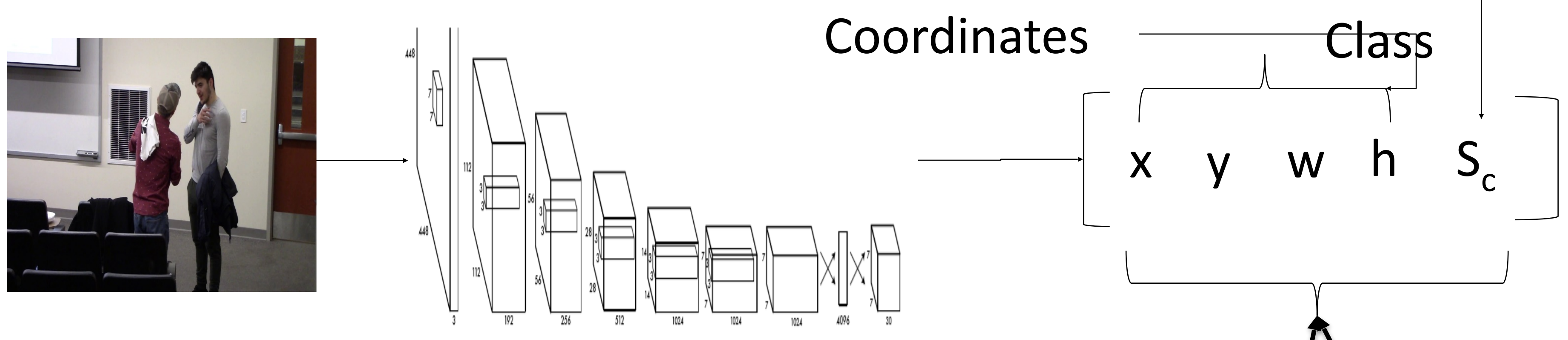
CERIAS
The Center for Education and Research in Information Assurance and Security

