Physical Adversarial Examples for Object Detectors

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li,
Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, Dawn Song

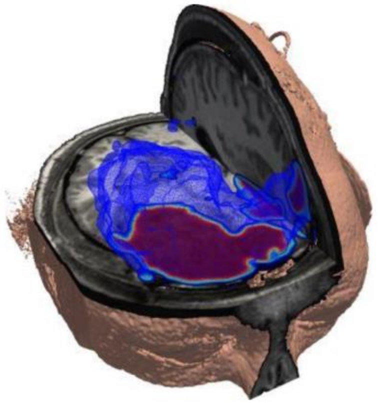
UNIVERSITY of WASHINGTON
UNIVERSITY OF MICHIGAN
Berkeley UNIVERSITY OF CALIFORNIA
Stanford University
Stony Brook University
SAMSUNG RESEARCH AMERICA

Deep Neural Networks are Useful

AlphaGo

Automated Game Playing

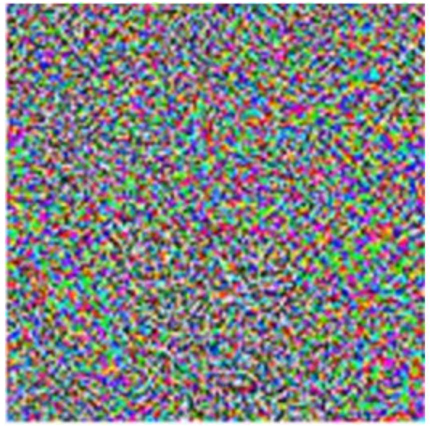
Speech Recognition, Machine XLAT



Fast Brain Lesion Segmentation
Image Courtesy: Nvidia/Imperial College

Deep Neural Networks are Useful, But Vulnerable

$+\ \varepsilon$



$=$

Image Credit: OpenAI

"panda"
57.7% confidence

"gibbon"
99.3% confidence

$f_\theta(x) = y$

$f_\theta(x') \neq y$

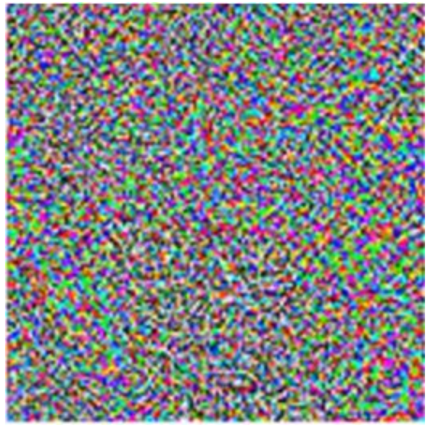
$f_\theta(x') = y^*$

where $x' = x + \delta$

Goodfellow et al., Explaining and Harnessing Adversarial Examples, arXiv 1412.6572, 2015

Deep Neural Networks are Useful, But Vulnerable

$+\ \varepsilon$



$=$

Image Credit: OpenAI

“panda”
57.7% confidence

“gibbon”
99.3% confidence

Can we **physically & robustly** perturb **real** objects,
in ways that cause misclassifications in a DNN?

Current State of Physical Attacks

Classification

What’s the dominant object in this image?

Our prior work

Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR 2018

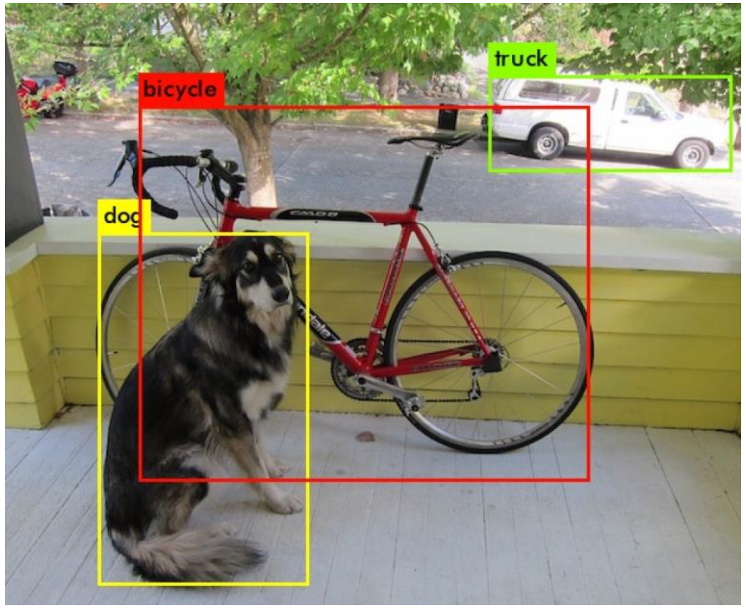
Kurakin et al., Adversarial Examples in the Physical World, arXiv 1607.02533, 2016
Athalye et al., Synthesizing Robust Adversarial Examples, ICML 2018
Brown et al., Adversarial Patch, arXiv 1712.09665
Sharif et al., Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, CCS 2016

