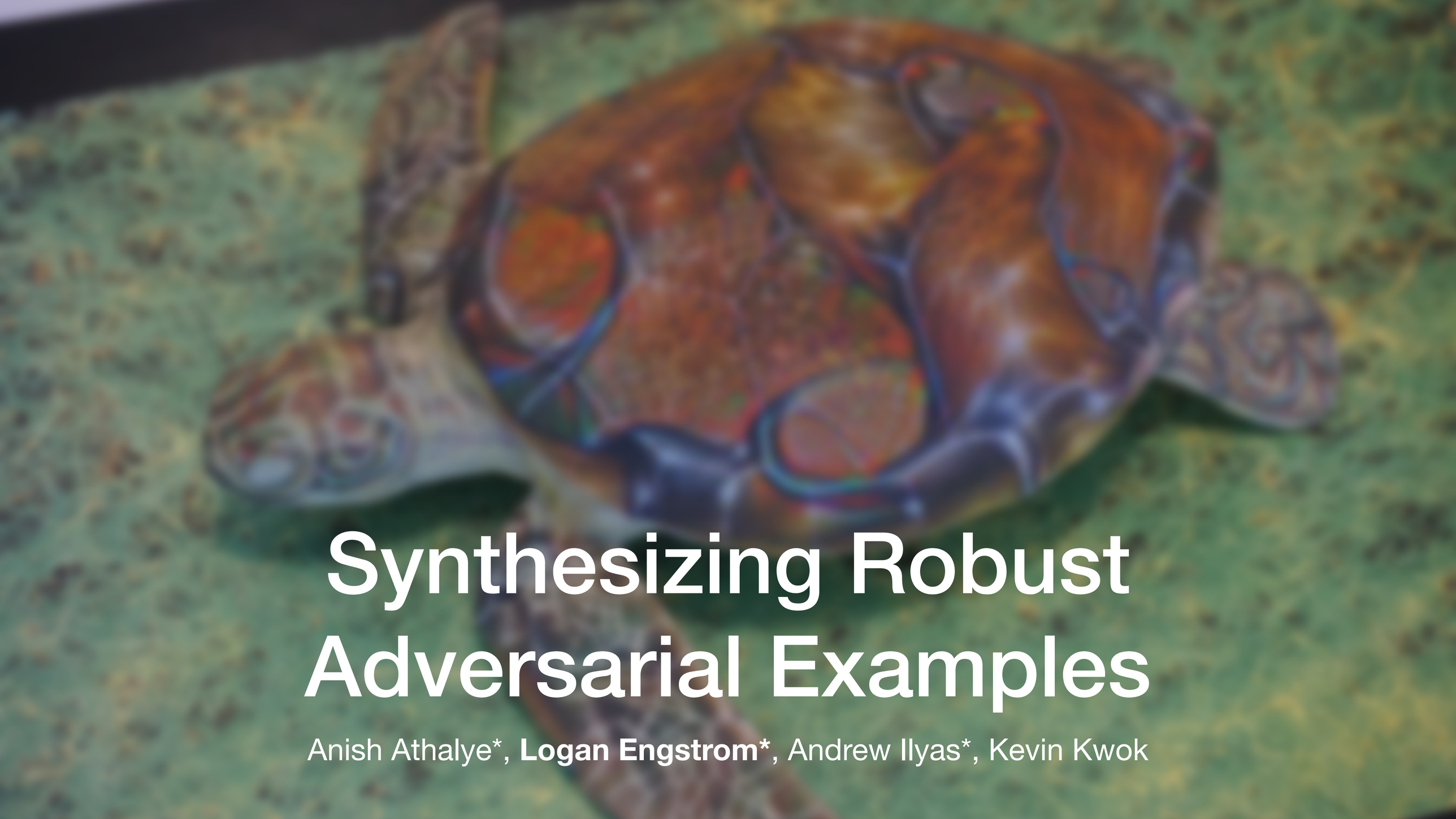


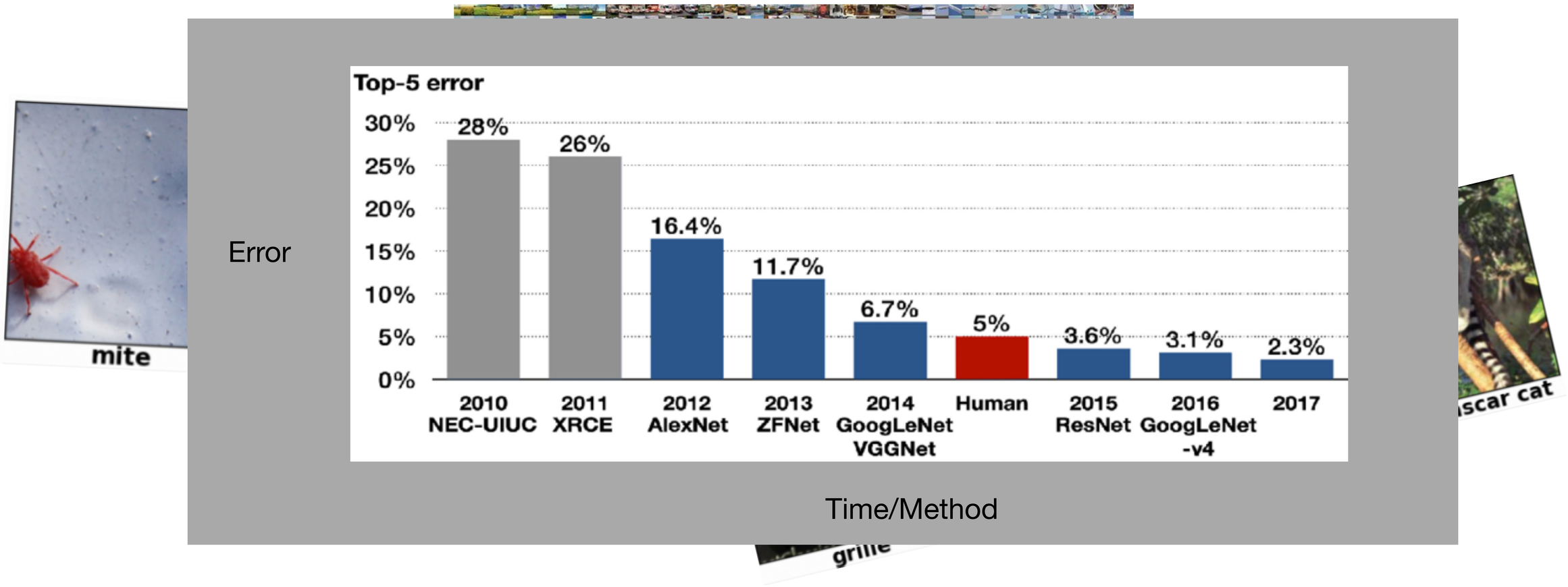
Synthesizing Robust Adversarial Examples

