Robust Assessment of Real-World Adversarial Examples

Brett Jefferson

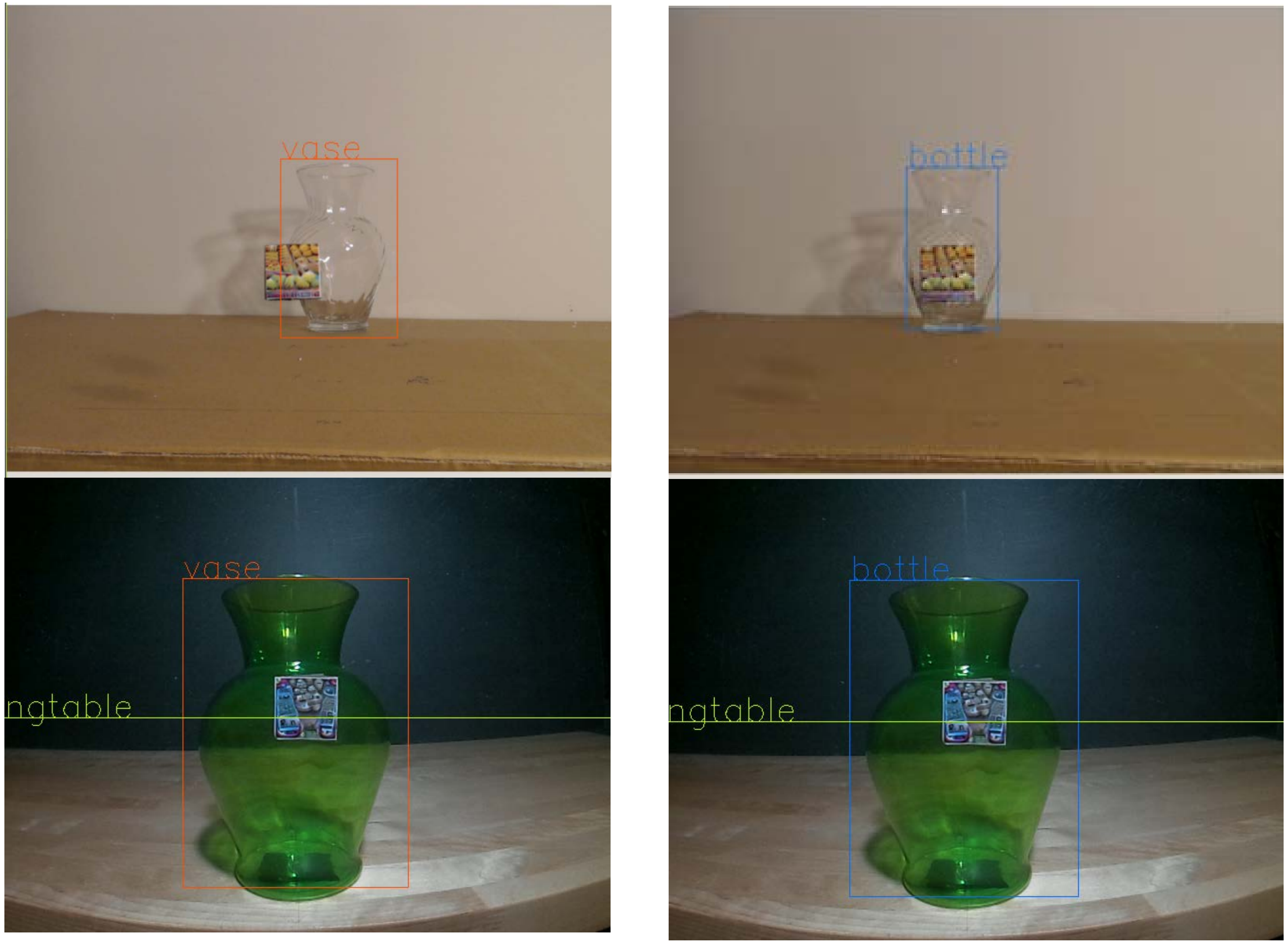
Carlos Ortiz Marrero

Pacific Northwest
National Laboratory

Richland, WA 99354

Pacific Northwest
NATIONAL LABORATORY

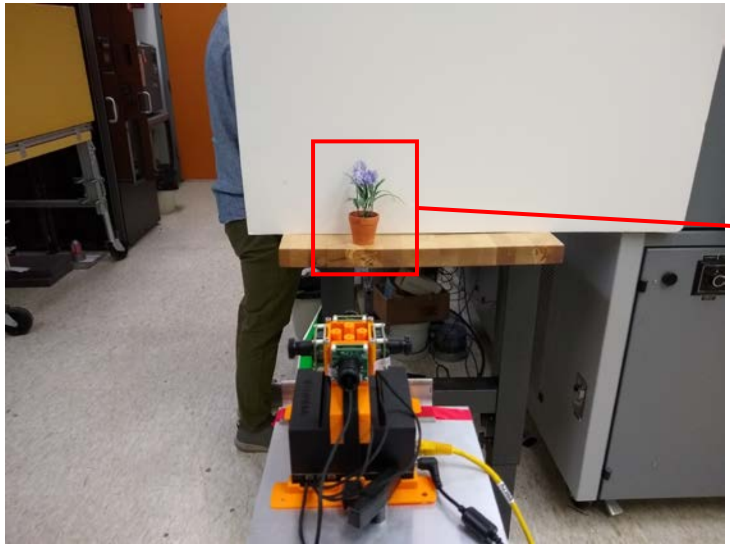
Adversarial Perturbations Can Be Brittle!

