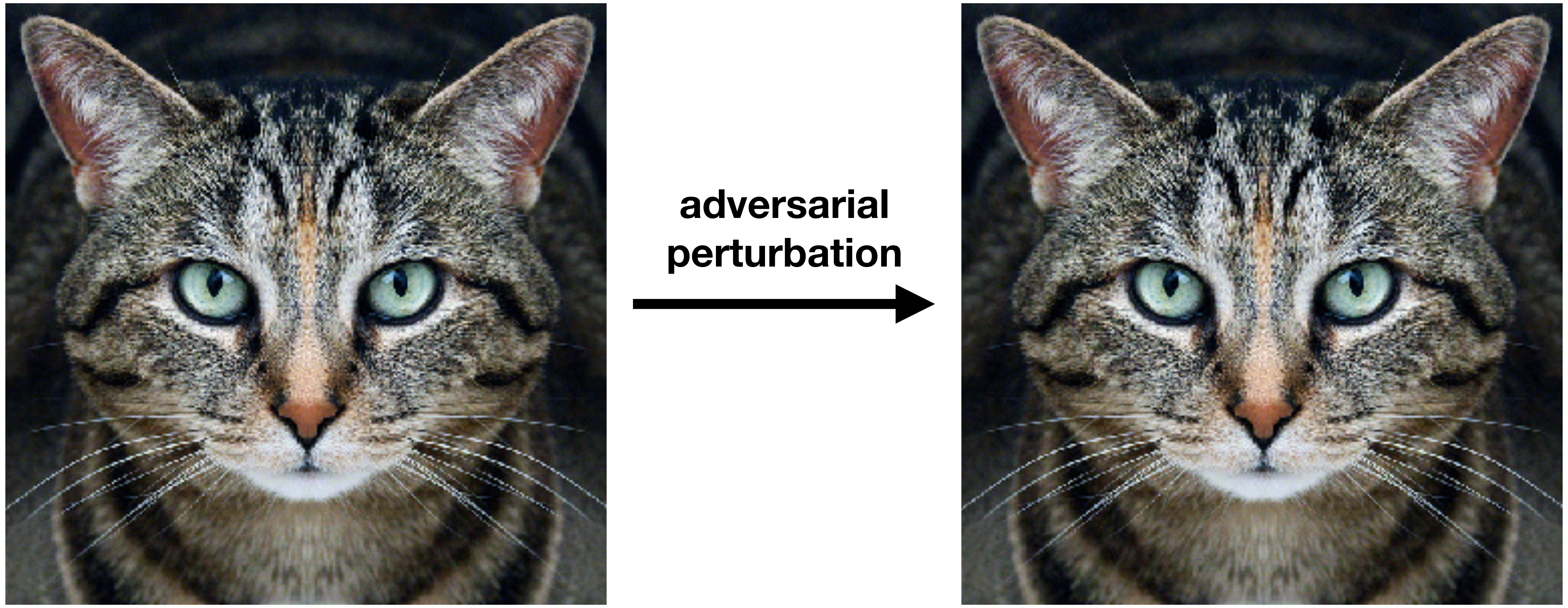
Synthesizing Robust Adversarial Examples

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

Adversarial examples

Adversarial examples

- Imperceptible perturbations to an input can change a neural network's prediction