Anish Athalye*, **Logan Engstrom***, Andrew Ilyas*, Kevin Kwok

Travel back in time to 2017!

2017: Unbelievable progress in machine learning!

ImageNet progress!



2017: Unbelievable progress in machine learning!

Self driving cars!

Robotaxi Revolutions Don't Come Cheap, So Zoox Boosts Funding By $500 Million

Alan Ohnsman Forbes Staff
I follow technology-driven changes that are reshaping transportation.

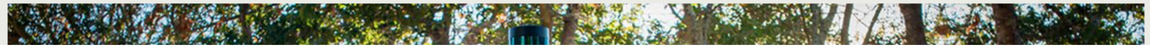
Follow

NEWS | TRANSPORTATION

Nuro Raises $92 Million for Adorable Autonomous Delivery Vehicles

Somewhere between a delivery truck and a sidewalk robot, Nuro's robotic vehicles want to deliver your groceries

BY EVAN ACKERMAN | 30 JAN 2018 | 6 MIN READ |



The New York Times

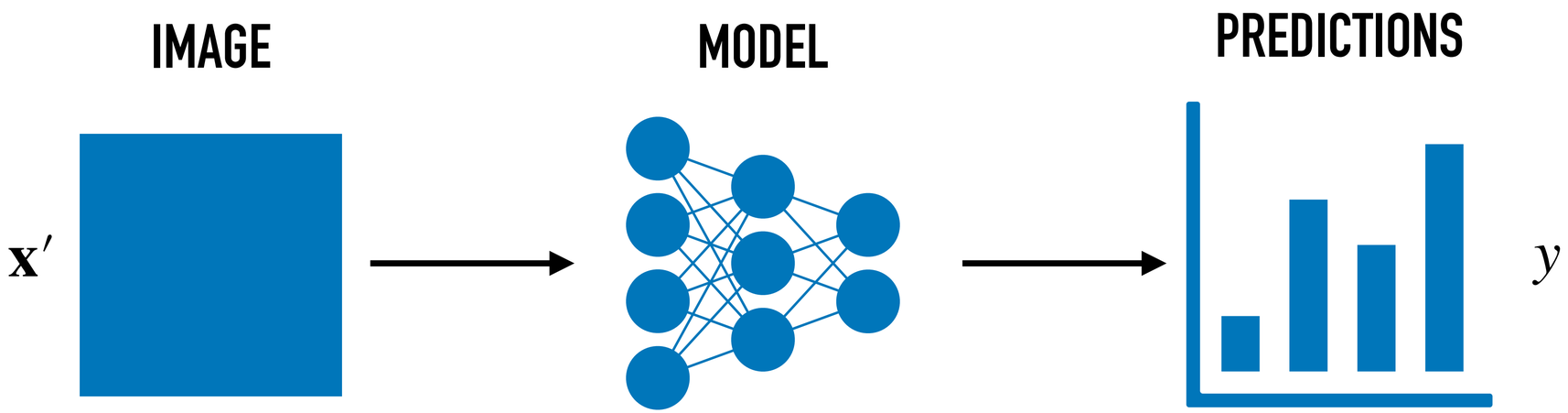
General Motors to Buy Cruise Automation in Push for Self-Driving Cars

Does ML really work that well yet?

Adversarial examples

Adversarial examples

- Suppose we have a ML model mapping inputs -> probabilities

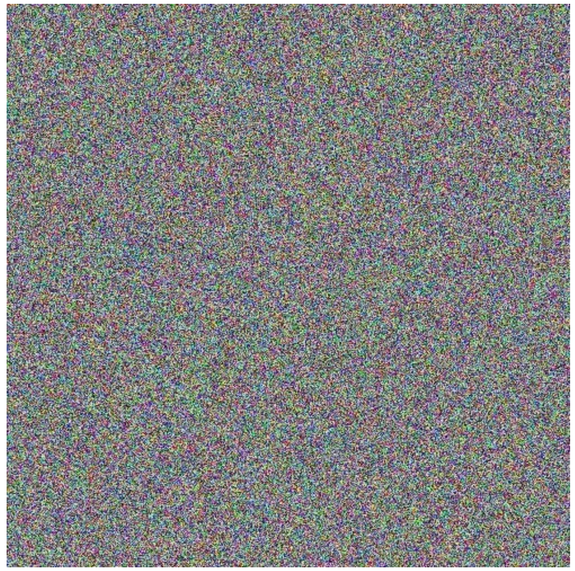


Adversarial examples

- Suppose we have a ML model mapping inputs -> probabilities
- Imperceptible perturbations to an input can change our neural network's prediction