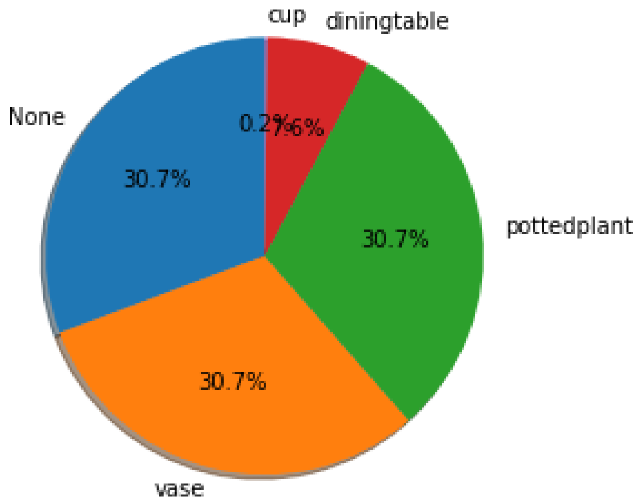


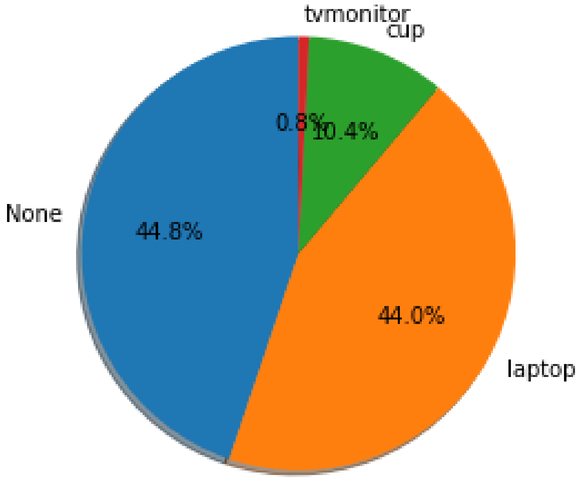
Real-World Assessments Have Different Challenges

Real-World Challenges:

- Camera focusing
- Auto exposure
- Small perturbations
- How to physically manifest adversaries
  - Fabric/ Materials
  - DPI
  - Color Quality
  - Scale Considerations





