Robust Physical-World Attacks on Machine Learning Models

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

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

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

***Abstract*—Deep neural network-based classifiers are known to be vulnerable to adversarial examples that can fool them into misclassifying their input through the addition of small-magnitude perturbations. However, recent studies have demonstrated that such adversarial examples are not very effective in the physical world—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 new attack algorithm—Robust Physical Perturbations (RP_2)—that generates perturbations by taking images under different conditions into account. Our algorithm can create spatially-constrained perturbations that mimic vandalism or art to reduce the likelihood of detection by a casual observer. We show that adversarial examples generated by RP_2 achieve high success rates under various conditions for real road sign recognition by using an evaluation methodology that captures physical world conditions. We physically realized and evaluated two attacks, one that causes a Stop sign to be misclassified as a Speed Limit sign in 100% of the testing conditions, and one that causes a Right Turn sign to be misclassified as either a Stop or Added Lane sign in 100% of the testing conditions.**

I. Introduction

Even though Deep Neural Networks (DNNs) have been applied with great success in a variety of areas ranging from speech processing [7] to medical diagnostics [4], recent work has demonstrated that they are vulnerable to adversarial perturbations [3], [6], [8], [10], [11], [17], [18], [21]. Such maliciously crafted changes to the input of DNNs cause them to misbehave in unexpected and potentially dangerous ways.

Yet the effectiveness of such attacks in the physical world has been disputed by concurrent and independent studies. On the one hand, Lu *et al.* study the fast gradient sign, the iterative, and the L-BFGS algorithms and claim that perturbations generated for road signs classifiers are not effective under varying viewing conditions (changing angles and distances) when printed on paper [13]. On the other hand, Athalye and Sutskever demonstrate that more sophisticated algorithms can produce perturbed images that are robust to variations of the camera's perspective when printed out [2].

Although prior studies make significant progress toward mounting attacks on classifiers in the physical world, several key gaps concerning physical realizability remain. First, it is infeasible to add perturbations to the background of an object. Second, it is harder to hide the perturbations in simple objects such as road signs than it is in the complex images used in concurrent work. Third, there are additional physical limits on the imperceptibility of perturbations—current approaches create such highly subtle perturbations that a camera might not capture them under widely varying conditions (long distances, and wide angles).

In this work, we explore the following key question: "Is it possible to create robust and subtle adversarial perturbations on real-world objects?" We focus on road sign classification due to their critical function in road safety and security. If an attacker can *physically* and *robustly* manipulate road signs in a way that, for example, causes a Stop sign to be interpreted as a Speed Limit sign by an ML-based vision system, then that can lead to severe consequences. Our goal is to dive deeply into the nuances of the physical world, and to show that physical adversarial perturbations can exist robustly under realistic assumptions.

Creating adversarial perturbations in the physical world requires overcoming several challenges due to environmental factors: (1) the distance between the camera in the vehicle and the road sign constantly varies depending on the specific road, the specific vehicle, the number of lanes, and the position of the vehicle with respect to the sign; (2) the angle of the camera varies with respect to the sign; (3) lighting; (4) debris either on the road sign or on the vehicle. Additionally, imperfections in the fabrication process of physical perturbations can reduce the effectiveness of an attack. Therefore, a robust and effective attack on DNNs used in vision for vehicles must be able to tolerate multiple sources of error.

Furthermore, current adversarial algorithms aim to create perturbations that are imperceptible to humans [6]. In the physical world, imperceptibility can affect the robustness of the perturbation—the modifications to a road sign can be so subtle that a camera might not be able to perceive those changes. Furthermore, the region where perturbations can be added is limited. In a digital image, an algorithm is free to add perturbations everywhere, including any background imagery that appears behind the road sign at the time the image was taken. In the physical world, it is infeasible to add perturbations to the background of a sign (as there is no fixed background),

*All authors are in alphabetical order.

therefore limiting perturbations to the road sign itself.

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, angles and resolutions of a camera. RP_2 only perturbs the region of a sign instead of background which varies in practice. We add a masking matrix to our objective function to realize such spatial constraints for perturbations.

Using the proposed algorithm, 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, (2) *sticker attacks*, where an attacker fabricates only the perturbations, and then sticks them to an existing road sign. In order to make our attacks less conspicuous to human observers, we generate sticker perturbations that look like relatively common forms of vandalism. For poster-printing attacks, we demonstrate subtle perturbations—small perturbations (that are noticeable only to very careful observers) that occupy the entire region of the sign (*not* including any background imagery). For sticker attacks, we demonstrate camouflage attacks with graffiti and abstract art. Both of our attack classes do not require special resources—only access to a color printer and a camera.