Different types of Deep Learning Models

Classification

What's the dominant object in this image?

Object Detection



What are the objects in this scene, and where are they?

Focus of this paper

Semantic Segmentation

What are the precise shapes and locations of objects?

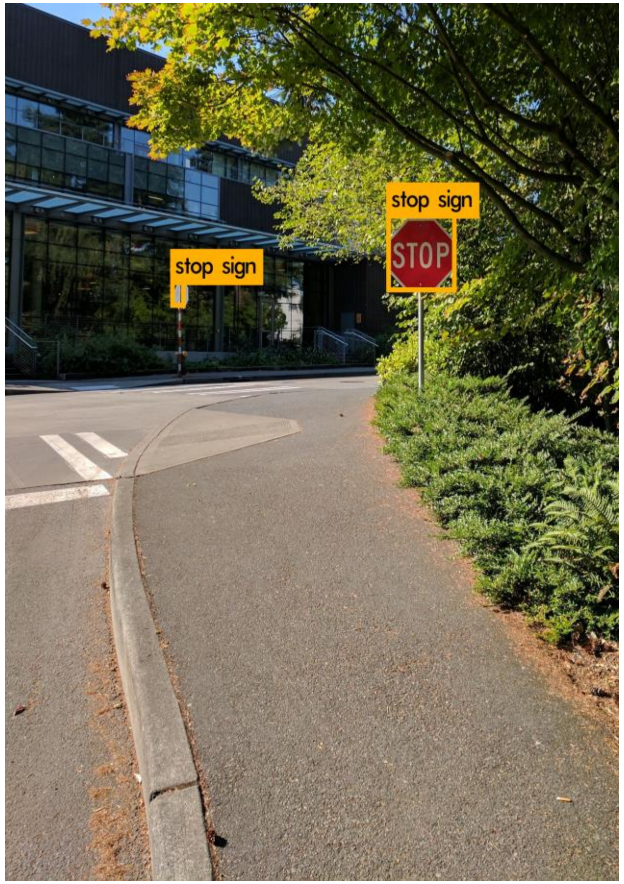
Challenges in Attacking Detectors

Detectors process entire scene, allowing them to use contextual information

Not limited to producing a single labeling, instead labels all objects in the scene

The location of the target object within the scene can vary widely

We will start with an algorithm to attack classifiers and modify it to attack detectors



Review: Robust Physical Perturbations (RP2)

$$\min \quad H(x+\delta, x), \quad \text{s.t.} \quad f_\theta(x+\delta) = y^*$$

A distance function

The target class

$$\underset{\delta}{\operatorname{argmin}} \; \lambda||\delta||_p + J(f_\theta(x+\delta), y^*)$$

Perturbation/Noise Matrix

Loss Function

Lp norm (L-0, L-1, L-2, …)

Challenge: How can we make the perturbations robust to **changing environmental conditions?**

Modeling the Effects of the Environment

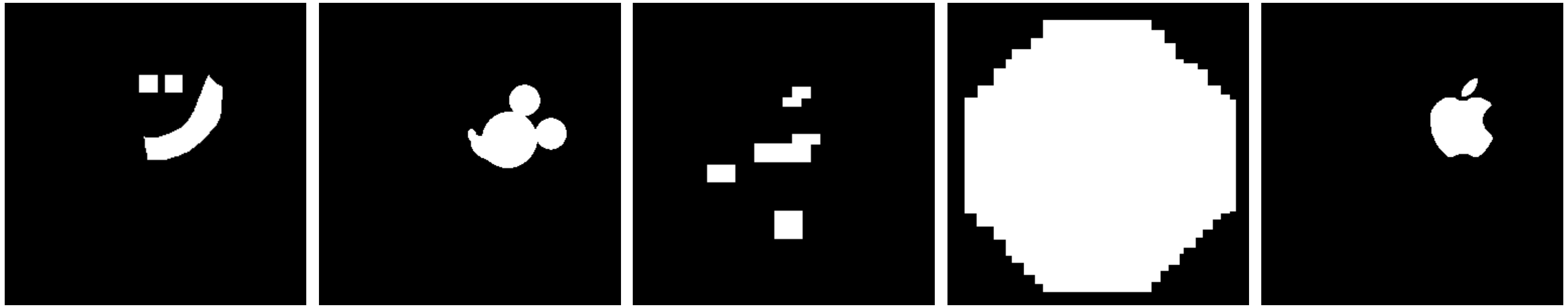
$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$