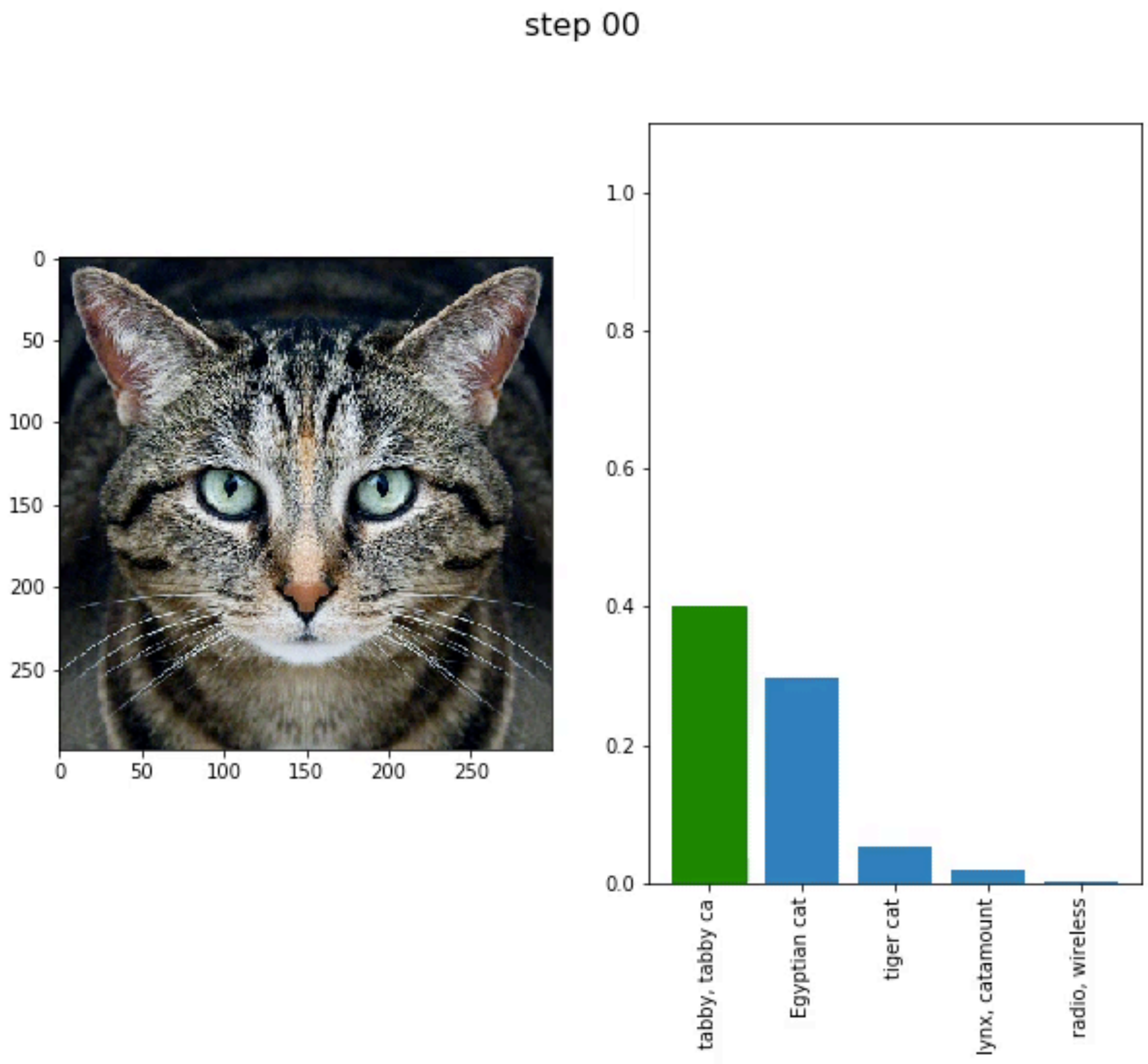
88% tabby cat **99% guacamole**

Adversarial examples

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

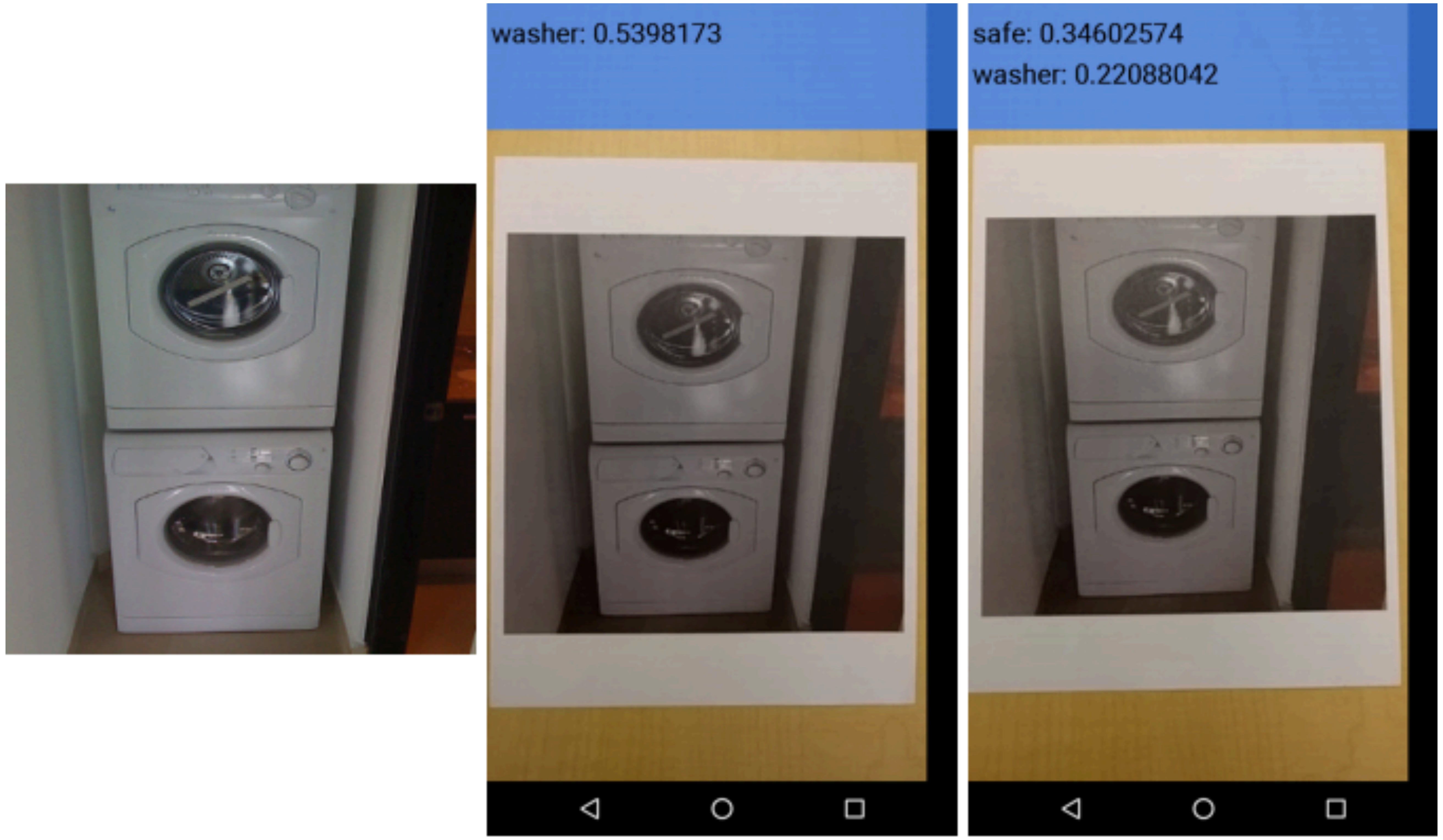
Optimize:

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


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?