88% tabby cat

+



adversarial perturbation

x 0.00001

99% guacamole

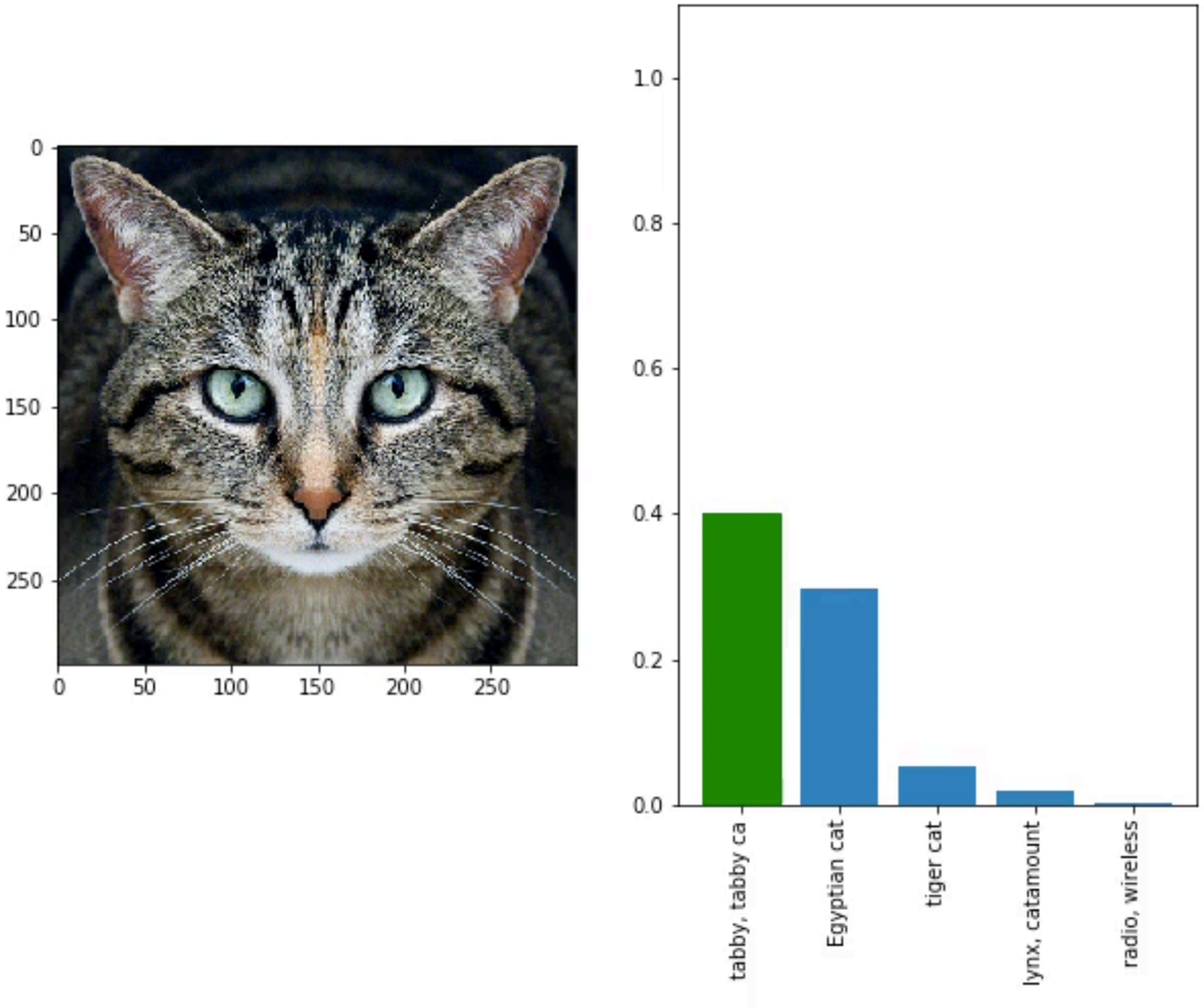
Adversarial examples

Given: Input image *x*, target label *y*

Optimize:

$$\begin{aligned} \arg\max_{\mathbf{x}'} \quad & P\left(y \mid \mathbf{x}'\right) \\ \text{subject to} \quad & d(\mathbf{x}, \mathbf{x}') < \epsilon \end{aligned}$$

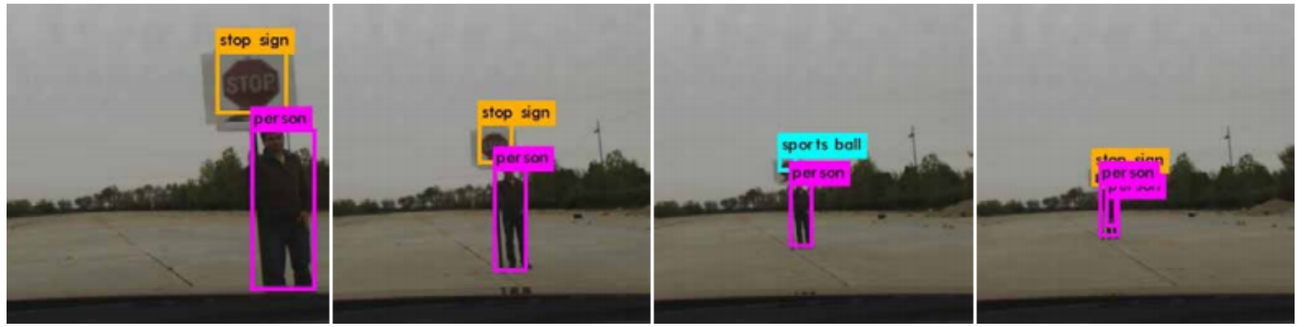
step 00


0
50
100
150
200
250
0
50
100
150
200
250
1.0
0.8
0.6
0.4
0.2
0.0
tabby, tabby ca
Egyptian cat
tiger cat
lynx, catamount
radio, wireless

