SYNTHESIZING ROBUST ADVERSARIAL EXAMPLES

Anish Athalye*, **Logan Engstrom***, **& Andrew Ilyas***
Massachusetts Institute of Technology
{aathalye,engstrom,ailyas}@mit.edu

Kevin Kwok
labsix
kevin@labsix.org

ABSTRACT

Adversarial examples generated with standard methods do not consistently fool a classifier in the physical world due to a combination of viewpoint shifts, camera noise, and other natural transformations. These examples require complete control over direct input to the classifier, which is impossible in many real-world systems. We introduce an algorithm for producing adversarial examples that remain adversarial under an attacker-chosen distribution. We first demonstrate its application in two dimensions, producing adversarial images that are robust to noise, distortion, and affine transformation, showing that these input distortions are ineffective against robust adversarial examples. Finally, we apply the algorithm to produce the first physical 3D-printed adversarial objects, demonstrating that our approach works end-to-end in the real world. Our results show that adversarial examples are a practical concern for real-world systems.

1 INTRODUCTION

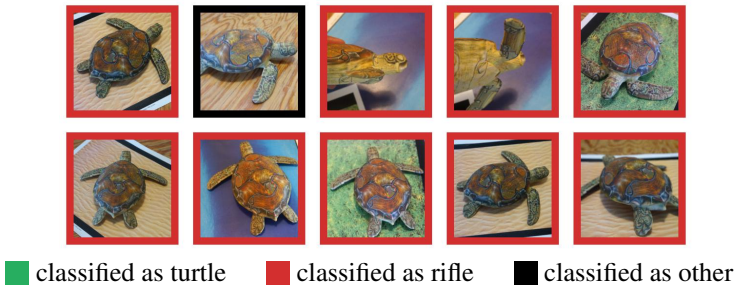


Figure 1: Randomly sampled poses of a **single** 3D-printed turtle adversarially perturbed to classify as a rifle at every viewpoint by an ImageNet classifier. The unperturbed model is classified correctly as a turtle 100% of the time. See https://youtu.be/qPxlhGSG0tc for a video where every frame is fed through the classifier: the turtle is consistently classified as a rifle.

We show that neural network-based classifiers are vulnerable to physical-world adversarial examples. We introduce a new algorithm for reliably producing physical 3D objects that are adversarial from every viewpoint. Figure 1 shows an example of an adversarial object constructed using our approach, where a 3D-printed turtle is consistently classified as rifle by an ImageNet classifier.

Prior attempts at real-world adversarial examples have had limited success in producing robust examples. While some progress has been made, current efforts have demonstrated single datapoints on nonstandard classifiers, and only in the two-dimensional case, with no clear generalization to three dimensions. In two dimensions, "viewpoints" can be approximated by an affine transformations of an original image. Constructing adversarial examples for the physical world requires the ability to generate entire 3D adversarial objects, which must remain adversarial in the face of complex transformations not applicable to 2D objects, such as 3D rotations and perspective projection.

*Equal contribution

1.1 CONTRIBUTIONS

In this work, we definitively show that adversarial examples pose a real threat in the physical world, and that input distortions of noise, scaling, rotation, translation do not effectively defend against adversarial attacks:

- We introduce Expectation Over Transformation (EOT), an algorithm that produces adversarial examples that are simultaneously adversarial over a distribution of transformations
- We consider the problem of constructing 3D adversarial examples under the EOT framework, viewing the 3D rendering process as part of the transformation, and we show that the approach successfully synthesizes adversarial objects
- We fabricate adversarial objects and show that they remain adversarial, demonstrating that our approach works end-to-end in the physical world, showing that *adversarial examples are of real concern in practical deep learning systems*

2 APPROACH

First, we present the Expectation Over Transformation (EOT) algorithm, a general framework allowing for the construction of adversarial examples that remain adversarial under any chosen transformation distribution T. We then describe our end-to-end approach for generating adversarial objects using a specialized application of EOT and a differentiable 3D renderer.

2.1 EXPECTATION OVER TRANSFORMATION

When constructing adversarial examples in the white-box case (that is, with access to a classifier and its gradient), we know in advance a set of possible classes Y and a space of valid inputs X to the classifier; we have access to the function $P(y|x)$ and its gradient $\nabla P(y|x)$, for any class $y \in Y$ and input $x \in X$. In the standard case, adversarial examples are produced by maximizing the log-likelihood of the target class over an ϵ-radius ball around the original image; however, these examples have been shown to be unable to remain adversarial even under minor perturbations inevitable in any real-world observation (Luo et al., 2016; Lu et al., 2017).

To address this issue, we introduce *Expectation over Transformation (EOT)*; the key insight behind EOT is to model such perturbations within the optimization procedure. In particular, rather than optimizing the log-likelihood of a single example, EOT uses a chosen distribution T of transformation functions t taking an input x' generated by the adversary to the "true" input $t(x')$ perceived by the classifier. Furthermore, rather than simply taking the norm of $x' - x$ to constrain the solution space, EOT instead aims to constrain the *effective distance* between the adversarial and original inputs, which we define as $\delta = \mathbb{E}_{t\sim T}[d(t(x) - t(x'))]$. Intuitively, this is how different we expect the true input to the classifer will be, given our new input. Then, EOT solves the following optimization:

$$\hat{x} = \arg\max_{x'} \mathbb{E}_{t\sim T}[\log P(y|t(x'))] \qquad \text{s.t. } \mathbb{E}_{t\sim T}[d(t(x'), t(x))] < \epsilon$$

EOT crucially generalizes beyond simple transformations; in particular, EOT finds examples robust under any perception distribution $Q(\cdot; x)$ parameterized by the generated example x as long as $\frac{d}{dx}Q(\cdot; x)$ is well-defined.