Adversarial examples in the physical world



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$

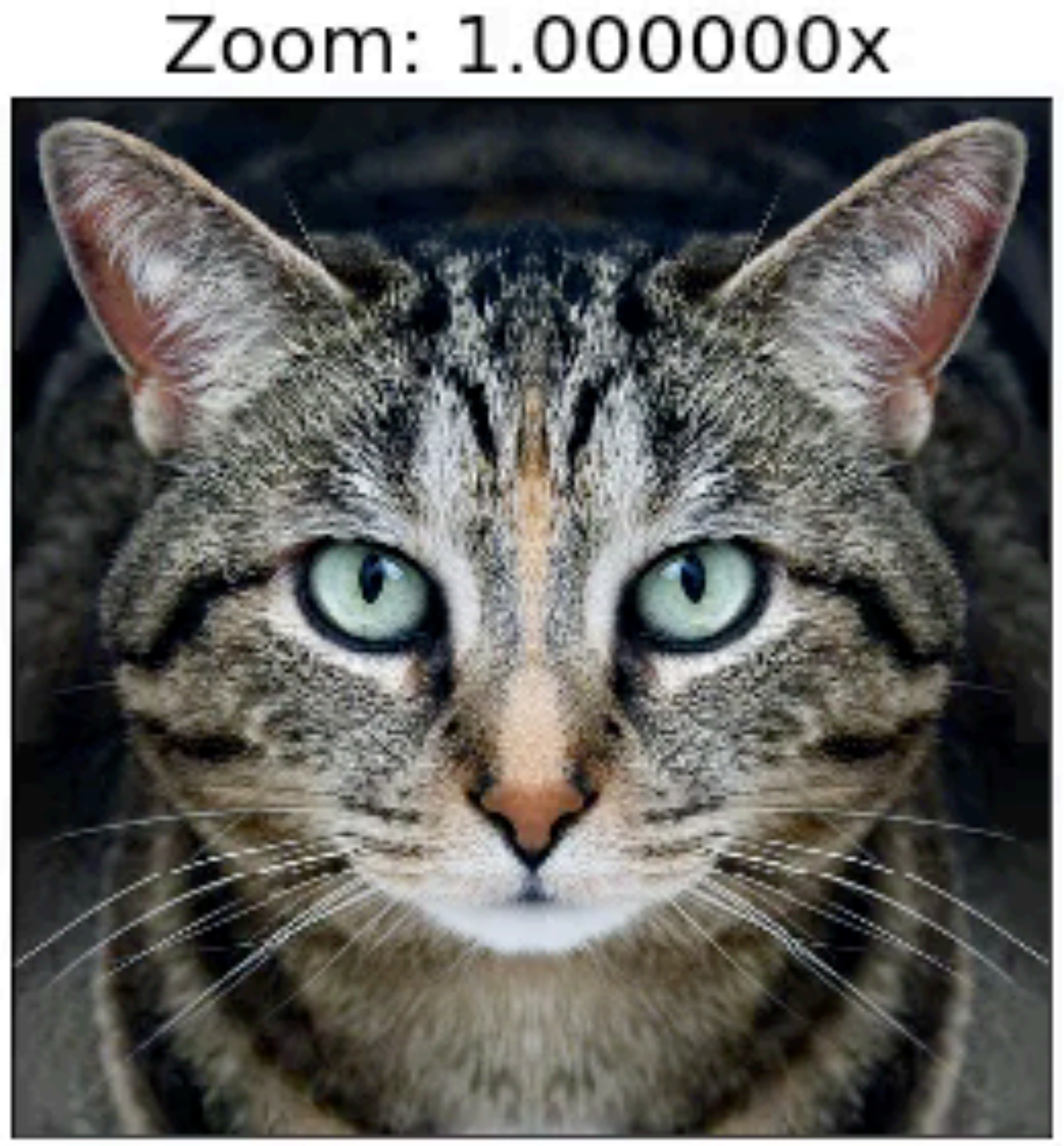
(Kurakin et al. 2016)

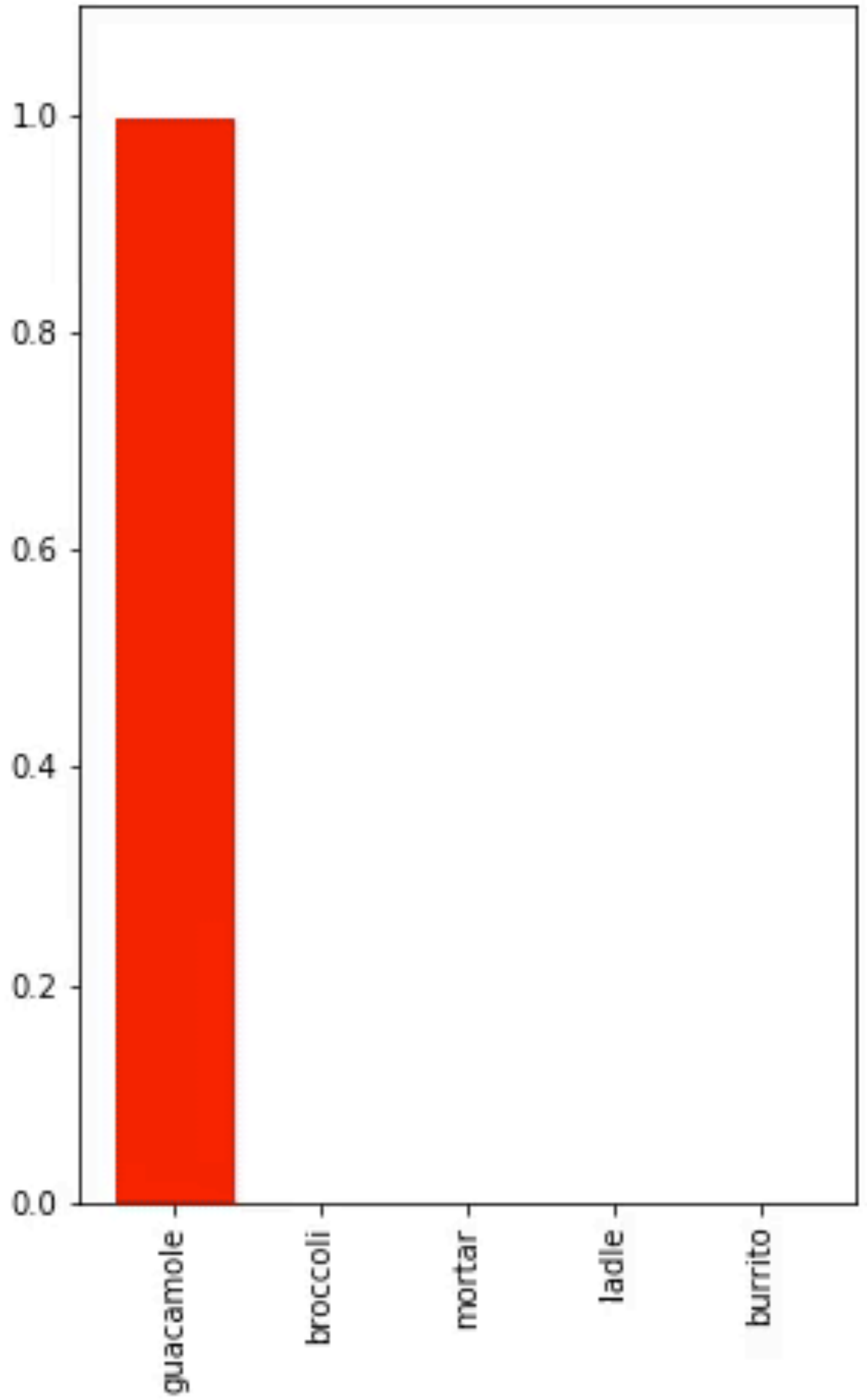
... or not?