- Sample from the distribution X^V by:
  - Taking real images varying physical conditions (e.g., distances and angles)
  - Using synthetic transformations

Optimizing Spatial Constraints

$$\underset{\delta}{\operatorname{argmin}}\ \lambda ||M_x \cdot \delta||_p + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$

Example Masks

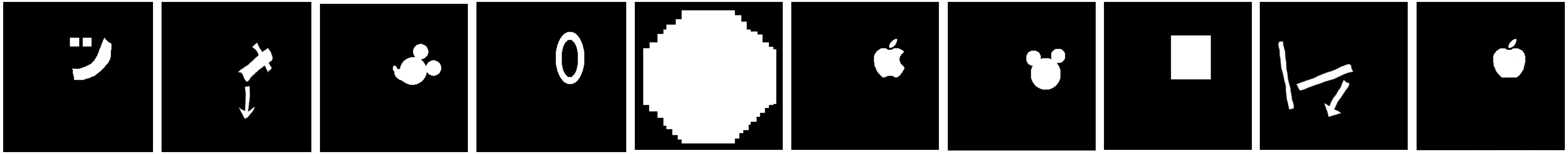


Subtle Poster

Camouflage Sticker

Approximate vandalism

How To Choose A Mask?

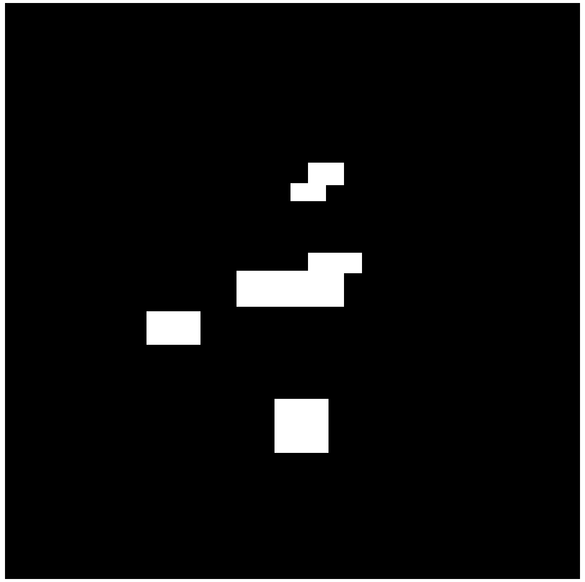


We had very good success with the octagonal mask

Possibility: Mask surface area should be large or should be focused on “sensitive” regions

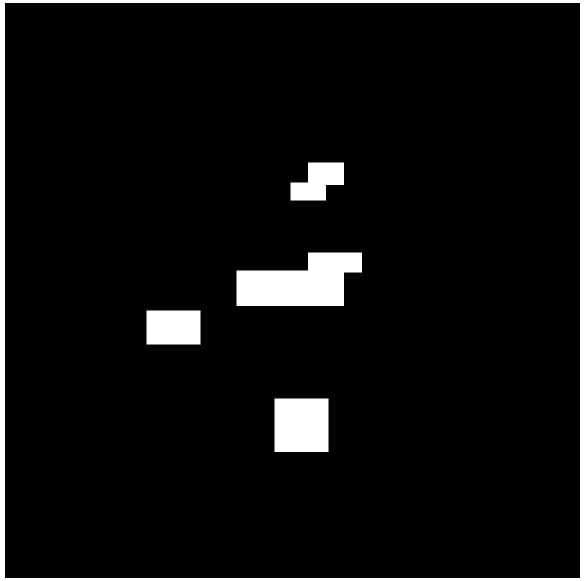
Use L-1

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p$$
$$+\mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$



