# Trustworthy AI: Adversarial Attack on Object Detection Model

## Subverting Perimeter surveillance Demo

## Securing The AI Attack Surface

# TRANSLATING EVASION ATTACKS TO THE REAL WORLD

## ADVERSARIAL PATCH ATTACKS

### Computer Vision Adversarial Attack

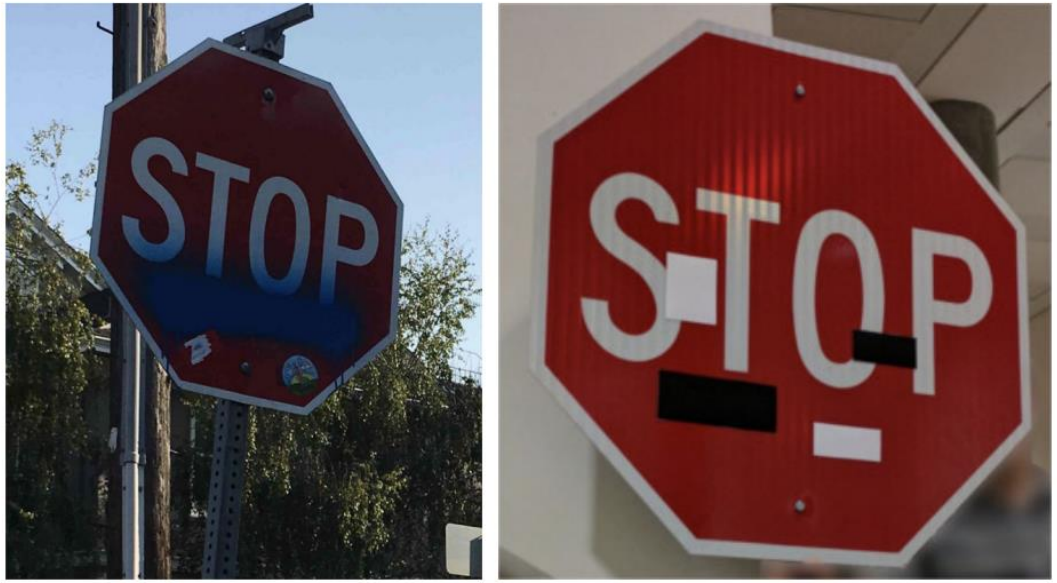
### Modifying digital examples is easy

- Can we generate adversarial noise that translates when reproduced physically?

### Minor modifications to adversarial evasion constraints can produce real world examples

- Emphasize printability and object rotation/augmentation
- Transfer Attacks often effective

Change a Stop sign to a Speed Limit Sign!



"Robust Physical-World Attacks on Deep Learning Visual Classification", Eykholt et. al
https://arxiv.org/pdf/1707.08945.pdf

