Oregon State University

DARTS: Deceiving Autonomous Cars with Toxic Signs

Submitted to ACM CCS 2018 , *18 Feb 2018*

Chawin Sitawarin, Arjun Nitin Bhagoji

Shahab Nikkhoo

COLLEGE OF ENGINEERING

School of Electrical Engineering and Computer Science

Evasion attack

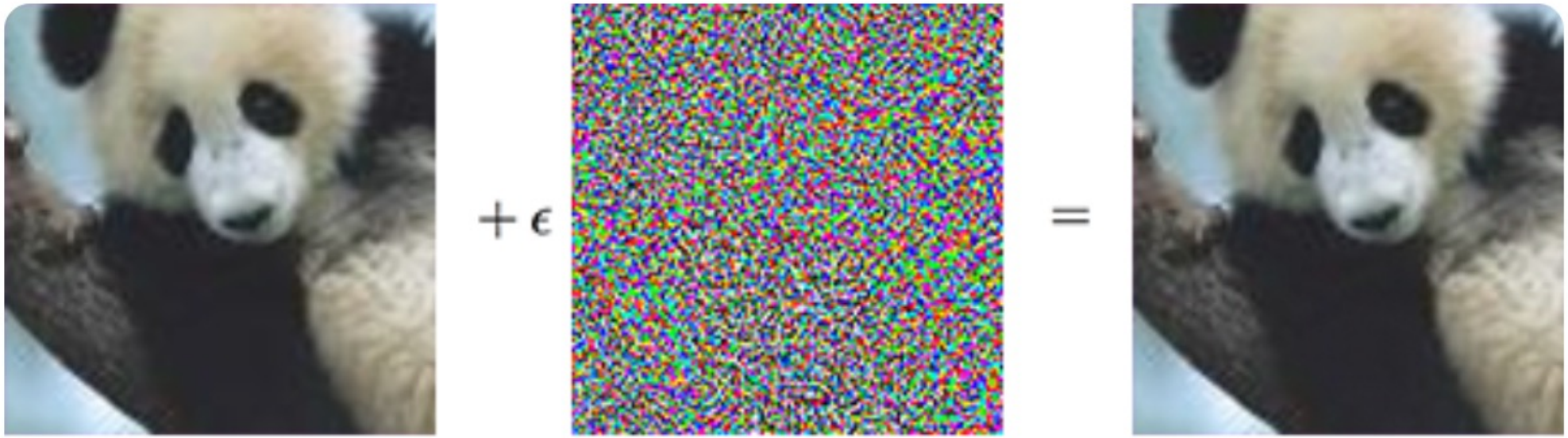
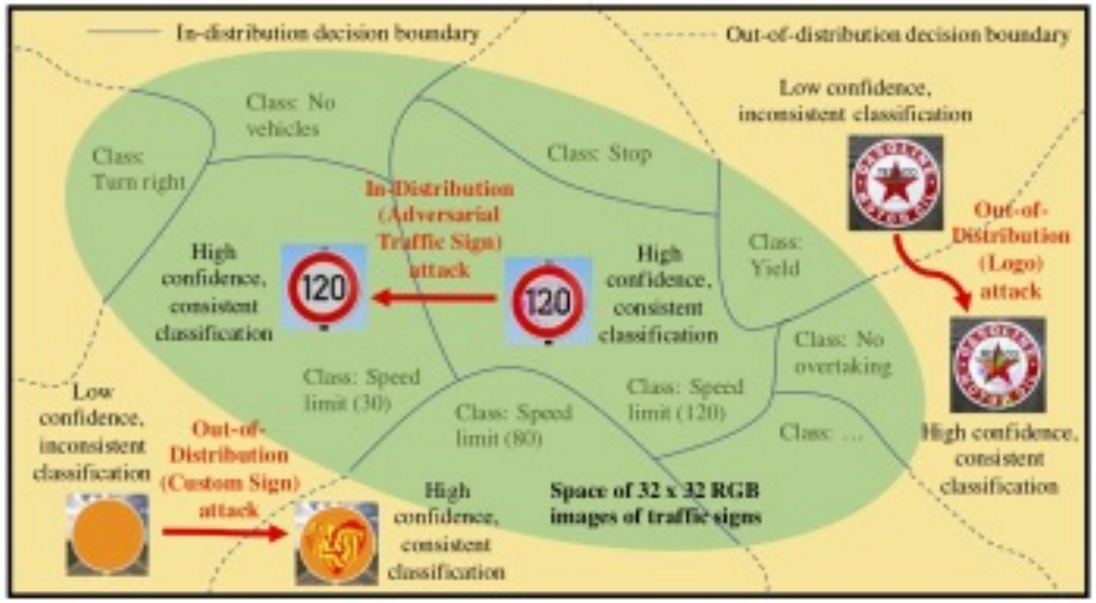
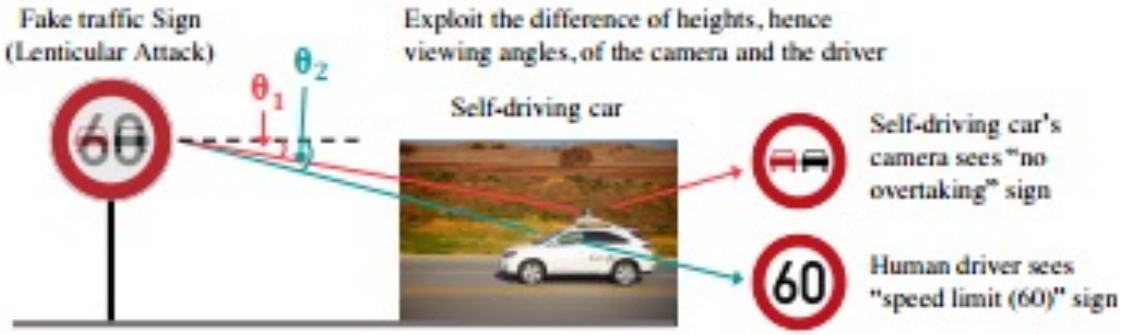


