function that outputs the sign of its operand, ϵ is a parameter controlling the magnitude of the perturbation, $\bigtriangledown_\theta$ is the gradient with respect to the model parameters, and $P(\cdot)$ is a projection function that maps each dimension of the feature vector x to the valid range of pixel values, i.e. [0, 255]. The loss function $J(x, y)$ computes the classification loss based on the feature vector x and the corresponding label y.

Another approach for generating adversarial perturbations uses an iterative optimization-based algorithm to search for

perturbations under certain constraints [3], [15]. This involves solving the following objective function:

$$\arg\min_{x'} \lambda d(x, x') + J(x, y) \tag{2}$$

where the function $d(x, x')$ serves as a distance function, such as Euclidean distance or L1 norm, and λ is a parameter controlling the contribution of that term. Both of the previously mentioned types of untargeted attack (FGS and optimization methods) can be modified to generate targeted attacks by simply minimizing the distance between the perturbed instance x' and target label y'.

Universal Perturbations is another untargeted attack for generating adversarial examples, but rather then generating an adversarial input, the attack generates a perturbation vector [19]. The vector can be applied to any image and fool a classifier with high probability. Since this algorithm seeks the nearest class for misclassification, it is not easy to generate universal perturbations for targeted attacks.

The relevant common properties of the mentioned existing algorithms for generating adversarial examples are: (1) They assume digital access to the input vectors of the DNN. This assumption might be too strong for a physical attack on a deep learning model. If an adversary can modify the digital pixel values, they can simply replace the input with an object of their choosing and do not need to resort to sophisticated algorithms. (2) They optimize for small magnitude perturbations so that the modified images are "invisible" to human perception. This assumption makes sense in digital image processing, as the values of the input are set precisely and the perturbations do not get destroyed when fed into the neural network. However, if such perturbations are fabricated in the physical world, losses at each stage of the process (fabrication imperfections, camera imperfections, environmental interference) can destroy the information contained in them. Indeed, recent work has demonstrated that three existing algorithms are ineffective in the physical world [16]. By contrast, we choose to create perturbations that are visible, but inconspicuous. Our algorithm is the first one to generate robust physical perturbations for real-wold applications, such as road signs recognition.

B. Physical-Realizability of Adversarial Perturbations

Recent work has examined the effectiveness of adversarial perturbations in the physical world. Sharif *et al.* attacked facial recognition systems by printing adversarial perturbations on the frames of eyeglasses [26]. Their work demonstrated successful physical attacks in relatively stable physical conditions with little variation in pose, lighting, distance from the camera, and camera angle. These assumptions might be applicable for facial recognition systems as an access control mechanism, but are not applicable to the majority of other critical real-world applications. For instance, in road sign recognition, the conditions under which an image is taken inherently include more variability. A car approaching a road sign may capture images of the sign from a diverse range of distances and angles. Both of these factors can lower the resolution of the resulting image and so distort the shape of the perturbation and reduce its effectiveness. In our work, we explicitly design our perturbations to be effective in the presence of these diverse physical-world conditions (large distances/angles and resolution changes).

Furthermore, Kurakin *et al.* showed that printed adversarial examples can be misclassified when viewed through a smartphone camera [11]. In this attack, a *digital* image's pixels are manipulated and then printed. When a picture of the printout is taken and fed into the classifier, the main object depicted in the image is incorrectly labeled even though there are no immediately obvious malicious changes. 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 transformations [1]. However, neither of these works modify physical objects—they are limited to printing out the entire image, including background scenery. Furthermore, both works: (1) Hide perturbations in the images of relatively complex objects. (2) Evaluate printed adversarial examples in controlled lab settings without much physical variance. In our work, we modify physical objects (road signs) that are relatively simple objects making it harder to conceal perturbations, and we perform real-world tests—*e.g.*, recording video while driving outside where we cannot control any of the environmental factors including the background scenery of the road signs.

Finally, Lu *et al.* performed experiments with physical adversarial examples of road sign images [16]. They computed a perturbation of the image that is imperceptible to humans and print the resulting image on A4 paper. They concluded that adversarial examples are not effective in the physical world. By contrast, our work demonstrates that it *is* possible to generate physical adversarial perturbations, implying that there are potential security and safety concerns in cases where DNNs are used for control of physical objects such as cars.

III. Problem Statement

Our goal is to demonstrate that it is possible to create robust physical perturbations of real-world objects that trick deep learning classifiers into returning incorrect class labels even when images are taken under different physical conditions, and even at extreme angles and distances. In this work, we focus on deep neural networks applied to road sign recognition because of the critical role of these objects in road safety and security.

A. U.S. Road Sign Classification

To the best of our knowledge, there is current currently no publicly available road-sign classifier for U.S. road signs. Therefore, we use the LISA dataset [18], a U.S. sign dataset comprised of 47 different road signs to train a DNN-based classifier. This dataset does not contain equal numbers of images for each sign. In order to balance our training data, we chose 17 common signs with the most number of training examples. Furthermore, since some of the signs dominate the dataset due to their prevalence in the real world (*e.g.*, Stop and Speed Limit 35), we limit the maximum number of examples used for training to 500 per sign. Our final dataset includes commonly used signs such as Stop, Speed Limits, Yield, and Turn Warnings. Finally, the original LISA dataset contains image resolutions ranging from 6×6 to 167×168 pixels. We resized all images to 32×32 pixels, a common input size for other well-known image datasets such as CIFAR10 [10].

TABLE I: Description of the subset of the LISA dataset used in our experiments.