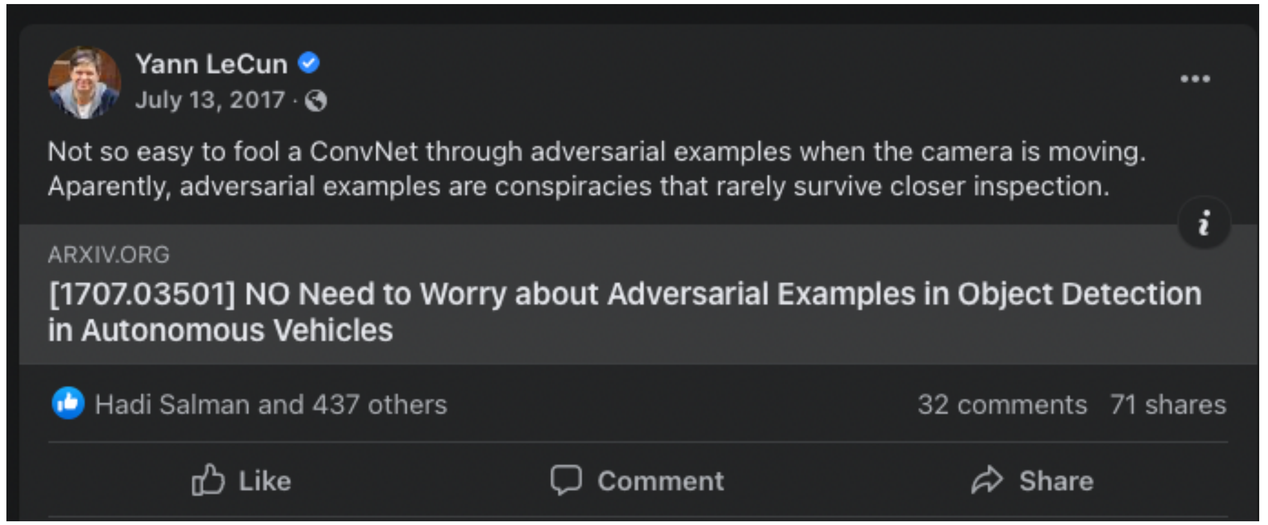
Do adversarial examples work in the physical world?

Maybe not?

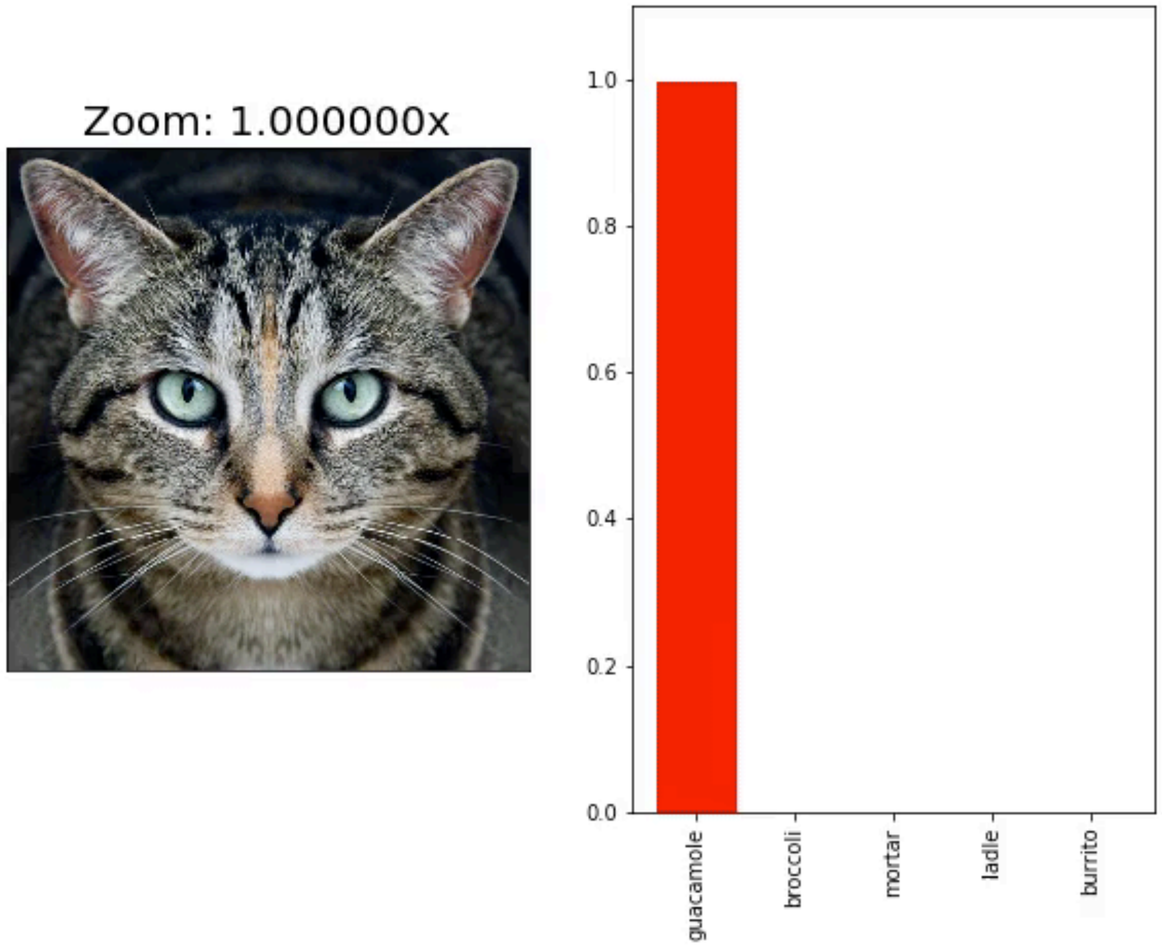
Foveation-based Mechanisms Alleviate Adversarial Examples (Luo et al. 2015)



NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles (Lu et al. 2017)

