

# SPAA: Stealthy Projector-based Adversarial Attacks on Deep Image Classifiers

## - Supplementary Materials -

Bingyao Huang\*

Haibin Ling†

### 1 INTRODUCTION

In this supplementary material, we provide additional quantitative comparisons in § 2. Then, we present more qualitative comparisons of stealthy projector-based adversarial attacks in § 3.

The source code, dataset and experimental results are made publicly available at <https://github.com/BingyaoHuang/SPAA>.

### 2 ADDITIONAL QUANTITATIVE COMPARISONS

In this section, we provide additional quantitative comparisons on different stealthiness loss functions in § 2.1.

#### 2.1 Different stealthiness loss functions

In Tab. 1, as a supplementary of the main paper’s Table 1, we show more SPAA’s projector-based attack results when using *different stealthiness loss functions* (main paper Equation 9). We compare three stealthiness loss functions:  $L_2$ ,  $\Delta E$  and  $\Delta E + L_2$ . (1) For attack success rates (averaged over three classifiers),  $L_2$  has the highest attack success rates when  $d_{\text{thr}} \leq 9$  and  $\Delta E + L_2$  provides the highest attack success rates when  $d_{\text{thr}} > 9$ ; (2) For perturbation sizes (averaged over three classifiers),  $L_2$  gives the largest perturbations for all  $d_{\text{thr}}$ , and  $\Delta E + L_2$  obtains the lowest perturbations when  $d_{\text{thr}} = 5$  and  $\Delta E$  has the lowest perturbations when  $d_{\text{thr}} > 5$ .

### 3 ADDITIONAL QUALITATIVE COMPARISONS

We show more qualitative comparisons as a supplementary of the main paper Figures 4-5. We show more *targeted* projector-based attacks in Fig. 1 to Fig. 13 and *untargeted* attacks in Fig. 14 to Fig. 26. For each figure, the 1<sup>st</sup> to the 3<sup>rd</sup> rows are our SPAA, PerC-AL + CompenNet++ [2, 6] and One-pixel DE [3], respectively. The 1<sup>st</sup> column shows the camera-capture scene under plain gray illumination. The 2<sup>nd</sup> column shows inferred projector input adversarial patterns. The 3<sup>rd</sup> column plots model inferred camera-captured images. The 4<sup>th</sup> column presents real captured scene under adversarial projection *i.e.*, the 2<sup>nd</sup> column projected onto the 1<sup>st</sup> column. The last column provides normalized differences between the 4<sup>th</sup> and 1<sup>st</sup> columns. On the top of each camera-captured image, we show the classifier’s predicted labels and probabilities. For the 2<sup>nd</sup> to 4<sup>th</sup> columns, we also show  $L_2$  norm of perturbations. Note that for One-pixel DE, the 3<sup>rd</sup> column is blank because it is an online method and no inference is available.

### REFERENCES

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [2] Bingyao Huang and Haibin Ling. Compennet++: End-to-end full projector compensation. In *ICCV*, 2019.

\*College of Computer and Information Science, Southwest University, Chongqing, China. E-mail: bhuang@swu.edu.cn

†Department of Computer Science, Stony Brook University, Stony Brook, NY 11794, USA. E-mail: hling@cs.stonybrook.edu

- [3] Nicole Nichols and Robert Jasper. Projecting trouble: Light based adversarial attacks on deep learning classifiers. In *AAAI Fall Symposium: ALEC*, 2018.
- [4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.
- [5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.
- [6] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *CVPR*, pages 1039–1048, 2020.

Table 1: Quantitative comparison of **different stealthiness loss functions** and perturbation thresholds of our SPAA. Three classifiers, *i.e.*, Inception v3 [5], ResNet-18 [1] and VGG-16 [4] are attacked, and the results are averaged on 13 setups. The four big sections show our SPAA results with different thresholds for perturbation size  $d_{\text{thr}}$  and stealthiness loss as mentioned in the main paper Alg. 1. The 4<sup>th</sup> to 6<sup>th</sup> columns are targeted (T) and untargeted (U) attack success rates, and the last four columns are stealthiness metrics.

