Synthesizing Robust Adversarial Examples

Supplementary Material

1 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 1), 3D (Table 2), and physical-world case (Table 3).

2 Robust 2D Adversarial Examples

We give a random sample out of our 1000 2D adversarial examples in Figures 1 and 2.

3 Robust 3D Adversarial Examples

We give a random sample out of our 200 3D adversarial examples in Figures 3 and 4 and 5. We give a histogram of adversariality (percent classified as the adversarial class) over all 200 examples in Figure 6.

4 Physical Adversarial Examples

Figure 7 gives all 100 photographs of our adversarial 3D-printed turtle, and Figure 8 gives all 100 photographs of our adversarial 3D-printed baseball.

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 1: 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 2: 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 3: Distribution of transformations for the physical-world 3D case, approximating rendering, physical-world phenomena, and printing error.

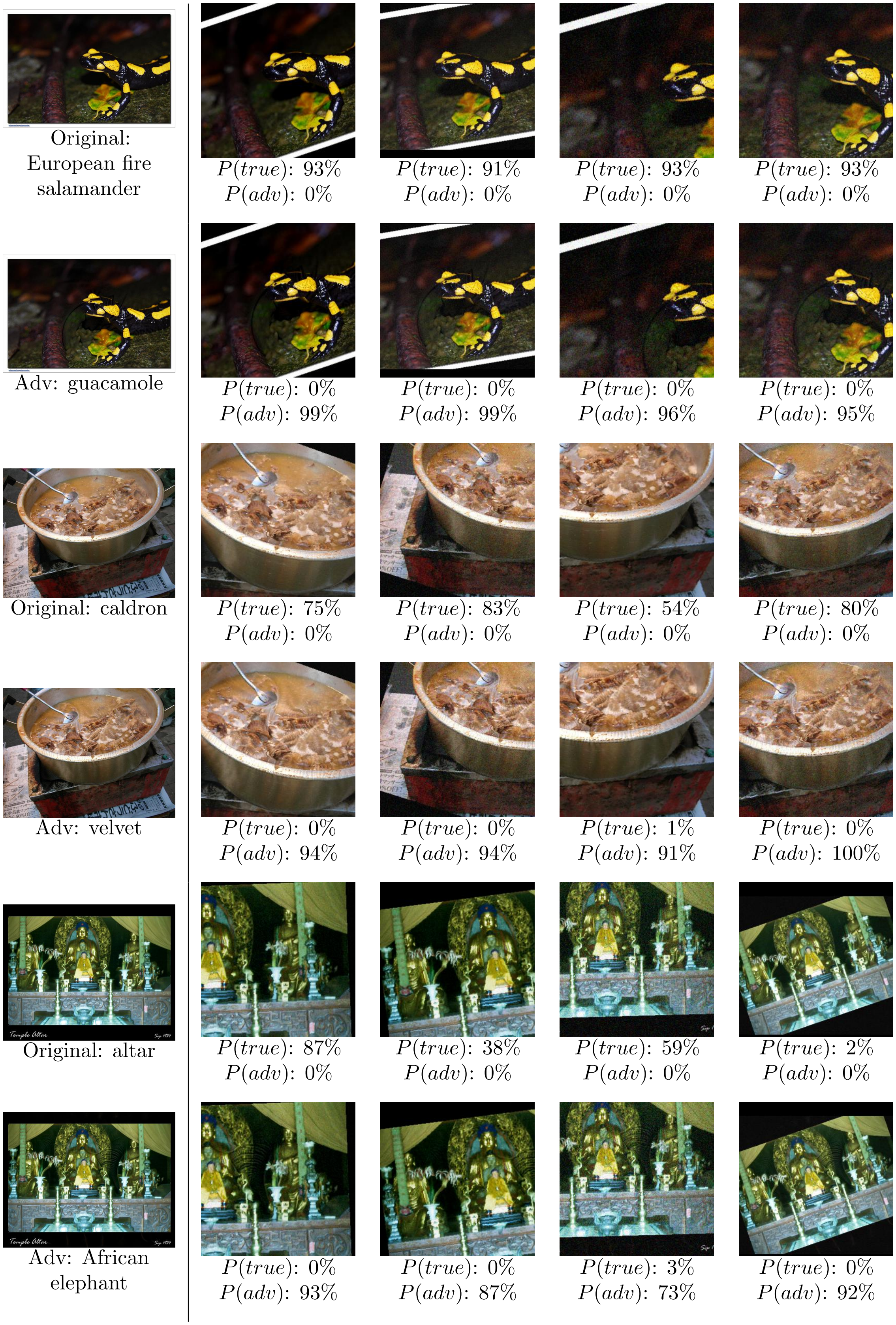

Figure 1: A random sample of 2D adversarial examples.

