

Supplemental material of “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

¹University of Michigan, Ann Arbor

²University of Washington

³ University of California, Berkeley

⁴Samsung Research America and Stony Brook University

1. NPS Formal Definition

Given a set of printable colors (RGB triples) P and a set $R(\delta)$ of (unique) RGB triples used in the perturbation that need to be printed out in physical world, the non-printability score is given by:

$$NPS(\delta) = \sum_{\hat{p} \in R(\delta)} \prod_{p' \in P} |\hat{p} - p'| \quad (1)$$

2. Inception-v3 experiments

We provide results demonstrating the generality of RP₂. We attacked the Inception-v3 classifier with a sticker causing it to misclassify a microwave as a phone and misclassify a coffee mug as a cash machine. We chose a sticker attack since poster printing an entirely new surface for the objects may raise suspicions. Note that for both attacks, we have reduced the range of distances used due to the smaller size of the cup and microwave compared to a road sign (*e.g.* Coffee Mug height- 11.2cm, Microwave height- 24cm, Right Turn sign height- 45cm, Stop Sign- 76cm). Table 1 summarizes our attack results on the microwave and Table 2 summarizes our attack results on the coffee mug. For the microwave, the targeted attack success rate is 90%. For the coffee mug, the targeted attack success rate is 71.4% and the untargeted success rate is 100%.

Table 1: Sticker perturbation attack on the Inception-v3 classifier. The original classification is microwave and the attacker’s target is phone. See example images in Table 3. Our targeted-attack success rate is 90%

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
2' 0°	Phone (0.78)	Microwave (0.03)
2' 15°	Phone (0.60)	Microwave (0.11)
5' 0°	Phone (0.71)	Microwave (0.07)
5' 15°	Phone (0.53)	Microwave (0.25)
7' 0°	Phone (0.47)	Microwave (0.26)
7' 15°	Phone (0.59)	Microwave (0.18)
10' 0°	Phone (0.70)	Microwave (0.09)
10' 15°	Phone (0.43)	Microwave (0.28)
15' 0°	Microwave (0.36)	Phone (0.20)
20' 0°	Phone (0.31)	Microwave (0.10)

Table 2: Sticker perturbation attack on the Inception-v3 classifier. The original classification is coffee mug and the attacker’s target is cash machine. See example images in Table 4. Our targeted-attack success rate is 71.4%.

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
8" 0°	Cash Machine (0.53)	Pitcher (0.33)
8" 15°	Cash Machine (0.94)	Vase (0.04)
12" 0°	Cash Machine (0.66)	Pitcher (0.25)
12" 15°	Cash Machine (0.99)	Vase (<0.01)
16" 0°	Cash Machine (0.62)	Pitcher (0.28)
16" 15°	Cash Machine (0.94)	Vase (0.01)
20" 0°	Cash Machine (0.84)	Pitcher (0.09)
20" 15°	Cash Machine (0.42)	Pitcher (0.38)
24" 0°	Cash Machine (0.70)	Pitcher (0.20)
24" 15°	Pitcher (0.38)	Water Jug (0.18)
28" 0°	Pitcher (0.59)	Cash Machine (0.09)
28" 15°	Cash Machine (0.23)	Pitcher (0.20)
32" 0°	Pitcher (0.50)	Cash Machine (0.15)
32" 15°	Pitcher (0.27)	Mug (0.14)

*These authors contributed equally.

Table 3: Uncropped images of the microwave with an adversarial sticker designed for Inception-v3.




Distance/Angle	Image	Distance/Angle	Image
2' 0°		2' 15°	
5' 0°		5' 15°	
7' 0°		7' 15°	
10' 0°		10' 15°	
15' 0°		20' 0°	

Table 4: Cropped Images of the coffee mug with an adversarial sticker designed for Inception-v3.

Distance/Angle	Image	Distance/Angle	Image
8"0°		8"15°	
12"0°		12"15°	
16"0°		16"15°	
20"0°		20"15°	
24"0°		24"15°	
28"0°		28"15°	
32"0°		32"15°	


3. Additional LISA-CNN Tables

In this section, we have included additional tables that could not be included in the main text due to space limitations. Table 5 shows the detailed results of the poster-printed perturbation attack on the Right Turn sign. Table 6 shows some example video frames from the drive-by evaluation of two attacks on the LISA-CNN classifier.

Table 5: 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 1 of the main text. 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)

Table 6: Drive-by testing summary for LISA-CNN. In our baseline test, all frames were correctly classified as a Stop sign. We have added the yellow boxes as a visual guide manually.

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