2.2 SYSTEM OVERVIEW

In the 2D case, we design T to approximate a realistic space of possible distortions involved in printing out an image and taking a natural picture of it. This amounts to a set of random transformations of the form $t(x) = Ax + \epsilon$, which are more thoroughly described in Section 3.

In the 3D case, each transformation t from a *texture* to a perceived image is a map between the texture to a rendering of the textured 3D model. The transformations map the texture to a given 3D model, and then simulate rendering, rotation, translation, and perspective projection of the simulated 3D object in addition to the perceptual mechanisms used in the 2D case. These 3D transformations required us to implement a differentiable 3D renderer as a sparse matrix multiplication.

Images	Accuracy			Adversariality			ℓ_2
	min	mean	max	min	mean	max	mean
Original	0.0%	70.0%	100%	0.0%	0.0%	9.1%	N/A
Adversarial	0.0%	0.9%	18.2%	80.1%	**96.4%**	100%	5.6×10^{-5}

Table 1: Evaluation of 1000 2D adversarial examples with random targets. We evaluate each example over 1000 randomly sampled transformations to calculate accuracy and adversariality (percent classified as the adversarial class).

We use the Lagrangian-relaxed form of the problem, as Carlini & Wagner (2017) do in the conventional (non-EOT, single-viewpoint) case. Then, in order to encourage imperceptibility of the generated images, we set $d(x', x)$ to be the ℓ_2 norm in the LAB color space, a perceptually uniform color space where Euclidean distance roughly corresponds with perceptual distance. Note that the $\mathbb{E}_{t\sim T}[||LAB(t(x) - t(\hat{x}))||_2^2]$ can be sampled and estimated in conjunction with $\mathbb{E}[P(y|t(x))]$; in general, the Lagrangian formulation gives EOT the ability to intricately constrain the search space (in our case, using LAB distance) at insignificant computational cost (without computing a complex projection). Our optimization, then, is:

$$\hat{x} = \arg\min_{x'} \mathbb{E}_{t\sim T}[-\log P(y|t(x')) + \lambda||LAB(t(x) - t(x'))||_2^2]$$

We use SGD to find the optimum, clipping to the set of valid inputs (e.g. $[0, 1]$ for images).

3 Evaluation

We show that we can reliably produce transformation-tolerant adversarial examples in both the 2D and 3D case. Furthermore, we show that we can synthesize and fabricate 3D adversarial objects, even those with complex shapes, in the physical world: these adversarial objects remain adversarial regardless of viewpoint, camera noise, and other similar real-world factors.

3.1 Robust 2D Adversarial Examples

In the 2D case, we consider the distribution that includes rescaling, rotation, lightening or darkening, Gaussian noise, and any in-bounds translation of the image.

We take the first 1000 images in the ImageNet validation set, randomly choose a target class for each image, and use EOT to synthesize an adversarial example that is robust over the chosen distribution. For each adversarial example, we evaluate over 1000 random transformations sampled from the distribution at evaluation time. Table 1 summarizes the results. On average, the adversarial examples we produce are **96.4% adversarial**, showing that our approach is highly effective in producing robust adversarial examples. Figure 2 shows an example of a synthesized adversarial example.

3.2 Robust 3D Adversarial Examples

We produce 3D adversarial examples by modeling the 3D rendering as a transformation under EOT. Given a textured 3D object, we optimize over the texture such that the rendering is adversarial from any viewpoint. We consider a distribution that incorporates different camera distances, lateral translation, rotation of the object, and solid background colors.

We consider 5 complex 3D models, choose 20 random target classes per model, and use EOT to synthesize adversarial textures for the models with minimal parameter search (four constant, pre-chosen λ values were tested across each model, target pair). For each of the 100 adversarial examples, we sample 100 random transformations from the distribution. Table 2 summarizes results, and Figure 3 shows renderings of drawn samples with classification probabilities.

The simulated adversarial object have an average adversariality of 84.0%, showing that EOT usually produces highly adversarial objects. See Appendix D for a plot of the distribution.

Images	Accuracy			Adversariality			ℓ_2
	min	mean	max	min	mean	max	mean
Original	28.0%	84.0%	100.0%	0	0	0	N/A
Adversarial	0.0%	1.7%	26.0%	0.0%	**84.0%**	100.0%	6.5×10^{-5}

Table 2: A 3D adversarial example with a random target.

Models	Adversarial	Misclassified	Correct
Turtle	**82%**	16%	2%
Baseball	**59%**	31%	10%

Table 3: Analysis of the adversarial objects, over 100 photos of each model over a wide distribution of viewpoints. Both models are classified as the adversarial target class in the majority of viewpoints.

3.3 Physical Adversarial Examples

To fabricate physical-world adversarial examples, beyond modeling the 3D rendering process, we must model physical-world phenomena such as lighting effects and camera noise. Furthermore, we must the 3D printing process: in our case, we use commercially-available full-color 3D printing. With the 3D printing technology we use, we find that color accuracy varies between prints, so we model printing errors as well. We approximate all of these phenomena by a distribution of transformations under EOT. In addition to the transformations considered for 3D in simulation, we consider camera noise, additive and multiplicative lighting, and per-channel color inaccuracies.