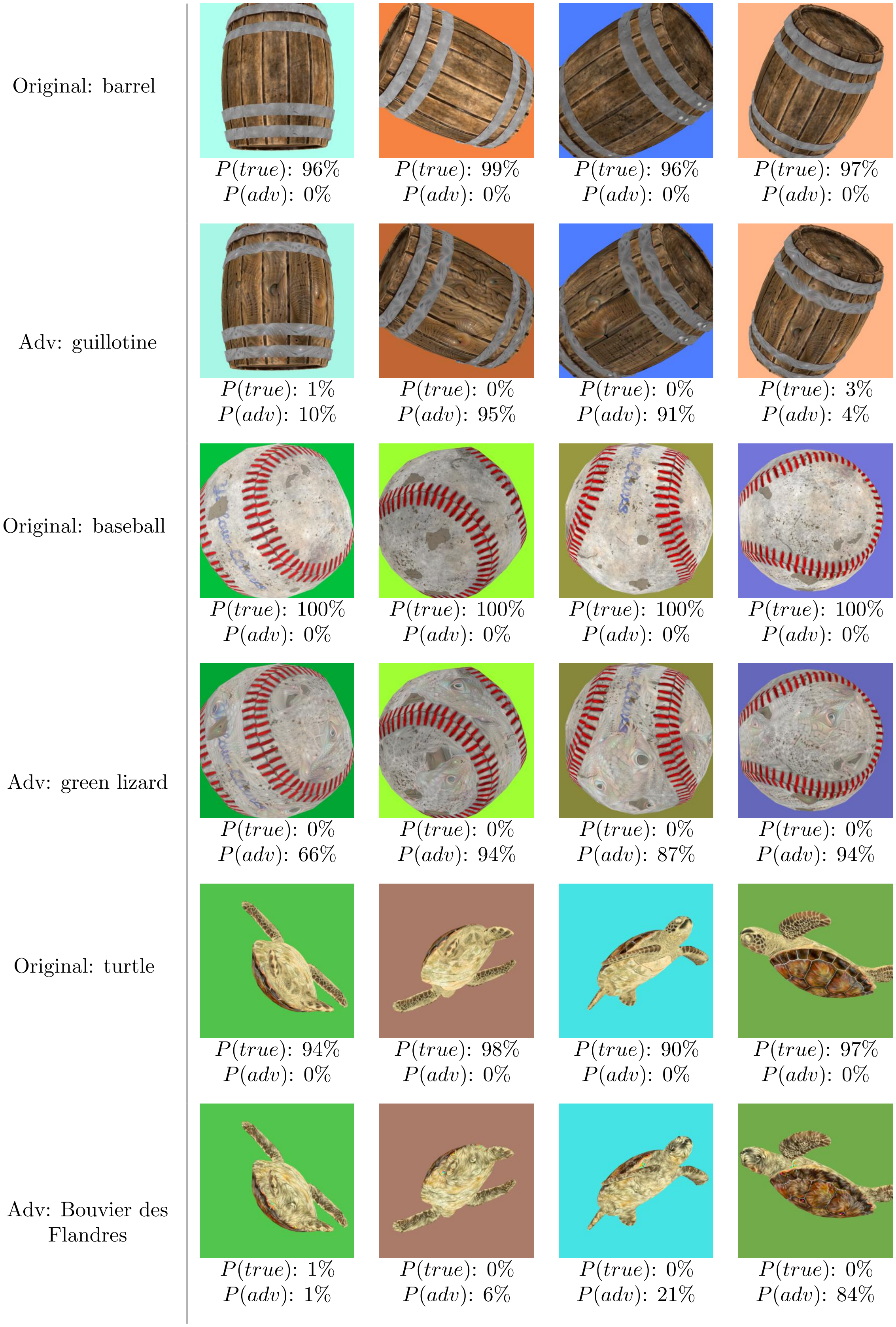
Original: barrel
$P(true)$: 96% $P(adv)$: 0%
$P(true)$: 99% $P(adv)$: 0%
$P(true)$: 96% $P(adv)$: 0%
$P(true)$: 97% $P(adv)$: 0%
Adv: guillotine
$P(true)$: 1% $P(adv)$: 10%
$P(true)$: 0% $P(adv)$: 95%
$P(true)$: 0% $P(adv)$: 91%
$P(true)$: 3% $P(adv)$: 4%
Original: baseball
$P(true)$: 100% $P(adv)$: 0%
$P(true)$: 100% $P(adv)$: 0%
$P(true)$: 100% $P(adv)$: 0%
$P(true)$: 100% $P(adv)$: 0%
Adv: green lizard
$P(true)$: 0% $P(adv)$: 66%
$P(true)$: 0% $P(adv)$: 94%
$P(true)$: 0% $P(adv)$: 87%
$P(true)$: 0% $P(adv)$: 94%
Original: turtle
$P(true)$: 94% $P(adv)$: 0%
$P(true)$: 98% $P(adv)$: 0%
$P(true)$: 90% $P(adv)$: 0%
$P(true)$: 97% $P(adv)$: 0%
Adv: Bouvier des Flandres
$P(true)$: 1% $P(adv)$: 1%
$P(true)$: 0% $P(adv)$: 6%
$P(true)$: 0% $P(adv)$: 21%
$P(true)$: 0% $P(adv)$: 84%

