Adversarial camera stickers:
A physical camera-based attack on deep learning systems

Juncheng B. Li, Frank R. Schmidt, J. Zico Kolter

Bosch Center for Artificial Intelligence

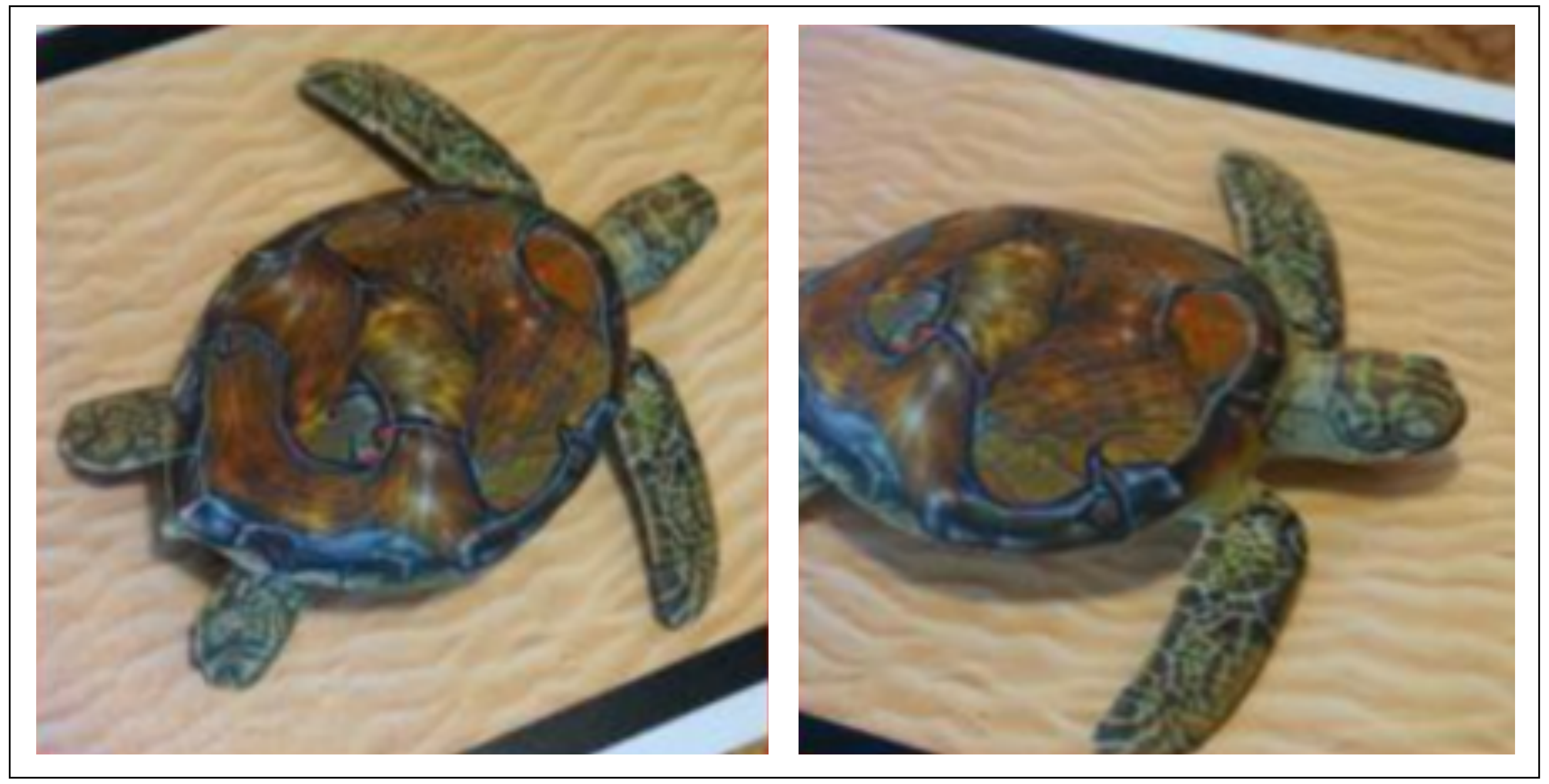
BOSCH

Adversarial attacks: not just a digital problem

All existing physical attacks modify the ***object***.