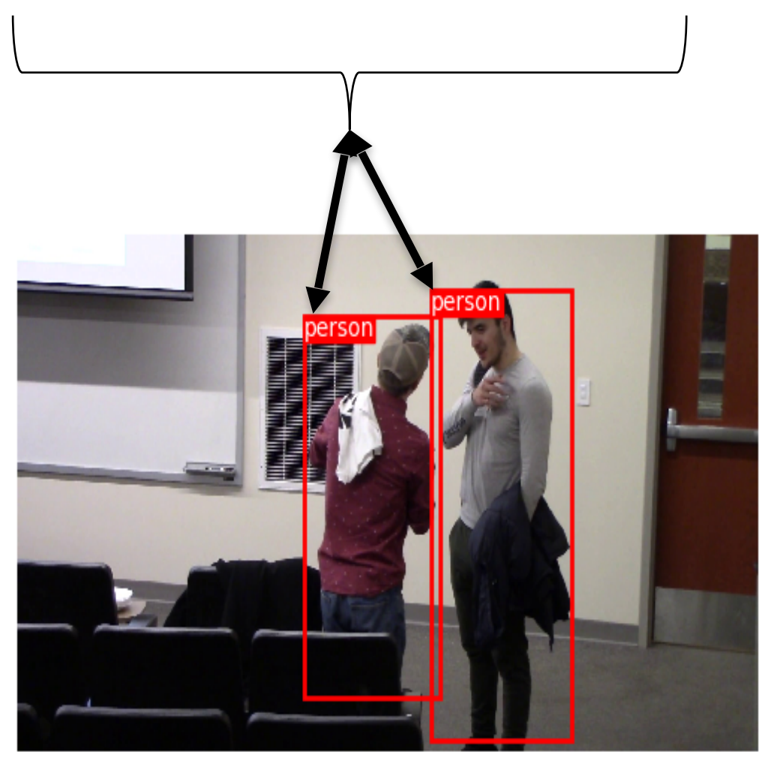
We Need Robust Neural Networks

Daniel Merrick, Karthik Maiya, Kirthi Sivamani, Rui Wang, Prof. Yung-Hsiang Lu

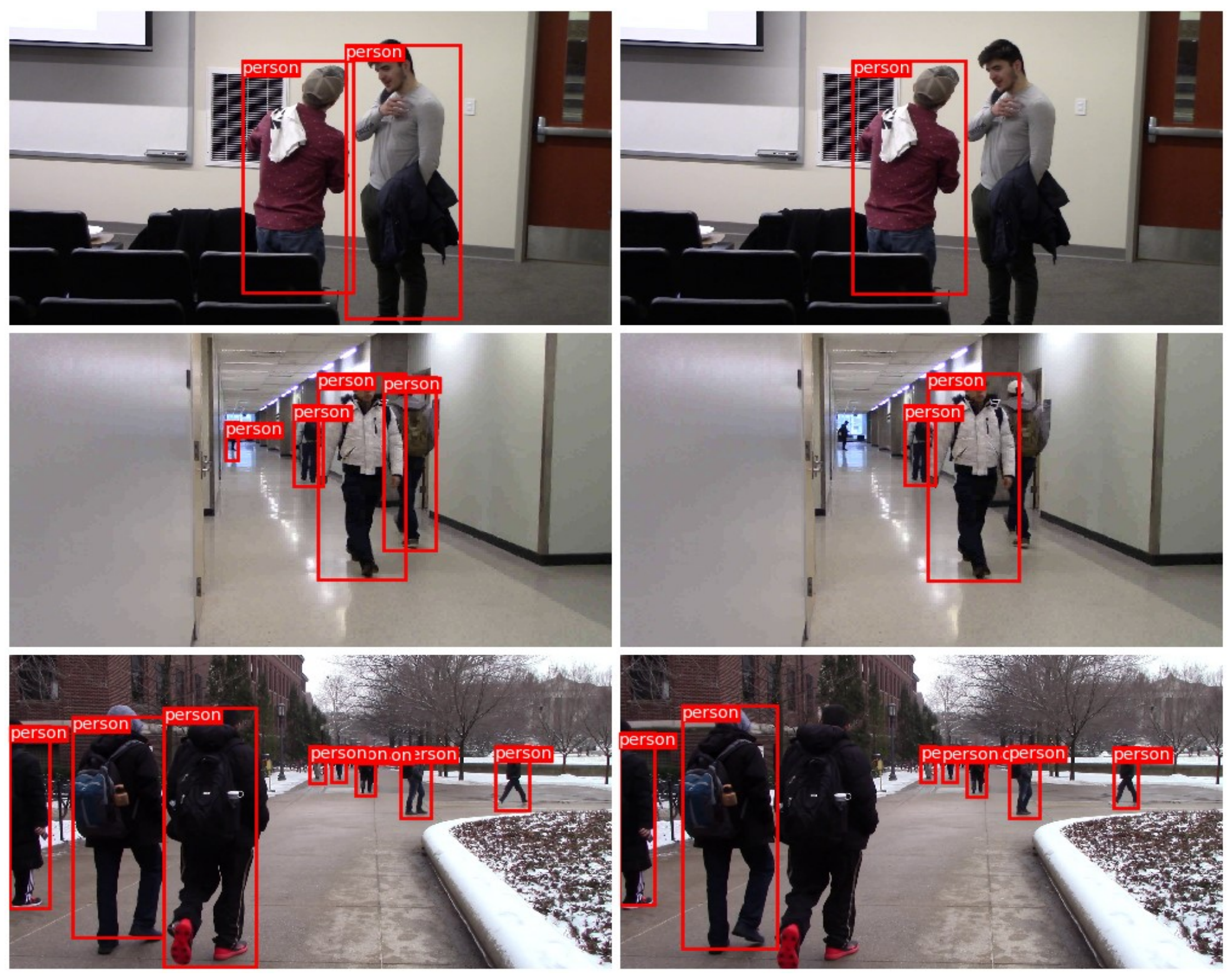
1. Background: Object Detection and Convolutional Neural Networks (CNN)



- Our work focuses on understanding and improving the stability of deep convolutional object detection algorithms. Object detection algorithms have two roles (1) localizing objects and (2) classifying objects.