$d_{\text{thr}}$	Stealthiness loss	Classifier	T. top-1 (%)	T. top-5 (%)	U. top-1 (%)	$L_2 \downarrow$	$L_\infty \downarrow$	$\Delta E \downarrow$	SSIM $\uparrow$
$d_{\text{thr}} = 5$	$L_2$	Inception v3	41.54	67.69	84.62	6.273	5.101	2.588	0.937
		ResNet-18	73.08	90.00	100.00	6.304	5.158	2.701	0.940
		VGG-16	69.23	83.85	100.00	6.629	5.428	2.824	0.934
		Average	<b>61.28</b>	<b>80.51</b>	<b>94.87</b>	<b>6.402</b>	<b>5.229</b>	<b>2.704</b>	<b>0.937</b>
	$\Delta E$	Inception v3	32.31	65.38	76.92	5.951	4.768	2.236	0.944
		ResNet-18	57.69	79.23	92.31	5.828	4.698	2.269	0.949
		VGG-16	46.92	79.23	92.31	6.464	5.215	2.493	0.938
		Average	<b>45.64</b>	<b>74.62</b>	<b>87.18</b>	<b>6.081</b>	<b>4.893</b>	<b>2.333</b>	<b>0.944</b>
	$\Delta E + L_2$	Inception v3	33.85	65.38	69.23	6.021	4.832	2.282	0.942
		ResNet-18	54.62	76.92	92.31	5.842	4.709	2.280	0.950
		VGG-16	52.31	76.92	92.31	6.243	5.028	2.407	0.941
		Average	<b>46.92</b>	<b>73.08</b>	<b>84.62</b>	<b>6.036</b>	<b>4.856</b>	<b>2.323</b>	<b>0.944</b>
$d_{\text{thr}} = 7$	$L_2$	Inception v3	67.69	84.62	100.00	7.603	6.199	3.135	0.904
		ResNet-18	92.31	94.62	100.00	7.786	6.396	3.349	0.907
		VGG-16	83.08	97.69	100.00	8.117	6.668	3.435	0.899
		Average	<b>81.03</b>	<b>92.31</b>	<b>100.00</b>	<b>7.835</b>	<b>6.421</b>	<b>3.306</b>	<b>0.903</b>
	$\Delta E$	Inception v3	53.08	83.08	92.31	7.272	5.806	2.586	0.913
		ResNet-18	88.46	93.08	100.00	7.426	5.946	2.686	0.913
		VGG-16	80.00	93.85	100.00	7.755	6.219	2.818	0.906
		Average	<b>73.85</b>	<b>90.00</b>	<b>97.44</b>	<b>7.484</b>	<b>5.990</b>	<b>2.697</b>	<b>0.911</b>
	$\Delta E + L_2$	Inception v3	56.15	80.77	92.31	7.285	5.826	2.612	0.913
		ResNet-18	90.77	94.62	100.00	7.381	5.914	2.681	0.914
		VGG-16	80.77	94.62	100.00	7.849	6.306	2.862	0.903
		Average	<b>75.90</b>	<b>90.00</b>	<b>97.44</b>	<b>7.505</b>	<b>6.015</b>	<b>2.718</b>	<b>0.910</b>
$d_{\text{thr}} = 9$	$L_2$	Inception v3	76.15	90.00	100.00	9.336	7.620	3.766	0.872
		ResNet-18	95.38	98.46	100.00	9.640	7.923	4.066	0.874
		VGG-16	90.00	99.23	100.00	9.978	8.211	4.156	0.864
		Average	<b>87.18</b>	<b>95.90</b>	<b>100.00</b>	<b>9.651</b>	<b>7.918</b>	<b>3.996</b>	<b>0.870</b>
	$\Delta E$	Inception v3	75.38	90.77	100.00	9.100	7.269	3.134	0.877
		ResNet-18	94.62	96.92	100.00	9.300	7.435	3.250	0.878
		VGG-16	88.46	99.23	100.00	9.526	7.630	3.351	0.871
		Average	<b>86.15</b>	<b>95.64</b>	<b>100.00</b>	<b>9.309</b>	<b>7.444</b>	<b>3.245</b>	<b>0.875</b>
	$\Delta E + L_2$	Inception v3	71.54	90.00	100.00	9.112	7.282	3.149	0.877
		ResNet-18	94.62	97.69	100.00	9.263	7.412	3.249	0.879
		VGG-16	90.77	100.00	100.00	9.763	7.832	3.448	0.867
		Average	<b>85.64</b>	<b>95.90</b>	<b>100.00</b>	<b>9.379</b>	<b>7.509</b>	<b>3.282</b>	<b>0.874</b>
$d_{\text{thr}} = 11$	$L_2$	Inception v3	76.92	92.31	100.00	11.190	9.156	4.386	0.843
		ResNet-18	97.69	100.00	100.00	11.605	9.545	4.785	0.846
		VGG-16	94.62	99.23	100.00	11.750	9.671	4.784	0.835
		Average	<b>89.74</b>	<b>97.18</b>	<b>100.00</b>	<b>11.515</b>	<b>9.457</b>	<b>4.652</b>	<b>0.841</b>
	$\Delta E$	Inception v3	80.77	92.31	100.00	11.044	8.921	3.909	0.845
		ResNet-18	96.15	100.00	100.00	11.392	9.176	4.058	0.848
		VGG-16	93.08	100.00	100.00	11.625	9.373	4.127	0.837
		Average	<b>90.00</b>	<b>97.44</b>	<b>100.00</b>	<b>11.353</b>	<b>9.157</b>	<b>4.031</b>	<b>0.843</b>
	$\Delta E + L_2$	Inception v3	82.31	93.08	100.00	11.046	8.927	3.921	0.845
		ResNet-18	95.38	100.00	100.00	11.361	9.157	4.059	0.847
		VGG-16	93.85	100.00	100.00	11.742	9.477	4.181	0.835
		Average	<b>90.51</b>	<b>97.69</b>	<b>100.00</b>	<b>11.383</b>	<b>9.187</b>	<b>4.054</b>	<b>0.842</b>