Sharif et al., 2016

Etimov et al., 2017



Athalye et al., 2017

QUESTION

All existing physical attacks modify the ***object***, but is it possible instead to fool deep classifiers by modifying the ***camera***?

This paper: A physical adversarial camera attack

- We show it is indeed possible to create visually inconspicuous modifications to a camera that fool deep classifiers

- Uses a small specially-crafted translucent sticker, placed upon camera lens

- The adversarial attack is ***universal***, meaning that a ***single*** perturbation can fool the classifier for a given object class over multiple viewpoints and scales

Juncheng B. Li, Frank R. Schmidt, J. Zico Kolter | Bosch Center for Artificial Intelligence
BOSCH

The challenge of physical sticker attacks

The challenge

- (Inconspicuous) physical stickers are ***extremely limited*** in their resolution (can only create blurry dots over images)
- Need to both learn a model of allowable perturbations ***and*** create the adversarial image

Our solution

- A differentiable model of sticker perturbations, based upon alpha blending of blurred image overlays
- Use gradient descent to both fit the perturbation model to observed data, and construct an adversarial attack

Methodology

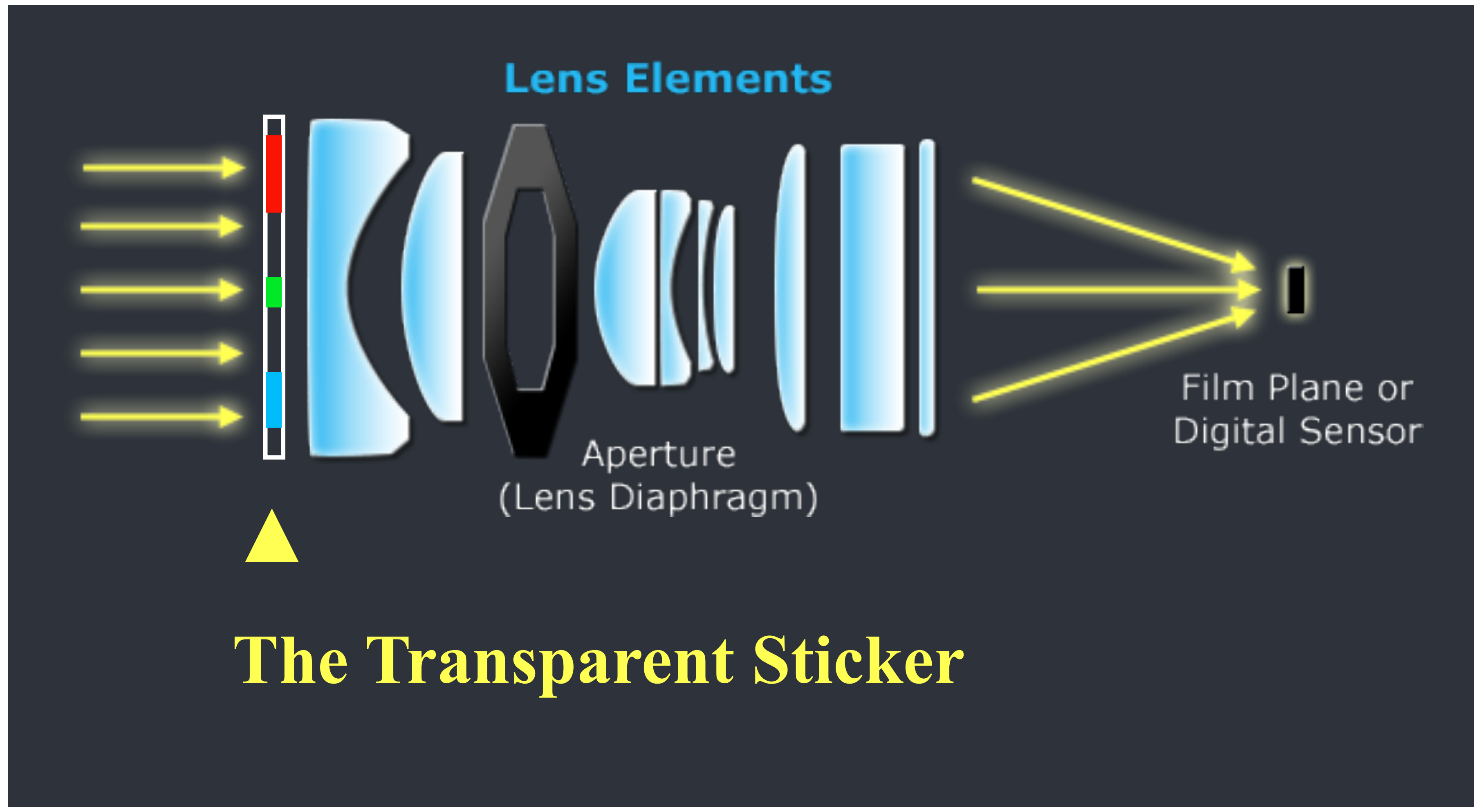
- Attack model consists of smoothed alpha blend between observed image and some fixed color (iterated to produce multiple dots)