2. Problem: Object Detectors are Inconsistent

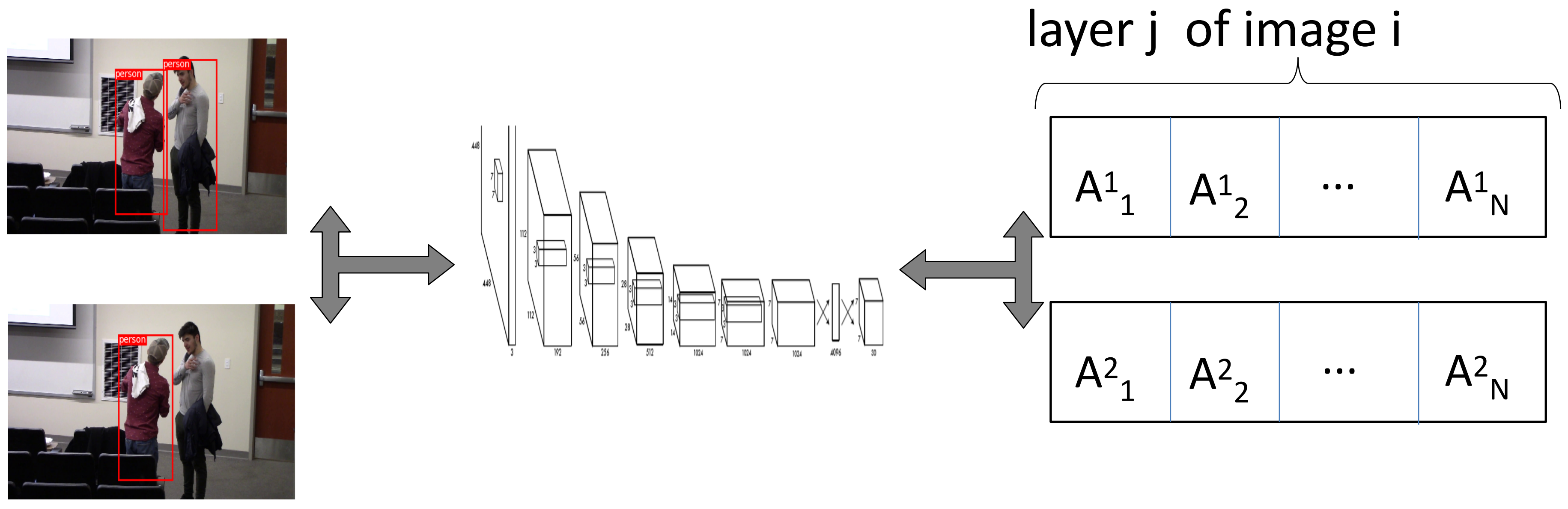


- Each pair of images are one frame apart
- Images in the left column are correctly detected
- Images in the right column have incorrect detections

