

Supplementary Material:

Invisible Perturbations: Physical Adversarial Examples Exploiting the Rolling Shutter Effect

Athena Sayles*, Ashish Hooda*, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes

University of Wisconsin–Madison

{esayles, hooda, mohitg, chatterjee, earlence}@cs.wisc.edu

1. Distributions of Transformations

To make our adversarial signal effective in a physical setting, we use the EOT framework. We choose a distribution of transformations. The optimization produces an adversarial example that is robust under the distribution of transformations. Table 1 describes the transformations.

Physical transformations. The relative translation involves moving the object in the image’s field of view. A translation value of 0 means the object is in the center of the

*Both authors contributed equally to this work.

Type	Transformation	Range
Physical	Rotation	$[0, 360^\circ]$
	Horizontal Flip	$\{0, 1\}$
	Vertical Flip	$\{0, 1\}$
	Relative translation	$[0, 0.7]$
	Relative Distance	$[1, 1.5]$
	Relative lighting	$[0.8, 1.2]$
Color Error (per channel)	Affine additive	$[-0.2, 0.2]$
	Affine multiplicative	$[0.7, 1.3]$

Table 1: Ranges for the transformation parameters used for generating and evaluating signals

image, while a value of 1 means the object is at the boundary of the image. The relative distance transform involves enlarging the object to emulate a closer distance. A distance value of 1 is the same as the original image, while for the value of 1.5, the object is enlarged to 1.5 times the original size.

Color correction. Moreover, we apply a multiplicative brightening transformation to the ambient light image to account for small changes in ambient light. To account for the color correction, we used an affine transform of the form $Ax + B$, where A and B are real values sampled from a uniform distribution independently for each color channel.

2. Additional Simulation Results

For evaluating the attack in a simulated setting, we select 5 classes from the ImageNet dataset. We select 7 target classes for each source class and report the results in Table 2. The attack generation and evaluation is the same as described previously. The attack success rate is calculated as the percentage of images classified as the target among 200 transformed images each averaged over all the possible signal offsets. Fig. 2, 1 and 3 give a random sample of 4 transformed images for 3 source classes. For each source class, we give attacked images for 3 target classes.

Source (confid.)	Affinity targets	Attack success	Target confidence (StdDev)
Coffee mug (83%)	Perfume	99%	82% (13%)
	Petri dish	98%	88% (15%)
	Candle	98%	85% (18%)
	Menu	97%	84% (16%)
	Lotion	91%	75% (17%)
	Ping-pong ball	79%	68% (27%)
	Pill bottle	23%	40% (17%)
Street sign (87%)	Monitor	99%	94% (12%)
	Park bench	99%	90% (13%)
	Lipstick	84%	78% (20%)
	Slot machine	48%	59% (19%)
	Carousel	41%	61% (25%)
	Pool table	34%	47% (19%)
	Bubble	26%	37% (22%)
Teddy bear (93%)	Tennis ball	92%	88% (19%)
	Sock	76%	57% (22%)
	Acorn	75%	72% (25%)
	Pencil box	69%	48% (20%)
	Comic book	67%	44% (18%)
	Hour glass	64%	53% (25%)
	Wooden spoon	62%	53% (22%)
Soccer ball (97%)	Pinwheel	96%	87% (15%)
	Goblet	78%	55% (17%)
	Helmet	66%	59% (22%)
	Vase	44%	44% (17%)
	Table lamp	43%	46% (14%)
	Soap dispenser	37%	34% (16%)
	Thimble	10%	15% (02%)
Rifle (96%)	Bow	76%	64% (24%)
	Microphone	74%	63% (22%)
	Tripod	65%	65% (22%)
	Tool kit	57%	56% (22%)
	Dumbbell	35%	44% (21%)
	Binoculars	35%	40% (18%)
	Space bar	17%	33% (17%)

Table 2: Performance of affinity targeting using our adversarial light signals on five classes from ImageNet. For each source class we note the top 7 affinity targets, their attack success rate, and average classifier confidence of the target class. (Average is taken over all offsets values for 200 randomly sampled transformations.)


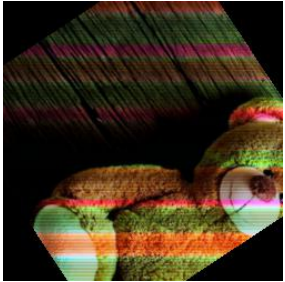
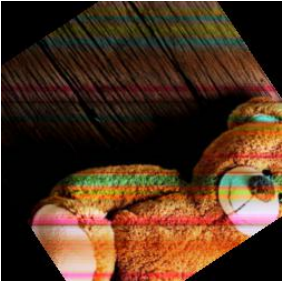
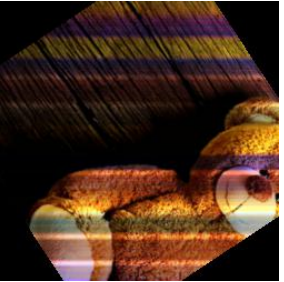





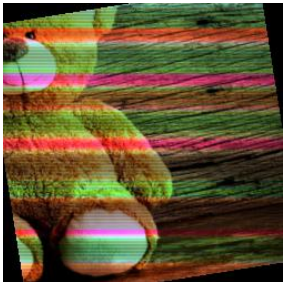

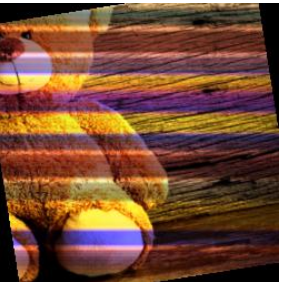



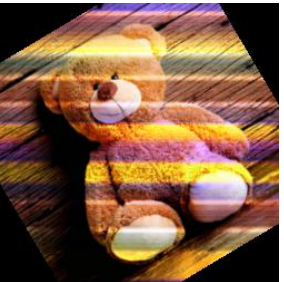
Original - Teddy Bear	Sock	Pencil box	Hour glass
 97%	 90%	 25%	 20%
 100%	 83%	 66%	 61%
 100%	 91%	 40%	 83%
 100%	 78%	 88%	 86%

Figure 1: A random sample of targeted attacks against class - Teddy Bear. The attack is robust to viewpoint, distance and small lighting changes. The numbers denote the confidence values for the respective classes.