We evaluate physical adversarial examples over two models: one of a turtle, and one of a baseball. Unperturbed models are correctly classified with 100% accuracy over a large number of samples. We choose target classes for each of the models at random — "rifle" for the turtle, and "espresso" for the baseball — and we use EOT to synthesize adversarial examples. We evaluate the 3D-printed adversarial objects by taking 100 photos of each object over a variety of viewpoints. Figure 5 shows a random sample of these images. Table 3 gives a quantitative analysis over all images.

4 Related Work

State of the art neural networks are vulnerable to adversarial examples (Szegedy et al., 2013). A number of methods exist for synthesizing adversarial examples in the white-box, single-viewpoint scenario where the adversary directly controls the input to the neural network, including the Fast Gradient Sign Method (Goodfellow et al., 2015), a Lagrangian relaxation formulation (Carlini & Wagner, 2017), and Projected Gradient Descent (Madry et al., 2017).

Kurakin et al. (2016) demonstrate the transferability of FGSM-generated adversarial misclassification on a printed page. Evtimov et al. (2017) proposed a potential method for generating robust physical-world adversarial examples in the 2D case. However, the approach is limited to generating 2D adversarial examples, with no clear translation to the 3D case. The method additionally requires the taking and preprocessing of a large quantity of photos to produce each adversarial example, and is limited to a single class of allowed perturbations.

5 Conclusion

This work shows that adversarial examples pose a practical concern to neural network-based image classifiers. By introducing EOT, a general-purpose algorithm for the creation of robust examples under any chosen distribution, and modeling 3D rendering and printing within the framework of EOT, we succeed in fabricating three-dimensional adversarial examples. In particular, with access only to low-cost commercially available 3D printing technology, we successfully print physical adversarial objects that are strongly classified as a desired target class over a variety of angles, viewpoints, and lighting conditions by a standard ImageNet classifier.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security & Privacy*, 2017.

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust Physical-World Attacks on Deep Learning Models. 2017. URL https://arxiv.org/abs/1707.08945.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. 2016. URL https://arxiv.org/abs/1607.02533.

Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. 2017. URL https://arxiv.org/abs/1707.03501.

Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. 2016. URL https://arxiv.org/abs/1511.06292.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. 2017. URL https://arxiv.org/abs/1706.06083.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. 2013. URL https://arxiv.org/abs/1312.6199.

A MAIN FIGURES

Here we present the figures referenced throughout the main body of the paper.

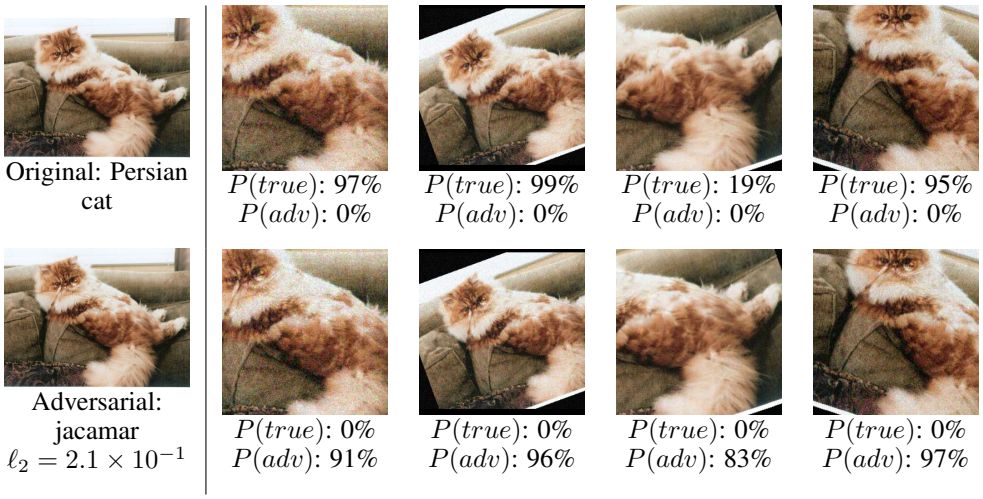


Figure 2: A 2D adversarial example, showing classifier confidence in true and adversarial classes for original and corresponding adversarial images over randomly sampled poses.

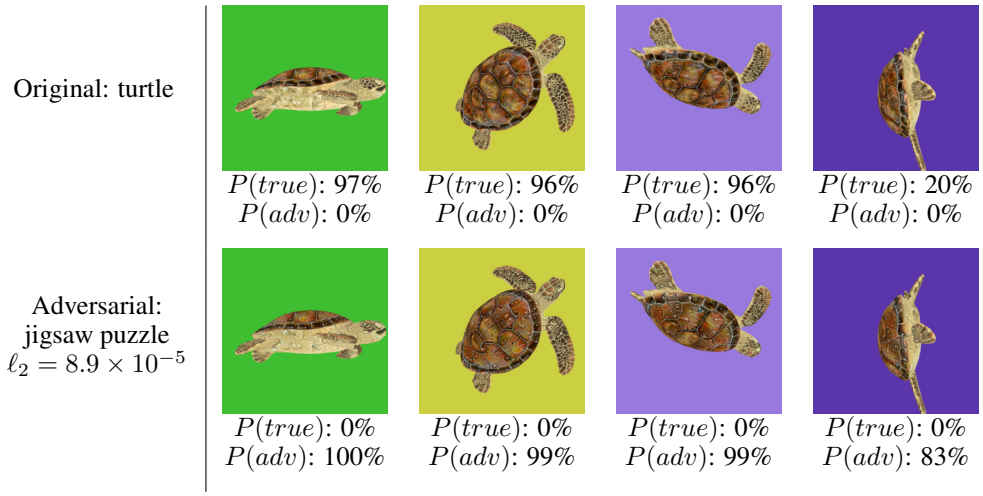


Figure 3: Random sample of 3D adversarial examples, showing classifier confidence in true and adversarial classes for original and corresponding adversarial images over 100 randomly sampled poses.