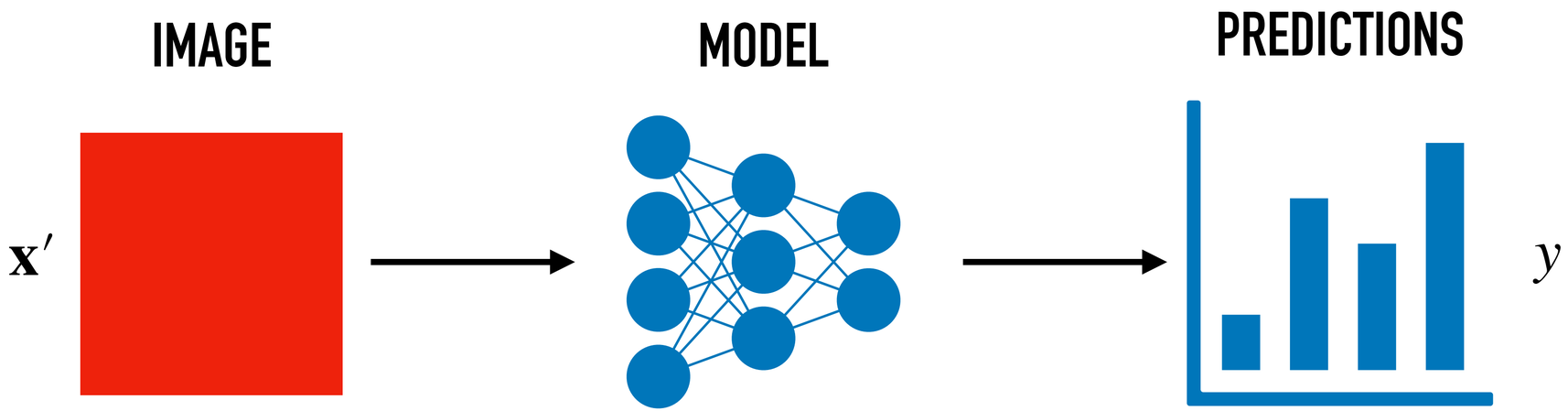


Standard examples are fragile



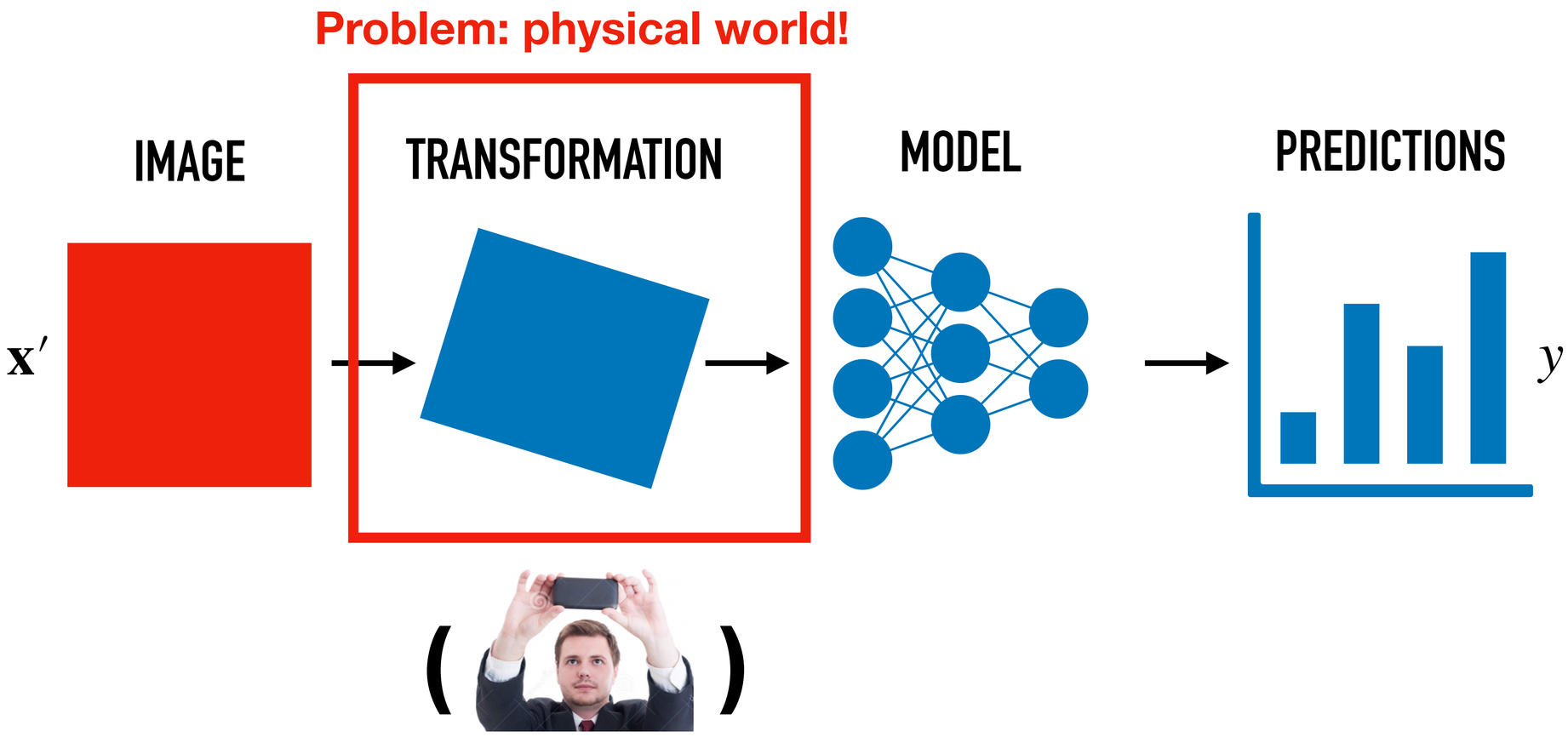
Are adversarial examples fundamentally fragile?

