STANDARD DETECTORS AREN'T (CURRENTLY) FOOLED BY PHYSICAL ADVERSARIAL STOP SIGNS

By: Jiajun Lu, Hussein Sibai, Evan Fabry, David Forsyth

Preprint: 2017 arXiv

Presented By: Brandon Ellis

SO, WHY ARE YOU HERE?

Standard Detectors Aren't (Currently) Fooled by Physical Adversarial Stop Signs

or

"The research paper equivalent of a diss track"

Background Information

The Paper

- Overview
- Attacking Classifiers
- Attacking Detectors
- Bad Paper is Bad
- Experiment
- Results

Discussion

STANDARD DETECTORS AREN’T (CURRENTLY) FOOLED BY PHYSICAL ADVERSARIAL STOP SIGNS

BACKGROUND INFORMATION

REFRESHER

Classifier

- Accepts Image
- Produces Label

Detector

- Identifies Boxes “Worth Labeling”
- Generates Labels
- “How boxes span objects in a detector is complex”

What is an adversarial example?

- “An example that has been adjusted to produce the wrong label when presented to a system at test time.”
- Done with small/easy adjustments.
- Evidence that it’s hard to tell if example is adversarial

STANDARD DETECTORS AREN'T (CURRENTLY) FOOLED BY PHYSICAL ADVERSARIAL STOP SIGNS

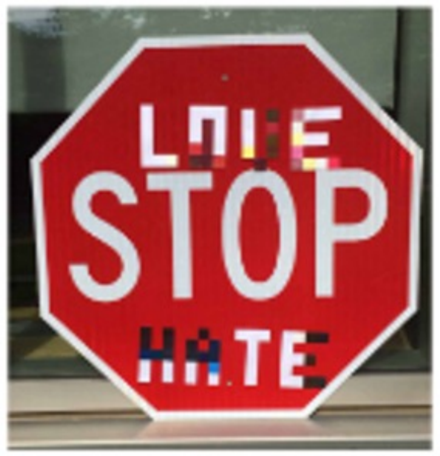
THE PAPER

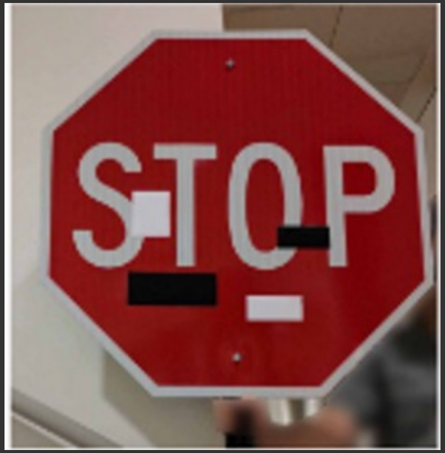
THE PAPER

- Attacking a Classifier
- Attacking a Detector
- “Think Before You Write”
- Prove Attack Fails Against Detector

ATTACKING A CLASSIFIER

- Road Sign Attack
- All Attacks are on Classifier
- But… is it useful?





ATTACKING A DETECTOR

- Detector Implementation

- Attacking a Detector is Difficult

"YOUR PAPER IS BAD, AND YOU SHOULD FEEL BAD" - JOHN A. ZOIDBERG, MD

"Robust physical-world attacks on machine learning models."

I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song.

arXiv preprint arXiv:1707.08945, 2017

- Demonstrated Misclassification of Stop Signs
- Attack Types
- Methodology
- Conclusion
- "Poor Proxy of a Detection System"

THE EXPERIMENT

- Standard Detectors
- Data
- Test Types

DID IT WORK?

Images

Poster Attack:

- YOLO detects about as well as true signs
- Faster RCNN detects about as well as true signs

Sticker Attack:

- YOLO detects about as well as true signs
- Faster RCNN detects about as well as true signs

- Faster RCNN detects signs more accurately than YOLO
- As sign shrinks, YOLO fails earlier than Faster RCNN

Video

Poster Attack:

- YOLO detects stop sign well
- Faster RCNN detects stop sign very well

Sticker Attack:

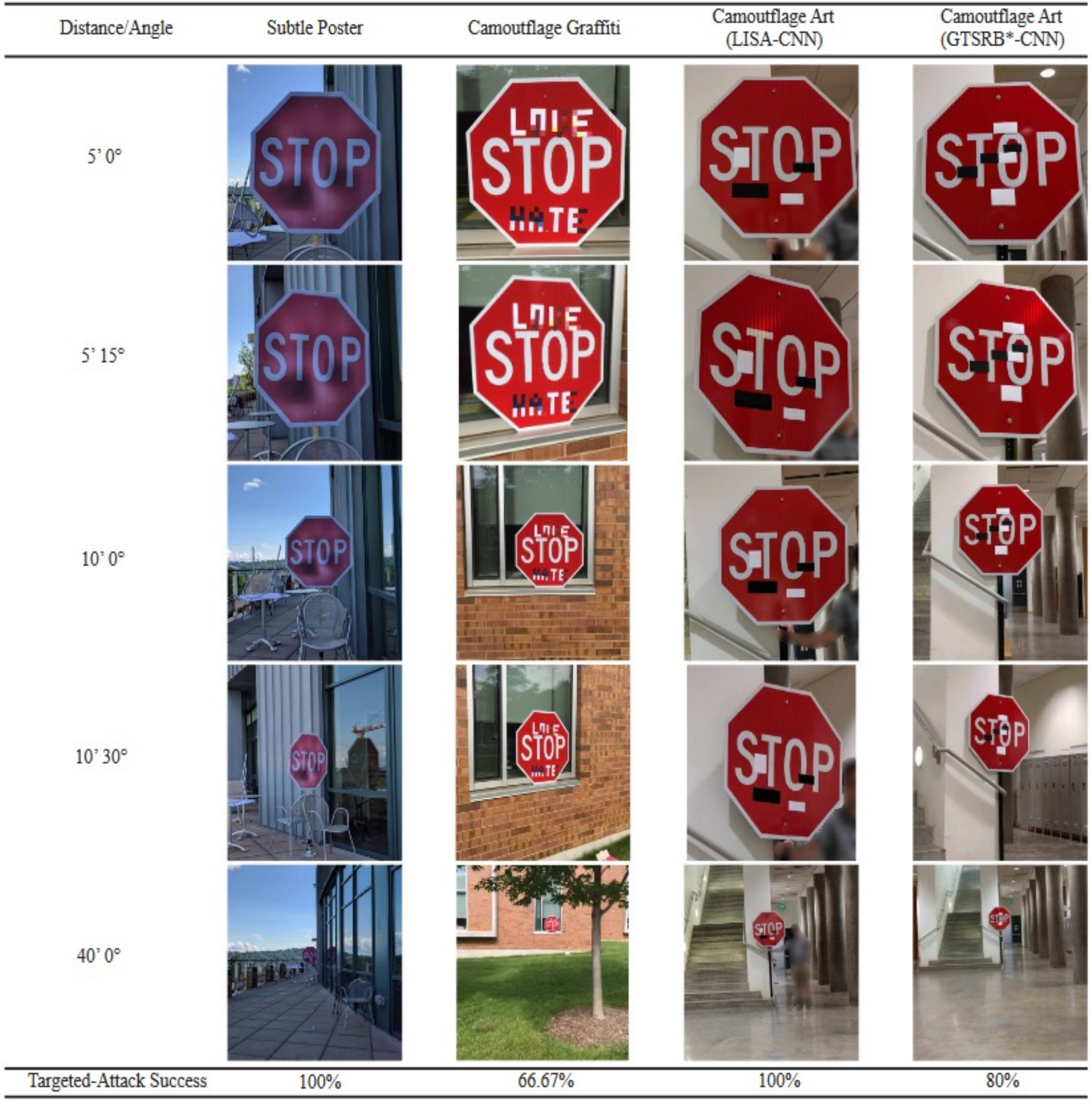
- YOLO detects stop sign
- Faster RCNN detects stop sign very well

- Faster RCNN detects sign more accurately than YOLO
- YOLO works better on higher res video
- Faster RCNN far/small signs accurately

“These effects are so strong that there is no point in significance testing, etc.”

DID IT WORK?

