Rogue Signs: Deceiving Traffic Sign Recognition with Malicious Ads and Logos

Chawin Sitawarin [1], Arjun Nitin Bhagoji [1], Arsalan Mosenia [1], Prateek Mittal [1], Mung Chiang [2]
[1] Princeton University [2] Purdue University

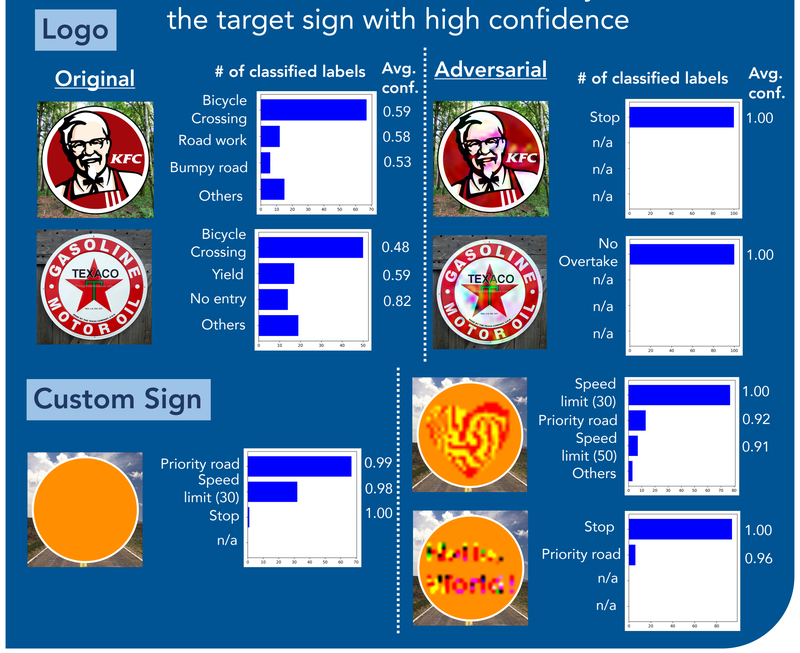
PRINCETON Electrical Engineering

Motivation

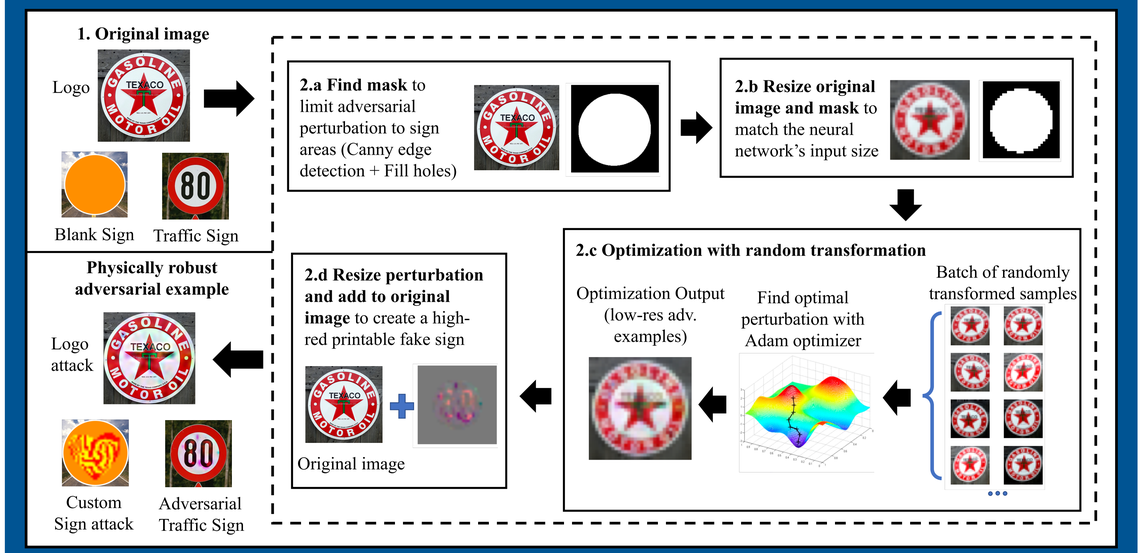
- Traffic sign recognition is an integral part of autonomous cars. Any misclassification of traffic signs can lead to accidents and/or large traffic interruption.
- Physically robust attacks on image recognition systems proposed in [1, 2] demonstrate several successful adversarial examples against a traffic sign classifier.
- We propose **Out-of-Distribution** (OOD) and **lenticular printing** as two new attack spaces and thoroughly test them in the real-world setting.