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

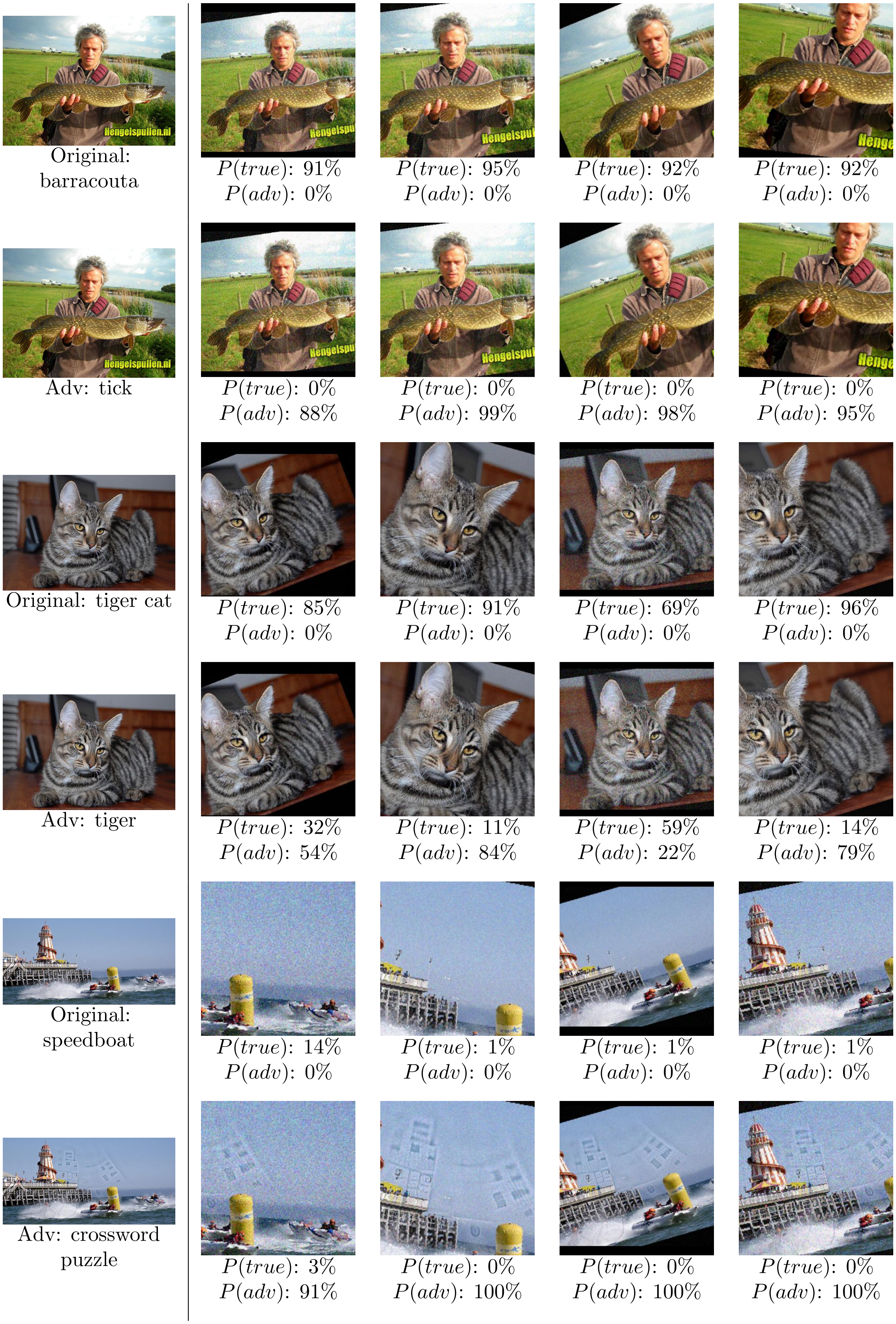

Figure 2: A random sample of 2D adversarial examples.