*Detections made by YOLOv3

3. Explaining Why: Measuring the Sensitivity of Learned Features

Average activations A^i_j at layer j of image i

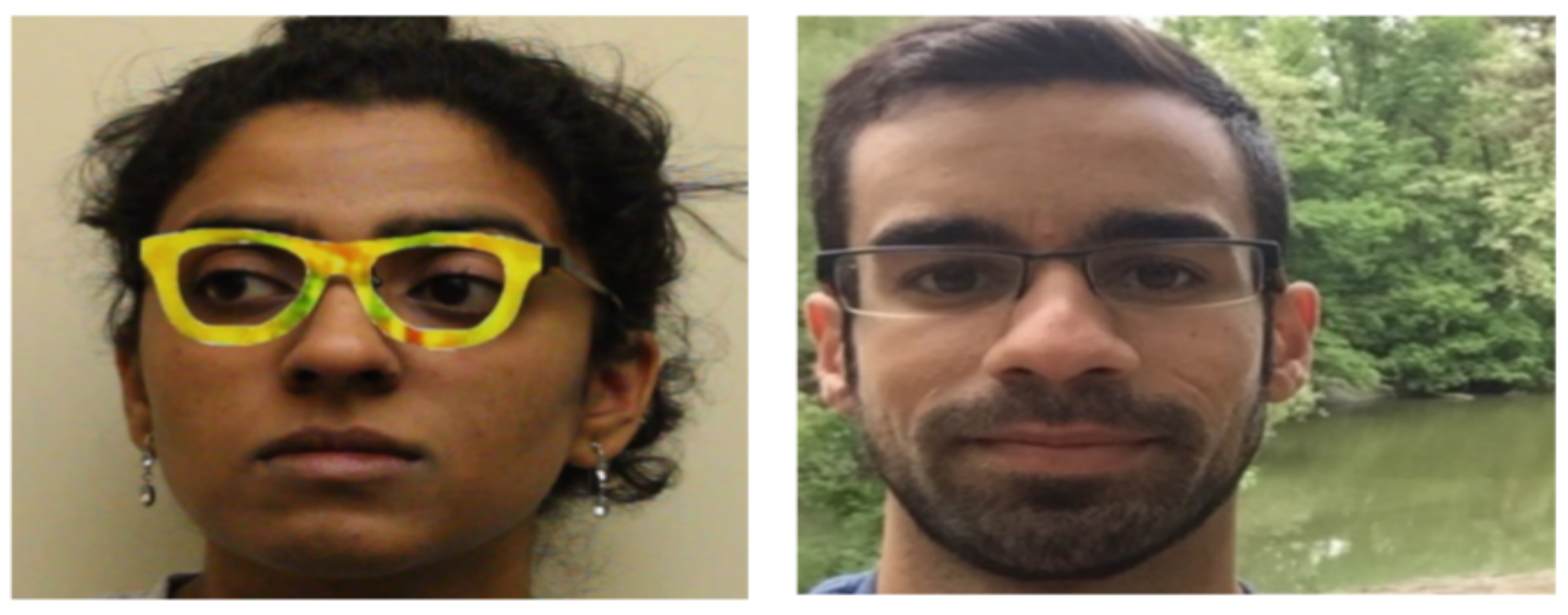


A^1_1	A^1_2	…	A^1_N
A^2_1	A^2_2	…	A^2_N

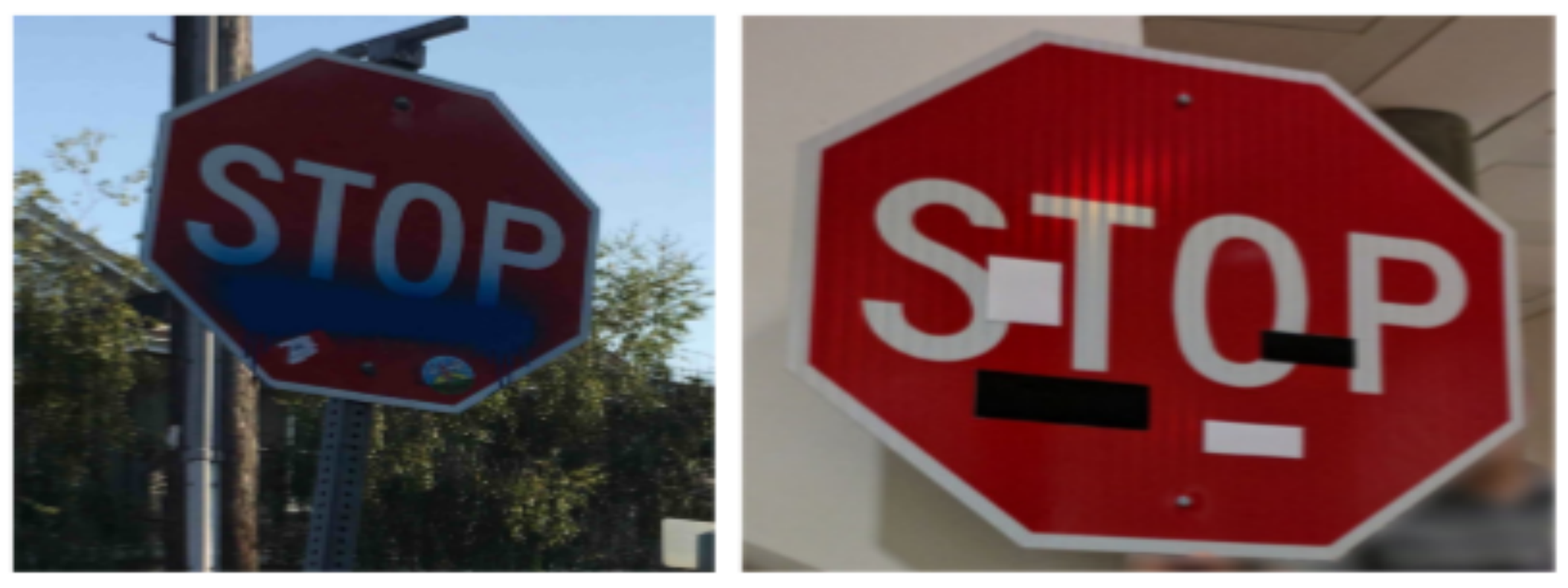
Sensitivity =

- Small scale tests have shown that particular features learned by the model