Our experimental results on a road sign classifier under different 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 our controlled testing conditions. RP_2, thus, answers a key, fundamental open question regarding the susceptibility of image classifiers to adversarial inputs in the real world.

Our Contributions.

- We design Robust Physical Perturbations (RP_2), the first algorithm that generates perturbations for road signs that are robust against varying conditions, such as distances, angles, and resolutions. Using this algorithm, 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*, 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.
- Given the lack of a standardized method in evaluating physical attacks, we propose an evaluation methodology to study the effectiveness of adversarial perturbations in the physical world. This methodology captures a sampling of real-world conditions an autonomous vehicle might experience (Section IV).
- We evaluate our attacks using the proposed methodology. We find that:
  - Poster-printed subtle perturbations cause the Stop sign to be misclassified as a Speed Limit sign in 100% of all test cases.
  - Sticker camouflage graffiti perturbations cause the Stop sign to be misclassified as a Speed Limit sign in 66.67% of all test cases.
  - Sticker camouflage art perturbations cause the Stop sign to be misclassified as a Speed Limit sign in 100% of all test cases.

Given our attacking success under different physical conditions, we believe this work can serve to inform future defense research.

II. Related Work

We survey related work in generating adversarial examples, where all techniques assume digital access to the input vectors. Specifically, given a classifier $f_\theta(x)$ and an input x with ground truth label y, 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 discussion on 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 [6]. The generated adversarial instance x' is computed as:

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

where $P(\cdot)$ is a projection function that maps each dimension of 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 feature vector x and the corresponding label y. Here, ϵ represents the magnitude of perturbation.

Another approach is to use an iterative optimization based algorithm to search for perturbations under certain constraints [3], [12]:

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

where, the distance function $d(x, x')$ serves as regularizer. Usually, different functions such as Euclidean distance or L1 norm will be selected based on the actual requirements. Both of the untargeted attack types (FGS and optimization methods) can be easily modified to generate targeted attacks by minimizing the distance between the perturbed instance x' and target label y'.

Universal Perturbations is an untargeted attack that creates perturbations that can be applied to any image [15] and is therefore useful as a black-box attack. Since this algorithm seeks the nearest class (y') for misclassification, it is not easy to generate universal perturbations for targeted attacks.

The relevant common properties of the above existing algorithms for generating adversarial examples are: (1) They assume digital access to the input vectors of the DNN. This

assumption can be too strong for an autonomous vehicle setting; (2) They require the magnitude of perturbations to be "invisible" to human perception. This assumption makes sense in digital image processing, as 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 very subtle perturbations. Indeed, recent work has demonstrated that three existing algorithms are ineffective in the physical world [13]. In contrast, we contribute the first algorithm that generates robust physical perturbations for road signs.

B. Physical-Realizability of Adversarial Perturbations

Recent work has examined whether adversarial perturbations are effective in the physical world. We group these existing perturbations into two types: (1) subtle perturbations, where the magnitude of perturbations is so small that it is generally imperceptible to casual human observers,[1] and (2) camouflage perturbations, where the perturbations do not necessarily have small magnitude, but are focused in defined regions, and take on inconspicuous shapes such as graffiti or art.

1) Subtle Perturbations: Kurakin *et al.* have shown how printed adversarial examples are misclassified when viewed through a smartphone camera [9]. This attack manipulates the *digital* representation of an image, and then prints that manipulated image, which is fed as input to a classifier through a smartphone camera. Physically realizing such an attack for road signs can raise concern in human observers because the approach requires perturbations to be added to the *background* of the road sign (in addition to adding perturbations to the main object in the image). See Figure 1 for an example of what we mean by background imagery. In the physical world, there is no fixed background to a road sign—it can change depending on the viewpoint of the observer. This means that a physical deployment of Kurakin's attack needs to have the background printed out on paper, in addition to the road sign itself. The background is a very conspicuous addition to the environment. In contrast, we introduce an optimization-based approach that spatially constrains the perturbation to regions within the road sign. Furthermore, our spatially-constrained perturbations can be applied to a printed-out road sign, and to an existing physical road sign as "sticker perturbations." Additionally, through our evaluation methodology, we show that our attack is robust to a wider range of varying distances and angles.