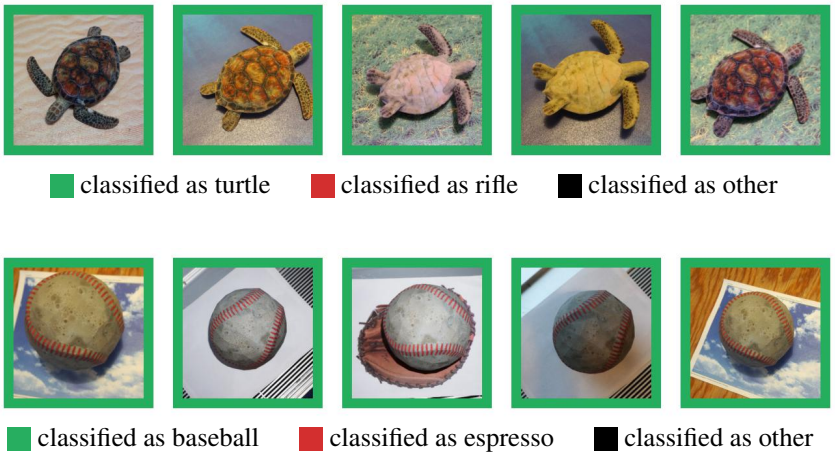
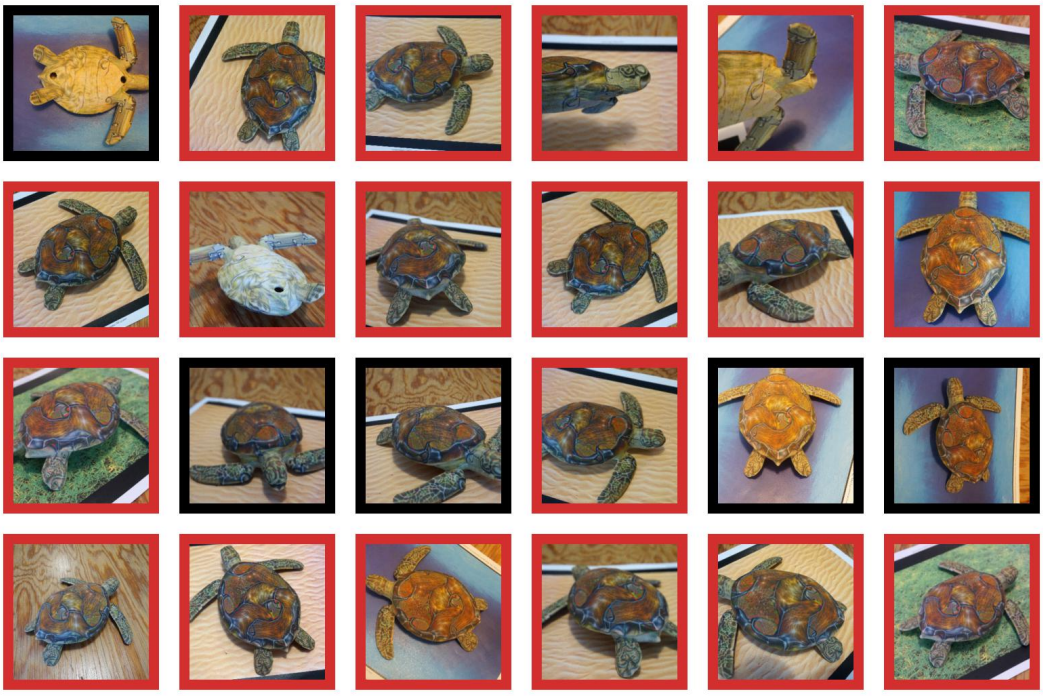


Figure 4: A sample of photos of **unperturbed** 3D prints. The unperturbed 3D-printed objects are consistently classified as the true class.