are hyper-sensitive to small input changes. This seems to be correlated with detection stability.

4. Security Implications: Adversarial Examples are Created to Attack Models

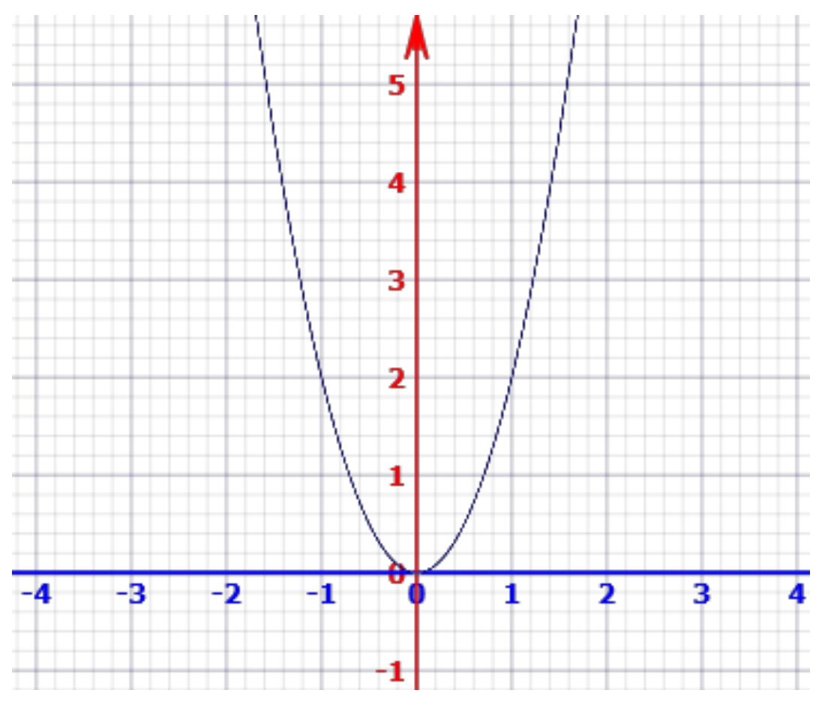


- An adversarial attack (glasses) causes the model to misclassify the lady on the left as the man on the right. (Sharif et al, 2016)