# ADVERSARIAL PATCH THREAT

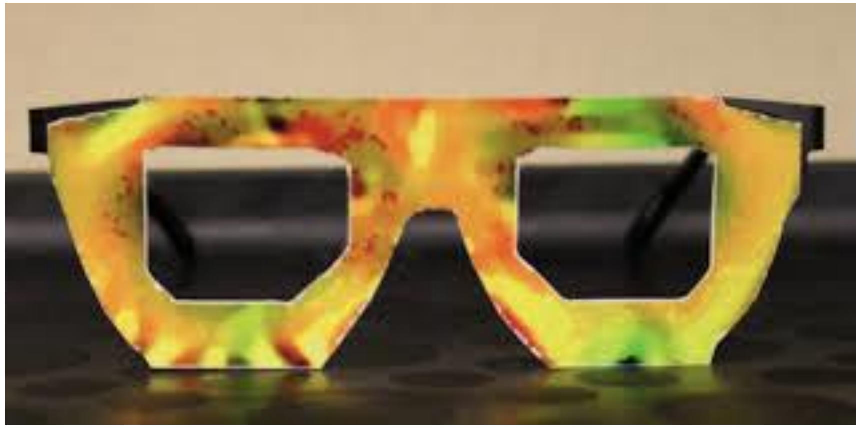
## POTENTIAL THREAT OUTCOMES

### Untargeted attacks

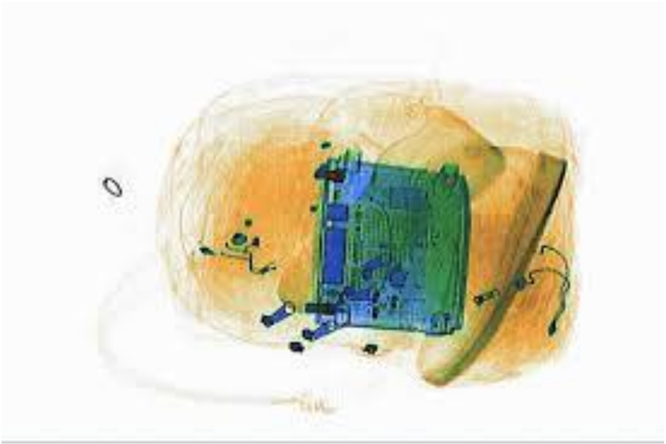
- Evade Perimeter Surveillance cameras
- Evade Crowd Surveillance firearm detection
- Evade X-ray scanning
- Disrupt Maintenance Sensor Monitors

### Targeted Attacks

- Disrupt Targeting Systems
- Cause Self Driving Car Misfunctions
- Deceive Facial Recognition Algorithms



"Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition", Sharif et. al
https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf



# SUBVERTING PERIMETER SURVEILLANCE DEMO

# SURVEILLANCE ADVERSARIAL PATCHES – REAL SCENARIO

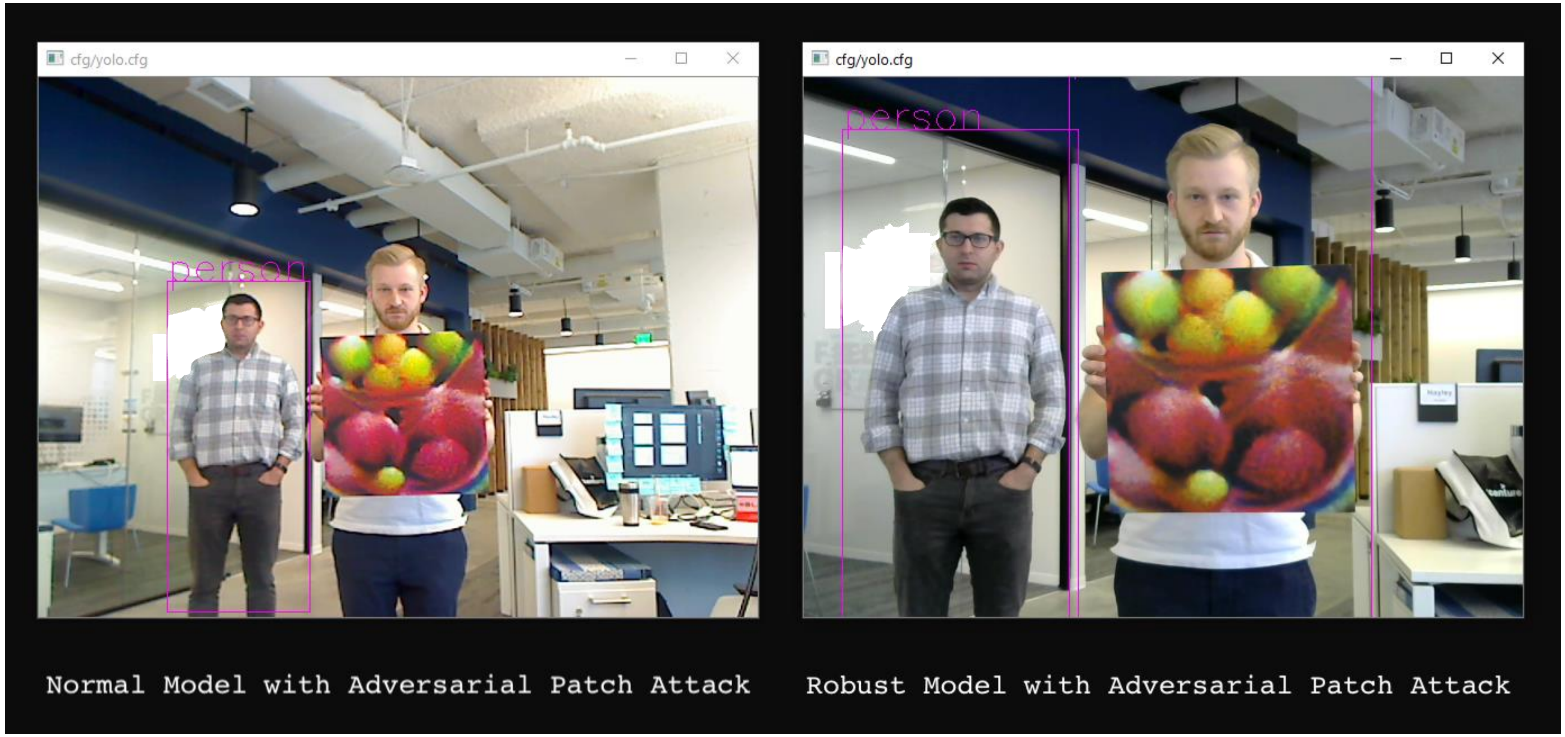
## CAN YOU MAKE A PERSON INVISIBLE TO CAMERAS WITH AN ADVERSARIAL PATCHES?