Standard adversarial examples



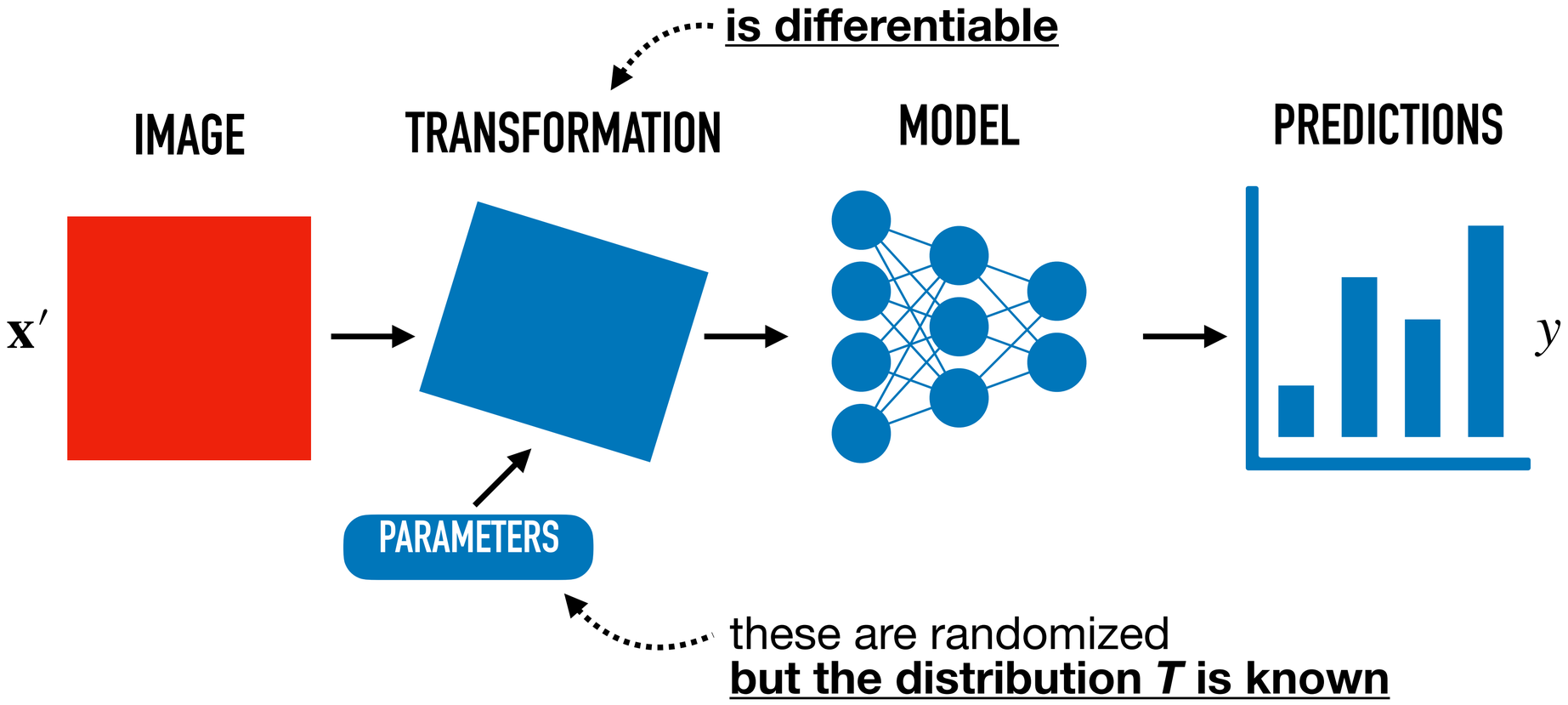
optimize $P(y \mid \mathbf{x}')$ using gradient descent

Physical world adversarial examples



Challenge: No direct control over model input

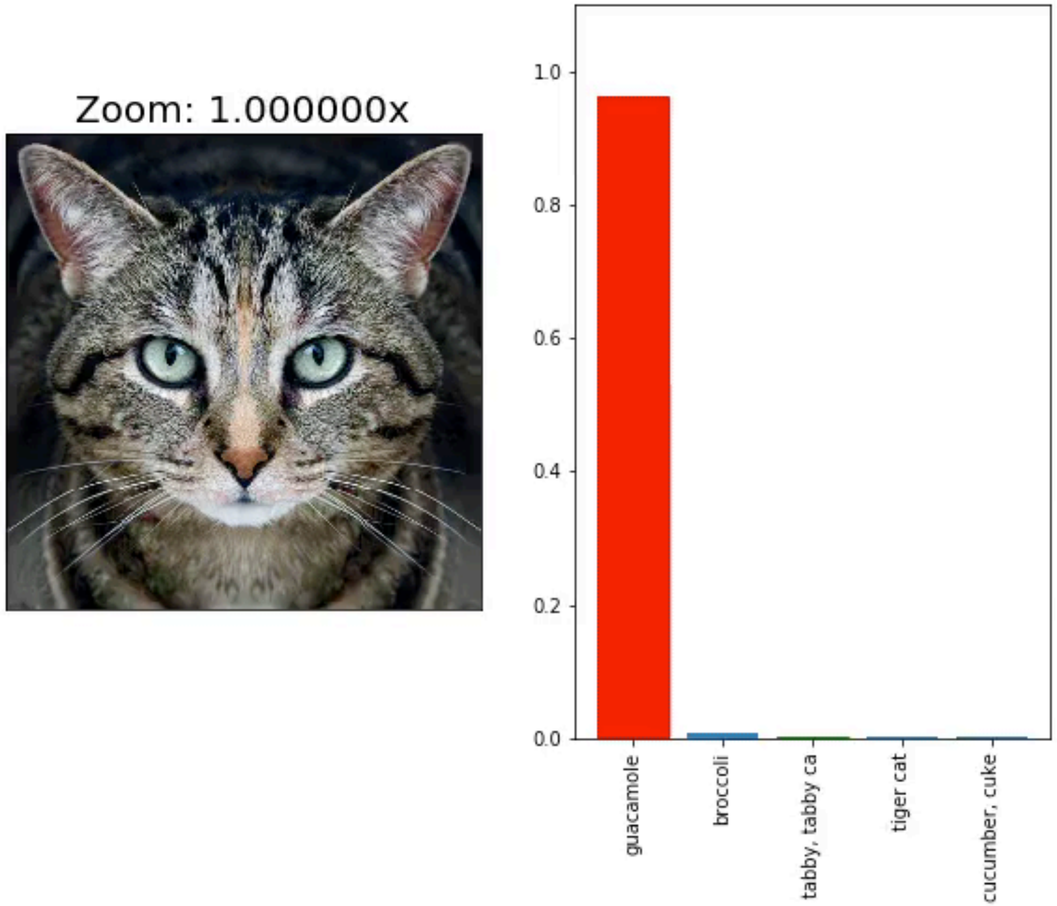
Solution: Expectation Over Transformation Attack



optimize $\mathbb{E}_{t\sim T}\left[P(y \mid t(\mathbf{x}'))\right]$ **using gradient descent**