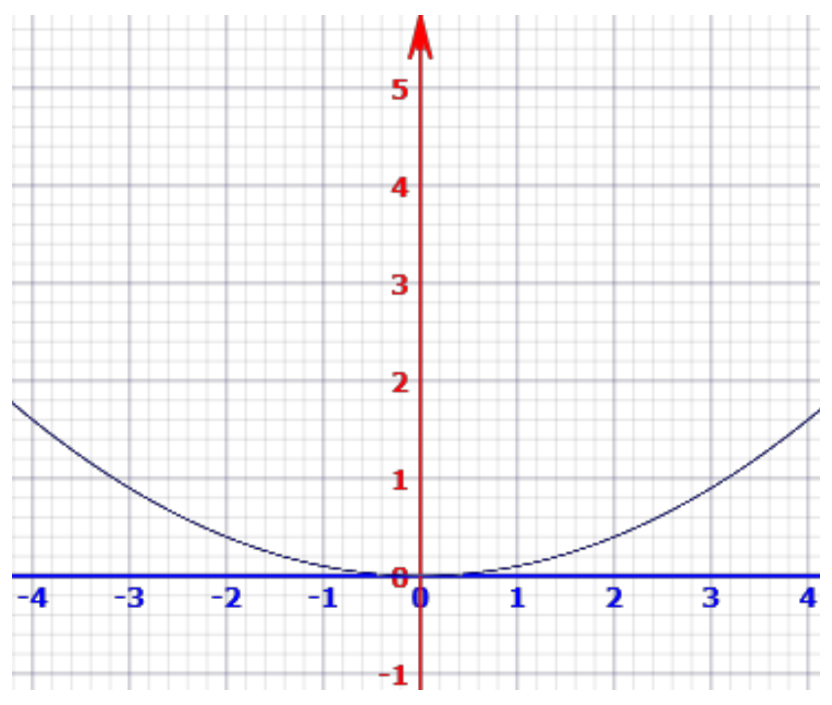
- The stop sign on the left is correctly detected. The stop sign on the right is incorrectly detected as a 45 mph speed limit sign. (Eykholt et al, 2018)

5. Solution: How to Improve Networks to be More Robust



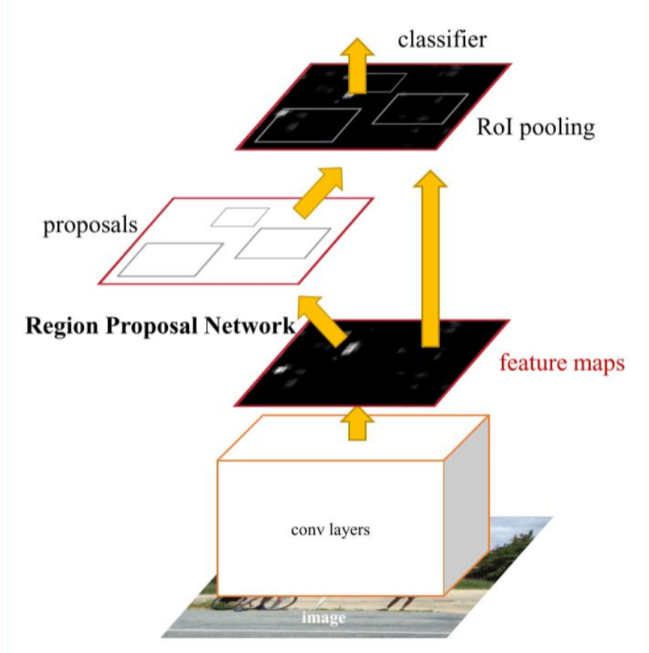
- Sensitive Function

- Early results show certain features our model learned are sensitive (left)
- We can learn smooth feature representations of the input space (right) with gradient regularization