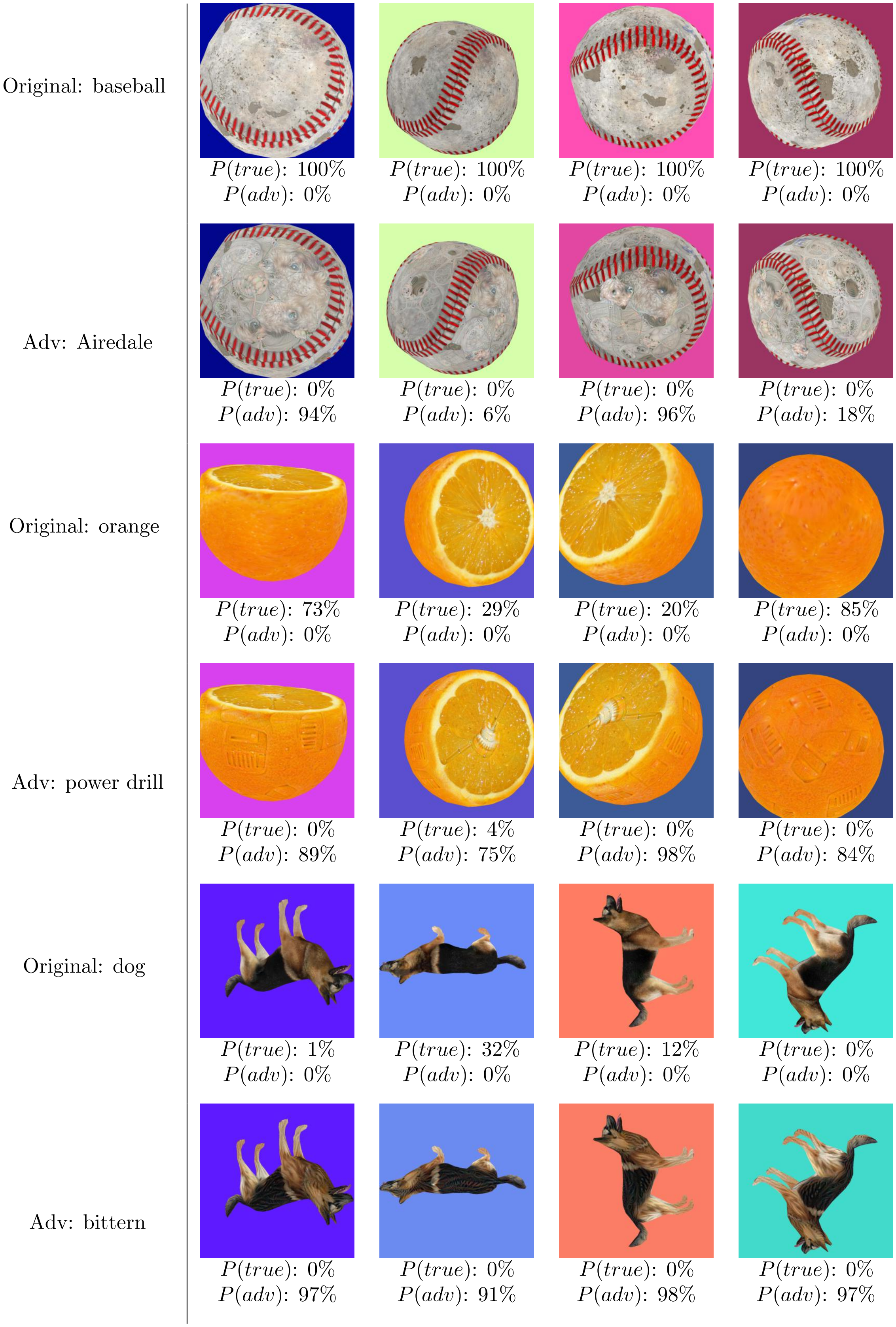

Figure 4: A random sample of 3D adversarial examples.

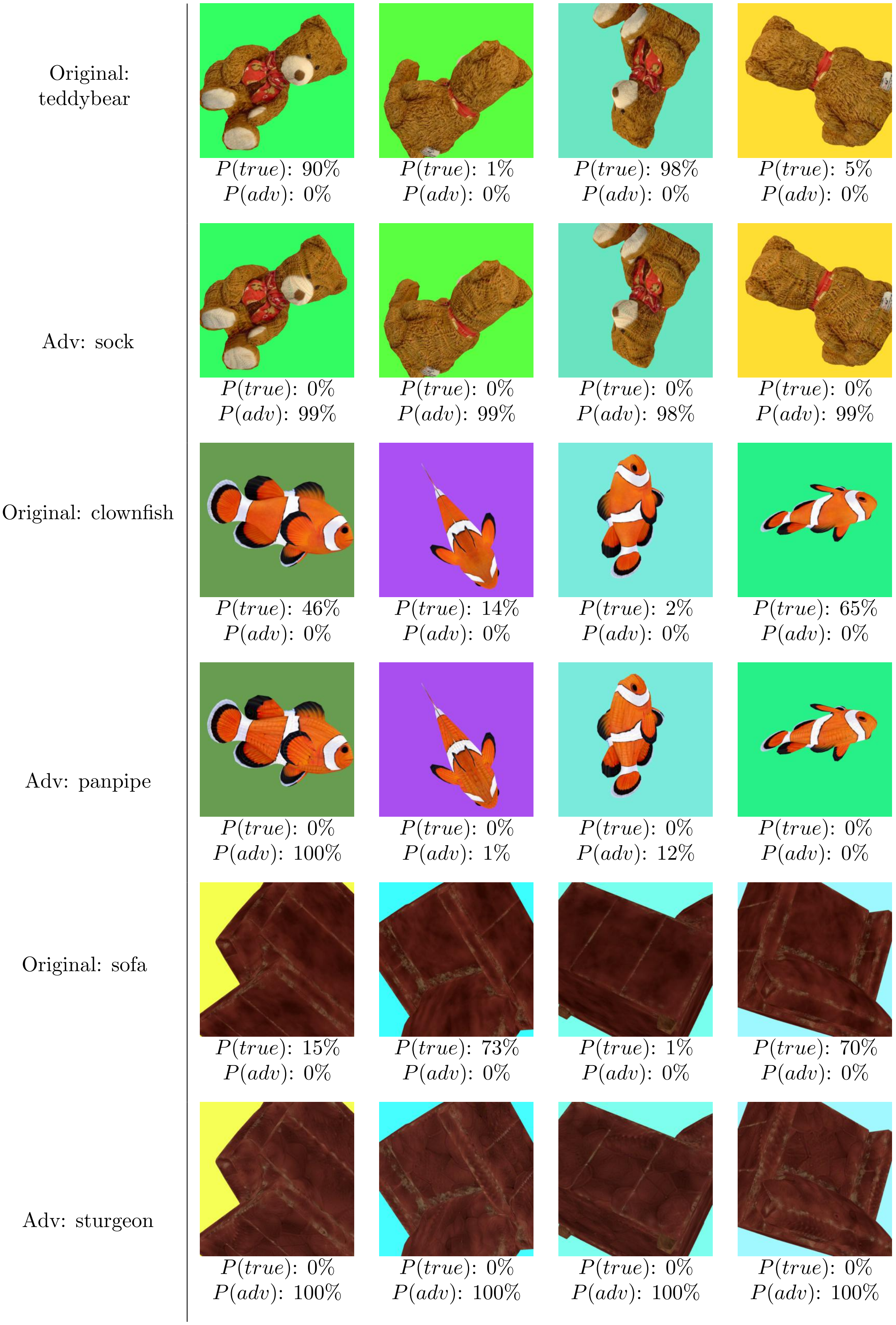
Original: teddybear
$P(true)$: 90% $P(adv)$: 0%
$P(true)$: 1% $P(adv)$: 0%
$P(true)$: 98% $P(adv)$: 0%
$P(true)$: 5% $P(adv)$: 0%
Adv: sock
$P(true)$: 0% $P(adv)$: 99%
$P(true)$: 0% $P(adv)$: 99%
$P(true)$: 0% $P(adv)$: 98%
$P(true)$: 0% $P(adv)$: 99%
Original: clownfish
$P(true)$: 46% $P(adv)$: 0%
$P(true)$: 14% $P(adv)$: 0%
$P(true)$: 2% $P(adv)$: 0%
$P(true)$: 65% $P(adv)$: 0%
Adv: panpipe
$P(true)$: 0% $P(adv)$: 100%
$P(true)$: 0% $P(adv)$: 1%
$P(true)$: 0% $P(adv)$: 12%
$P(true)$: 0% $P(adv)$: 0%
Original: sofa
$P(true)$: 15% $P(adv)$: 0%
$P(true)$: 73% $P(adv)$: 0%
$P(true)$: 1% $P(adv)$: 0%
$P(true)$: 70% $P(adv)$: 0%
Adv: sturgeon
$P(true)$: 0% $P(adv)$: 100%
$P(true)$: 0% $P(adv)$: 100%
$P(true)$: 0% $P(adv)$: 100%
$P(true)$: 0% $P(adv)$: 100%