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)

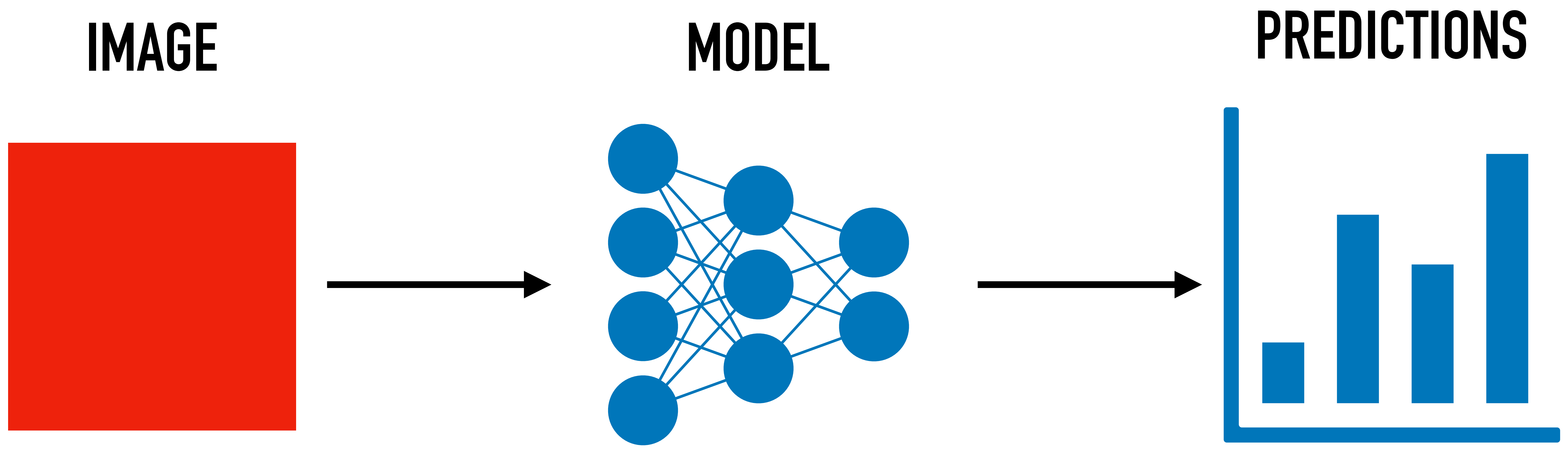
Standard examples are fragile





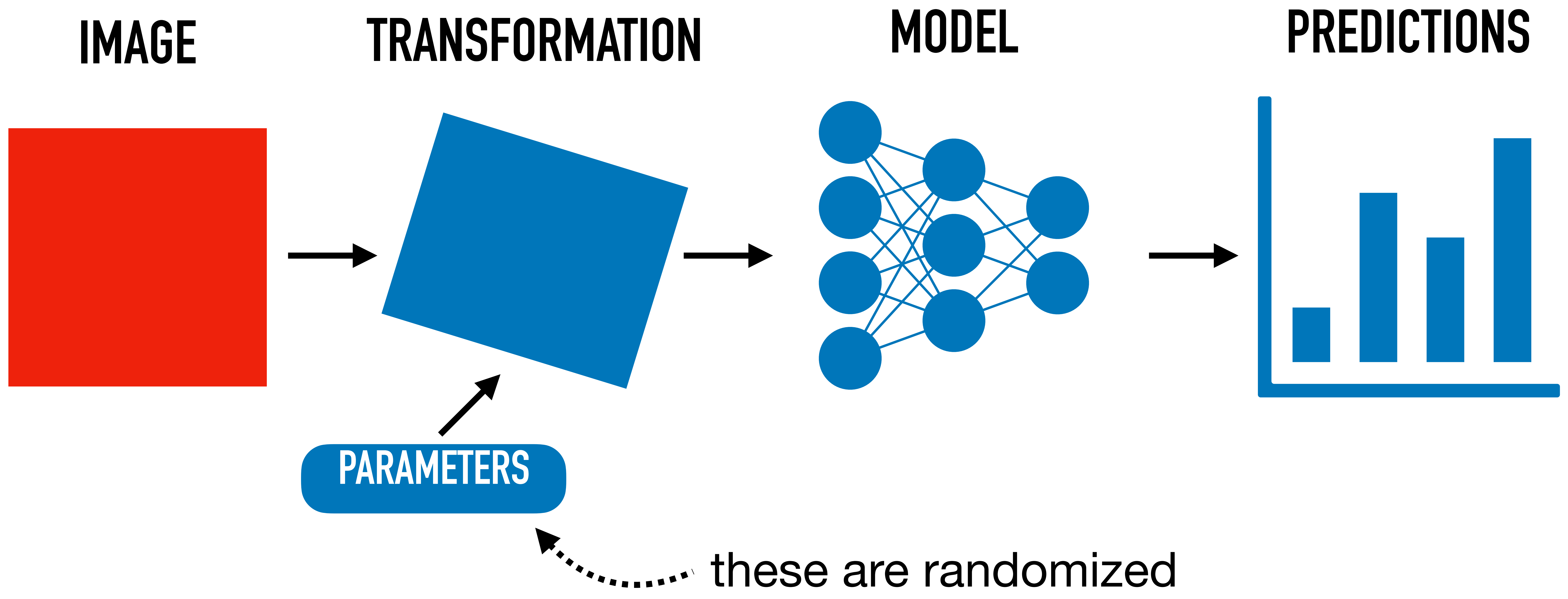
Are adversarial examples fundamentally fragile?