Sign Type	Training Example Size	Test Example Size
addedLane	242	52
keepRight	275	56
laneEnds	175	35
merge	213	53
pedestrianCrossing	500	209
school	104	29
schoolSpeedLimit25	86	19
signalAhead	500	163
speedLimit25	275	74
speedLimit30	99	41
speedLimit35	500	103
speedLimit45	114	27
speedLimitUrdbl	109	23
stop	500	398
stopAhead	124	44
turnRight	70	22
yield	197	39
Total	4083	1387

TABLE II: Architecture description of LISA-CNN. The network reads in $32 \times 32 \times 3$ images as input.

Layer Type	Number of Channels	Filter Size	Stride	Activation
conv	64	8x8	2	ReLU
conv	128	6x6	2	ReLU
conv	128	5x5	1	ReLU
FC	17	-	-	Softmax

Table I summarizes our final training and testing datasets.[2]

We set up and trained our road sign classifier in TensorFlow using this refined dataset. The network we used was originally defined in the Cleverhans library [20] and consists of three convolutional layers followed by a fully connected layer (see Table II). Our final classifier accuracy is 91% on the test dataset. For the rest of the paper, we refer to this classifier as LISA-CNN for simplicity.

B. Improving the Classifier

In order to test the versitility of RP_2, we also run our attack against a classifier trained with a larger road sign dataset, the German Traffic Sign Recognition Benchmark (GTSRB) [27]. For this purpose, we use a publicly available implementation [29] of a multi-scale CNN architecture that has been known to perform well on road sign recognition [24]. Our goal is to guarantee that our attack is effective across different training datasets and network architectures. Because we did not have access to German Stop signs for our experiments, we replace the German Stop signs in GTSRB with the entire set of U.S. Stop sign images in LISA. After training, our classifier achieves 95.7% accuracy on the test set, which also had the German Stop signs replaced with U.S. Stop signs.[3] For the rest of the paper, we refer to this classifier as GTSRB*-CNN. The network architecture is described in Table III.

TABLE III: Architecture description of GTSRB*-CNN. The network reads in $32 \times 32 \times 3$ images as input.

Layer Type	Number of Channels	Filter Size	Stride	Activation
conv	3	1x1	1	ReLU
conv	32	5x5	1	ReLU
conv	32	5x5	1	ReLU
maxpool	32	2x2	2	-
conv	64	5x5	1	ReLU
conv	64	5x5	1	ReLU
maxpool	64	2x2	2	-
conv	128	5x5	1	ReLU
conv	128	5x5	1	ReLU
maxpool	128	2x2	2	-
FC	1024	-	-	ReLU
FC	1024	-	-	ReLU
FC	43	-	-	Softmax

C. Threat Model

In contrast to prior work, we seek to physically modify an existing road sign in a way that causes a road sign classifier to output a misclassification while keeping those modifications inconspicuous to human observers. Here, we focus on evasion attacks where attackers can only modify the testing data and do not have access to the training data (as they would in poisoning attacks).

We assume that attackers do not have digital access to the computer systems running the classifiers. If attackers have this superior level of access, then there is no need for adversarial perturbations—they can simply feed malicious input data directly into the model to mislead the system as they want, or they can compromise other control software, completely bypassing all classifiers.

Following Kerckhoffs' principle [25], it is often desirable to construct defenses that are robust in the presence of white-box attackers. As one of the broader goals of our work is to inform future defense research, we assume a strong attacker with white-box access, *i.e.*, an attacker gains access to the classifier after it has been trained [22]. Therefore, although the attacker can only change existing physical road signs, they have full knowledge of the classifier and its architecture. Finally, due to the recent discovery of transferability [21], black-box attacks can be carried out using perturbations computed with white-box access on a different model. As our goal is to inform future defenses, we will focus on white-box attacks in this paper.

Specific to the domain of autonomous vehicles, future vehicles might not face this threat as they might not need to depend on road signs. There could be databases containing the location of each sign. Moreover, with complete autonomy, vehicles might be able to manage navigating complex traffic flow regions by means of vehicle-to-vehicle communication alone. However, these are not perfect solutions and some have yet to be developed. An autonomous vehicle might not be able to rely solely on a database of road sign locations. Such records might not always be kept up-to-date and there might be unexpected traffic events such as construction work and detours due to accidents that necessitate traffic regulation with

[2]speedLimitUrdbl stands for unreadable speed limit. This means the sign had an additional sign attached and the annotator could not read due to low image quality

[3]This test set consists of all GTSRB test images except the German Stop sign images. In addition, we include our own images of 181 US Stop signs. None of the images in the test set are present in either the training set or the validation set of the network. We will release the set of 181 real-world US Stop sign images we took after publication.

temporary road signs. We demonstrate that physically modifying real-world objects to fool classifiers is a real threat, but we consider the rest of the control pipeline of autonomous vehicles to be outside the scope of our work.

Attack Generation Pipeline. Based on our threat model, the attack pipeline proceeds as follows:

1) Obtain several clean image of the target road sign without any adversarial perturbation under different conditions, including various distances and angles.
2) Use those images after appropriate pre-processing as input to RP_2 and generate adversarial examples.
3) Reproduce the resulting perturbation physically by printing out either the entire modified image in the case of poster-printing attacks or just the relevant modified regions in the case of sticker attacks.
4) Apply the physically reproduced perturbation to the targeted physical road signs.