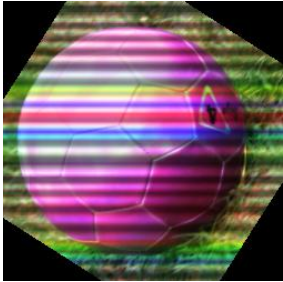
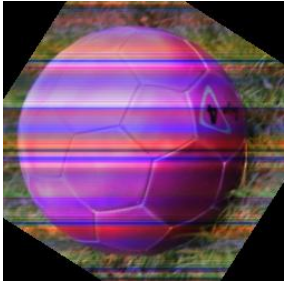
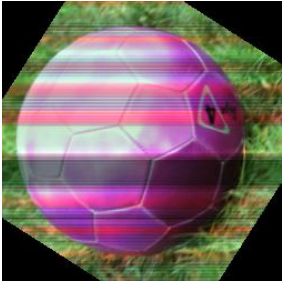

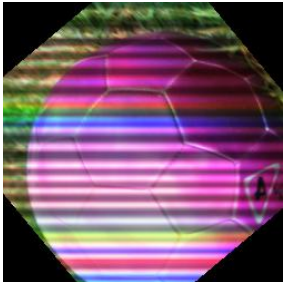
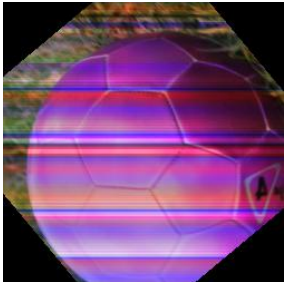
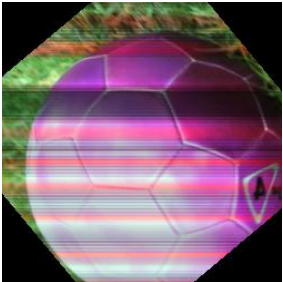

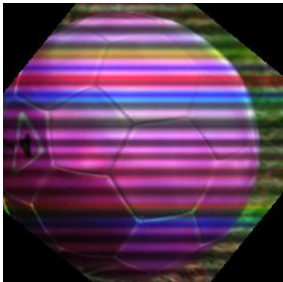
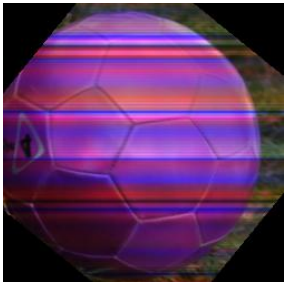
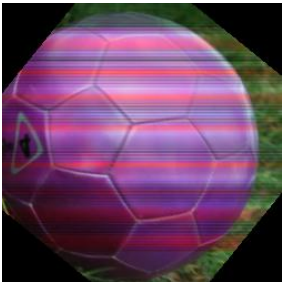

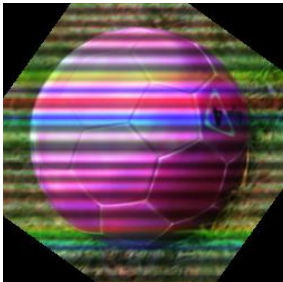
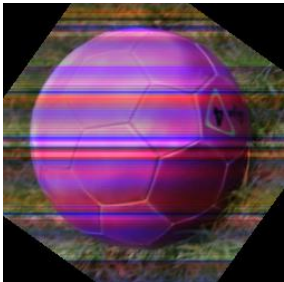
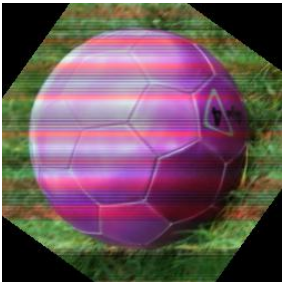
Original - Soccer ball	Pinwheel	Goblet	Helmet
 100%	 96%	 54%	 70%
 98%	 98%	 73%	 58%
 90%	 83%	 32%	 40%
 99%	 88%	 55%	 24%

Figure 2: A random sample of targeted attacks against class - Soccer ball. The attack is robust to viewpoint, distance and small lighting changes. The numbers denote the confidence values for the respective classes.



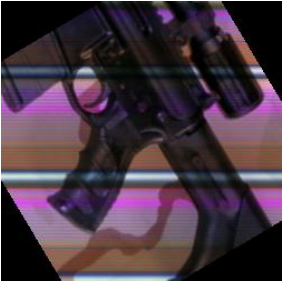













Original - Rifle	Bow	Microphone	Tool kit
			
81%	94%	32%	70%
			
77%	100%	87%	50%
			
66%	98%	56%	72%
			
65%	100%	29%	77%

Figure 3: A random sample of targeted attacks against class - Rifle. The attack is robust to viewpoint, distance and small lighting changes. The numbers denote the confidence values for the respective classes.