Image processing pipeline



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

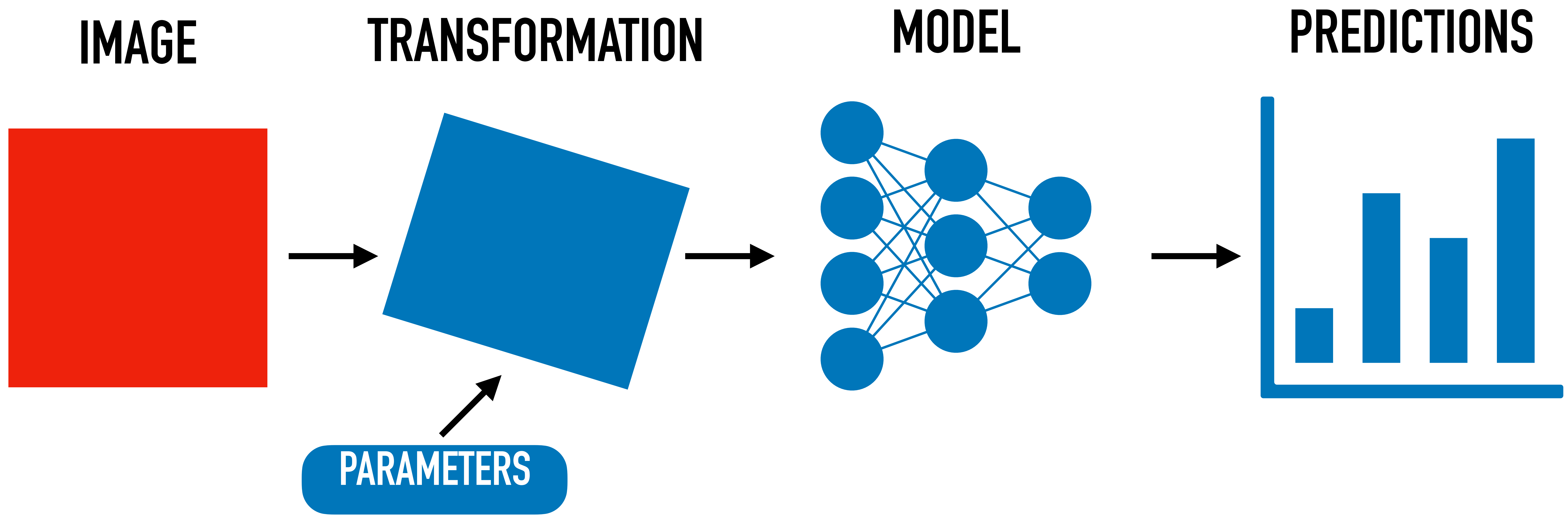
Physical world processing pipeline



Challenge: No direct control over model input

Attack: Expectation Over Transformation

is differentiable



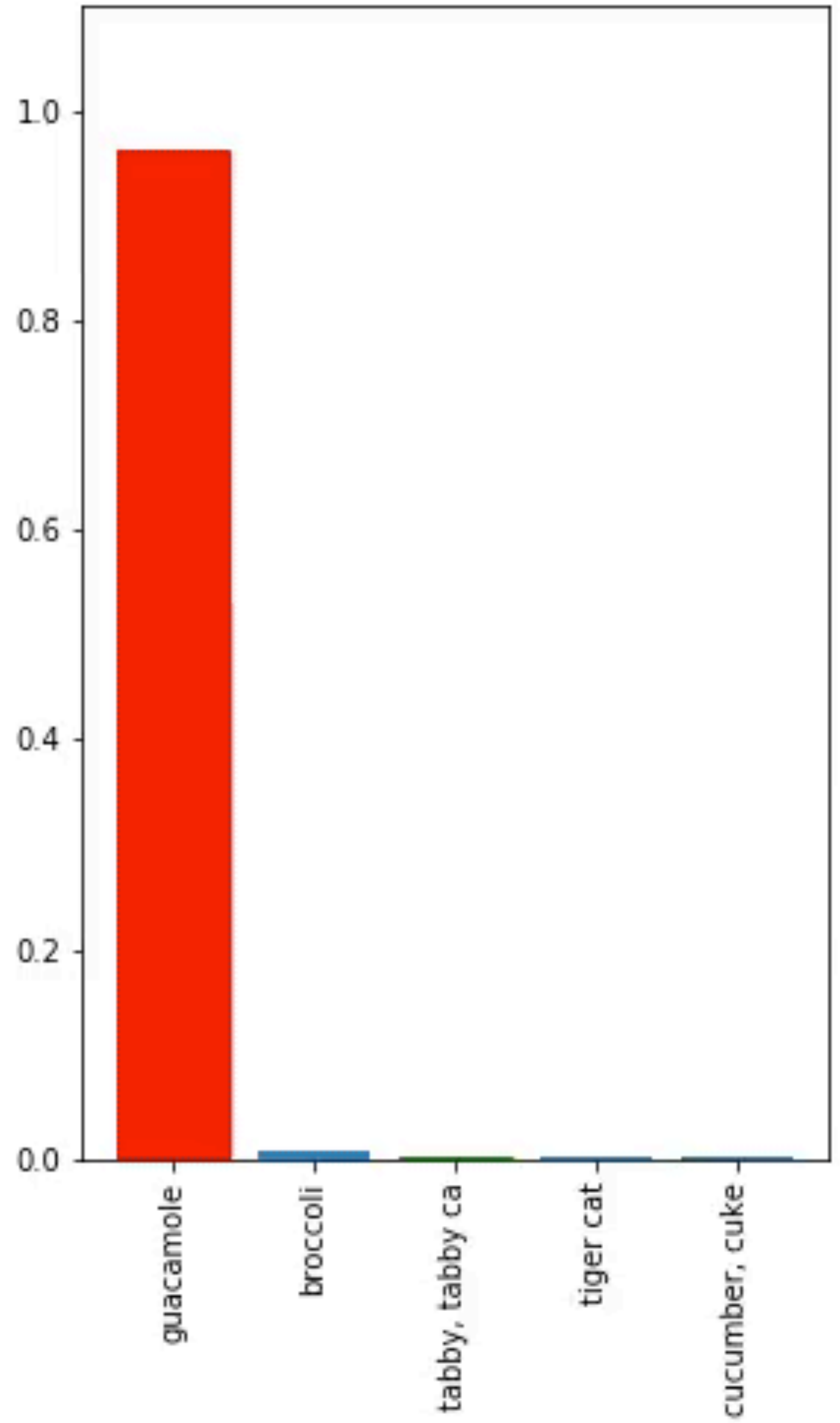
these are randomized
but the distribution *T* is known