Image reference:
https://www.plugandplaytechcenter.com/resources/machine-learning-security-3-risks-be-aware/

Oregon State University
College of Engineering

Attack methods

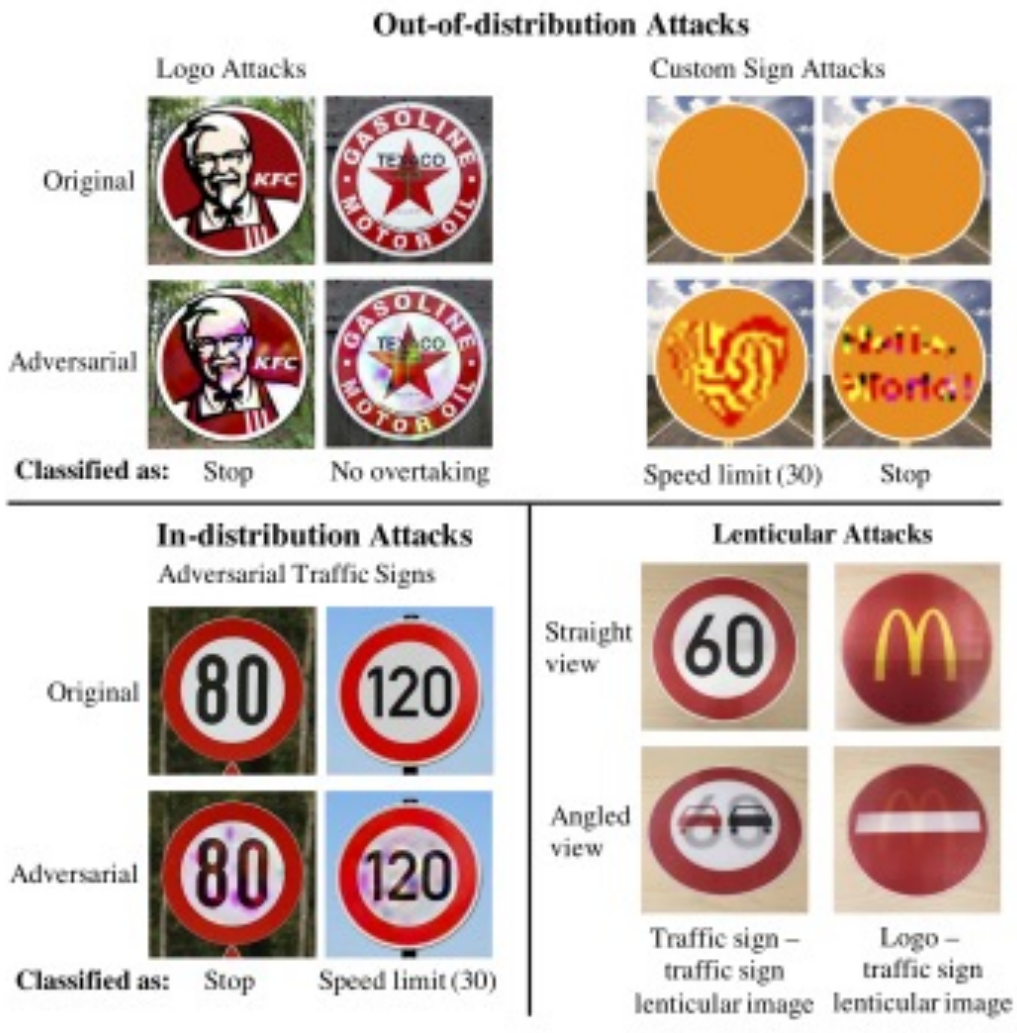




Oregon State University
College of Engineering

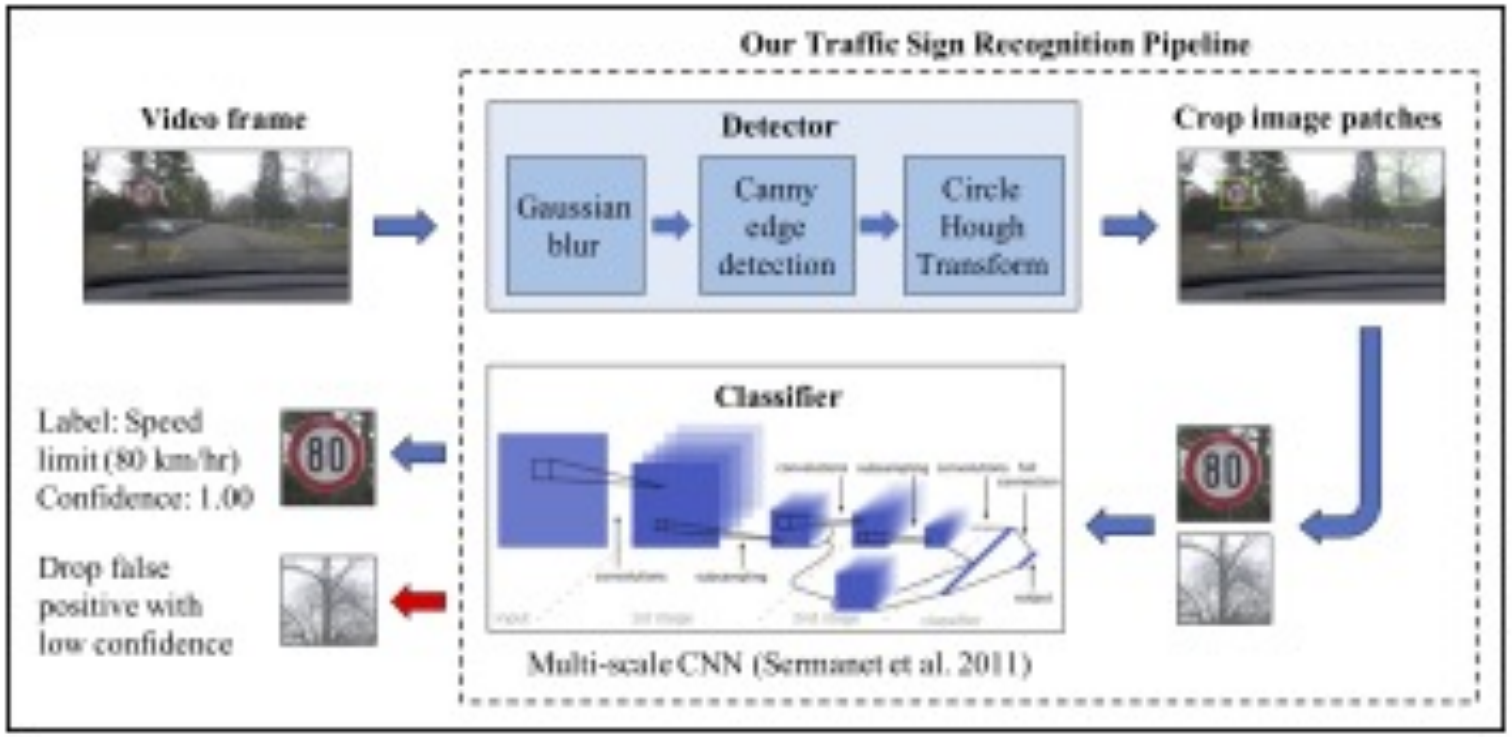
attack methods

- In distribution
- Out of distribution
- Lenticular

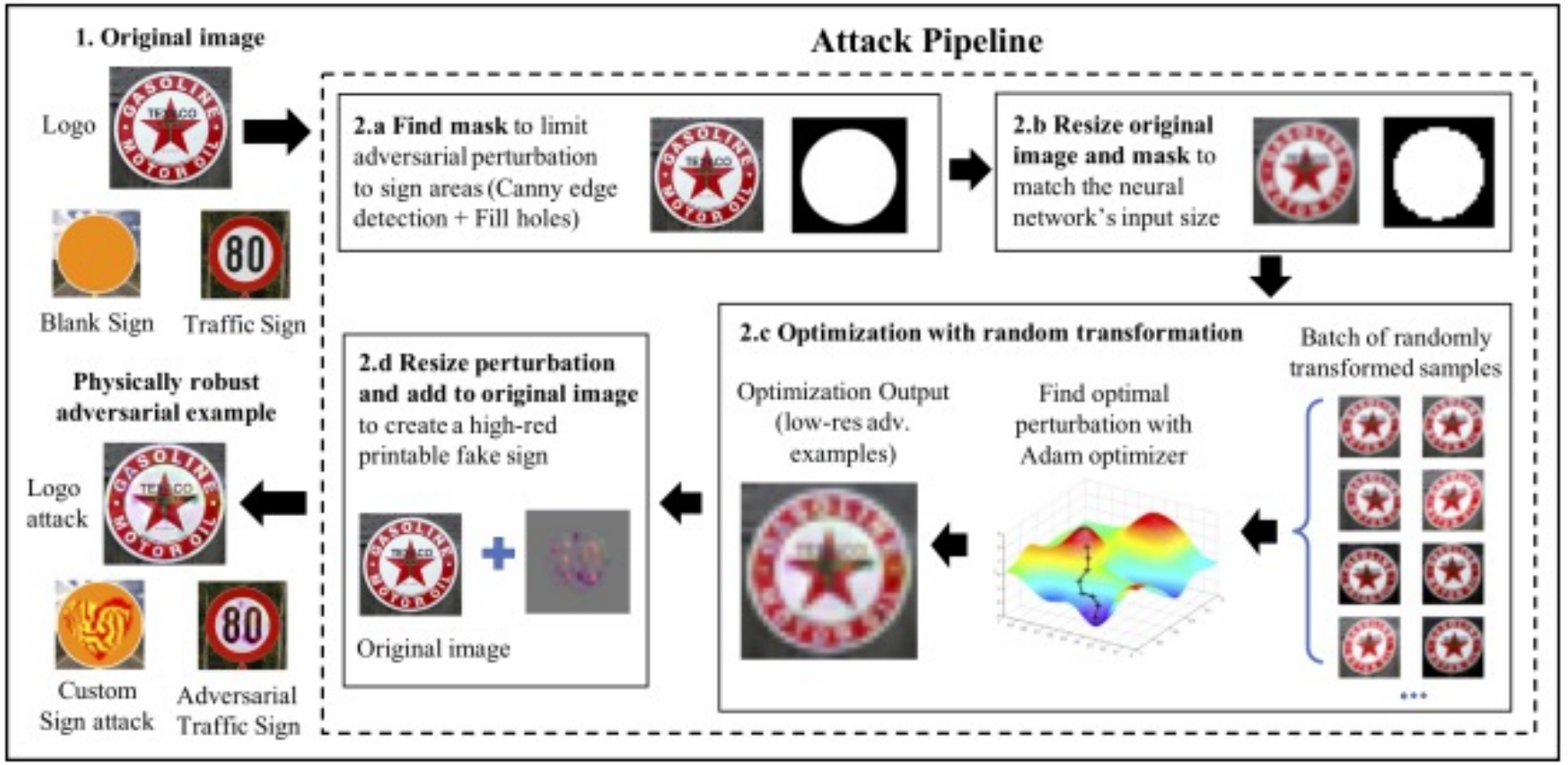


Classification pipeline

- Detection
- classification

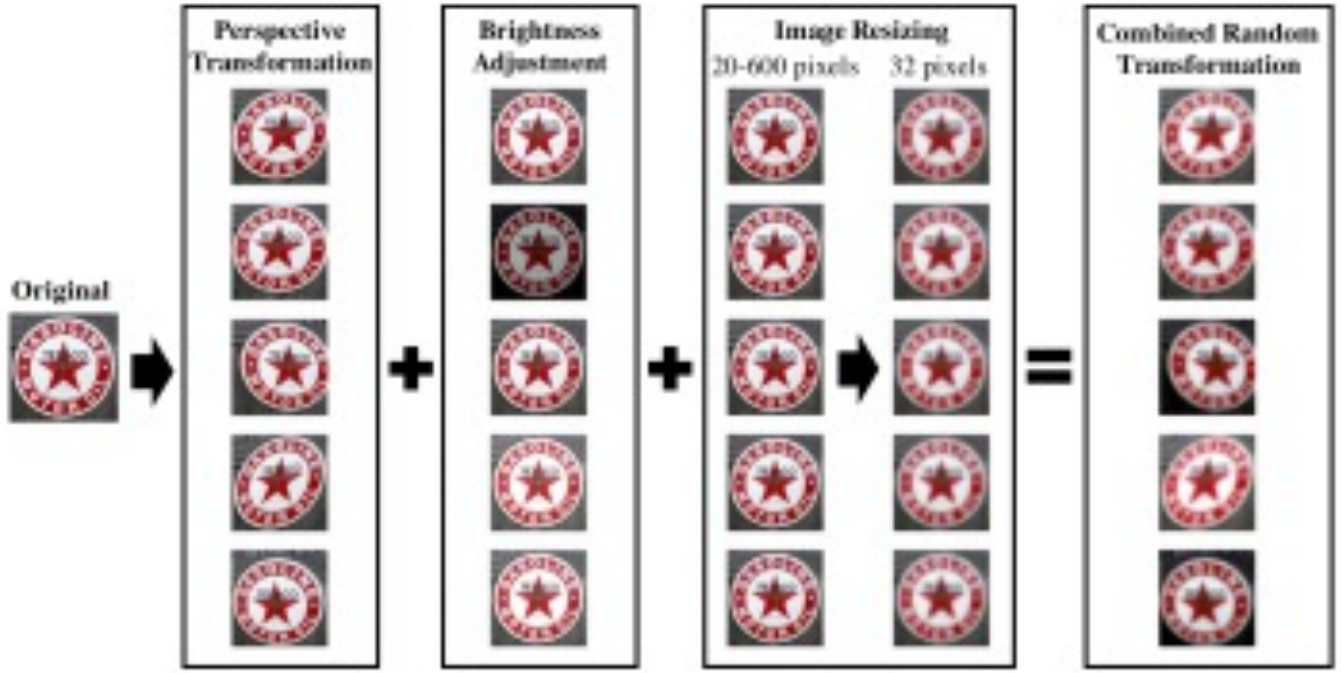