Figure 5: A random sample of 3D adversarial examples.

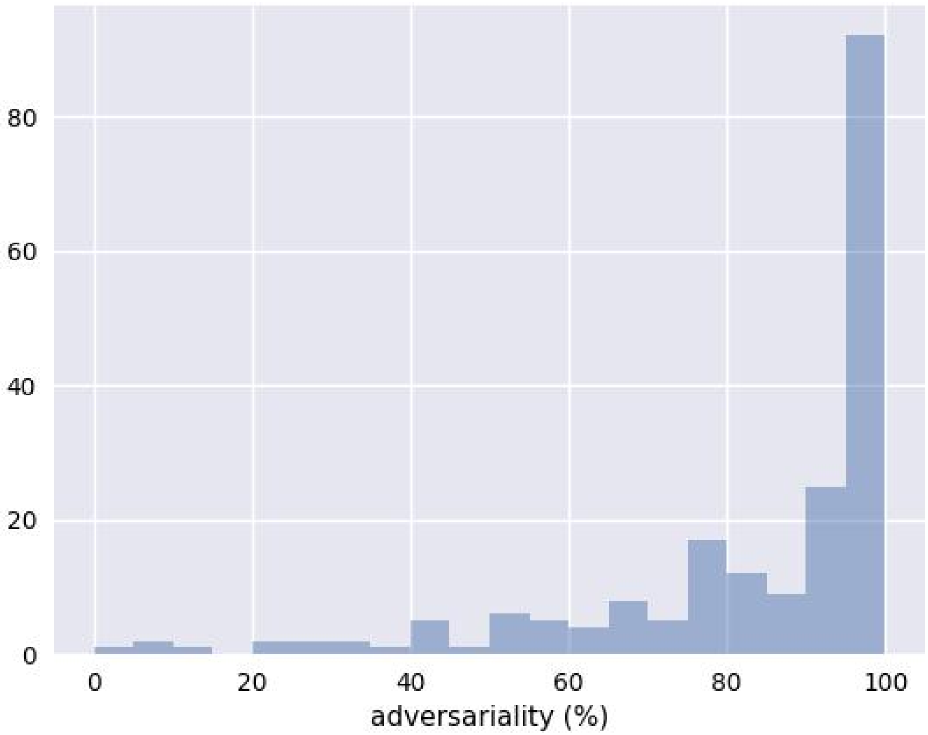

Figure 6: A histogram of adversariality (percent of 100 samples classified as the adversarial class) across the 200 3D adversarial examples.