Result of Evtimov’s Experiment

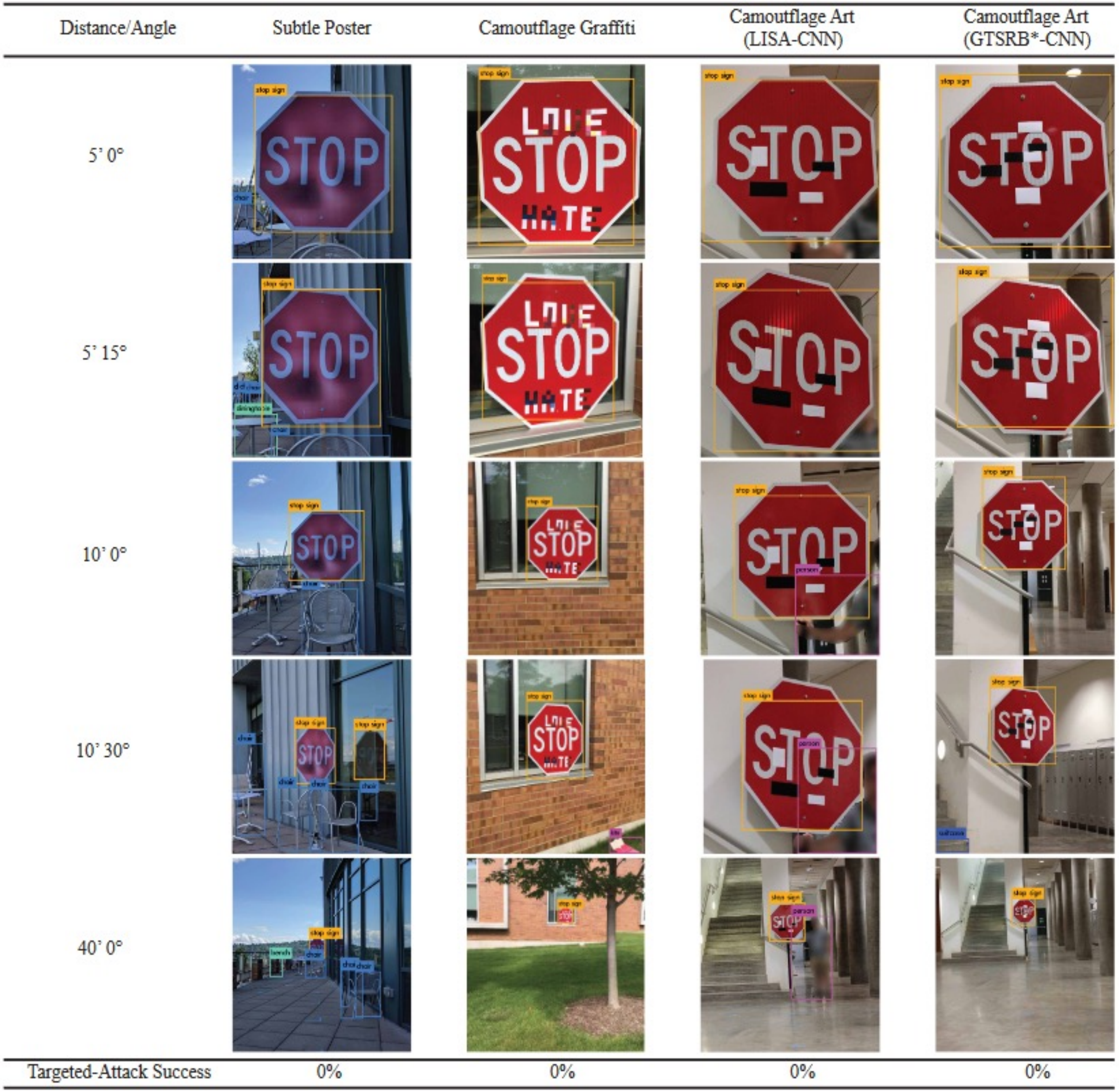


Distance/Angle	Subtle Poster	Camoutflage Graffiti	Camoutflage Art (LISA-CNN)	Camoutflage Art (GTSRB*-CNN)
5’ 0°				
5’ 15°				
10’ 0°				
10’ 30°				
40’ 0°				
Targeted-Attack Success	100%	66.67%	100%	80%

STOP
LOVE
HATE

DID IT WORK?