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

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

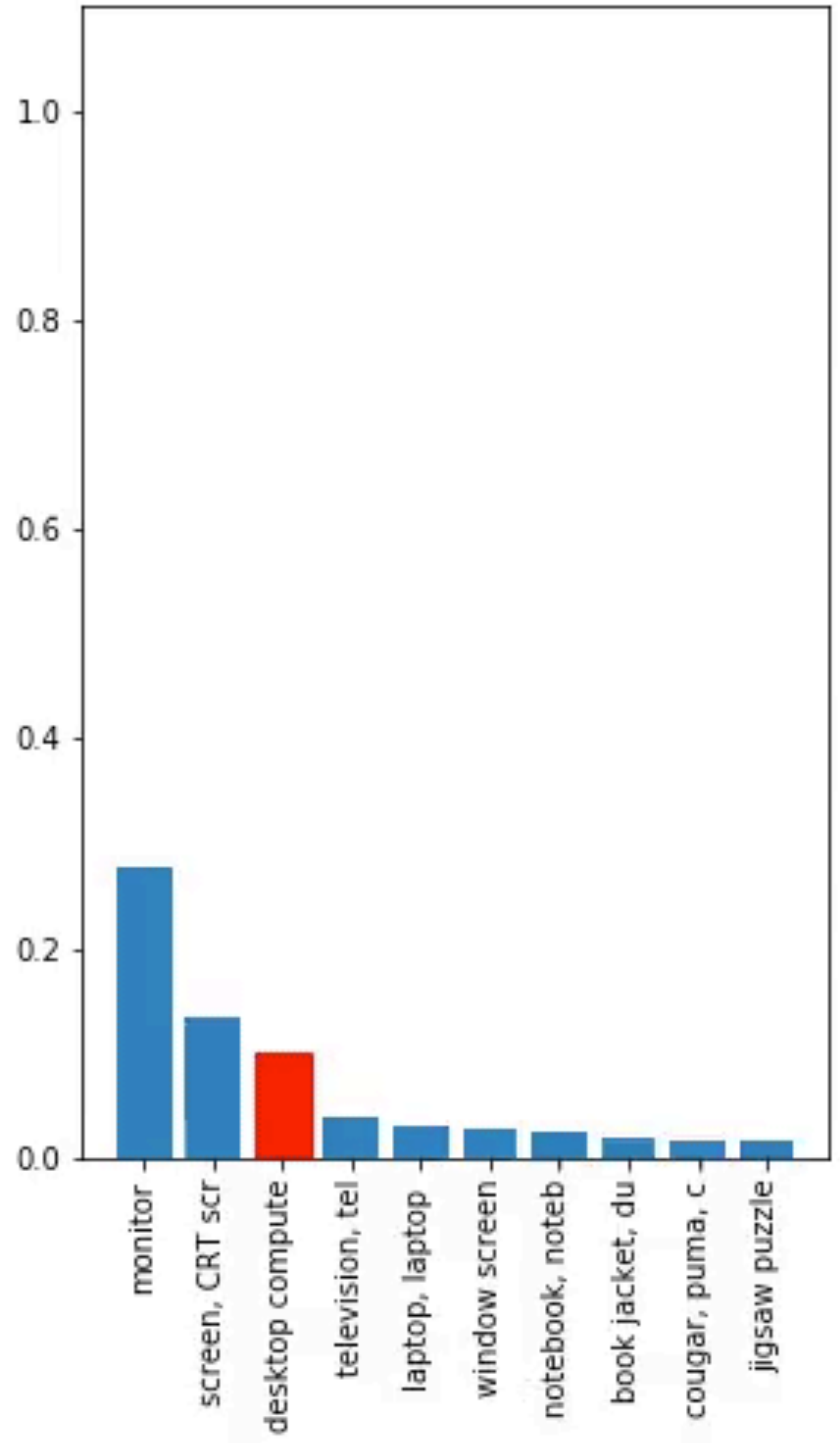