$$I^{out}_{x,y} = (1 - \alpha_{x,y}) \cdot I^{in}_{x,y} + \alpha_{x,y} \cdot c, \qquad \alpha_{x,y} = \exp\left(-\frac{(x - x_c)^2 + (y - y_c)^2}{2\sigma^2}\right)$$

- Parameters of attack include color c, dot position (x_c, y_c) and bandwidth σ

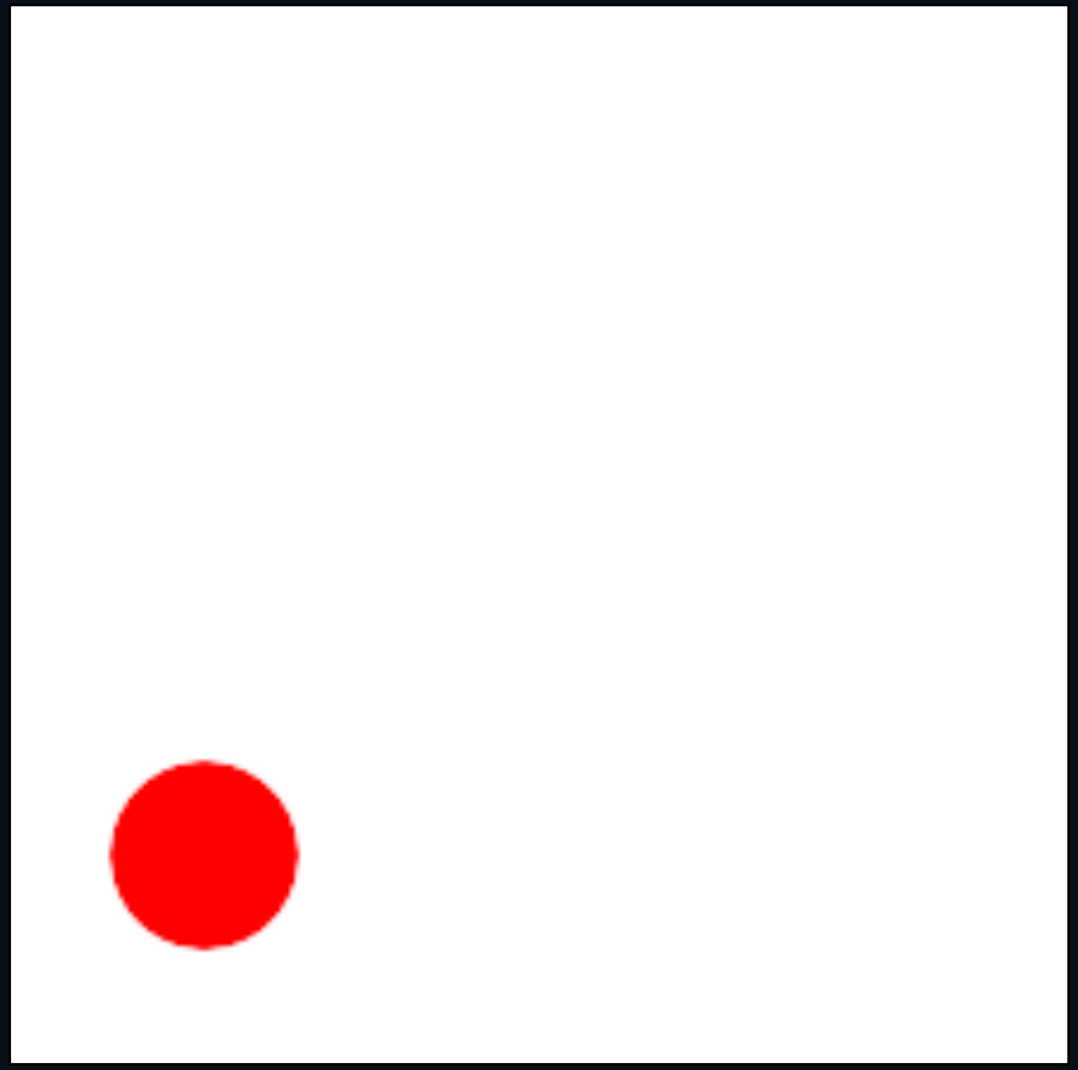
- **Key idea:** use gradient descent over some parameters (e.g., color, bandwidth) to fit model to observed physical images, over other parameters (e.g. location) to maximize loss