Process of Creating a Useful Sticker Attack

L-1 Perturbation



Result Mask

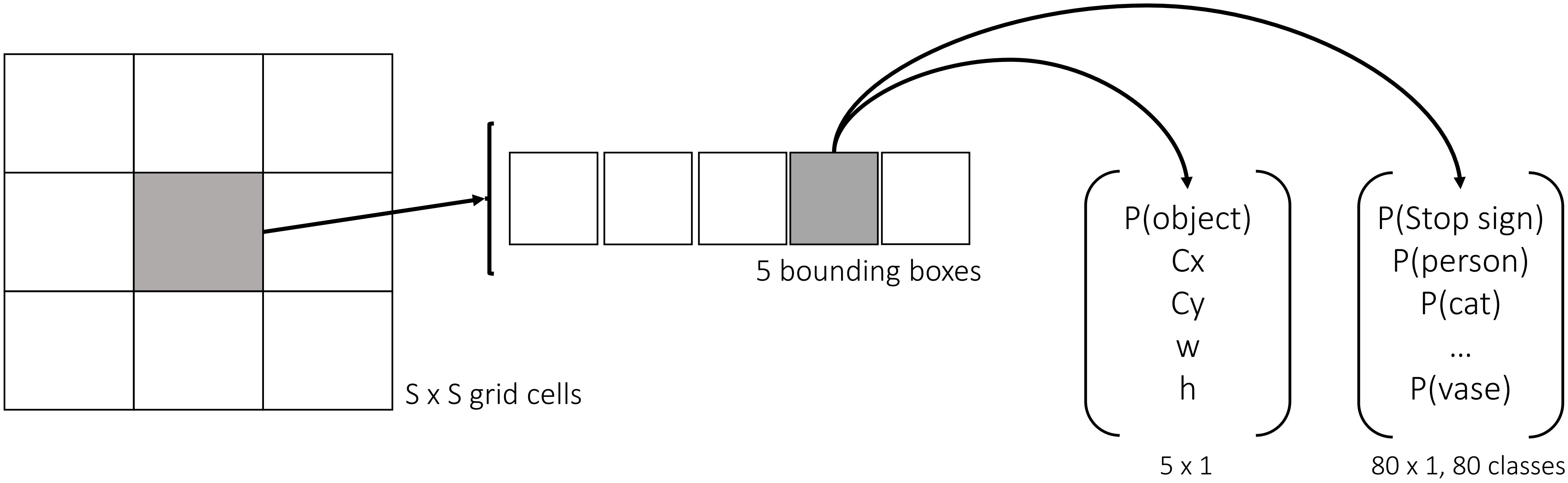
Sticker Attack!

Adapting RP2: Translational Invariance

…

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$

Adapting RP2: Adversarial Loss Function



$$J_d(x,y) = \max_{s \in S^2, b \in B} P(s, b, y, f_\theta(x))$$

Input scene

Prob. of object being class ‘y’

Output of YOLO, 19 x 19 x 425 tensor

Minimize the probability of “Stop” sign among all predictions

Poster and Sticker Attack

Evaluation Method & Data

- Record a video while moving towards a sign
- Sample video frames
- Count number of frames in which Stop sign was NOT detected