Lu *et al.* perform experiments with printed road signs on poster paper [13]. Their approach as well adds subtle perturbations to the background of the road signs. They hypothesize that varying distances and angles affect the ability of adversarial examples and therefore such adversarial examples cannot perform physical attacks. They report on experiments that support this statement. Our work shows that it *is* possible to generate such kind of physical adversarial perturbations, implying that there are immediate security and safety concerns in cases where DNNs are used for control of physical objects such as cars.

[1]This is the most common type of perturbation in existing work.

Fig. 1: Current adversarial deep learning algorithms perturb the background imagery (squiggled area) as well as the main object (the stop sign).

2) Camouflage Perturbations: Sharif *et al.* have shown how spectacles with adversarial perturbations printed on them can be used to fool face recognition systems [1], [20]. This attack is similar to our goals—fabricating a physical object that can be added to an existing object in order to confuse DNNs. The main difference with our work is the environmental differences. Face recognition systems operate in more stable conditions (distance of face from camera, angle, lighting, etc.) than an autonomous vehicle. A face is generally well-lit and positioned at a similar distance and angle with respect to the camera each time an image is taken [20]. In an autonomous vehicle, the magnitude of distance/angle deviations is much larger, making it harder to realize robust physical attacks. A second difference is that face recognition typically occurs in a private environment with limited observers—thus there is no strong requirement for the perturbation to be inconspicuous. However, in the road setting, the road sign is in public view and there maybe an unlimited number of human observers, making it vital that the adversarial perturbation be as inconspicuous as possible.

III. Problem Statement

Typically, a vision subsystem for an autonomous vehicle consists of an object detector, which helps it detect pedestrians, road lights, road signs, other vehicles, etc., and a classifier that characterizes a detected road sign as a Stop or a Yield, for instance. In this work, we focus on the robustness of the classifier itself and show that it is possible to attack the classifier in physical world [9], [13]. We leave attacking object detectors to future work [5], [22].

A. Baseline Road Sign Classifier

Without loss of generality, we focus our work on U.S. road signs. To the best of our knowledge, currently there is no publicly available road-sign classifier for U.S. road signs. Therefore, we used the LISA dataset [14], a U.S. sign dataset comprised of 47 different road signs and trained 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 how common they are (*e.g.*, Stop or Speed Limit 35), we limited the maximum number examples used per sign to 500. Our final dataset includes commonly used signs such as Stop, Speed Limits, Yield, and Turn Indicators. Finally, the original

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

Sign	Number of Examples
addedLane	294
keepRight	331
laneEnds	210
merge	266
pedestrianCrossing	500
school	133
schoolSpeedLimit25	105
signalAhead	500
speedLimit25	349
speedLimit30	140
speedLimit35	500
speedLimit45	141
speedLimitUrdbl	132
stop	500
stopAhead	168
turnRight	92
yield	236
Total	4597

LISA dataset contained image resolutions ranging from 6x6 to 167x168 pixels. We resized all images to 32x32 pixels, a common input size for other well-known image datasets such as CIFAR10 and MNIST. Table I summarizes our final dataset.[2]

We trained a road sign classifier in TensorFlow using our refined dataset. Our model consists of three convolution layers followed by a fully connected layer. We use on the default CNN model provided by the Cleverhans library [16]. We trained the classifier for 100 epochs using AdaDelta, and an initial learning rate of 0.1. Our final classifier accuracy was 91% on the test dataset.

B. Threat Model

Unlike prior work, we seek to physically modify an existing road sign in ways that cause 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 instead of training data (poisoning attack). Performing poisoning attacks in an autonomous vehicle would mean that the attacker already has a superior level of control over the vehicle running the model, thus negating the need for any kind of adversarial perturbations to the model—the attacker can simply feed malicious data and mislead the model to cause harm. Therefore, we assume that the attacker *cannot* digitally perturb inputs.

In evasion attacks, an attacker can only change existing physical road signs. Here we assume that an attacker gains access to the classifier after it has been trained ("white-box" access) [17]. This assumption is practical since even without access to the actual model itself, by probing the system, attackers can usually figure out a similar surrogate model based on feedback [3]. Besides, based on Kerckhoffs's principle, we need to evaluate the most powerful attacker in order to inform future defenses that guarantee system robustness [19].