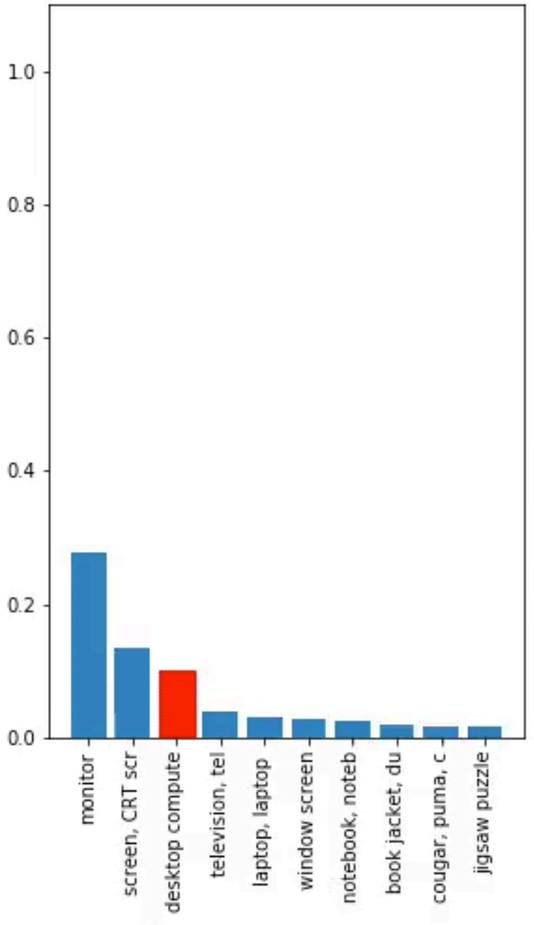
(sampling, chain rule, differentiating through *t*)

Attack produces robust examples



T = {rescale from 1x to 5x}

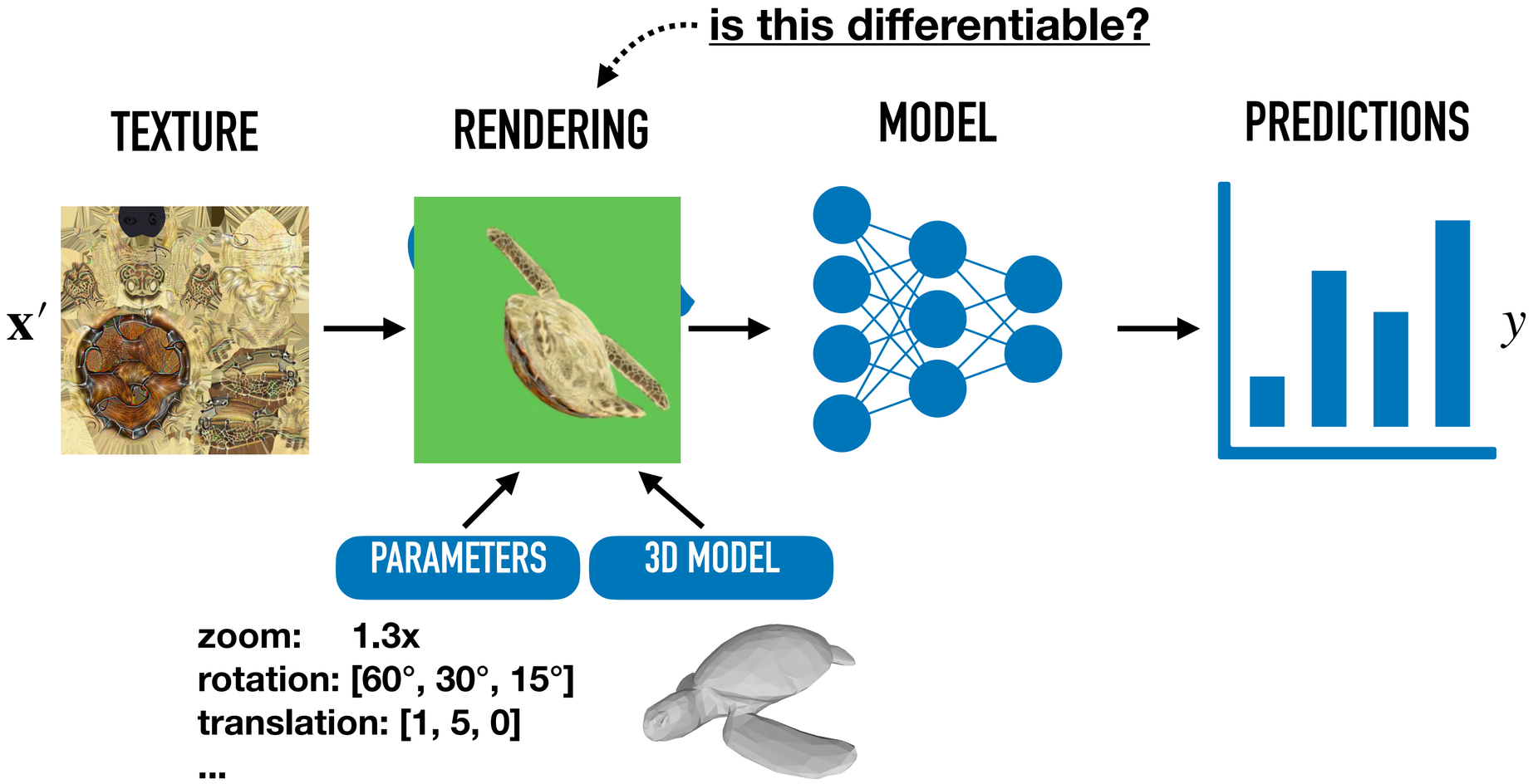
EOT produces robust physical-world examples