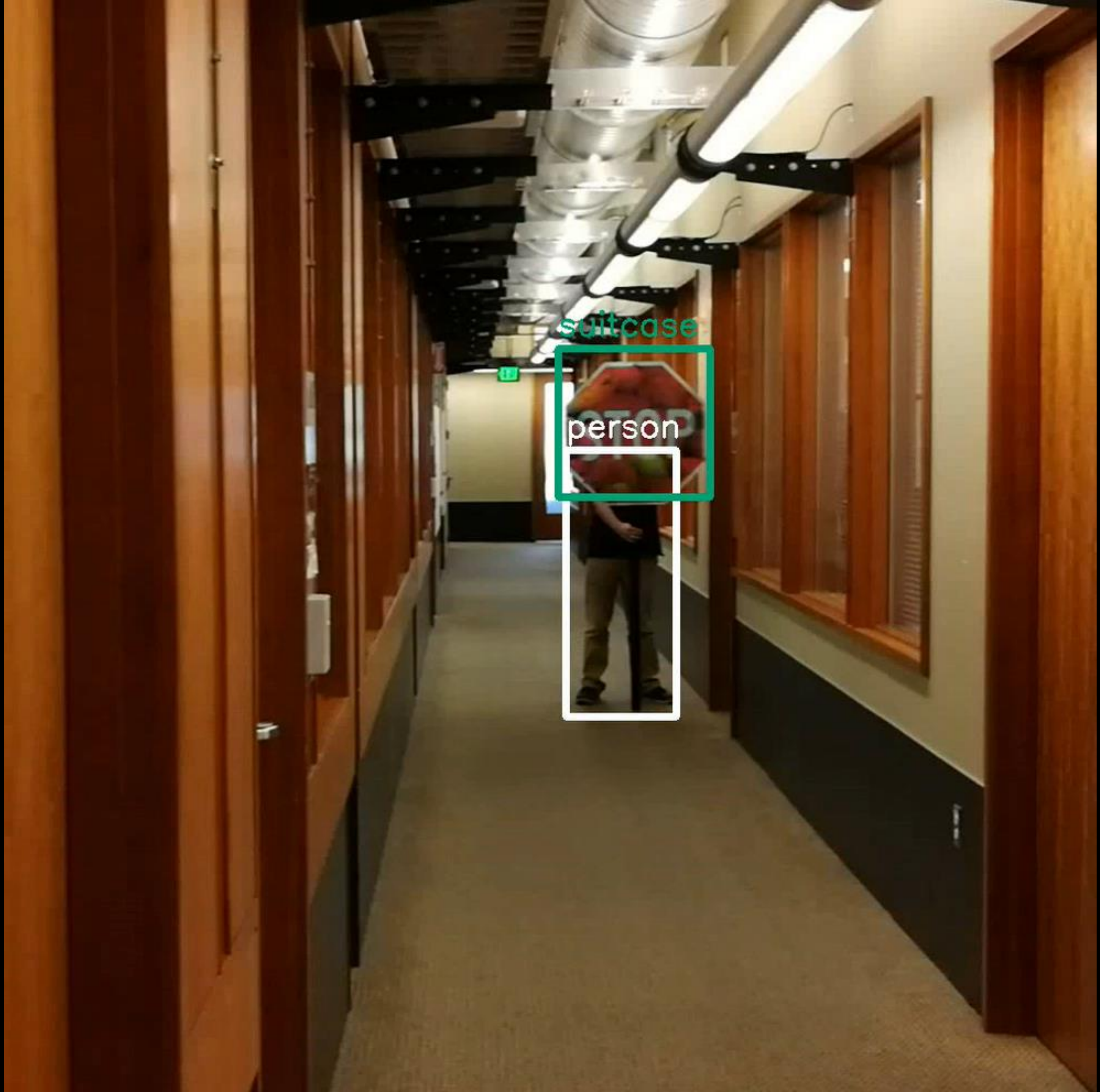
White-box

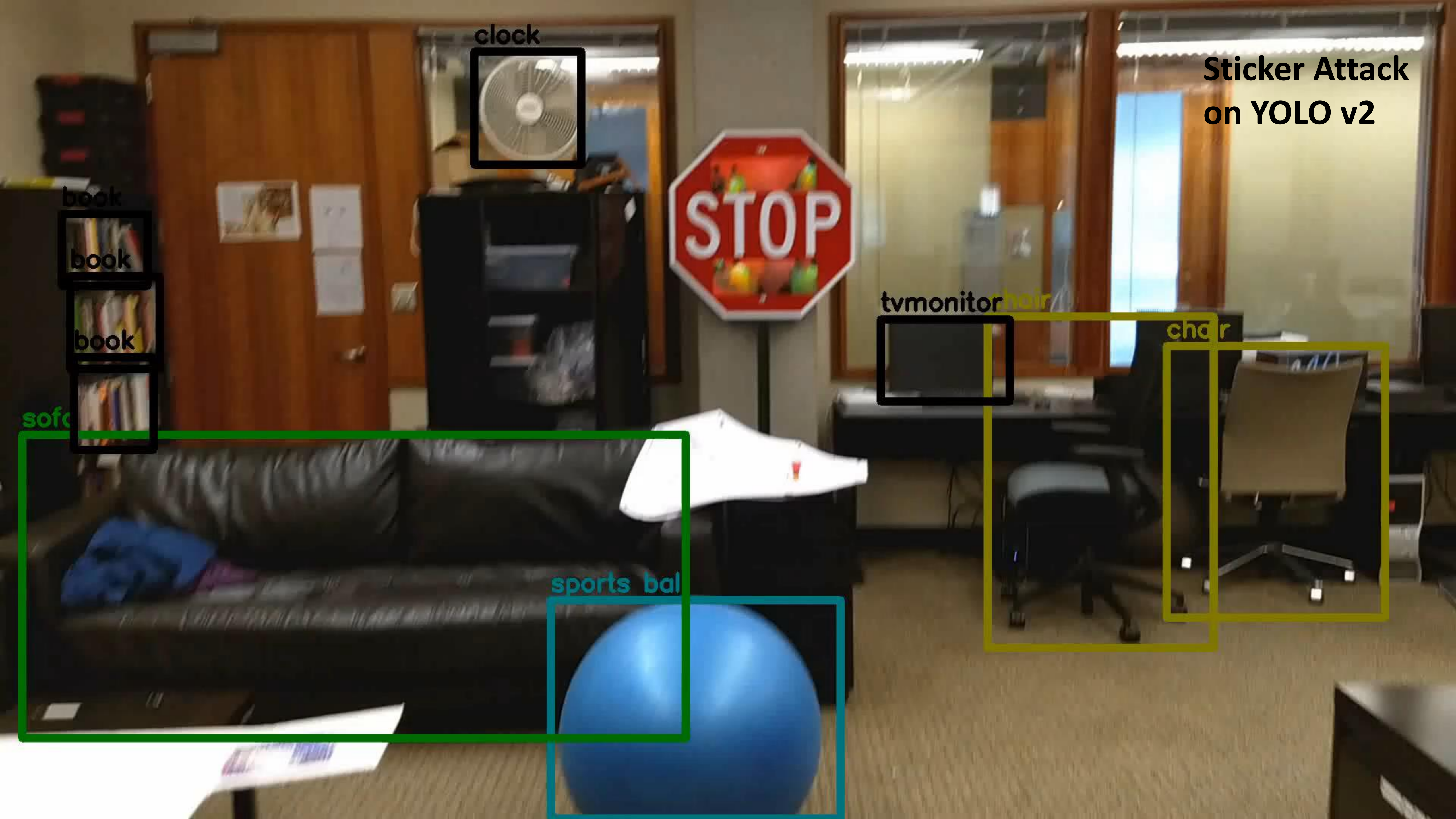
YOLO v2	Poster	Sticker
Indoors	202/236 (85.6%)	210/247 (85.0%)
Outdoors	156/215 (72.5%)	146/230 (63.5%)

Black-box

FR-CNN	Poster	Sticker
Indoors	189/220 (85.9%)	146/248 (58.9%)
Outdoors	84/209 (40.2%)	47/249 (18.9%)

Poster Attack on YOLO v2




Sticker Attack
on YOLO v2
clock
book
book
book
STOP
tvmonitor
sports bal

Black-box
transfer
to
Faster-RCNN

Creation Attacks

- Cause the detector to hallucinate
- A meaningless-to-humans object but detected as an attacker-chosen class

Threshold after which we stop optimizing for box confidence, set to 0.2

$$\text{object} = P_{\text{box}}(s, b, f_\theta(x)) > \tau$$

$$J_c(x, y) = \text{object} + (1 - \text{object}) \cdot P(s, b, y, f_\theta(x))$$

Y is the class that the attacker wants the detector to see

YOLO labels this as a Stop sign

Takeaways

- Adversarial examples generalize to varied environmental conditions
- With modest changes to our prior work on attacking classifiers, adversarial examples generalize to the richer class of object detection models
  - We introduced a new type of adversarial loss (disappearance, creation)
  - We also introduced the Translational Invariance property
- Our evaluation based on the state-of-the-art YOLO v2 detector shows that physical attacks are possible up to distances of ~30 feet
- Do these attacks have system wide effects?

Earlence Fernandes,
earlence@cs.washington.edu,
earlence.com

Thank you!

- Adversarial examples generalize to varied environmental conditions
- With modest changes to our prior work on attacking classifiers, adversarial examples generalize to the richer class of object detection models
  - We introduced a new type of adversarial loss (disappearance, creation)
  - We also introduced the Translational Invariance property
- Our evaluation based on the state-of-the-art YOLO v2 detector shows that physical attacks are possible up to distances of ~30 feet
- Do these attacks have system wide effects?