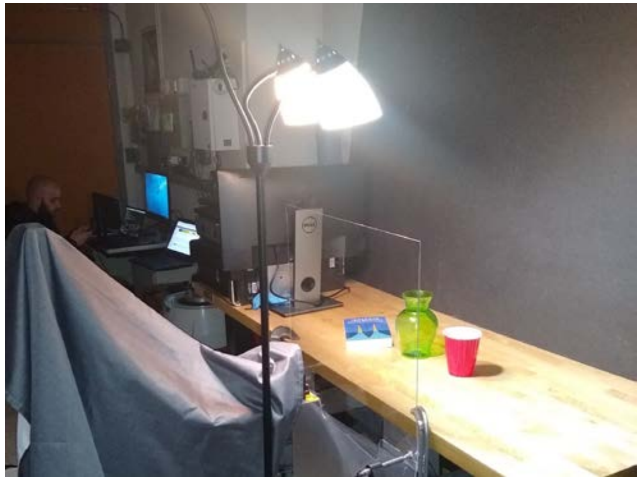


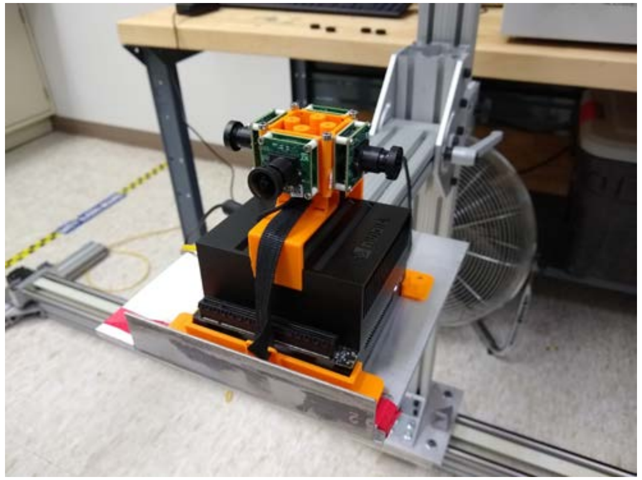
Even Background alone is noisy

We Sought

1. A measure that communicated in a single value of how well an adversary performed while accounting for
   a) Noisy frame-by-frame variation
   b) How the model performs in the absence of an adversary
2. A testing procedure that allowed for controlled, yet realistic variability
3. To understand (and quantify) how various objective functions and training sets impact adversarial success