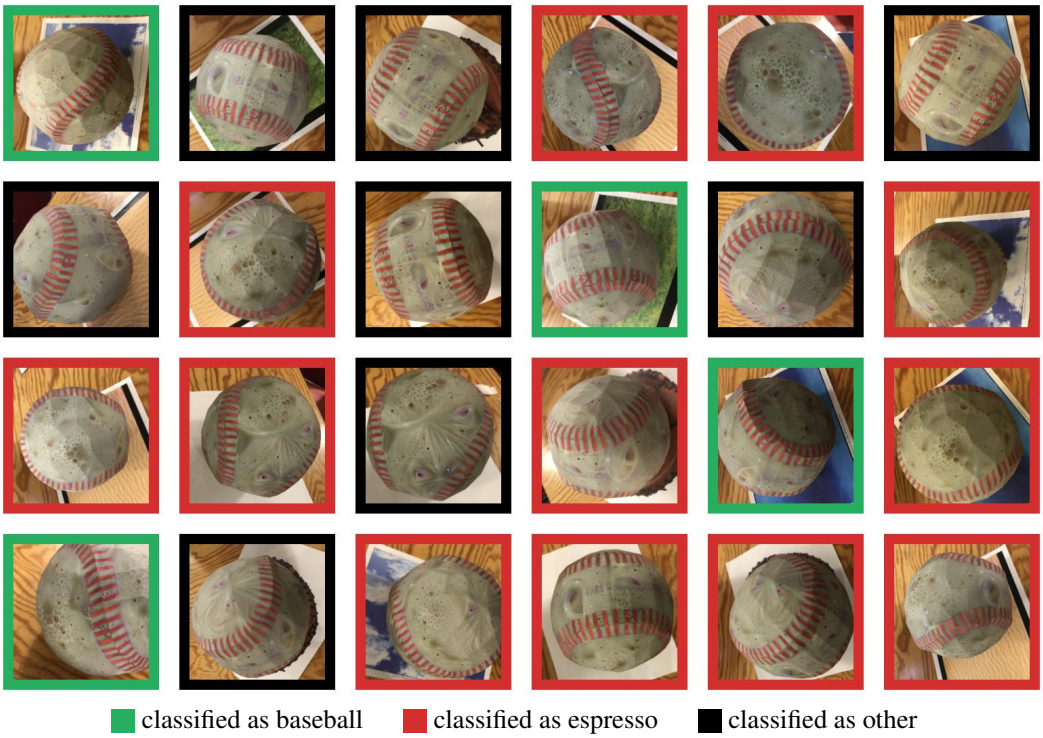


Figure 5: Random sample of photographs of the **two 3D-printed adversarial objects**. The 3D-printed adversarial objects are strongly adversarial over a wide distribution of viewpoints.

Figure 6: Three pictures of the same adversarial turtle (all classified as rifles), demonstrating the need for a wide distribution, and the efficacy of EOT in finding examples robust across wide distributions of physical-world effects like lighting.

B DISTRIBUTIONS OF TRANSFORMATIONS

Under the EOT framework, we must choose a distribution of transformations, and the optimization produces an adversarial example that is robust under the distribution of transformations. Here, we give the specific parameters we chose in the 2D (Table 4), 3D (Table 5), and physical-world case (Table 6).

Transformation	**Minimum**	**Maximum**
Scale	0.9	1.4
Rotation	$-22.5°$	$22.5°$
Lighten / Darken	-0.05	0.05
Gaussian Noise (stdev)	0.0	0.1
Translation	any in-bounds	

Table 4: Distribution of transformations for the 2D case, where each parameter is sampled uniformly at random from the specified range.

Transformation	**Minimum**	**Maximum**
Camera distance	2.5	3.0
X/Y translation	-0.05	0.05
Rotation	any	
Background	(0.1, 0.1, 0.1)	(1.0, 1.0, 1.0)

Table 5: Distribution of transformations for the 3D case when working in simulation, where each parameter is sampled uniformly at random from the specified range.

Transformation	**Minimum**	**Maximum**
Camera distance	2.5	3.0
X/Y translation	-0.05	0.05
Rotation	any	
Background	(0.1, 0.1, 0.1)	(1.0, 1.0, 1.0)
Lighten / Darken (additive)	-0.15	0.15
Lighten / Darken (multiplicative)	0.5	2.0
Per-channel (additive)	-0.15	0.15
Per-channel (multiplicative)	0.7	1.3
Gaussian Noise (stdev)	0.0	0.1

Table 6: Distribution of transformations for the physical-world 3D case, approximating rendering, physical-world phenomena, and printing error.

C Robust 2D Adversarial Examples

We give a random sample out of our 1000 2D adversarial examples in Figures 7 and 8.