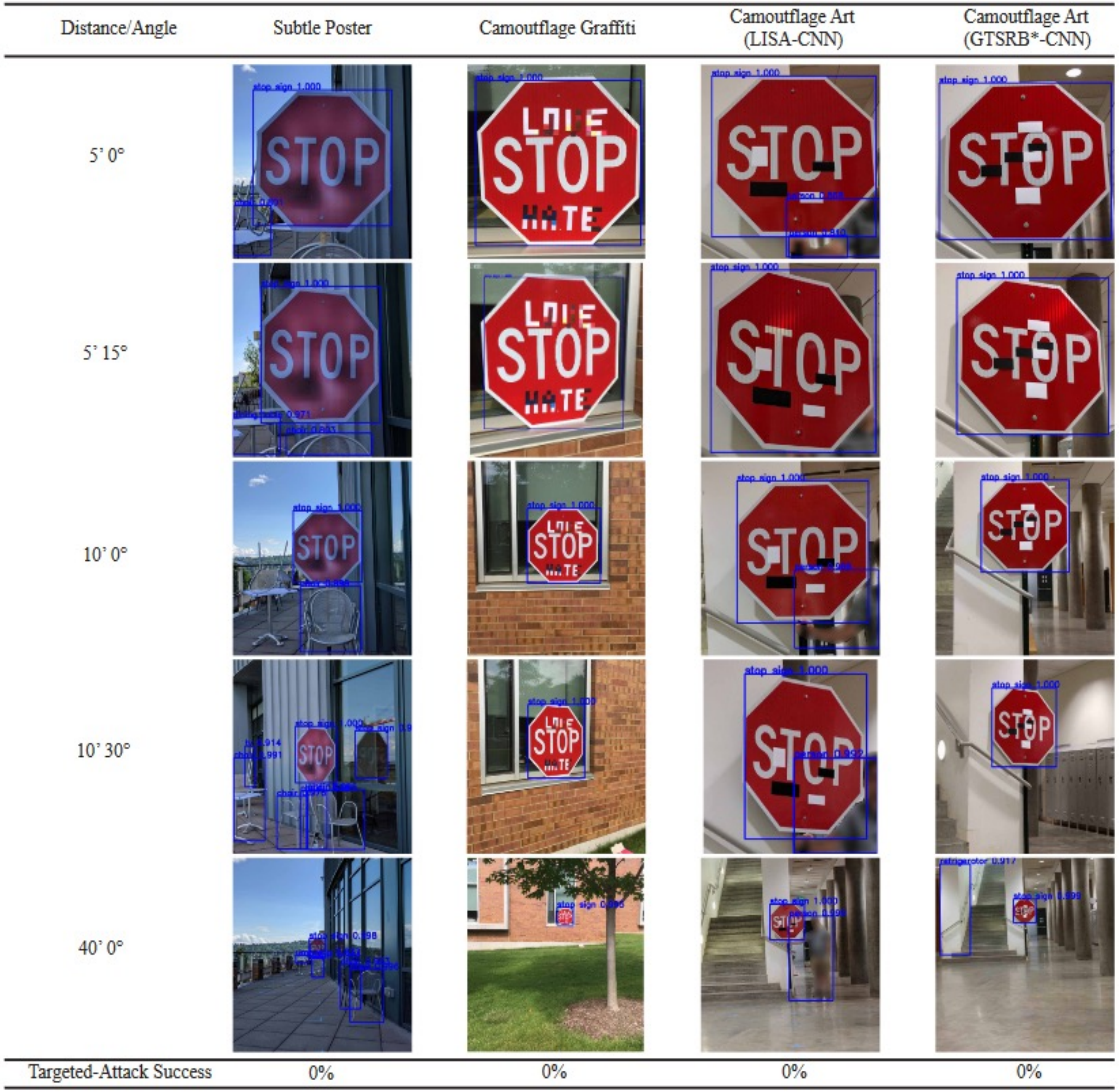
Evtimov's Study Images with YOLO Detector



Distance/Angle	Subtle Poster	Camouflage Graffiti	Camoutflage Art (LISA-CNN)	Camoutflage Art (GTSRB*-CNN)
5' 0°				
5' 15°				
10' 0°				
10' 30°				
40' 0°				
Targeted-Attack Success	0%	0%	0%	0%

DID IT WORK?