$$\min_{c,\sigma} \text{SSIM}(I^{out}, I^{real}), \qquad \max_{x_c, y_c} \sum_{i=1}^{m} \text{Loss}\left(I^{out}, y_{true}, y_{target}\right)$$



How does a dot look like through camera lense?

Clean Camera View



Red Dot

Resulting Blur

Simulated Blur

Results: Virtual Evaluation

Table 1. Performance of our 6-dot attacks on ImageNet test set

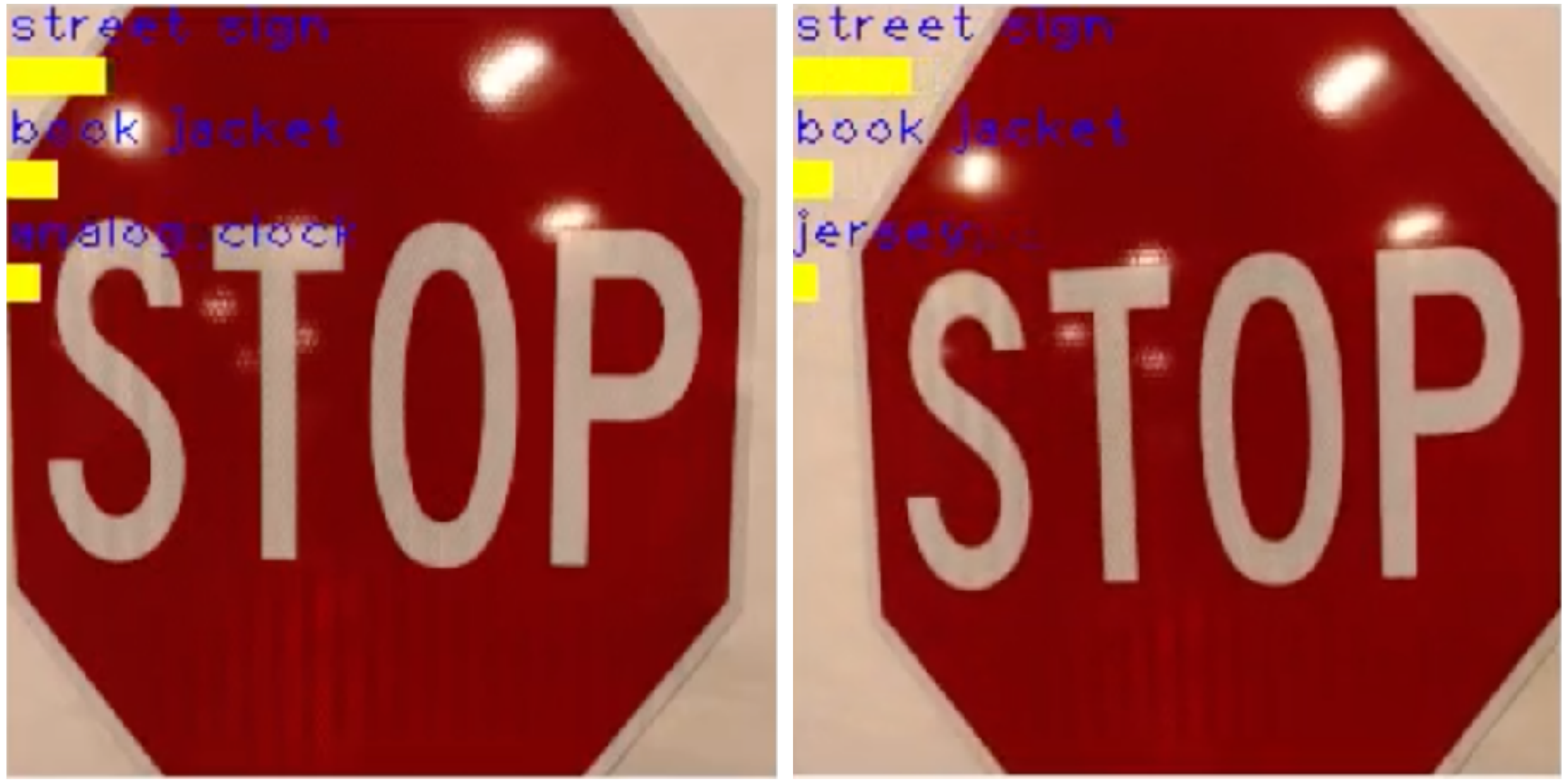
Class	Attack	Prediction: Correct	Target	Other
Keyboard	No	85%		15%
▼ Mouse	Yes	48%	36%	16%
Street sign	No	64%		36%
▼ Guitar Pick	Yes	32%	34%	34%
Street sign	No	64%		36%
▼ 50 random classes	Yes	18%	33%	49%
50 random classes	No	74%		26%
▼ 50 random classes	Yes	42%	31%	27%

Juncheng B. Li, Frank R. Schmidt, J. Zico Kolter | Bosch Center for Artificial Intelligence