IV. Robust Physical Perturbations

We present an algorithm to generate physical adversarial examples that are robust to varying distances and angles. We take an optimization-based approach to produce such a perturbation. However, unlike prior work, we formulate the objective function such that the perturbation does not affect the background of the real-world target object. RP_2 achieves this using a masking matrix. Based on the shape of that masking matrix, we discuss two types of physical adversarial examples.

A. RP_2

Our algorithm searches in a specific physical region for a small perturbation δ to be added to the original input $x' = x+\delta$ such that x' is misclassified by the model. This can be achieved by taking an optimization-based approach [3], [15]. Specifically, to generate untargeted adversarial examples, we approximate the solution to the following objective function:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p - J(f_\theta(x+\delta), y) \tag{3}$$

To generate targeted adversarial examples, we modify the objective function as follows:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + J(f_\theta(x+\delta), y^*) \tag{4}$$

Here, $||\delta||_p$ denotes the ℓ_p norm of δ, which is a metric to quantify the magnitude of the perturbation δ. y is the ground truth of x, and y^* denotes the label target class.

In the image domain, for a two-dimensional perturbation vector $\delta = \left[\delta_{(1,1)}, \ldots, \delta_{(H,W)}\right]$, the ℓ_p norm for $p > 0$ is given by:

$$||\delta||_p = \Big(\sum_{i,j} |\delta_{(i,j)}|^p\Big)^{1/p} \tag{5}$$

By convention, the ℓ_0 norm is the total number of perturbed pixels while the ℓ_∞ norm is the magnitude of the maximum perturbation.

$J(\cdot\ ,\ \cdot)$ is the loss function, which measures the difference between the model's prediction and either the ground truth label (non-targeted case) or the adversarial label (targeted case). λ is a hyperparameter that controls the regularization of the distortion.

To ensure that that the generated adversarial perturbation works robustly in the physical world, we collect a set of clean images of road signs under varying environmental conditions (distances, angles, lighting). Our final loss function is calculated based on all of this data as shown below. For untargeted attacks, the objective function is:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i+\delta), y) \tag{6}$$

Similarly, for targeted attacks, our objective function is:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i+\delta), y^*) \tag{7}$$

Computing Perturbation Masks. The mask is a region within which our algorithm constrains adversarial perturbation. We define masks in the shape of graffiti and abstract art so that the modifications to road signs are inconspicuous, and so that they lead observers to dismiss those changes as vandalism. By controlling the size and shape of the mask, the perturbation can be made relatively subtle. 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.

Optimizing Spatially-Constrained Perturbations. Having computed a mask, we modify the formulation 7 of the objective function to generate targeted, adversarial examples. Specifically, we restrict the terms of the objective function to operate exclusively on masked pixels. In our implementation, we use the Adam optimizer to minimize the following function:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x\cdot\delta||_p + NPS(M_x\cdot\delta) + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i + M_x\cdot\delta), y^*) \tag{8}$$

To improve the printability of the adversarial perturbation we can add an additional term to the objective function. Using the approach outlined by Sharif *et al.* [26], we can compute the non-printability score (NPS) of an adversarial perturbation. Given a perturbation vector, δ, and a set of printable tuples, P, the non-printability score is given by:

$$NPS(\delta) = \sum_{\hat{p}\in\delta}\prod_{p'\in P} |\hat{p} - p'| \tag{9}$$

B. Types of physical adversarial examples

Although both our attack classes use masks, depending on the size/shape of a given mask, we define two types of physical adversarial examples:

Subtle Poster. This type of physical adversarial example is a true-size poster printed version of a road sign that can be overlayed on a real road sign in the physical world. The mask is in the shape of the road sign. For example, the mask is an octagon for a Stop sign. This type requires a true-sized poster because the adversarial perturbations are situated across

the entire surface area of a sign. In practice, it is harder to physically realize such perturbations because it would either require manufacturing a sign with irregularities in its coloration, or destructive modifications to an existing sign (we plan on exploring these options in future work). Therefore, we choose to use poster paper as a low-cost means of achieving physical realizability.

Camouflage Sticker. This type of physical adversarial example is a set of stickers that can be applied to an existing road sign in the physical world. The mask is smaller than the sign itself, and takes the shape of common vandalism. For example, we show attacks with stickers in the form of abstract art, and graffiti. This type of physical adversarial example is the easiest to fabricate as it only requires access to a standard-format printer.

A common property for both types of physical adversarial examples in this paper is that they are inconspicuous to human observers. A *digital* adversarial example can be created such that it is very hard, if not impossible, for humans to detect the modification (see [5] for an example of an adversarial panda). In such cases, a camera might similarly not perceive the change in colors when the perturbation is physically realized. We handle this using two approaches: (1) We discover through our experiments that constraining the perturbation to the whole area of the sign results in low-magnitude and inconspicuous adversarial modifications. We believe that this is due to the fact that more pixels can be perturbed in this case as compared to when the perturbation area is further restricted. (2) Our sticker (graffiti and art) masks are specifically designed to "hide in the human psyche." As vandalism is common, humans tend to not perceive such visible additions to a road sign as indicative of an attack.

V. Evaluating Physical Adversarial Perturbations

To our best knowledge, there is currently no standardized methodology of evaluating physical adversarial perturbations. Therefore, we first discuss possible factors that affect perturbations in the physical world in the context of autonomous vehicles. Motivated by that discussion, we introduce an evaluation methodology that captures a range of physical-world factors with the goal of effectively testing physical attacks on deep learning models.