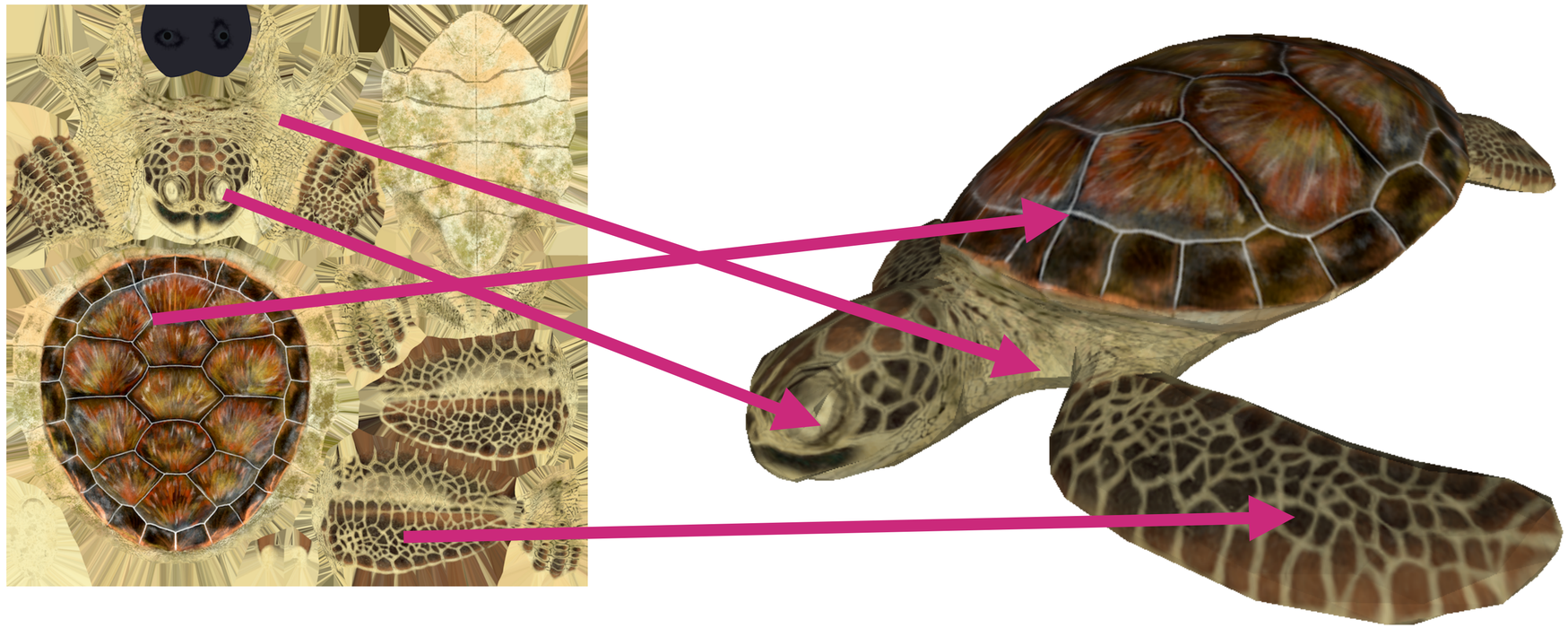
T = {rescale + rotate + translate + skew}

Can we make this work with 3D objects?

Physical world 3D processing pipeline

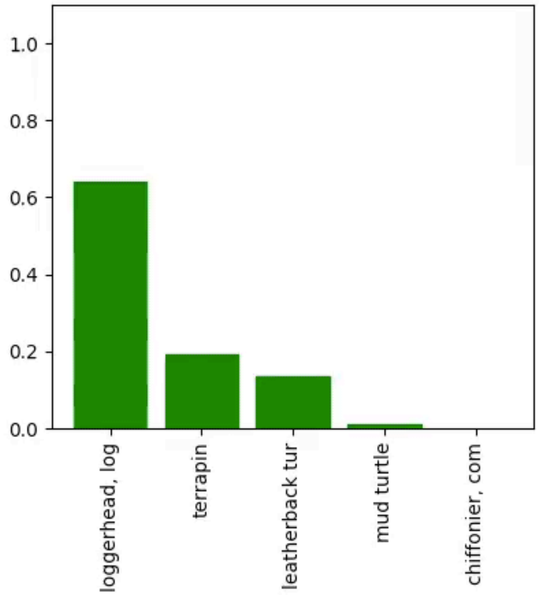


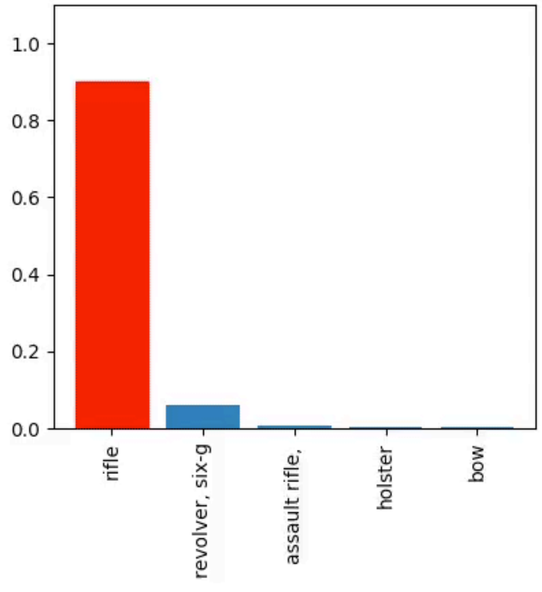
Differentiable rendering



- For any pose, 3D rendering is differentiable with respect to texture
- Simplest renderer: linear transformation of texture

EOT produces 3D adversarial objects






cliff, drop,
megalith, me
agama


rifle
shield, buck
revolver, si

EOT reliably produces 3D adversarial objects

Inputs		Classification accuracy	Attacker success rate	Distortion (l2)
2D	Original	70%	N/A	0
	Adversarial	0.9%	**96.4%**	5.6×10^{-5}
3D	Original	84%	N/A	0
	Adversarial	1.7%	**84.0%**	6.5×10^{-5}

Implications

- Defenses based on randomized input transformations are insecure
- Adversarial examples / objects are a physical-world concern