Based on our threat model, the attack pipeline proceeds as follows:

1) Obtain at least one clean image of the target road sign without any adversarial perturbations.
2) The images(s) of the road sign are pre-processed to be input into the classifier. First, the sign is extracted from the image(s) either through the use of a sign detection algorithm or cropping software. Then, the dimensions of the extracted sign are resized to match the input dimensions of the classifier.
3) The classifier and the extracted sign image(s) are processed by the attack algorithm, which outputs an adversarial image. Here the adversarial perturbation consists of only the road sign region instead of the background as in Figure 1 to meet practical requirements.
4) Given a digital adversarial perturbation, a mapping function outputs the corresponding physical locations on the road sign indicating where to apply the perturbation.
5) The digital adversarial perturbation is fabricated and applied to the road sign based on the output of the mapping function.

IV. Evaluation Methodology for Physical Adversarial Perturbations

We discuss possible factors that affect perturbations in the physical world. Then, we introduce our evaluation methodology that captures a range of physical-world factors.

A. Evaluation Components

Autonomous vehicles experience a range of varying conditions in the physical world—changing distances, angles, lighting, and debris. 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 main physical components that can affect the ability of adversarial perturbations.

- **Environmental Conditions:** The distance and angle of a camera in an autonomous vehicle with respect to a road sign continuously varies. 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 on the image, as well as other environmental factors including changes in lighting conditions, and 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

[2]Urdbl stands for unreadable. This means the sign had an additional sign attached, that the annotator could not read due to low image quality

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 [20].

- **Resolution Changes:** The camera in a vehicle will not necessarily produce images whose dimensions are the same as the size of input of the DNN—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 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.

Given these real world challenges, an attacker should be able to account for the above changes in physical conditions while computing perturbations, in order to successfully physically attack existing road sign classifiers. In our evaluation methodology, we focus on three major components that impact how a road sign is classified by, say, a self-driving car.

Distance First, 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 leading to it perceiving the sign from a sharp angle or it could be on the lane closest to the shoulder leading to it perceiving 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 under a range of valid resolutions. Since classifiers typically work on a fixed input size, no image taken by a self-driving car camera will be fed to the model without this modification. This might cause the image to stretch or bend and the colors to mix. Besides, different weather or physical conditions could also cause such resolution changes. So for evaluation purposes, we test the resolution condition by resizing the images in a range of valid resolutions to make sure it can attack robustly.

B. Evaluation Methodology

We develop our evaluation methodology to account for the possible physical variations. We obtain images with a (smartphone) camera from various distances and angles. Our methodology proceeds as follows:

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.[3] See Figure 2 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 success rate of the physical perturbation using the following formula:

$$\frac{\sum_{c \in C} \mathbb{1}_{\{f_\theta(\mathcal{A}(c)^{d,g}) = y' \;\&\&\; f_\theta(c^{d,g}) = y\}}}{\sum_{c \in C} \mathbb{1}_{\{f_\theta(c^{d,g}) = y\}}} \tag{3}$$

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

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.

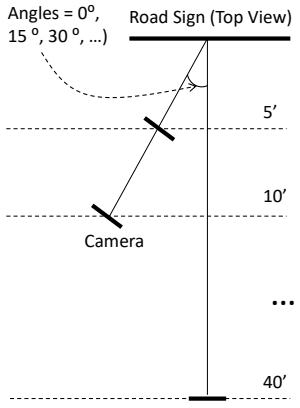

Fig. 2: For all our physical realizability experiments, we test perturbation by varying the angles and distances of the camera from the road sign. Distances vary from 5' to 40', and angles vary from 0°to 60°at 5', 0°to 30°at 10', 0°to 15°at 15' and 20'. From 25' onwards, we do not vary angle. All angle variations are to the left of the sign (U.S. driving).

V. Robust Physical Perturbations

We present our algorithm to generate physically-realizable perturbations that are robust to varying distances and angles. We

[3]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.

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

take an optimization-based approach to generate perturbations. However, unlike prior work, we formulate the objective function to ensure that perturbations are not added to the background of the road sign. We discuss how we physically realize such "bounded" attacks by fashioning "adversarial" signs out of poster paper as well as adding adversarial sticker perturbations to existing physical signs. Without loss of generality, our discussions focus on a Stop sign due to its critical role in safety. Our approach is general, and can be applied to other types of road signs. Later in this section, we discuss an attack on a Turn Right sign.

RP$_2$ To generate adversarial perturbations in a specific physical region, we search for a small perturbation δ to be added to the original input $x' = x + \delta$ so that x' is misclassified by the model. This can be achieved by taking an optimization-based approach [3], [21]. Specifically, to generate untargeted adversarial examples, we approximate the solution to the following objective function:

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

To generate targeted adversarial examples, we modify the objective function as below.