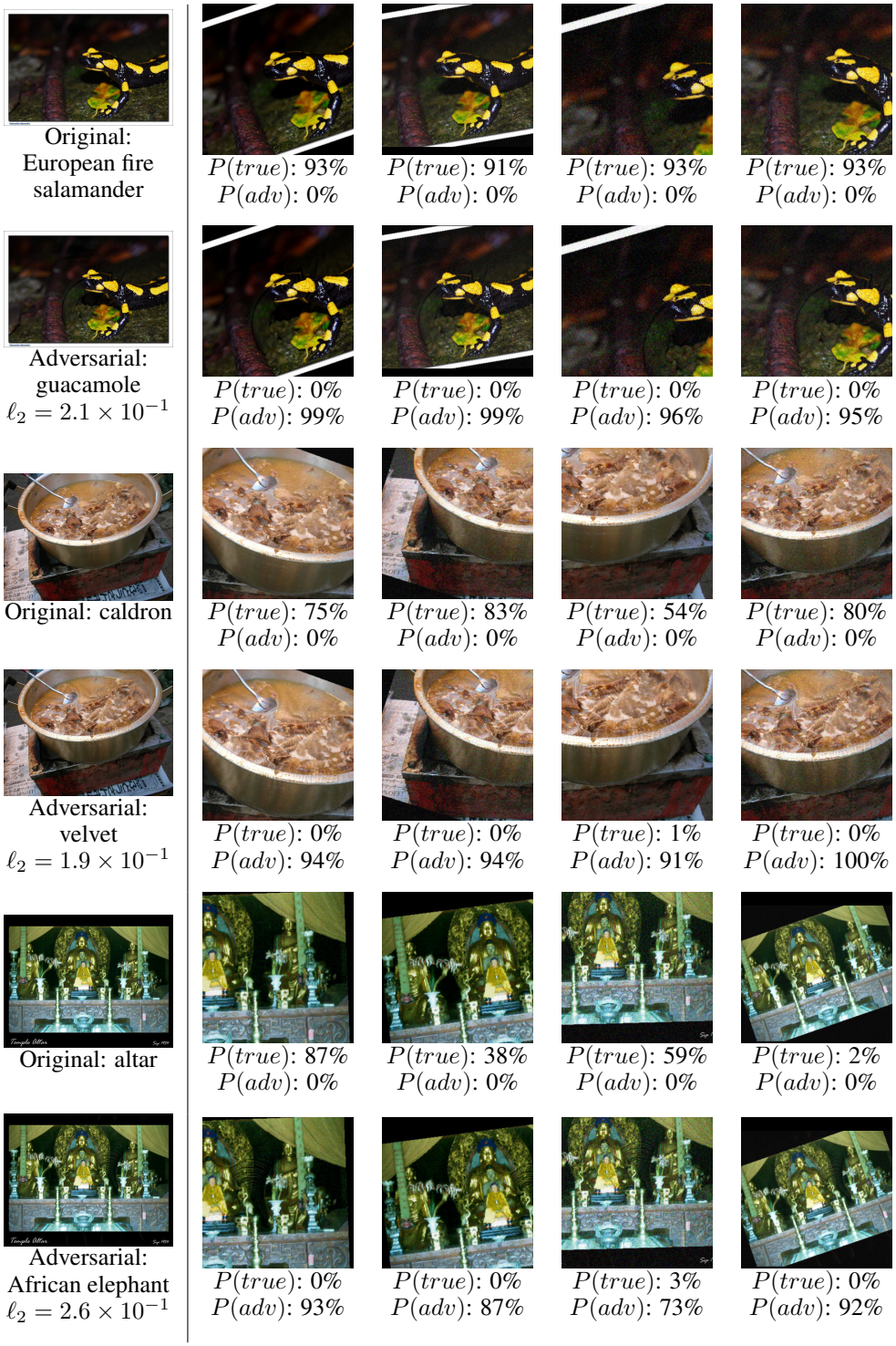


Figure 7: A random sample of 2D adversarial examples.

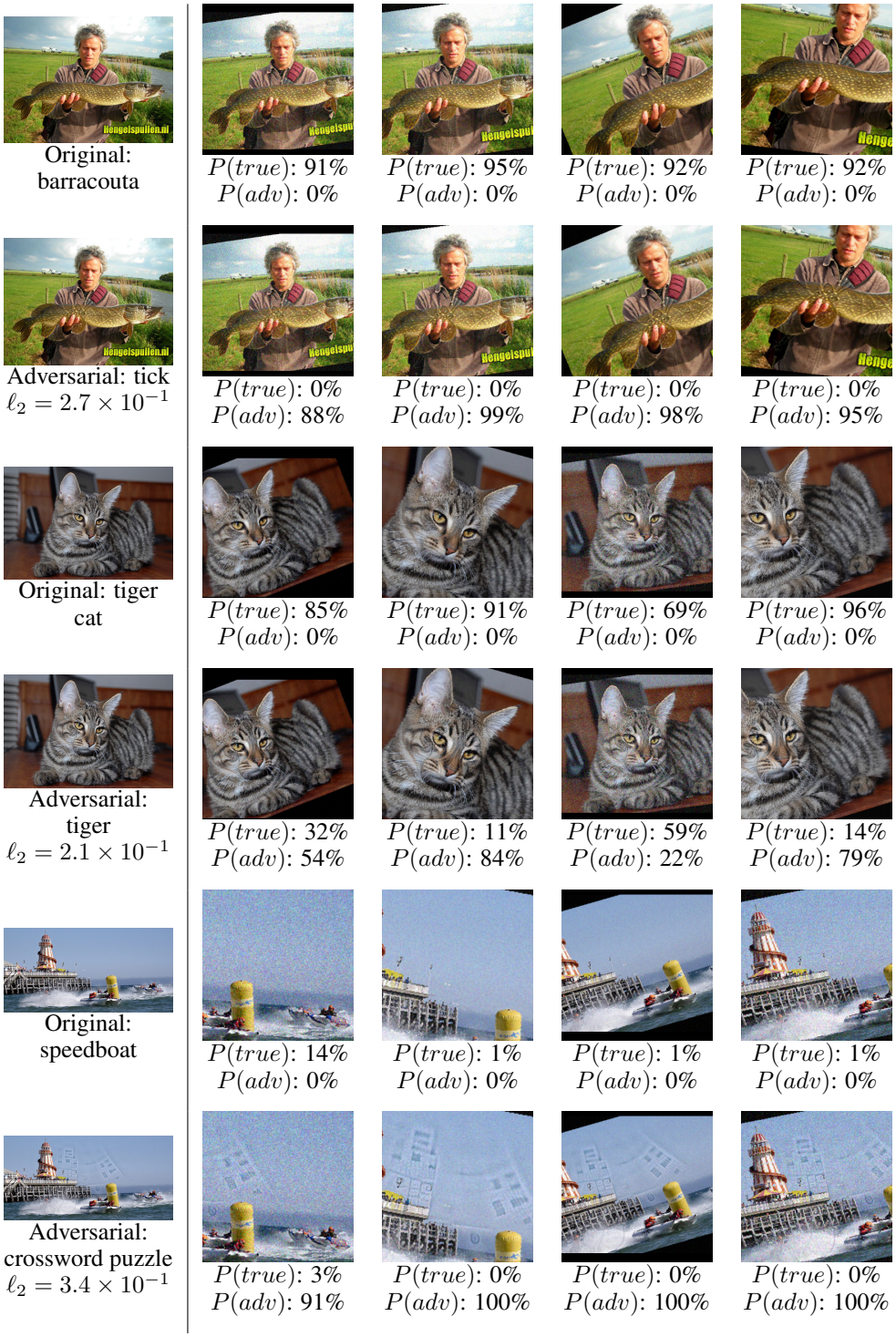


Figure 8: A random sample of 2D adversarial examples.