Testbed

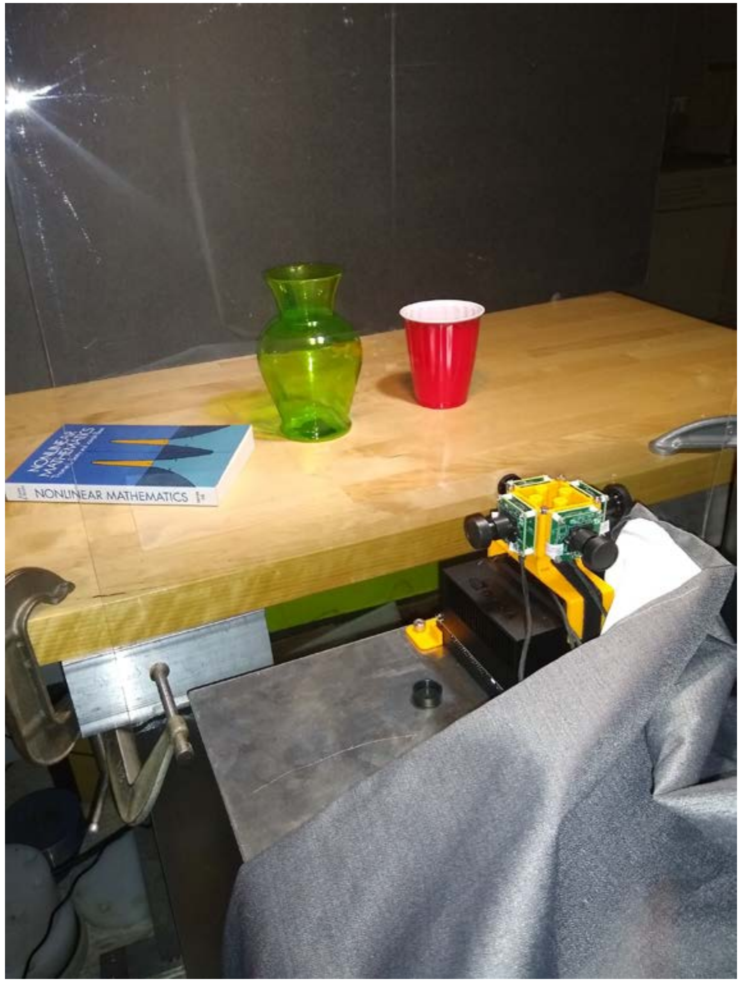




Desired a controlled environment to quantify frame-by-frame variation over many fixed (but slightly perturbed) scenes

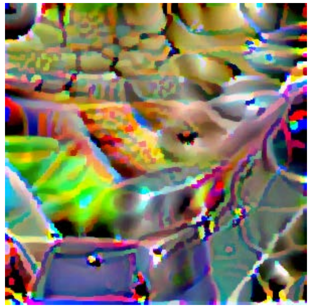
Primary evaluation platform/ testbed:

- Xavier Jetson
- e-CAM130_CUXVR camera
- Two Independent light sources
- Custom mounting system for controlled viewing

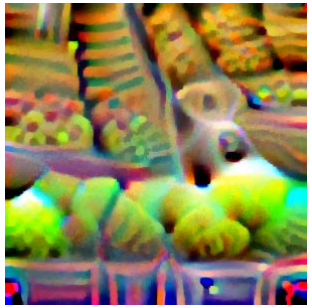


Adversarial Patches

Database: ImageNet
Optimization: Class Score x Objectness Score
Name: **“ImageNet(CxO)”**



Database: ImageNet
Optimization: Objectness Score
Name: **“ImageNet(O)”**



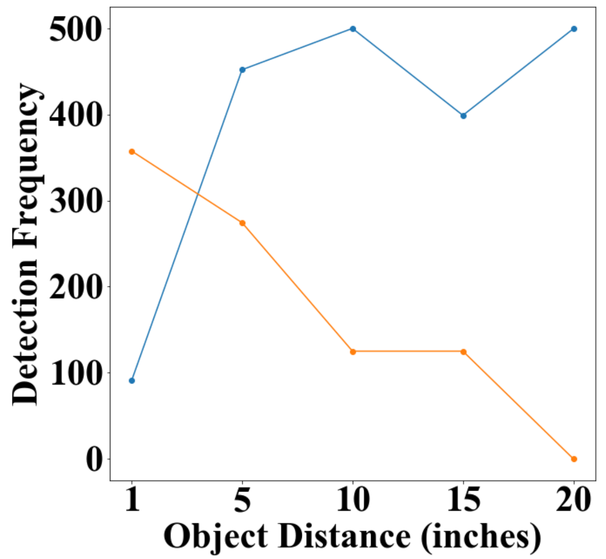
Database: ImageNet and OpenImages
Optimization: Class Score x Objectness Score
Name: **“Composite(CxO)”**