### Researchers from Belgium attacked the popular yolo2 algorithm for object detection

- First detects where there are objects, draws boxes around them
- Secondarily classifies objects into categories like person, dog, bike
- Objective is to defeat the first detection method, and fool the model into not registering any detections

### Show how to develop defenses to make model more robust

- Adversarial Training



Normal Model with Adversarial Patch Attack

Robust Model with Adversarial Patch Attack

"Fooling automated surveillance cameras: adversarial patches to attack person detection", Thys et. al
https://arxiv.org/pdf/1904.08653.pdf

# DEFENDING AGAINST ADVERSARIAL PATCHES

## RELATIVELY NEW FIELD, NOT A TON OF RESEARCH YET…

### Input Transformation

- Local Gradient Smoothing

### Adversarial Training

- Generalized robustness

### Feature Explanability

- Sentinet

### Image Partitioning/Voting

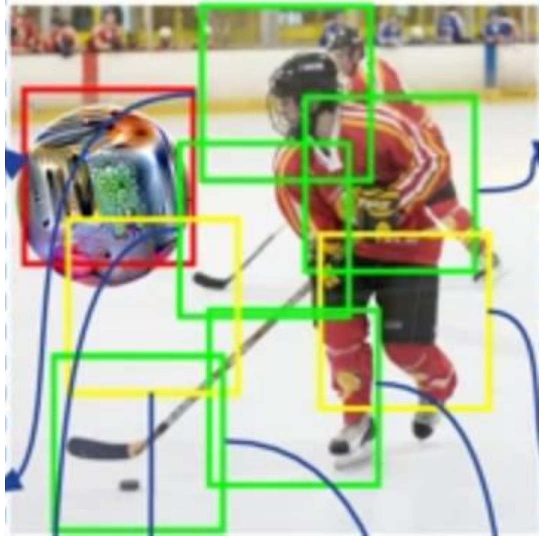
- Ally Patches

### Model Re-Architecture

- Interval Bound Propogation

### Other untested Adversarial Evasion techniques

"Sentinet: Detecting physical attacks against Deep Learning Systems", Chou et. al
https://arxiv.org/pdf/1812.00292.pdf



"Ally patches for spoliation of adversarial patches", Abdel-Hakim et. al
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0213-4

# GOING FORWARD

- **Prepare your AI attack surface**
  - Know your models
  - Know you threat outcomes
  - Know the attacks you are vulnerable to
  - Know the defenses for those attacks

# QUESTIONS/ COMMENTS

## Please Contact for more Information:

Iman Zabett – Data Scientist

Louis DiValentin – Data Scientist Manager