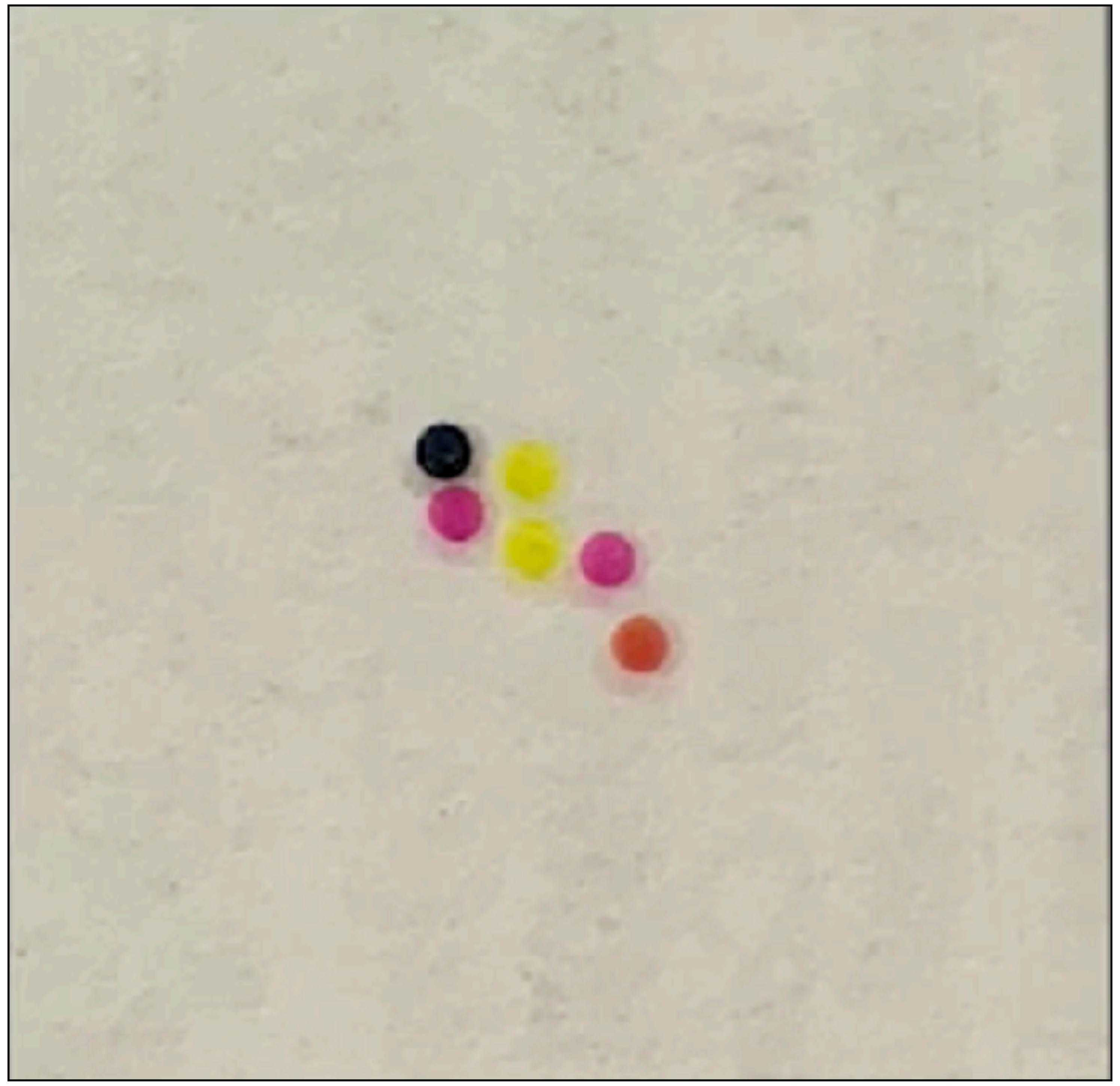

Pro M254dw

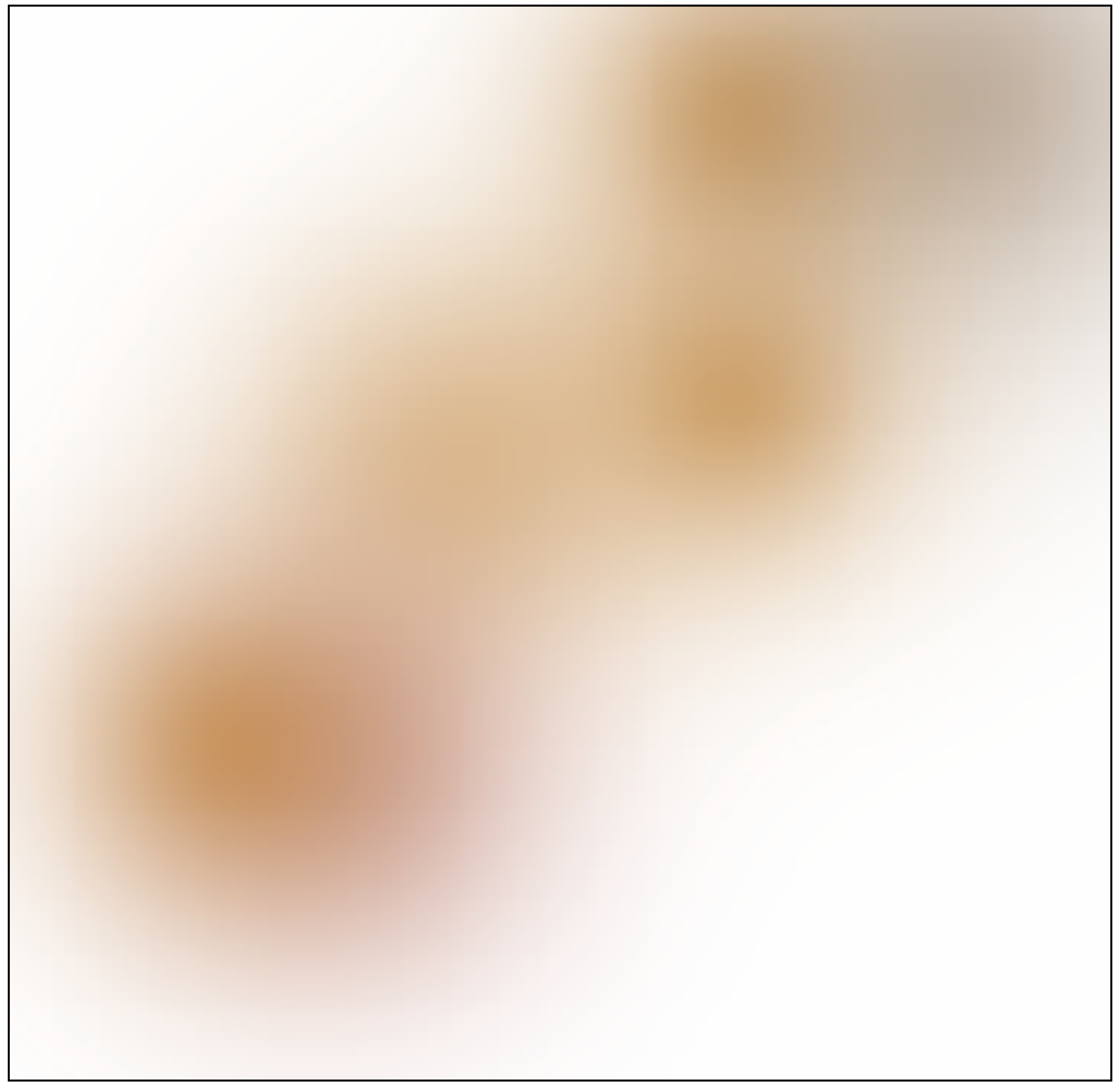
STAINLESS


This is a **ResNet-50** Model implemented with **pyTorch** deployed on a **Logitech C920** Webcam with **clear lense.**



This is a ResNet-50 model implemented with PyTorch deployed on a Logitech C920 WebCam with clear lense. It can recognize street sign at different angles with only minor errors.