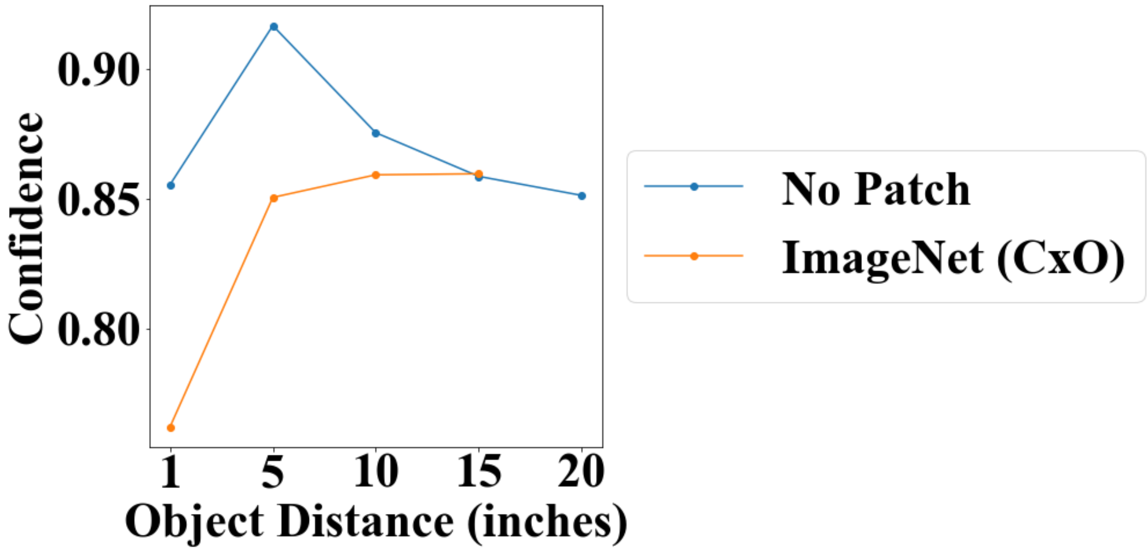
White Sticker

No Sticker

Thys S, Van Ranst W, Goedemé T. “Fooling automated surveillance cameras: adversarial patches to attack person detection”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2019

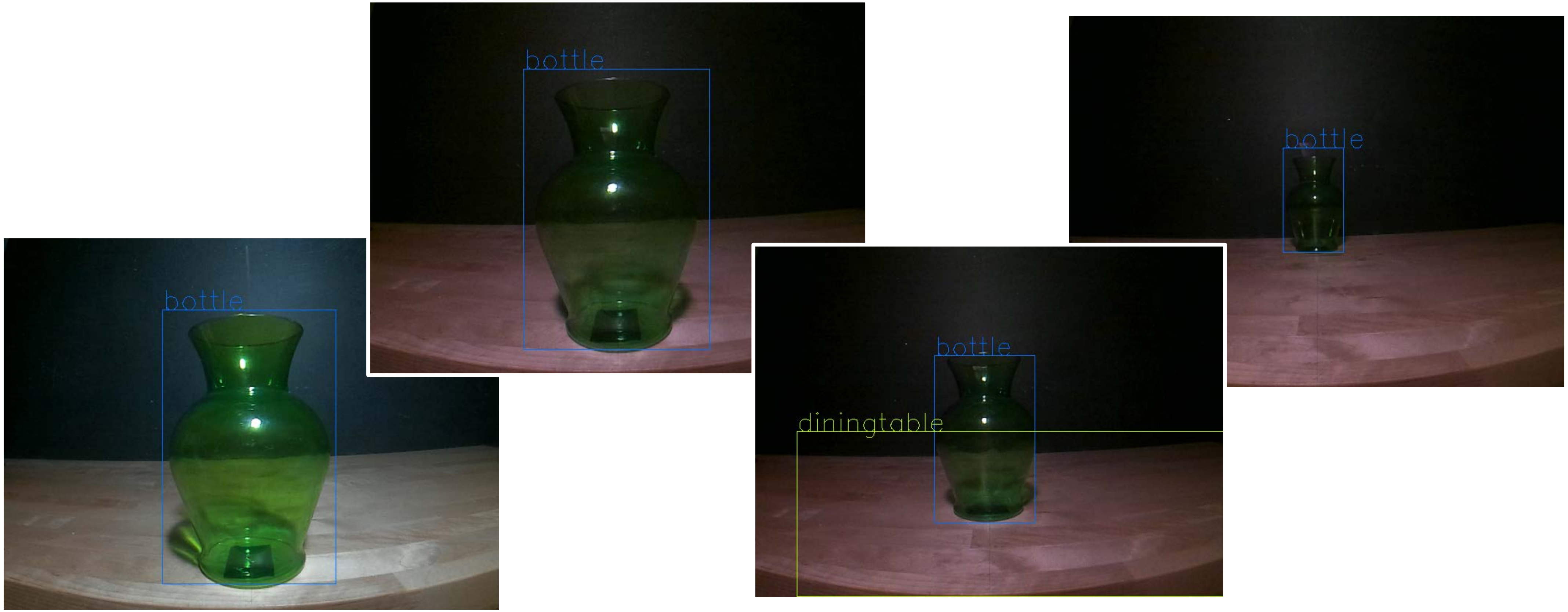
Lesson 1: Frequencies and confidences are not enough to understand adversarial (or baseline model) performance.