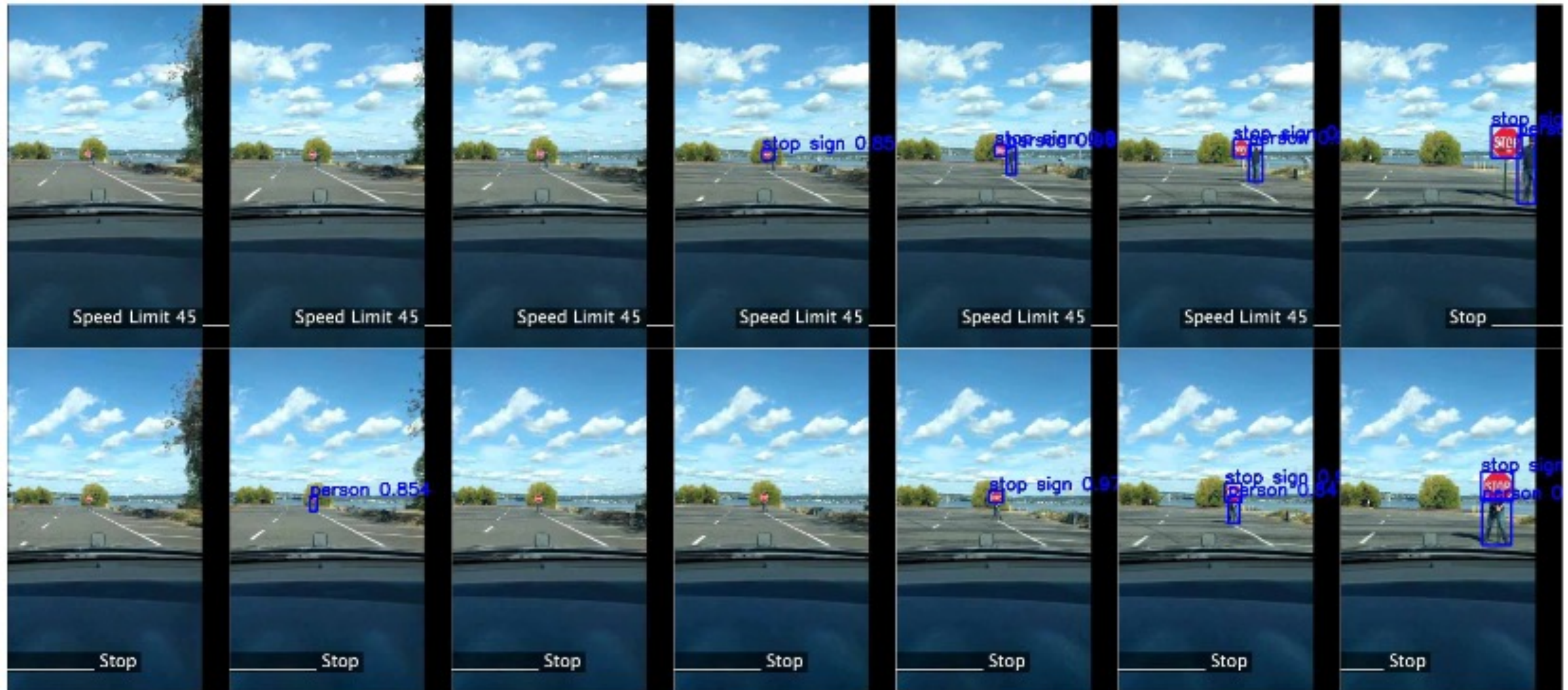
Evtimov's Study Images with Faster RCNN Detector



Distance/Angle	Subtle Poster	Camoutflage Graffiti	Camoutflage Art (LISA-CNN)	Camoutflage Art (GTSRB*-CNN)
5' 0°				
5' 15°				
10' 0°				
10' 30°				
40' 0°				
Targeted-Attack Success	0%	0%	0%	0%

DID IT WORK?