EOT produces robust examples

Zoom: 1.000000x



T = {rescale from 1x to 5x}

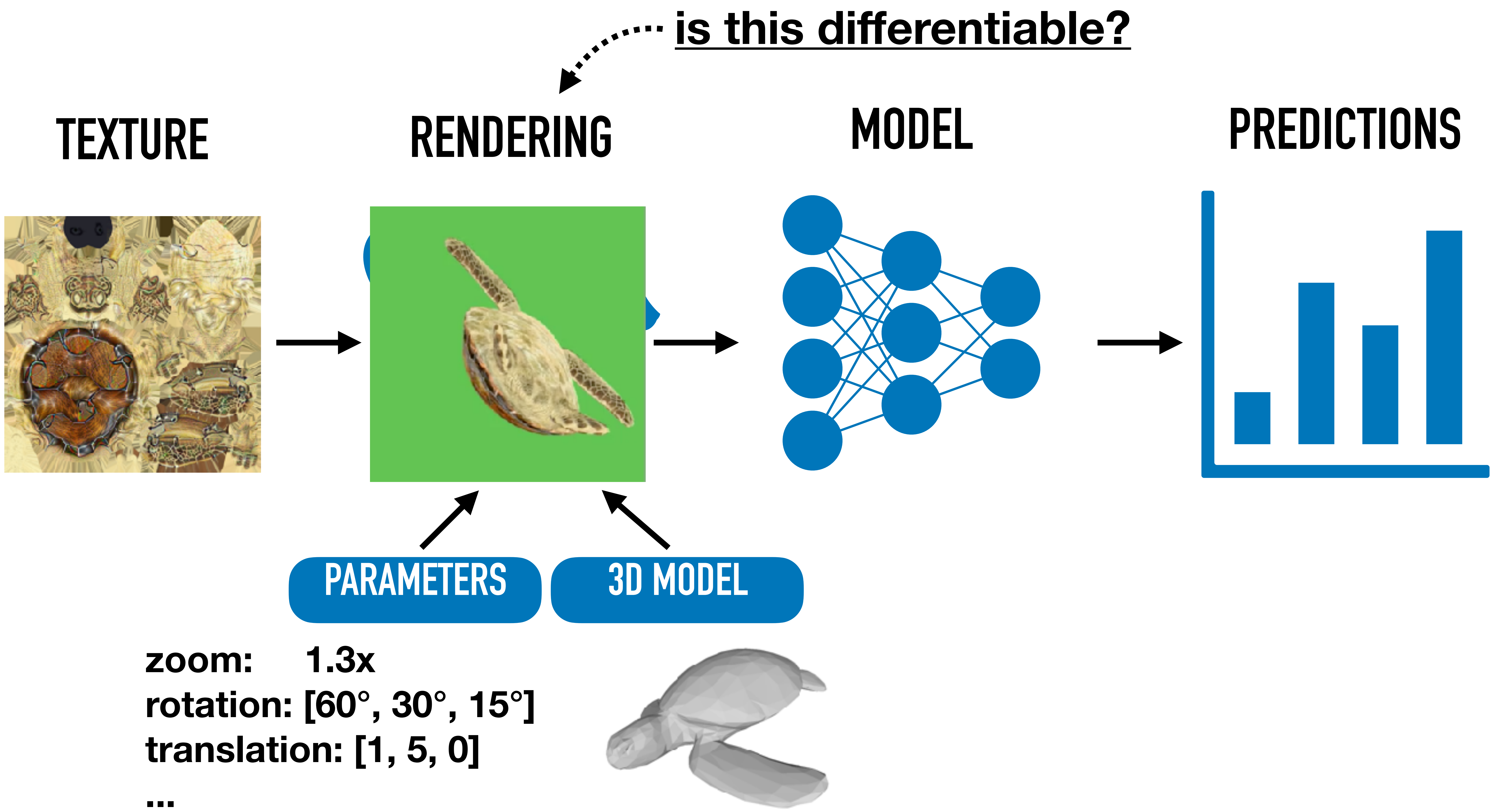
EOT produces robust physical-world examples