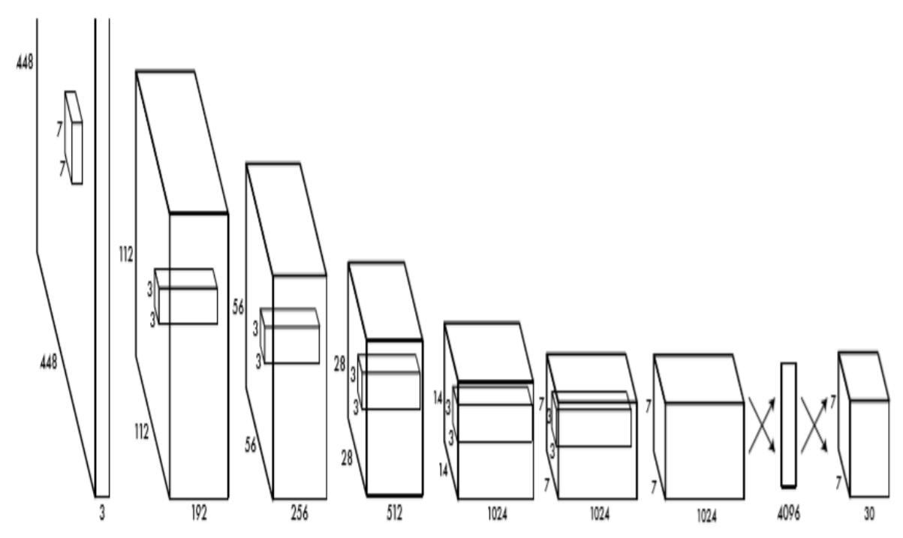
- Similar but less sensitive function

6. Looking Ahead: Research Questions We Must Address

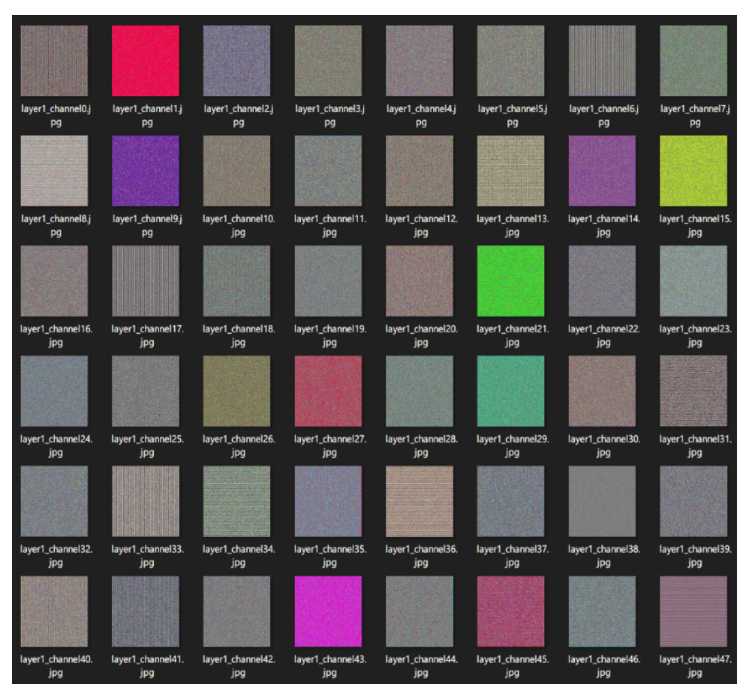


*Faster-RCNN

- 1. Are two-stage detectors (left) just as inconsistent as one-stage detectors (right)? Why?



*YOLOv3



- 2. What types of learned features (left) are most sensitive? Color, Texture, Shapes?

PURDUE UNIVERSITY Discovery Park
CERIAS