A. Evaluation Components

Cyber-physical systems can experience a range of varying conditions in the physical world. Specific to our domain, autonomous vehicles experience varying distances, angles, and lighting, and occlusions due to debris and weather. A physical attack on a road sign must be able to survive such changing conditions and still be effective at fooling the classifier. Additionally, the attacker is restricted to only manipulating the road signs themselves. Such restrictions break assumptions that current adversarial deep learning algorithms make about their level of access to the digital inputs of the classifier, and the kinds of values the adversarial perturbation might contain. Therefore, we list a few physical components that can impact the effectiveness of adversarial perturbations.

- **Environmental Conditions:** The distance and angle of a camera in an autonomous vehicle with respect to a road sign vary continuously. The resulting images that are fed into a road sign 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. This is further complicated by other environmental factors such as changes in lighting conditions and the presence of debris on the camera or on the road sign.
- **Spatial Constraints:** Current algorithms work on digital images, and they add adversarial perturbations to all parts of the image, including any 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.
- **Fabrication Error:** In order 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 [26].
- **Resolution Changes:** The camera in a vehicle will not necessarily produce images whose dimensions are the same as the size of input of the DNNs recognition system—some upscaling or downscaling can occur. Adversarial perturbations would need to survive such resolution changes and be correctly mapped to their corresponding physical locations.
- **Physical Limits on Imperceptibility:** An attractive feature of current adversarial deep learning algorithms is that their perturbations to a digital image are often so minor 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 that an autonomous vehicle may use.

In order to successfully and physically attack deep learning models, an attacker should account for these real world challenges when computing perturbations. In our evaluation methodology, we focus on three major components that impact how physical objects are classified by DNNs in cyber-physical systems in our chosen domain of road sign classification for autonomous vehicles.

Distance. First, a camera in a vehicle approaching a sign will take a series of images at regular intervals. As the distance changes, so does the level of detail captured in an image. Any successful physical perturbation must cause targeted misclassification in a range of distances.

Angle. Next, a camera can have a wide range of angles relative to the sign. For instance, the vehicle might be on a lane far away from the shoulder of the road and thus perceive the sign from a sharp angle. Alternatively, it could be on the lane closest to the shoulder and thus perceive the sign straight-on. Successful

attacks need to be able to fool the classifier for images taken from various angles.

Resolution. Finally, the perturbations need to cause misclassification even when the images are downsized for performance optimization. Cyber-physical systems in general and autonomous vehicles in particular impose strict requirements on how long operations should take. Classifiers operating on high-resolution inputs might not fulfill those goals. Thus, DNNs used in control decisions potentially favor smaller input sizes. As a result, physically applied perturbations might be distorted in the pre-processing pipeline; they might stretch or shrink and their colors might mix with those of neighboring pixels. For a physical attack to succeed, these modifications need to be taken into account.

B. Evaluation Methodology

Given the lack of a standardized evaluation process for physical attacks on deep learning models, we develop an evaluation methodology to account for varying physical conditions based on our analysis above. Our methodology contains two evalutation steps: (1) lab tests, where images are obtained in a controlled lab setting, while the environmental conditions are varied, and (2) field tests, where images are obtained in the field based on the specific applications when environmental conditions are not explicitly controlled. In field testing, a broader range of environmental conditions may be present that were not encountered in the lab tests.

For our chosen application, road sign classification, we refer to our lab tests as *stationary tests*. We obtain images using a stationary camera and vary the distance and viewing angle of the camera relative to a sign. For field testing, we perform *drive-by tests*. We mount a camera to a car and record a video of the drive. Afterwards, we extract the video frames and use them for classification. We now describe both evaluation steps in detail in the context of traffic sign classification.

Stationary Tests. This occurs in a stationary setting. Obtain images of a road sign at a variety of distances and angles.

1) Obtain a set of clean images C at varying distances $d \in D$, a set of distances, and at varying angles $g \in G$, a set of angles to the right relative to the normal vector of the plane of the sign. 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. We choose $D = \{5', 10', 15', 20', 25', 30', 40'\}$, and $G = \{0°, 15°, 30°, 45°, 60°\}$. We only use the full set G for $d = 5$' We choose $\{0°, 15°, 30°\}$ for $d = 10$', $\{0°, 15°\}$ for $d = 15$, 20' and $\{0°\}$ for all remaining d. We believe that these angles capture the possible angle/distance combinations in the real world.[4] See Figure 3 for a graphical representation of this setup.
2) Obtain a set of physically perturbed sign images for evaluation using the same angles and distances as when creating C. For $c \in C$, the corresponding adversarial image is represented by $\mathcal{A}(c)$
3) Crop and resize the images in c and corresponding $\mathcal{A}(c)$ to the input size for the model being used for classification.
4) Compute the attack success rate of the physical perturbation using the following formula:

$$\frac{\sum_{c \in C} \mathbb{1}_{\{f_\theta(\mathcal{A}(c)^{d,g}) = y^* \;\&\&\; f_\theta(c^{d,g}) = y\}}}{\sum_{c \in C} \mathbb{1}_{\{f_\theta(c^{d,g}) = y\}}} \tag{10}$$

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.[5]

Note that an image $\mathcal{A}(c)$ that causes misclassification is considered as successful attack only if the original image c with the same camera distance and angle is correctly classified.