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + J(f_\theta(x+\delta), y') \tag{5}$$

Here, $||\delta||_p$ denotes the ℓ_p norm of δ, which is a metric to quantify the magnitude of the perturbation δ.

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

$$||\delta||_p = \Big(\sum_{i,j} |\delta_{(i,j)}|^p\Big)^{1/p} \tag{6}$$

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 perturbations work robustly in the physical world, we collect a set of clean images of road signs by varying environmental conditions (distances, angles, lighting). Therefore, our final loss function is calculated considering all the these data as shown below. For untargeted attacks, the objective function would be:

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

Similarly, for targeted attacks, our objective function is:

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

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 the modifications to road signs as vandalism. By controlling the size and shape of the mask, perturbations 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 that indicate regions where no perturbation is added, and ones that indicate regions where a perturbation is added during optimization.

Optimizing Spatially-Constrained Perturbations. Having computed a mask, we modify the above formulation 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:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x\cdot\delta||_p + NPS(M_x\cdot\delta) + \sum_{i=1}^{k} J(f_\theta(x_i + M_x\cdot\delta), y^*) \tag{9}$$

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.* [20], 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:

$$NPS(\delta) = \sum_{\hat{p}\in\delta}\prod_{p'\in P} |\hat{p} - p'| \tag{10}$$

VI. Experimental Results

Using three different perturbations (subtle, camouflage graffiti, and camouflage art) generated by RP$_2$, we experiment with two different types of physically realizable perturbations for both targeted and untargeted attacks. 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 misclassification target, in 100% of test cases. Poster-printed camouflage graffiti caused the Right Turn sign to to be misclassified as a Stop sign, the misclassification target, 66.67% of the time. In sticker attacks, we print the perturbations on paper, cut them out, and stick them 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 misclassification rate.

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.

(a) (b) (c) (d) (e)

Fig. 3: Sample experimental images for the subtle poster experiments detailed in table II at a selection of distances and angles. From right to left: (a) 5' 0°(b) 5' 15°(c) 10' 0°(d) 10' 30°(e) 40' 0°

(a) (b) (c) (d) (e)

Fig. 4: Sample experimental images for the Right Turn sign experiments detailed in table II at a selection of distances and angles. From right to left: (a) 5' 0°(b) 5' 15°(c) 10' 0°(d) 10' 30°(e) 40' 0°

(a) (b) (c) (d) (e)

Fig. 5: Sample experimental images for the camouflage graffiti sticker experiments detailed in table II at a selection of distances and angles. From right to left: (a) 5' 0°(b) 5' 15°(c) 10' 0°(d) 10' 30°(e) 40' 0°

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 (9) 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 IV 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.

(a) (b) (c) (d) (e)

Fig. 6: Sample experimental images for the camouflage art sticker experiments detailed in table II at a selection of distances and angles. From right to left: (a) 5' 0°(b) 5' 15°(c) 10' 0°(d) 10' 30°(e) 40' 0°

For the Stop sign, we choose a mask that exactly covers the area of the original sign in order to avoid background distraction. This results in perturbation that is similar to existing work [9] and we hypothesize that it is imperceptible to the casual observer (see Figure 3 for an example).[5] In contrast to some findings in prior work, this attack is very effective in the physical world. The Stop sign is misclassified into our 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 (9) to guarantee small magnitude of perturbations. See Figure 4 for the result of applying the perturbations and printing the image. Table III shows the results of our experiments.

Our attack reports a 100% success rate for misclassification with 66.67% of the images classified as a Stop sign and 33.7% of the images classified as an Added Lane sign. It is interesting to note that in only 1 of the test cases was the Turn Right in the top two classes. In most other cases, a different warning sign was present. We hypothesize that given the similar appearance of warning signs, small perturbations are sufficient to confuse the classifier. In future work, we plan to explore this hypothesis with targeted classification attacks on other warning signs.

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.

Figures 5 and 6 show the result of the above steps for two types of perturbations. We achieve a 66.67% misclassification rate into our target class for the graffiti sticker attack and a 100% targeted misclassification rate for the abstract art sticker 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 (Figure 5). Table II 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 misclassification 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%.

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 noticed 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 compensated for this issue by increasing the width of the perturbations physically. In this final test, we achieve a 100% misclassification rate into our target class, with an average confidence for the target of 62.4% and standard deviation of 14.7% (See Figure 6 for an example image).