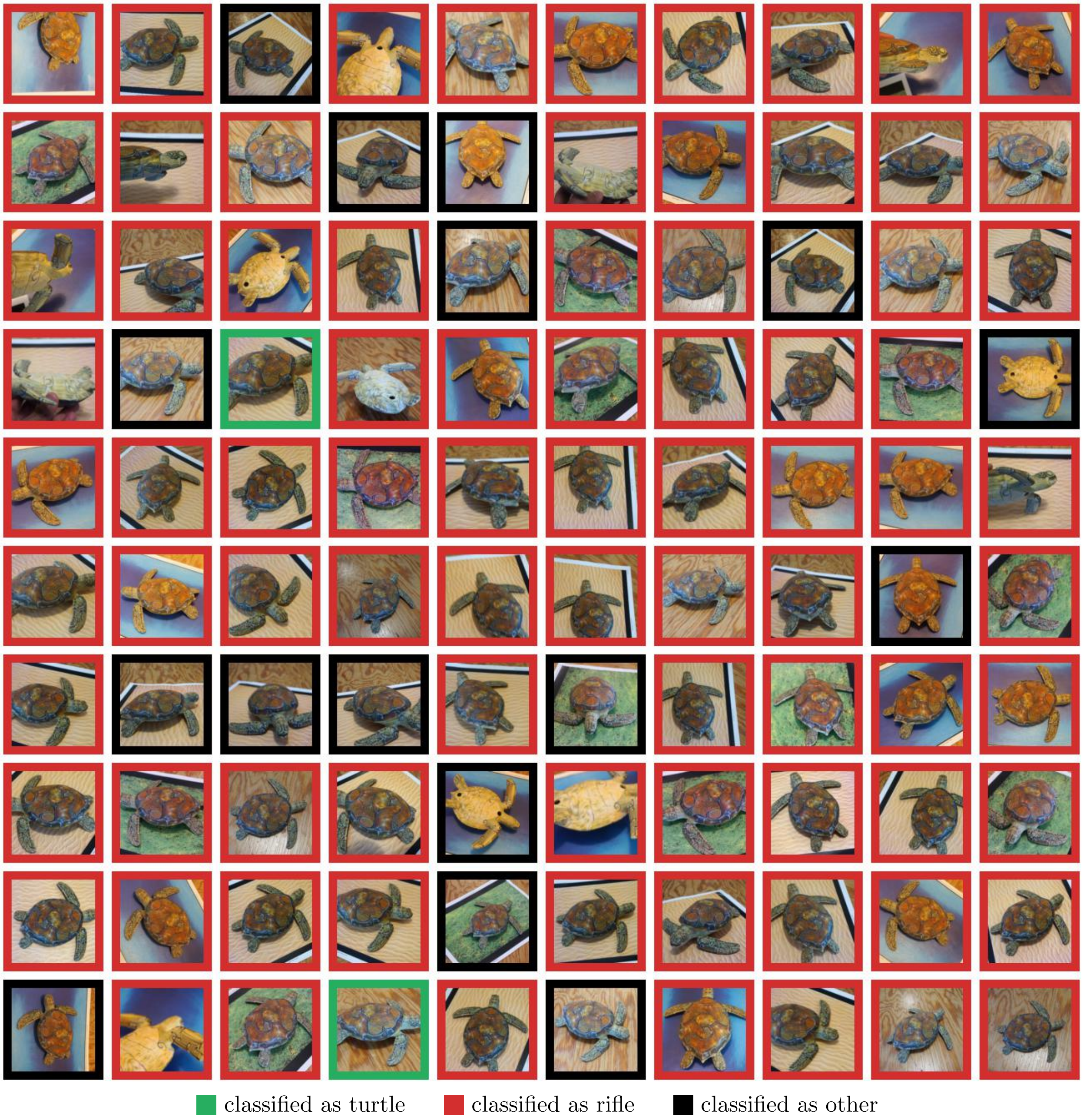

Figure 7: All 100 photographs of our physical-world 3D adversarial turtle.