D ROBUST 3D ADVERSARIAL EXAMPLES

We give a histogram of adversariality (percent classified as the adversarial class) over all 100 examples in Figure 9.

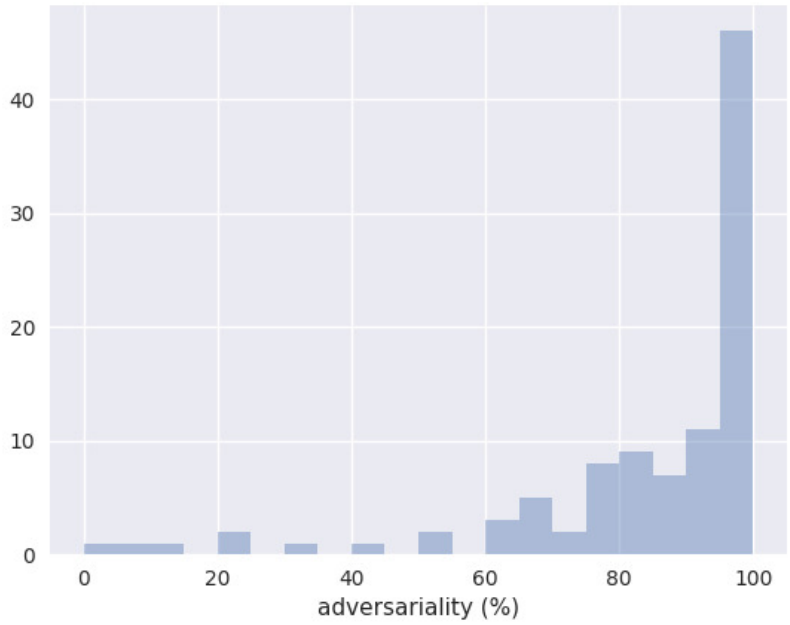


Figure 9: A histogram of adversariality (percent of samples classified as the adversarial class) across the 100 3D adversarial examples.

We give a random sample out of our 100 3D adversarial examples in Figures 10 and 11.