Attack pipeline



How to make it real?

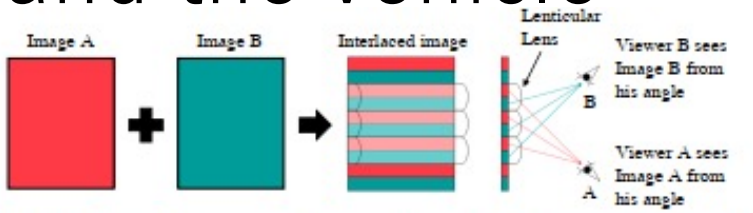
- Generate physical robust adversarial examples



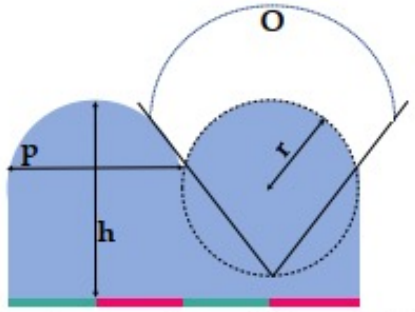
Oregon State University
College of Engineering

Lenticular attack

- Lenticular Printing attack is motivated by the key insight that the human driver and the vehicle mounted camera observe



(a) Illustration of the process of generating a lenticular image and its angle-dependent appearance.



(b) Several parameters (ρ, r, and h) determine the full angle of observation (O)

Oregon State University
College of Engineering

results

Attacks	Virtual attack success (VAS)	Simulated physical attack success (SPAS)	Avg. norm (L_1)	Avg. confidence
In-Distribution (auxiliary traffic data)	54.34 %	36.65 %	37.71	0.9721
Out-of-Distribution (Logo)	85.71%	65.07%	34.89	0.9753
Out-of-Distribution (Custom Sign)	29.44%	18.72%	N.A.	0.9508

Attacks	White-box (Avg. confidence)	Black-box (Avg. confidence)
In-Distribution (auxiliary traffic data)	92.82% (0.9632)	96.68% (0.9256)
Out-of-Distribution (Logo)	52.50% (0.9524)	32.73% (0.9172)
Out-of-Distribution (Custom Sign)	96.51% (0.9476)	97.71% (0.9161)

Oregon State University
College of Engineering

What about defense?

Adversarial training cannot defend against Lenticular Printing and Out-of-Distribution

Discussion points

- They are using their own classifier for detection and recognizing the signs.
- Where are the logos ? Are they looks like the signs?
- This method is useful for AR?