As future work, we envision adding and evaluating a perturbation-size compensation step to our attack generation pipeline.

VII. Conclusion

In this paper, we introduced Robust Physical Perturbations (RP_2), an algorithm that generates robust, physically realizable

[5]We envision that future work conduct a set of user studies to determine whether observer suspicion is raised when they see such road signs.

TABLE II: Summary of Targeted Physical Perturbation Experiments with a Poster-Printed Stop sign, and a real Stop sign. In all our attacks, we attempted to mis-classify the Stop sign into a Speed Limit 45 sign, and we report on the top-2 classes along with confidence values. The graffiti attack adds perturbations in the form of the characters "LOVE HATE." The abstract art attack adds two black and two white rectangles to the sign. The subtle attack spreads the perturbation across the entire area of the sign. Each attack column lists the top two predictions made by the classifier along with the confidence of the prediction. 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 III: Poster printed perturbation (faded arrow) on a Turn Right sign at varying distances and angles. Our attack is successful in 66.7% percent of the test conditions.

Distance & Angle	Top Class (Confid.)	Second Class (Confiden.)
5' 0°	Stop (0.42)	Added Lane (0.16)
5' 15°	Stop (0.44)	Added Lane (0.13)
5' 30°	Stop (0.28)	Added Lane (0.19)
5' 45°	Stop (0.15)	Stop Ahead (0.12)
5' 60°	Added Lane (0.15)	Turn Right (0.12)
10' 0°	Stop (0.40)	Added Lane (0.22)
10' 15°	Stop (0.51)	Added Lane (0.07)
10' 30°	Stop (0.21)	Added Lane (0.19)
15' 0°	Stop (0.33)	Added Lane (0.30)
15' 15°	Stop (0.40)	Added Lane (0.12)
20' 0°	Added Lane (0.63)	Merge (0.09)
20' 15°	Stop (0.32)	Stop Ahead (0.15)
25' 0°	Added Lane (0.50)	Merge (0.15)
30' 0°	Added lane (0.49)	Stop (0.20)
40' 0°	Added Lane (0.40)	Stop (0.13)

adversarial perturbations. Previous algorithms assume that the inputs of DNNs can be modified digitally to achieve misclassification, but such an assumption is infeasible, as an attacker with control over DNN inputs can simply replace it with an input of his choice. Therefore, adversarial attack algorithms must apply perturbations physically, and in doing so, need to account for new challenges such as a changing viewpoint due to distances, camera angles, different lighting conditions, and occlusion of the sign. Furthermore, fabrication of a perturbation introduces a new source of error due to a limited color gamut in printers.

We use RP_2 to create two types of perturbations: *subtle perturbations*, which are small, undetectable changes to the entire sign, and *camouflage perturbations*, which are visible perturbations in the shape of graffiti or art. When the Stop sign was overlayed with a print out, subtle perturbations fooled the classifier 100% of the time under different physical conditions. When only the perturbations were added to the sign, the classifier was fooled by camouflage graffiti and art perturbations 66.7% and 100% of the time respectively under different physical conditions. Finally, when an untargeted poster-printed camouflage perturbation was overlayed on a Right Turn sign, the classifier was fooled 100% of the time. In future work, we plan to test our algorithm further by varying some of the other conditions we did not consider in this paper, such as sign occlusion.

ACKNOWLEDGEMENTS

We thank Kimberly Ruth for providing thoughtful comments on this paper. This work was supported in part by the UW Tech Policy Lab; the Short-Dooley Professorship; NSF grant CNS-1565252; BDD (Berkeley Deep Drive); the CLTC (Center for Long-Term Cybersecurity); FORCES (Foundations Of Resilient CybEr-Physical Systems), which receives support from the National Science Foundation (NSF award numbers CNS-1238959, CNS-1238962, CNS-1239054, CNS-1239166);

and the National Science Foundation under Grant No. TWC-1409915.

REFERENCES

[1] "Best practices for developing with kairos," 2017. [Online]. Available: https://www.kairos.com/docs/api/best-practices

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

[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] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

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

[7] 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.

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

[9] 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

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

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

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

[13] 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.

[14] 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

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

[16] 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.

[17] 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.

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

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

[20] 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*, Oct. 2016. [Online]. Available: https://www.ece.cmu.edu/~lbauer/papers/2016/ccs2016-face-recognition.pdf

[21] 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.

[22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.