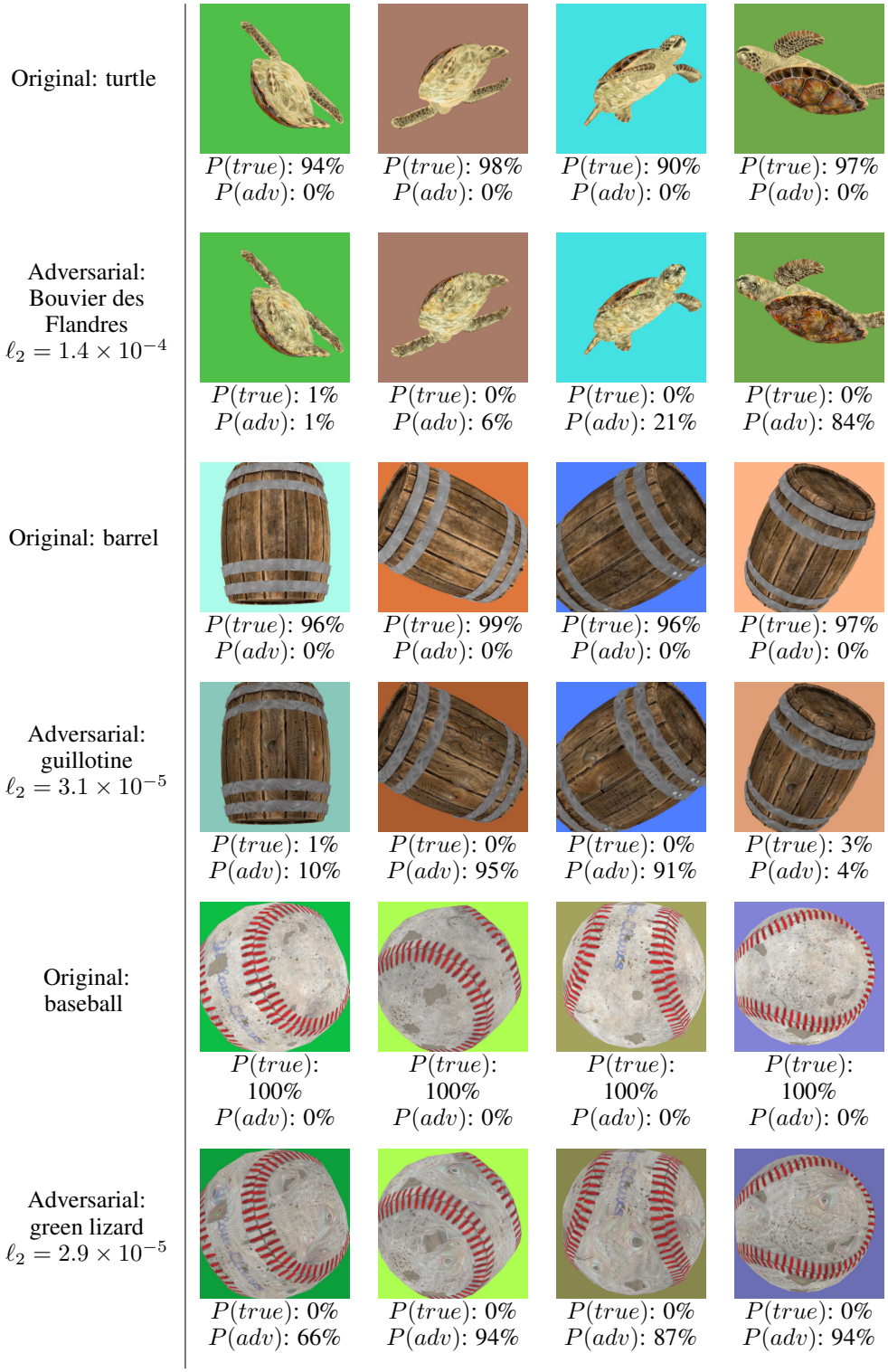


Figure 10: A random sample of 3D adversarial examples.

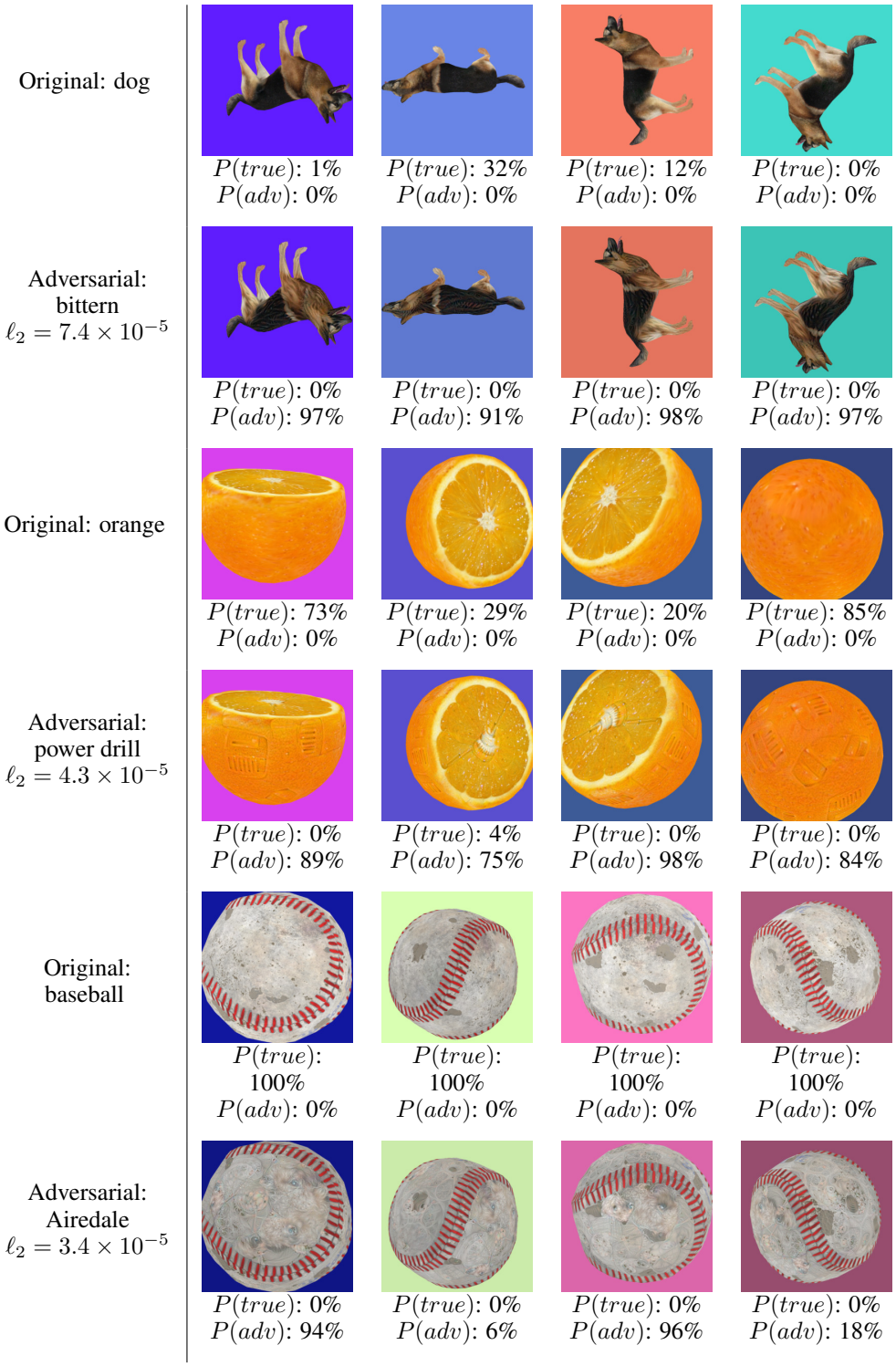


Figure 11: A random sample of 3D adversarial examples.

E PHYSICAL ADVERSARIAL EXAMPLES

Figure 12 gives all 100 photographs of our adversarial 3D-printed turtle, and Figure 13 gives all 100 photographs of our adversarial 3D-printed baseball.

Figure 12: All 100 photographs of our physical-world 3D adversarial turtle.

Figure 13: All 100 photographs of our physical-world 3D adversarial baseball.