

Robust Physical-World Attacks on Deep Learning Visual Classification

Kevin Eykholt*¹, Ivan Evtimov*², Earlence Fernandes², Bo Li³,

Amir Rahmati⁴, Chaowei Xiao¹, Atul Prakash¹, Tadayoshi Kohno², and Dawn Song³

Presented by Danning Ma

Background

- **deep neural networks (DNNs) are vulnerable to adversarial examples**
Because small-magnitude perturbations added to the input
- **Authors want to understand the adversarial examples to create resilient learning algorithms**
- **a general attack algorithm, Robust Physical Perturbations (RP2) by using the real-world case of road sign classification and a perturbation in the form of only black and white stickers**

- Robust Physical Perturbations (RP2)

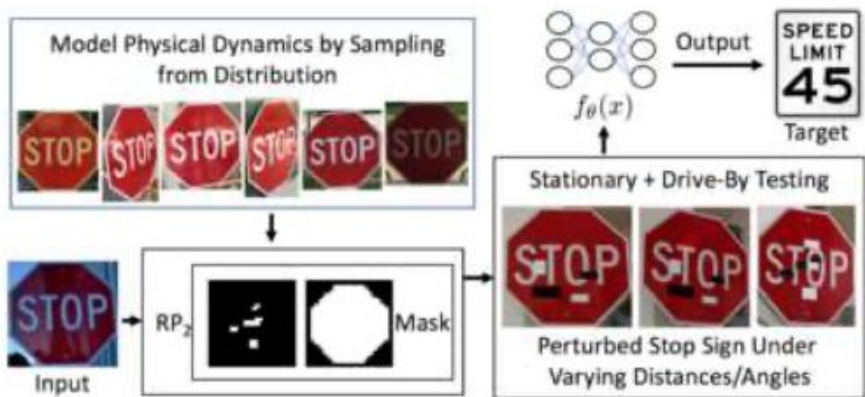
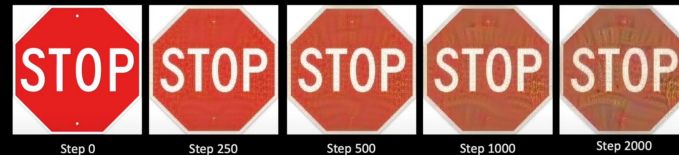


Figure 2: RP₂ pipeline overview. The input is the target Stop sign. RP₂ samples from a distribution that models physical dynamics (in this case, varying distances and angles), and uses a mask to project computed perturbations to a shape that resembles graffiti. The adversary prints out the resulting perturbations and sticks them to the target Stop sign.

Generation Process



Procedure to generate adversarial examples

a classifier $f_{\theta}(\cdot)$ with parameters θ and an input x with ground truth label y for x

an adversarial example x' is generated so that it is close to x in terms of certain distance, such as L_p norm distance.

x' will also cause the classifier to make an incorrect prediction

as $f_{\theta}(x') \neq y$ (untargeted attacks),

or $f_{\theta}(x') = y^*$ (targeted attacks) for a specific $y^* \neq y$.



a two-stage experiment design

- (1) A lab test where the viewing camera is kept at various distance/angle configurations.
 - (2) A field test where we drive a car towards an intersection in uncontrolled conditions to simulate an autonomous vehicle.
-

Challenges:

- **Environmental Conditions.**
- **Spatial Constraints.**
- **Physical Limits on Imperceptibility.**
- **Fabrication Error.**



Robust Physical Perturbation

single-image optimization problem searches for perturbation δ to be added to the input x ,

perturbed instance $x' = x + \delta$ is misclassified by the target classifier $f_{\theta}(\cdot)$:

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

where H is a chosen distance function, and y^* is the target class.

Lagrangian-relaxed form

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

Here $J(\cdot, \cdot)$ is the loss function, which measures the difference between the model's prediction and the target label y^* .

λ is a hyper-parameter that controls the regularization of the distortion.

We specify the distance function H as $\|\delta\|_p$, denoting the l_p norm of δ .

physical and digital transformations X^V .

We sample different instances x_i drawn from X^V . A physical perturbation can only be added to a specific object o within x_i .


To account for fabrication error: **Non-Printability Score (NPS)**: models printer color reproduction errors.

$T_i(\cdot)$ to denote the alignment function that maps transformations on the object to transformations on the perturbation

final robust spatially- constrained perturbation is thus optimized as:

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

two classifiers

- 1. LISA-CNN uses LISA, a U.S. traffic sign dataset containing 47 different road signs**
 - a.** consists of three convolutional layers and an FC layer.
 - b.** It has an accuracy of 91% on the test set
 - 2. GTSRB-CNN, that is trained on the German Traffic Sign Recognition Benchmark**
 - a.** GTSRB-CNN achieves 95.7% accuracy on the test set.
- 

Experiment design- Stationary(Lab) test

- a. a set of clean images C and a set of adversarially perturbed images ($\{A(c)\}, \forall c \in C$) at varying distances $d \in D$, and varying angles $g \in G$.
 - i. use $c^{(d,g)}$ here to denote the image taken from distance d and angle g .
 - ii. camera's vertical elevation should be kept approximately constant.
- b. Compute the attack success rate of the physical perturbation

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

Drive-By (Field) Tests

place a camera on a moving platform, and obtain data at realistic driving speeds

1. Begin recording video at **approximately 250 ft away from the sign.**
2. Perform video recording as above for a “clean” sign and for a sign with perturbations applied

Results for LISA-CNN

observe high attack success rates with high confidence

Object-Constrained Poster-Printing Attacks:

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

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

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

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

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

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

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

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

. Results for Inception-v3



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

Discussion

An autonomous vehicle will likely not run classification on every frame due to performance constraints, but rather, **would classify every j th frame, and then perform simple majority voting.**

Hence, an open question is to determine **whether the choice of frame (j) affects attack accuracy.**

Different companies developed different classifier and do you think the results of this paper apply to every classifiers?



Paper Critique

- Background information: DNNs,
- formulas to calculate the the minimum distance





Thank you