OOD Attacks

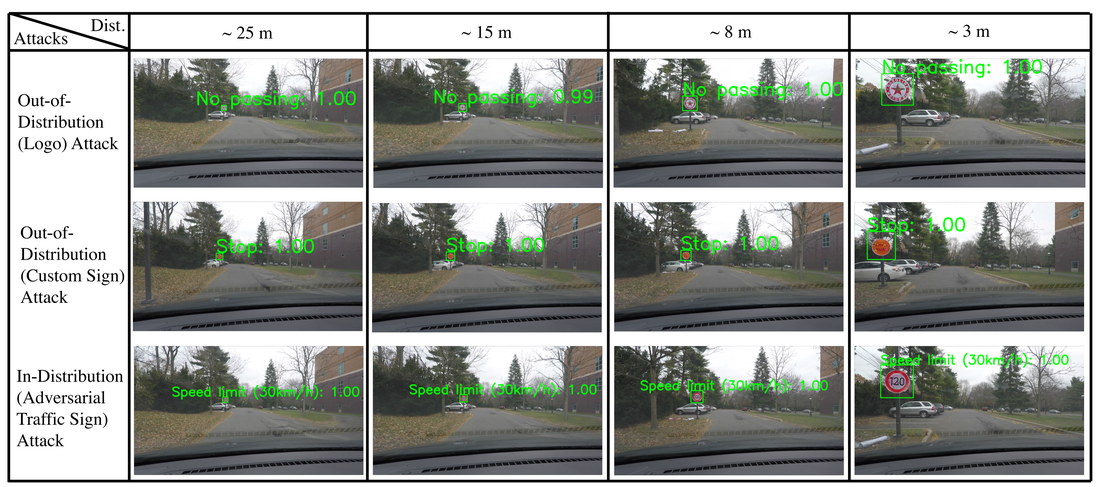
Our OOD attacks are consistently classified as the target sign with high confidence



OOD Attack Pipeline



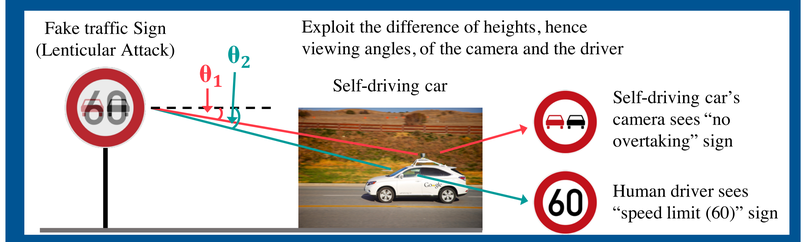
Drive-by Test



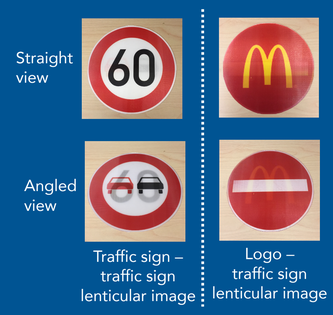
Success Rate (Average Confidence)

White-Box	Black-Box
52.50% (0.9524)	32.73% (0.9172)
96.51% (0.9476)	97.71% (0.9161)
92.82% (0.9632)	96.68% (0.9256)

Lenticular Printing



- Lenticular printing interlaces 2 images. Viewing angles determine which image shows up.
- Assume that a driver and a camera see signs at different angles.



References

[1] Evtimov et al., Robust physical-world attacks on machine learning models. CVPR 2018.
[2] Athalye et al. Synthesizing robust adversarial examples. ICLR 2018.

Conclusion

We propose two novel attacks, OOD and lenticular printing, and extensively evaluate them in both virtual and real-world settings. We carry out experiments in white-box and black-box scenarios as well as against adversarially trained model and find that our OOD attacks succeed with high probability and cause misclassification with high confidence.