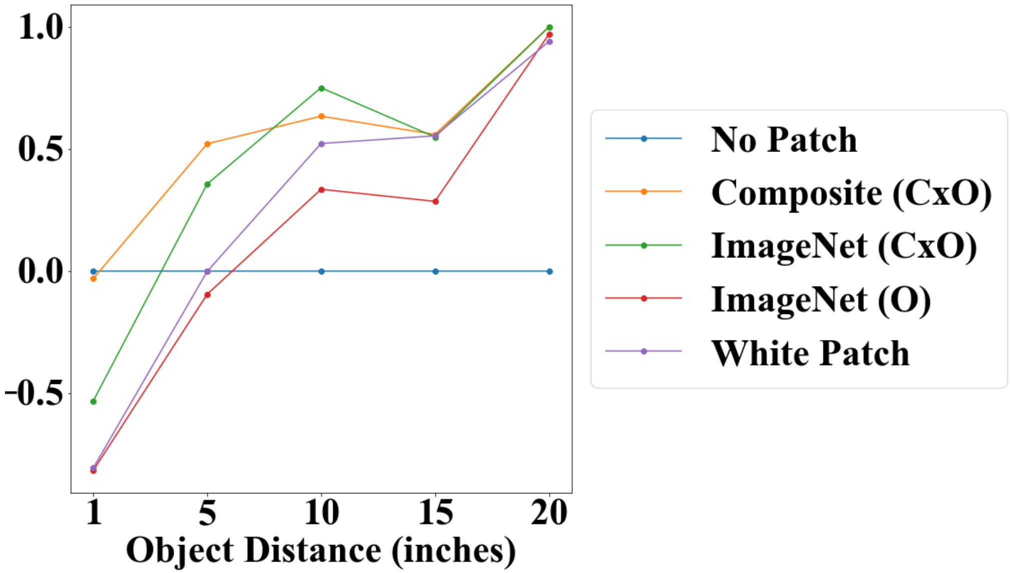
Lesson 2: Baseline performance matters!

('Vase' is a class in YOLOv2)



Lesson 3: A global score that accounts for baseline

$$S(P, E) = \frac{1}{|E|}\sum_{e\in E}\frac{(f_{\{\emptyset,e\}} - f_{\{P,e\}})}{\# \, Frames}$$



Normalized frequency differences per distance condition

Patch	Score
No Patch	0
♦Composite (CxO)	**0.536**
ImageNet (CxO)	0.424
White Patch	0.241
ImageNet (O)	0.135