Drive-By Tests. This occurs in a dynamic setting. 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 feet away from the sign. Our driving track was straight without curves.
2) 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.
3) Screen grab frames from the resulting video. An open question is to determine whether to process all frames, or to grab frames every k frames, and whether this choice of k affects the attack accuracy. In our experiments, we use $k = 10$. We also tried $k = 15$ and did not observe any significant change in terms of the attack success rates. In general, autonomous vehicles will likely use $k > 1$ rather than $k = 1$ in order to optimize performance, but this makes no significant difference in our evaluation results.
4) Crop and resize each sampled frame from Step 3 to the input size of the model.
5) Repeat the steps for a "clean" sign and for a sign with perturbations applied.
6) Calculate the success rate using the following formula:

$$\frac{\sum_{v \in V} \mathbb{1}_{\{f_\theta(\mathcal{A}(v^k)) = y^* \;\&\&\; f_\theta(v^k) = y\}}}{\sum_{v \in V} \mathbb{1}_{\{f_\theta(v^k) = y\}}} \tag{11}$$

where v^k represent the kth sampled frame here.

We hypothesize that if both types of tests produce high success rates, the attack is likely to be successful in most cases under real driving conditions.

We note that our current testing methodology does not control lighting variations or occlusions. As is evident from experimental data discussed later (Section VI), our lighting conditions vary from indoor light to natural light. We do ensure that neither the road sign nor the camera are occluded in any way (beyond the natural effects of a car's windshield and the

[4]The minimum stopping distance for a car traveling at 10 mph is about 5'. The minimum stopping distance for a car traveling at 30 mph is about 45'. Changes in the camera angle relative the the sign will normally occur when the car is turning, changing lanes, or following a curved road.

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

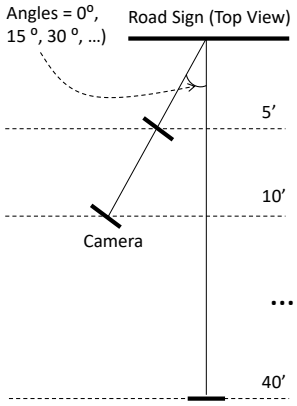


Fig. 3: For all our physical realizability experiments, we test perturbation by varying the angles and distances of the camera from the road sign. Distances vary from 5' to 40', and angles vary from 0°to 60°at 5', 0°to 30°at 10', 0°to 15°at 15' and 20'. From 25' onwards, we do not vary angle. All angle variations are to the left of the sign (U.S. driving).

atmosphere). As future work, we envision additions to our evaluation methodology that examine the effect of lighting and occlusion more thoroughly.

Safety while performing a drive-by test. We perform our experiments in a mostly empty parking lot. We take care to not perform experiments while other moving vehicles are present. We also take care to not disrupt the normal traffic flow, if any, in the parking lot during the course of our data collection. We do not modify any road signs belonging to the city. We only modify signs that we own. Furthermore, we do not fix the signs to the road. Rather, they are held in position by a team member. Finally, we ensure that testing is carried in a way that could not cause harm to pedestrians or other parking lot users.

VI. Experimental Results

We begin by comprehensively evaluating our attacks on the LISA-CNN using our proposed evaluation methodology. First, we run stage 1 of the evaluation in a stationary setting. We test the effectiveness of two classes of attacks (poster-printing and sticker) using three types of perturbations (subtle, camouflage graffiti, and camouflage art). In poster-printing attacks, we print a digitally perturbed true-sized image of either a Stop sign or a Right Turn sign, cut the print into the shape of the sign, and overlay it on a physical road sign. Subtle perturbations cause the Stop sign to be misclassified as a Speed Limit 45 sign, the attack's target class, in 100% of test cases. Poster-printed camouflage graffiti caused the Right Turn sign to to be misclassified as a Stop sign, the attack's target class, 73.33% of the time. In sticker attacks, we print the perturbation on paper, cut it out, and stick it to a Stop sign. Sticker camouflage graffiti attacks caused the Stop sign to be misclassified as a Speed Limit 45 sign 66.67% of the time and sticker camouflage art attacks resulted in a 100% targeted-attack success rate.

Next, we perform drive-by testing for attacks on the LISA-CNN. Our attack is able to fool the classifier into believing that a Stop sign is a Speed Limit 45 sign in 100% of the extracted frames for a subtle poster attack, and in 84.8% of the extracted frames for a camouflage abstract art attack.

Finally, we also run our full evaluation methodology for camouflage abstract art attacks targeting the GTSRB*-CNN. Our attack is able to fool the classifier into believing that a Stop sign is a Speed Limit 80 sign in 80% of the testing conditions for a stationary test, and in 87.5% of the extracted frames for a drive-by test.

A. Poster-Printing Attacks

We first show that an attacker can overlay a true-sized poster-printed perturbed road sign over a real-world sign and achieve misclassification into a target class of her choosing. The attack has the following steps:

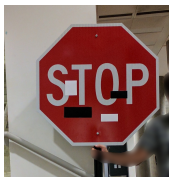
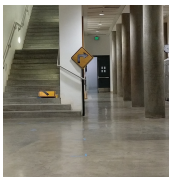
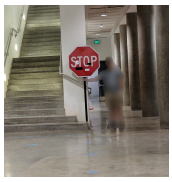
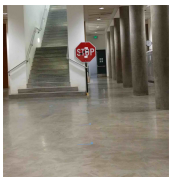
Step 1. The attacker obtains a series of high resolution images of the sign under varying angles, distances, and lighting conditions. We use 34 such images in our experiments. None of these images were present in the datasets used to train and evaluate the baseline classifier.
Step 2. The attacker then crops, rescales, and feeds the images into RP_2 and uses equation (8) as the objective function. She takes the generated perturbation, scales it up to the dimensions of the sign being attacked, and digitally applies it to an image of the sign.
Step 3. The attacker then prints the sign (with the applied perturbation) on poster paper such that the resulting print's physical dimensions match that of a physical sign. In our attacks, we printed $30'' \times 30''$ Stop signs and $18'' \times 18''$ Right Turn signs.
Step 4. The attacker cuts the printed sign to the shape of the physical sign (octagon or diamond), and overlays it on top of the original physical sign.