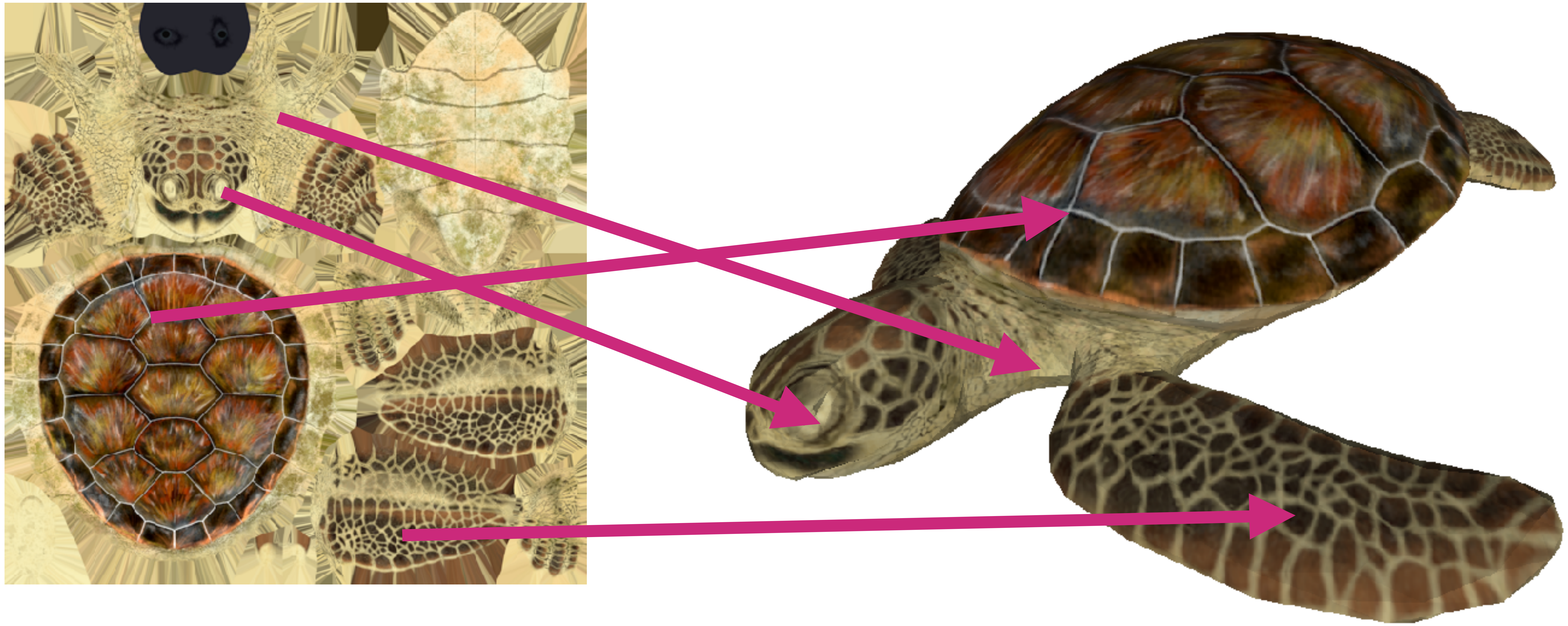
T = {rescale + rotate + translate + skew}

Can we make this
work with 3D objects?

Physical world 3D processing pipeline

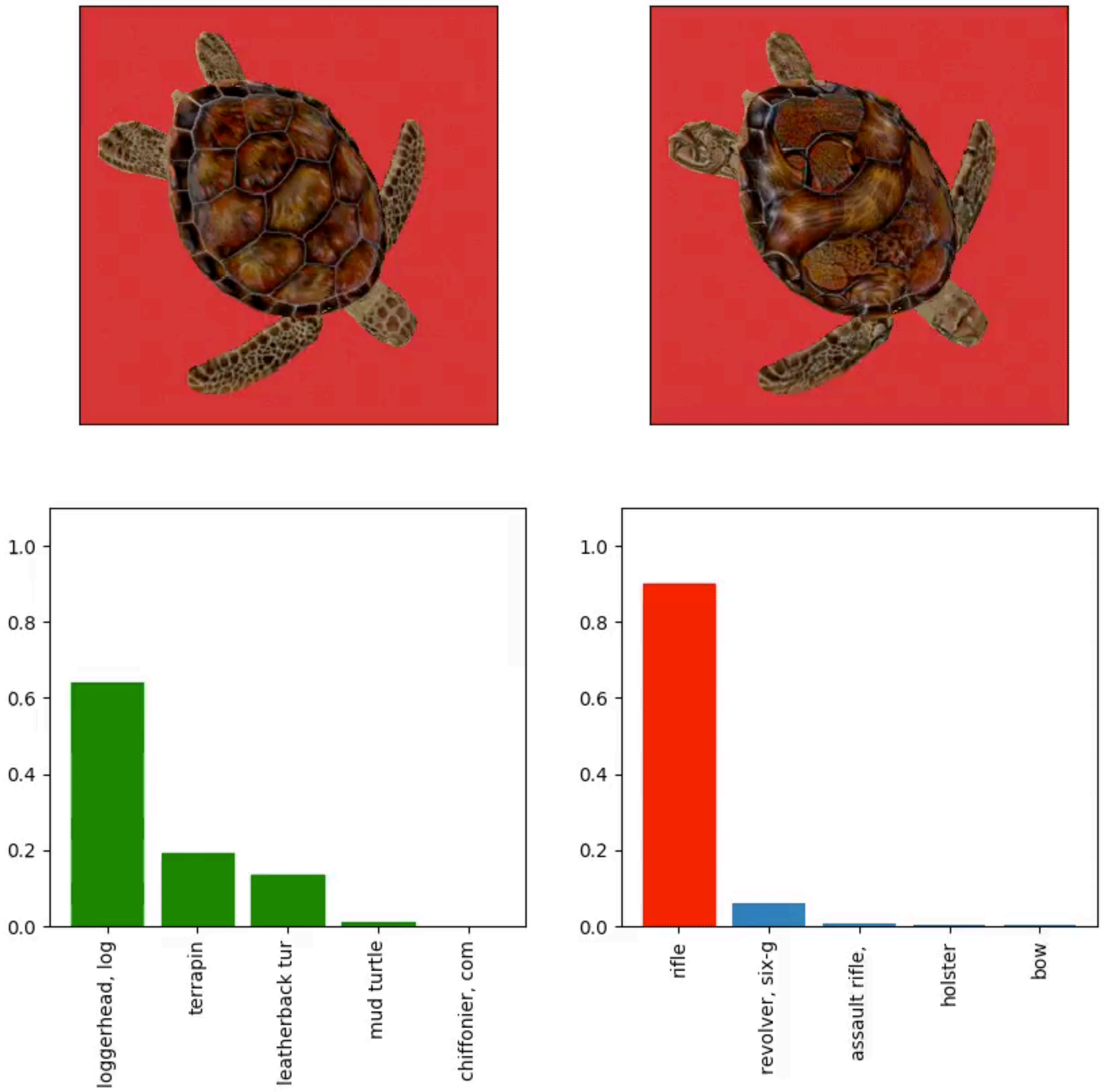


Differentiable rendering



- 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

Poster (and live demo): 6:15 – 9:00pm @ Hall B #73