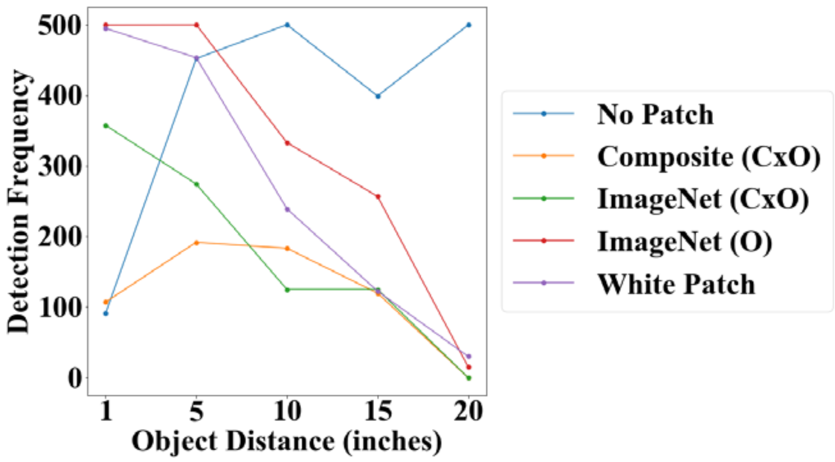
♦ Best Performing Patch

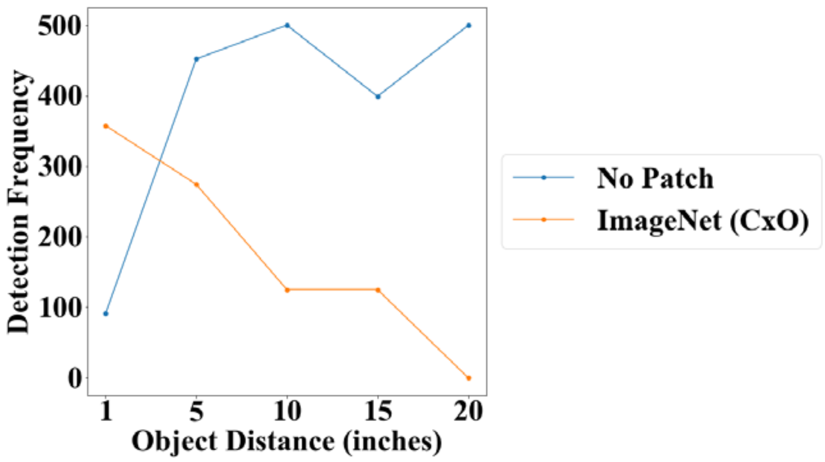
In Summary

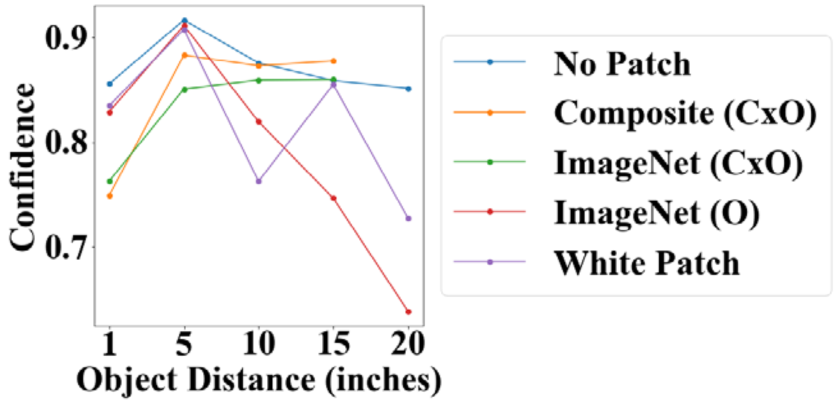
1. We provide an example of testing robustness of adversarial examples that
   a) can generalize to other physically developed examples
   b) accounts for natural changes and baseline performance
2. We provide a measure of robustness that is practical and evaluates real-world adversarial examples