We use our methodology from Section V to evaluate the effectiveness of such an attack. In order to control for the performance of the classifier on clean input, we also take images of a real-size printout of a non-perturbed image of the sign for each experiment. We observe that all such baseline images lead to correct classification in all experiments.

For the Stop sign, we choose a mask that exactly covers the area of the original sign in order to avoid background distraction. This choice results in a perturbation that is similar to that in existing work [11] and we hypothesize that it is imperceptible to the casual observer (see the second column of Table IV for an example). In contrast to some findings in prior work, this attack is very effective in the physical world. 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 the target class is 80.51% with a standard deviation of 10.67%.

For the Right Turn sign, we choose a mask that covers only the arrow since we intend to generate subtle perturbations and we hypothesize that they can only be hidden in this region of the sign. In order to achieve this goal, we increase the regularization parameter λ in equation (8) to demonstrate small magnitude of perturbations. We achieve a 73.33% targeted-attack success rate; see the third column of Table IV for detailed results. It is interesting to note that in the cases where the physically perturbed sign was not classified into the target, it was still misclassified. In 3 of these 4 cases, a different warning sign was present. We hypothesize that given the similar appearance

TABLE IV: Sample experimental images for the attacks detailed in Table V and Table VI at a selection of distances and angles.

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%

of warning signs, small perturbations are sufficient to confuse the classifier.

Remarks. For subtle poster attacks on Stop signs, RP_2 achieves a 100% targeted-attack success rate in stationary testing conditions. For subtle poster attacks on Right Turn signs, RP_2 achieves a 73.33% targeted-attack success rate, and a 100% untargeted-attack success rate (with 33.7% of the images classified as an Added Lane sign instead of the target label of Stop). We conclude that subtle poster attacks in general are robust in the physical world in stationary settings.

B. Sticker Attacks

Next, we demonstrate how an attacker can generate perturbations that are easier to apply on a real-world sign in the form of a sticker by constraining the modifications to a region resembling graffiti or art. The steps for this type of attack are:

Step 1. The attacker generates the perturbations digitally by using RP_2 just as in Section VI-A.
Step 2. The attacker prints out the Stop sign in its original size on a poster printer and cuts out the regions that the perturbations occupy.
Step 3. The attacker applies the cutouts to the sign by using the remainder of the printed sign as a stencil.

The fourth and fifth columns of Table IV show the result of the above steps for two types of perturbations. In the stationary setting, we achieve a 66.67% targeted-attack success rate into our target class for the graffiti sticker attack and a 100% targeted-attack success rate for the sticker camouflage art attack.

1) Camouflage Graffiti Attack: Following the process outlined above, we generate perturbations in the shape of the text "LOVE HATE" and physically apply them on a real Stop sign (see the fourth column of Table IV). Table V shows the results of this experiment. This attack succeeds in causing 73.33% of the images to be misclassified. Of the misclassified images, only one image was classified as a Yield sign rather than a Speed Limit 45 sign, thus resulting in a 66.67% targeted-attack success rate with an average confidence of 47.9% and a standard deviation of 18.4%. For a baseline comparison, we took pictures of the Stop sign under the same conditions without any sticker perturbation. The classifier correctly labels the clean sign as a Stop for all of the images with an average confidence of 96.8% and a standard deviation of 3%.

TABLE V: Summary of targeted physical perturbation experiments on LISA-CNN with a poster-printed Stop sign (subtle attacks) and with a real Stop sign (camouflage graffiti attacks, camouflage art attacks). This table lists the top two confidence values for the classification of images in phase 1 of the evaluation and the misclassification target was Speed Limit 45. See Table IV for examples 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 VI: Poster-printed perturbation (faded arrow) attack against the LISA-CNN for a Right Turn sign at varying distances and angles. See example images in Table IV. Our targeted-attack success rate is 73.33%.

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
5' 0°	Stop (0.39)	Speed Limit 45 (0.10)
5' 15°	Yield (0.20)	Stop (0.18)
5' 30°	Stop (0.13)	Yield (0.13)
5' 45°	Stop (0.25)	Yield (0.18)
5' 60°	Added Lane (0.15)	Stop (0.13)
10' 0°	Stop (0.29)	Added Lane (0.16)
10' 15°	Stop (0.43)	Added Lane (0.09)
10' 30°	Added Lane (0.19)	Speed limit 45 (0.16)
15' 0°	Stop (0.33)	Added Lane (0.19)
15' 15°	Stop (0.52)	Right Turn (0.08)
20' 0°	Stop (0.39)	Added Lane (0.15)
20' 15°	Stop (0.38)	Right Turn (0.11)
25' 0°	Stop (0.23)	Added Lane (0.12)
30' 0°	Stop (0.23)	Added Lane (0.15)
40' 0°	Added Lane (0.18)	Stop (0.16)

2) Camouflage Abstract Art Attack: Finally, we execute a sticker attack that applies the perturbations resembling abstract art physically to a real-world sign. While executing this particular attack, we notice that after a resize operation, the perturbation regions were shortened in width at higher angles. This possibly occurs in other attacks as well, but it has a more pronounced effect here because the perturbations are physically smaller on average than the other types. We compensate for this issue by increasing the width of the perturbations physically. In this final test, we achieve a 100% targeted-attack success rate, with an average confidence for the target of 62.4% and standard deviation of 14.7% (See the fifth column of Table IV for an example image).