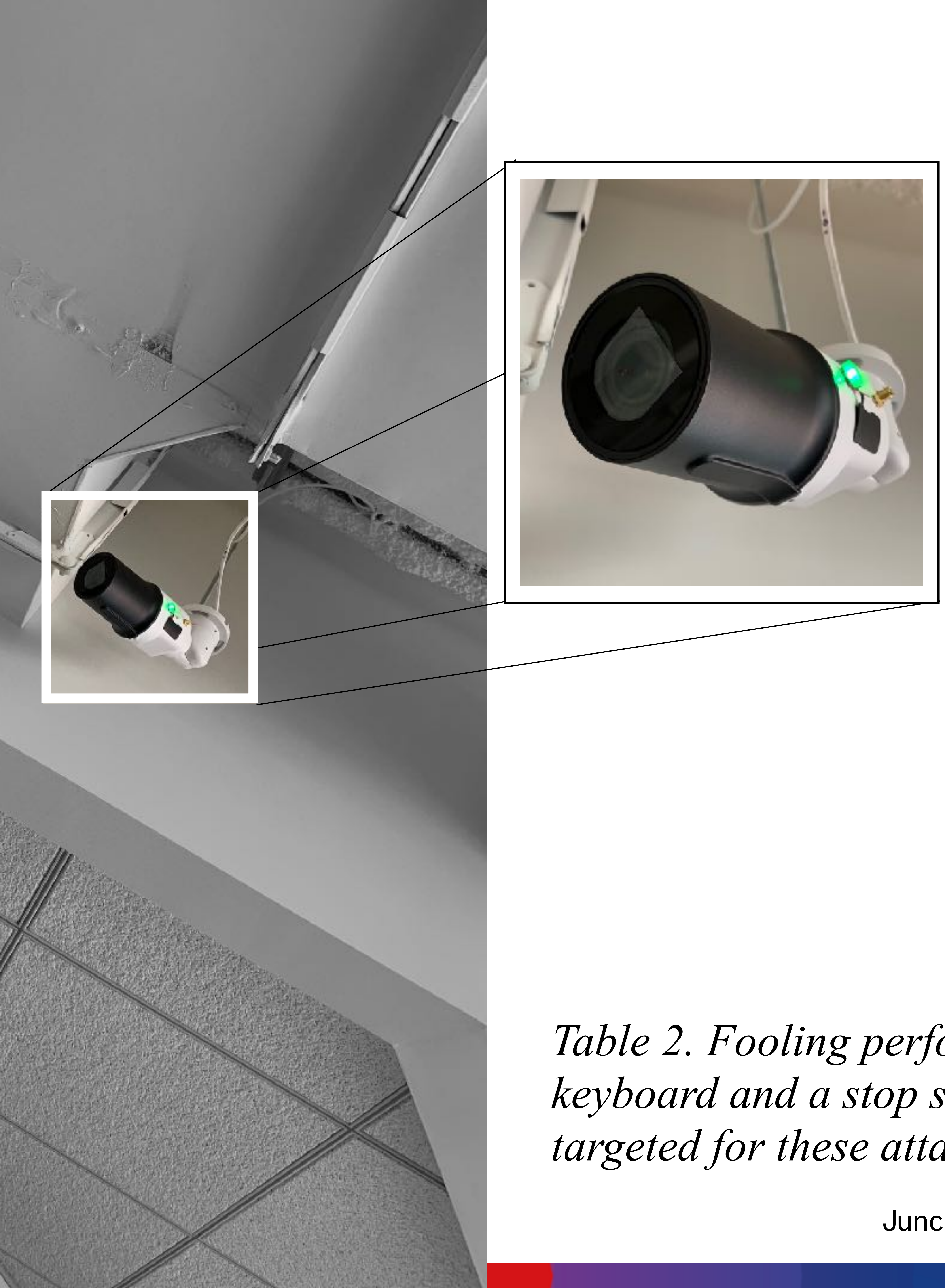


Now we cover the camera with our adversarial sticker made by our proposed method to achieve the targeted attack.
This should make a “street sign” misclassified as a “guitar pick”.

Juncheng B. Li, Frank R. Schmidt, J. Zico Kolter | Bosch Center for Artificial Intelligence

The Sticker results in very inconspicuous blurs in the view. We can achieve targeted attack most of the time at different angles and with different distances.

Results: Real World Evaluation



Original Class	Target class	Prediction		
		Correct	Target	Other
Keyboard	Mouse	271	**548**	181
	Space bar	320	**522**	158
Street sign	Guitar Pick	194	**605**	201
	Envelope	222	**525**	253
Coffee mug	Candle	330	**427**	243

Table 2. Fooling performance of the our method on two 1000 frame videos of a computer keyboard and a stop sign, viewed through a camera with an adversarial sticker placed on it targeted for these attacks.

Summary

- Adversarial attacks don't need to modify every object in the world to fool a deployed deep classifier, they just need to modify the camera
- Implications in self-driving cars, security systems, many other domains

To find out more, come see our poster

at Pacific Ballroom #65

on Tuesday, Jun 11th 06:30-09:00