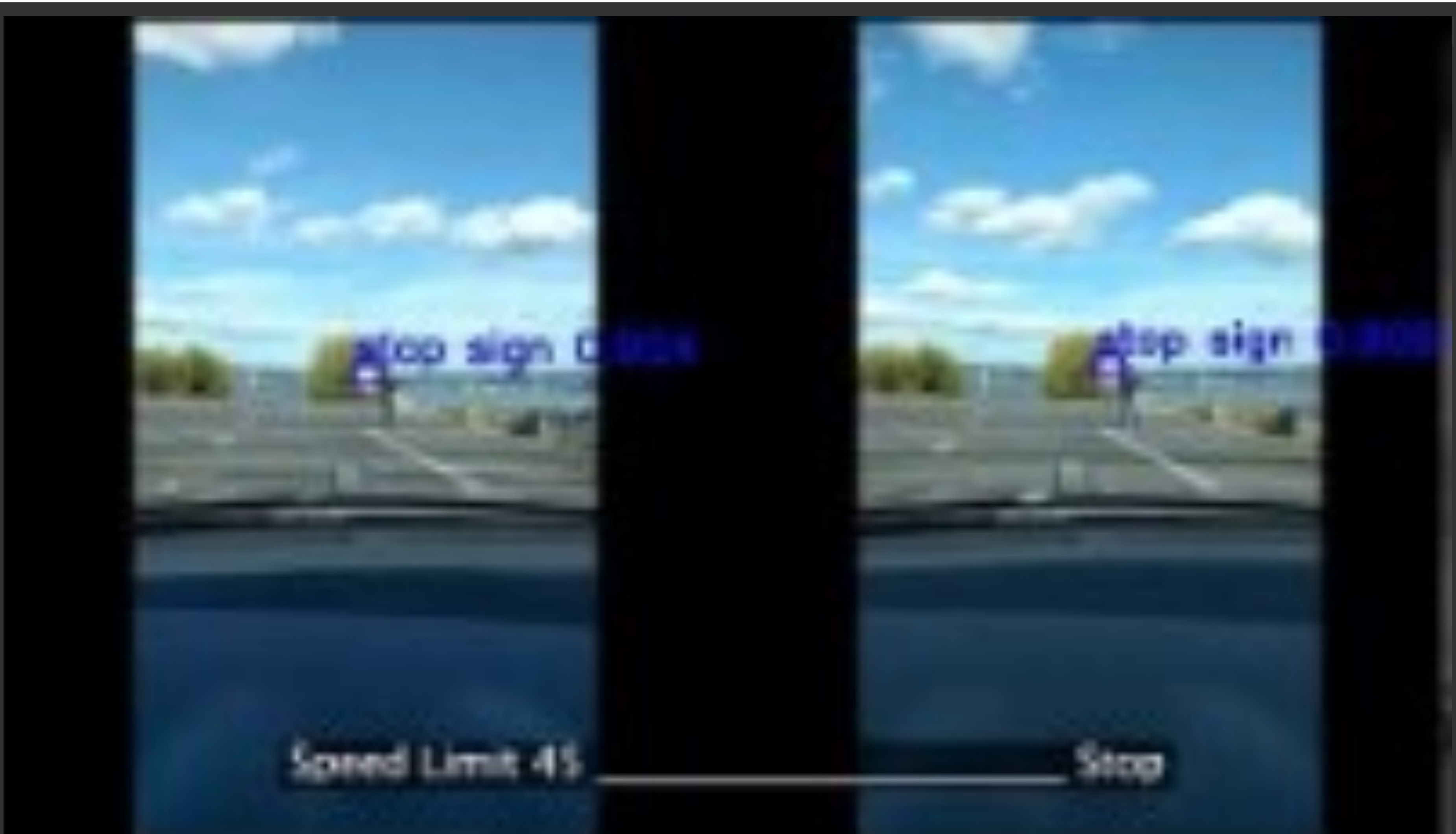
Faster RCNN at Low Resolution



Sticker Attack

Control

DID IT WORK?



CONTRIBUTIONS

- “We do not claim that detectors are necessarily immune to physical adversarial examples. Instead, we claim that there is no evidence as of writing that a physical adversarial example can be constructed that fools a detector. “
- “It is quite natural to study road sign classifiers because image classification remains difficult and academic studies of feature constructions are important. But there is no particular threat posed by an attack on a road sign classifier.”
- Explained issue with Evtimov’s work
- Explained why attack on detector is difficult

CONCLUSIONS

- Fooling Detector != Fooling Classifier
- Attacking Detector is Difficult

STANDARD DETECTORS AREN'T (CURRENTLY) FOOLED BY PHYSICAL ADVERSARIAL STOP SIGNS

DISCUSSION

DISCUSSION

Would you accept?

DISCUSSION

- Preprint vs Published
- Preprinted: 2017, Video Perturbation Attack: 2019