Remarks. For sticker attacks, we observe a 66.67% targeted-attack success rate when the perturbations are shaped as letters, and a 100% targeted-attack success rate when the perturbations are shaped as abstract art rectangles in stationary settings. We believe that the reason for the increased success rate in the abstract art case is due to the perturbations being physically situated over "vulnerable" regions of the sign that we computed using the L1 norm (see Section VII for more discussion). We conclude that sticker attacks are robust to a range of varying physical conditions in stationary settings.

C. Drive-By Testing Results

Per our evaluation methodology, we conduct drive-by testing of the perturbation of a Stop sign. In our baseline test where we record two consecutive videos of a clean Stop sign from a moving vehicle and perform frame grabs at $k = 10$, the Stop sign is correctly classified in all frames. We test subtle and abstract art perturbations for our LISA-CNN. We grab a frame every 10 frames ($k = 10$, for two consecutive video recordings), resize the images to the input size of the classifier, and obtain a classification. Table VII summarizes the results—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. Upon a close examination of the cases in which the abstract art attack failed, we observe that images taken at distances of approximately less than 5 feet from the sign are classified as Stop instead of our target of Speed Limit 45. This does not reduce the severity of the attack as it is already too late for a car to take corrective action.

Remarks. For the drive-by test with attacks on LISA-CNN, we observe high targeted-attack success rates: 100% for the subtle poster attack, and 84.8% for the camouflage abstract art attack. We conclude that our most successful targeted-attack categories (in stationary settings), are also successful in a more dynamic setting involving a moving vehicle.

D. Attacks on the GTSRB*-CNN

To show the versatility of our attack algorithms, we create and test attacks for a GTSRB*-CNN (see Section III-A). Based

TABLE VII: Drive-by testing summary for LISA-CNN. In our baseline test, all frames were correctly classified as a Stop sign. In all attack cases, the perturbations are the same as in Table V.

Perturbation	Attack Success	A Subset of Sampled Frames $k = 10$
Subtle poster	100%	
Camouflage abstract art	84.8%	

on our high success rates with the camouflage-art attacks, we decide to create an abstract art sticker perturbation. The last column of Table IV shows a subset of experimental images for a Stop sign with the perturbation applied as stickers. Table VIII summarizes our attack results—our attack tricks 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 tricks the classifier 87.5% of the time. Upon closer examination of the cases in which the drive-by attack failed, we observed that the classifier correctly recognizes the sign as Stop at very close distances to the sign (approximately less than 5 feet away). This is reasonable behavior and does not reduce the severity of the attack for two reasons: (1) It is too late for the car to take corrective action at such a short distance to the sign; (2) Our attack was created for distances larger than 5 feet away from the sign. Additional images of signs at distances smaller than 5 feet can be used during attack optimization to create perturbations to handle such cases.

Remarks. We observe that a camouflage-art attack results in GTSRB*-CNN misclassifying a Stop sign as a Speed Limit 80 sign in 80% of the stationary testing conditions, and in 87.5% of the drive-by testing conditions. We conclude that for a different classifier and a different training dataset, RP_2 is effective at generating robust physical adversarial examples that achieve high targeted-attack success rates in a stationary and drive-by test.

TABLE VIII: A camouflage art attack on GTSRB*-CNN. See example image in Table IV. The targeted-attack success rate is 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)

VII. Discussion

Choice of Masks. During our experiments, we observed that the choice of mask for RP_2 had an impact on the success of the attack and on the imperceptibility of the generated perturbation. The most effective adversarial examples for the Stop sign were produced by choosing a mask that exactly covers the sign, but not its background (these subtle attacks achieved 100% success in both the stationary and the drive-by experiments). We believe that this effectiveness is a result of the numerous pixels modifications. This size of the mask area allows a large fraction of the perturbation to survive re-scaling and stretching when printed and perceived at an angle or from a large distance by a camera. However, the necessity to overlay a poster on the real sign might cause more careful observers to notice this attack (even though it is likely to remain imperceptible to casual observers).

Hence, we chose masks that constrain the perturbation in common graffiti shapes like text. This modification had mixed success; for example, the camouflage graffiti perturbation only achieved a success rate of 66.67%. We hypothesized that these masks constrain the perturbation to regions of the sign that are not necessarily vulnerable (*i.e.*, do not greatly influence classification of the sign). To discover where these sensitive regions might be located, we used a mask in the shape of the sign as before (an octagon-shaped mask for a Stop sign), but we replaced the regularization term in equation 7 of RP_2 with the L1 loss function. This change penalizes the number of modified pixels instead of the magnitude of the modification, thus shrinking the perturbation area. As a result, we observed several relatively small regions where the resulting perturbation was concentrated. Based on our observation, we chose to shape our masks as rectangles that covered those regions only. We picked the rectangular shape so that the resulting perturbation would resemble abstract art and fail to raise suspicion even

in careful observers. The resulting adversarial examples for the LISA-CNN were solid black and white rectangles (the camouflage art attacks). We believe that this can be attributed to the necessity for large modifications in the constrained region—the black corresponds to a large pixel value being subtracted and the white corresponds to a large pixel value being added.