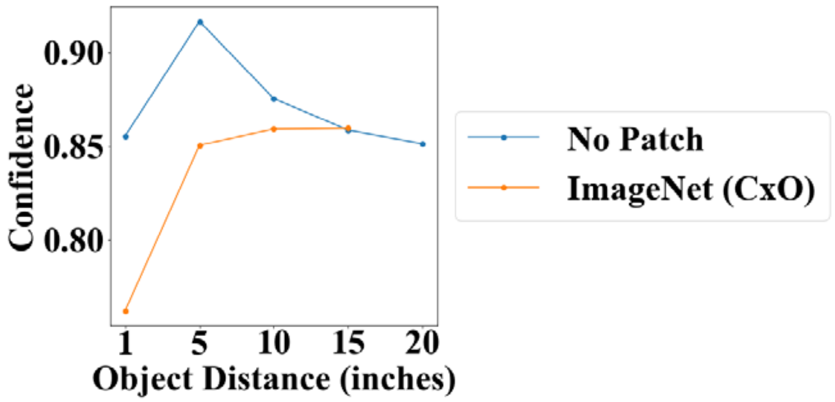
Thank You

Appendix







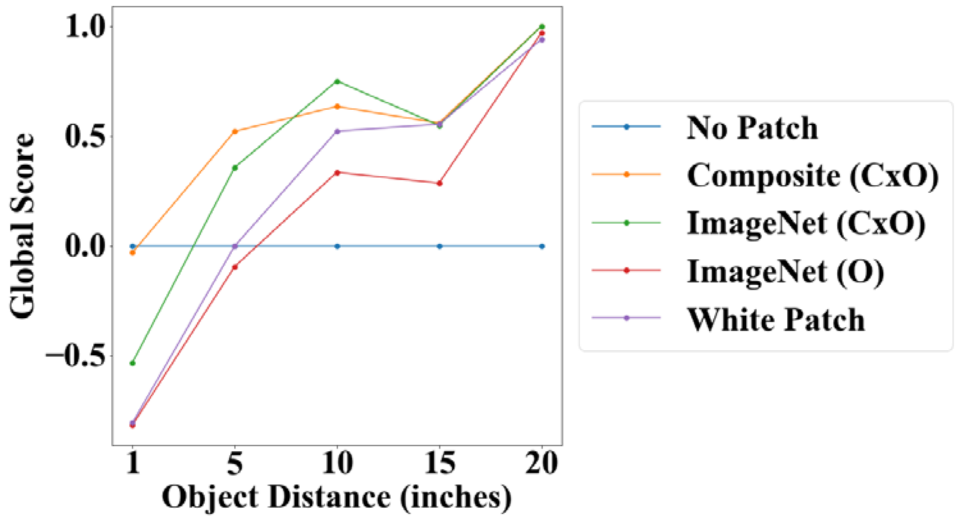


While detection frequency decreases with distance, confidence remains relatively high

At 15 inches, we cannot tell if detection decrease is due to adversarial activity or poor model performance

All patches show similar trend

Global Score



Patch	Score
No Patch	0
♦Composite (CxO)	**0.536**
ImageNet (CxO)	0.424
White Patch	0.241
ImageNet (O)	0.135

$$S(P, E) = \frac{1}{|E|} \sum_{e \in E} \frac{(f_{\{\emptyset,e\}} - f_{\{P,e\}})}{\# \, Frames}$$

- E := Environments/ Small perturbations
- Detection frequency difference between a no-adversary condition and adversary condition, normalized over perturbations of the scene.
- Provides an intuitive measure of adversarial performance over controlled environmental factors.
- Adversarial patch trained on two image datasets and minimizing the product of ‘class score’ and ‘objectness score’ out performs patches trained in less diverse way.

Pacific Northwest
NATIONAL LABORATORY

Class-To-Class

Understanding class-to-class misclassifications is certainly valuable for quantifying adversarial performance in a more complete way. However, the current study provides a coarse first approach to fast scoring with small scene changes