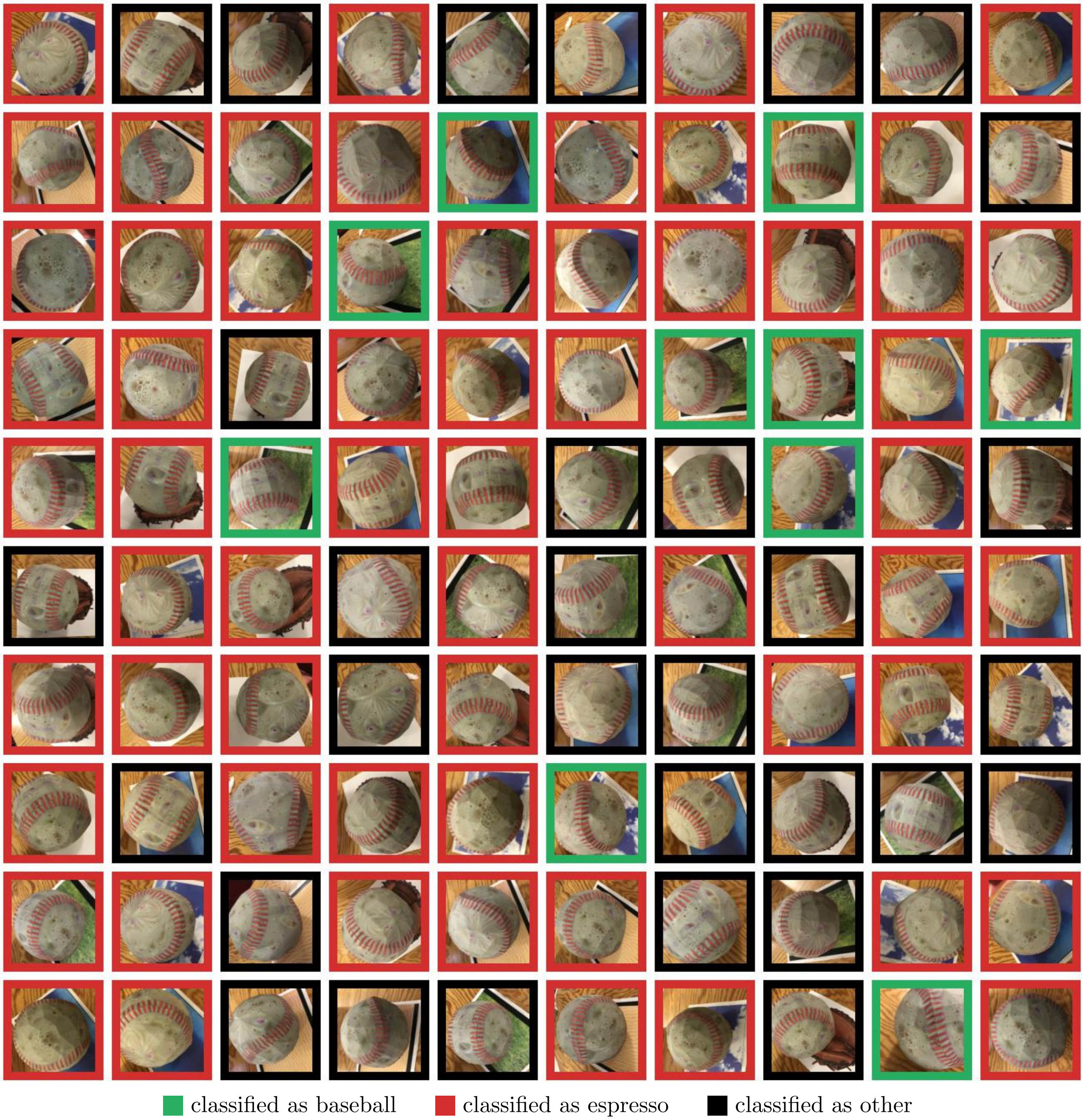

Figure 8: All 100 photographs of our physical-world 3D adversarial baseball.