More importantly, our experiments led us to discover two improvements in the physical fabrication process. First, we noticed that the white color of the perturbation does not get perceived as its full (255, 255, 255) RGB value when the physical modifications are fabricated out of regular paper.[6] In order to maximize the probability that white gets perceived closer to its true RGB value, we chose to fabricate the regions out photo paper, a material that reflects more light than normal paper, which increased the intensity of the white perturbation and improved the effectiveness of the attack. Second, we observed that resizing the images taken at large angles and at large distances caused the white to be mixed with other colors from its neighboring pixels.[7] In order to ensure that more of the white pixels get preserved during resizing, we physically enlarged the white stickers. A similar observation for the black regions led us to physically enlarge all stickers relative to the size that was mapped from the digital images produced by RP_2. These human-in-the-loop optimizations led to the reported attack success rates.

Future Work. Our work answers a fundamental question: Can deep learning classifiers be fooled by adversarial modifications to physical objects? However, several open questions remain. One of them is whether casual observers fail to distinguish our attacks from regular vandalism with high confidence. Due to the prevalence of defaced signs (especially in urban areas), we believe that it is reasonable to assume that this is indeed the case. However, a set of user studies may need to be conducted in future work to confirm or reject this hypothesis.

A second open question is whether black-box attacks can be physically realized. Our work demonstrates that a new algorithm and some human-in-the-loop optimizations lead to powerful white-box attacks. However, it is still unclear if existing black-box algorithms can be successful without the assumption of digital access to the inputs of the classifier. Similarly, the transferability of physical adversarial modifications needs to be explored further. Since it is common practice to attack a substitute model trained to resemble a classifier whose internals are hidden in the digital realm, future work should study whether this is feasible for generating physical adversarial examples. If this is the case, it is likely that powerful attacks might be mounted against safety-critical applications of machine learning in the future.

Finally, even though our research makes another step toward understanding the potential threats posed by adversarial modifications in the physical world, further work is necessary in this field for defenses to be developed. Current work has explored countermeasures in the scope of digital adversarial examples, but an effective solution remains elusive [17]. Moreover, we show that carefully crafted physical modifications can even survive large environmental variations such as extreme viewing angles and large distances. We envision that future defense work might utilize the extra limitations of adversarial perturbations in the physical world in order to create a deep learning network robust to such attacks.

[6]This is due to the varying white balance settings in modern digital cameras, which are outside the attacker's control. By contrast, black objects absorb all wavelengths of light instead of reflecting them and so they are impacted less by the camera's white balance. Hence, we used solid blocks of black ink printed on regular paper for the black regions of the perturbation.

[7]Even though this may be remedied by a careful choice of a resizing algorithm, we believe that it is unreasonable to give the attacker the power to pick that algorithm.

VIII. Conclusion

In this paper, we introduced Robust Physical Perturbations (RP_2), an algorithm that generates robust, physically realizable adversarial perturbations. Previous algorithms assumed that the inputs of DNNs can be modified digitally to achieve misclassification, but such an assumption is too strong in many cyber-physical scenarios. An attacker with control over the inputs of the model can simply replace them with an input of her choice. Therefore, attacks on DNNs in cyber-physical systems must apply perturbations to the physical object. The success of these attacks depends on the distance and angle of the object relative to the sensor, the amount of fabrication error of the generated perturbation, and on other environmental variations. By developing and thoroughly evaluating RP_2, we demonstrated that these challenges can be overcome by attackers and that defenses should not rely on any such physical limitations.

In particular, we used RP_2 to create two types of perturbations: *subtle perturbations*, which are small, slightly visible changes to the entire sign, and *camouflage perturbations*, which are visible but inconspicuous perturbations in the shape of graffiti or art. We also trained two different classifiers, LISA-CNN and GTSRB*-CNN, one using the LISA dataset and one using a slightly modified version of the GTSRB datasets.

When evaluated against LISA-CNN, subtle poster perturbations achieved a 100% targeted-attack success rate in both our lab and field tests. When only the perturbations were added to the sign, camouflage art perturbations achieved 100% targeted-attack success during lab testing, and an 84.8% targeted-attack success rate during field tests.

To demonstrate the versatility of our attack algorithm, we performed further evaluation on GTSRB*-CNN. For the attack, we created camouflage art due to its high success rate and small scale of necessary changes to the sign. The camouflage art perturbation achieved an 80% targeted-attack success rate during lab tests, and an 87.5% targeted-attack success rate during field tests.

References

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

[2] 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*. IEEE, 2010, pp. 2130–2135.

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

[4] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

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

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[7] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

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

[9] J. Kos and D. Song, "Delving into adversarial attacks on deep policies," *arXiv preprint arXiv:1705.06452*, 2017.

[10] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[11] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533

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

[13] ——, "Scalable optimization of randomized operational decisions in adversarial classification settings." in *AISTATS*, 2015.

[14] 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.

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

[16] 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.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[18] 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.*, vol. 13, no. 4, pp. 1484–1497, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1109/TITS.2012.2209421

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

[20] 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.

[21] 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.

[22] 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*. IEEE, 2016, pp. 372–387.

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

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

[25] C. E. Shannon, "Communication theory of secrecy systems," *Bell Labs Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[26] 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 23rd ACM SIGSAC Conference on Computer and Communications Security*, 2016. [Online]. Available: https://www.ece.cmu.edu/~lbauer/papers/2016/ccs2016-face-recognition.pdf

[27] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608012000457

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

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

[30] 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.