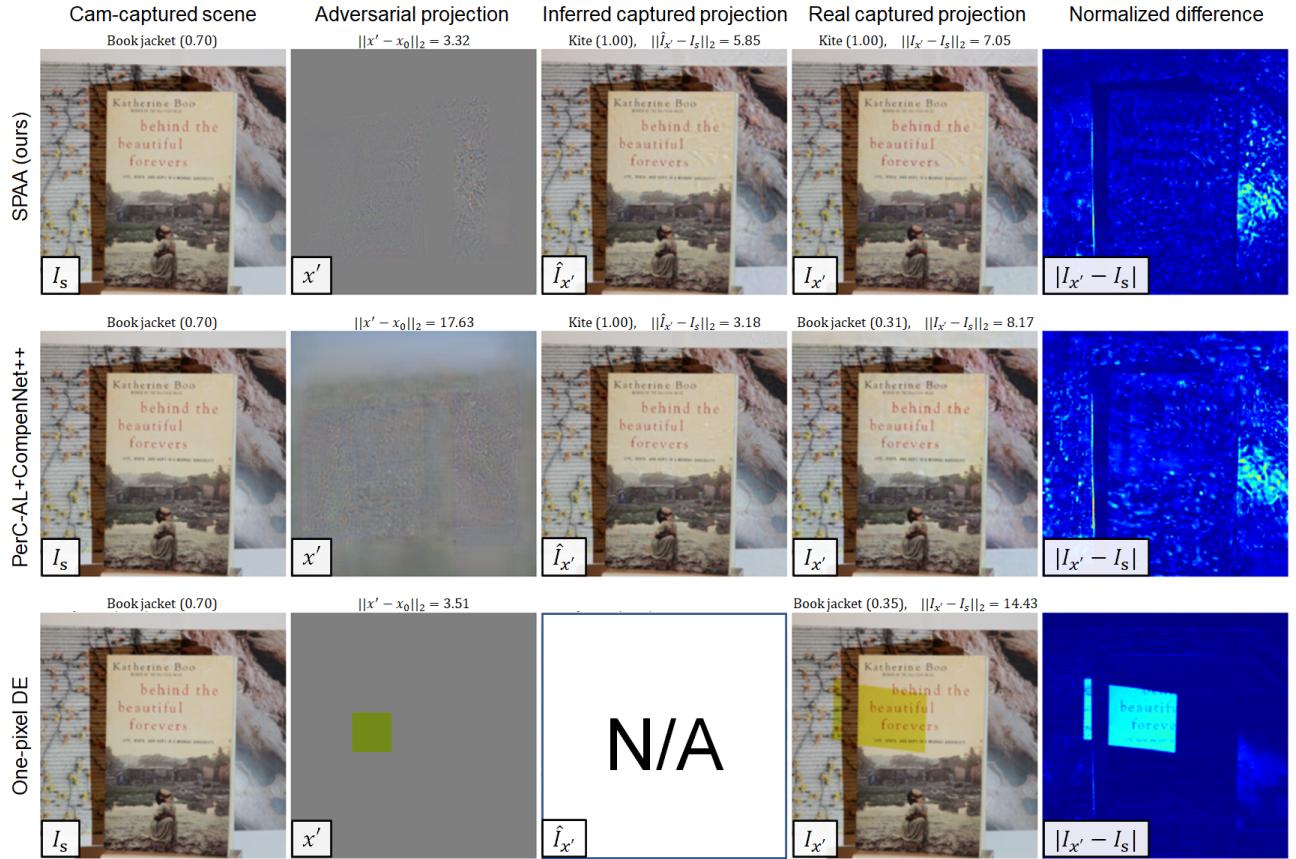


Figure 1: **Targeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection as kite.

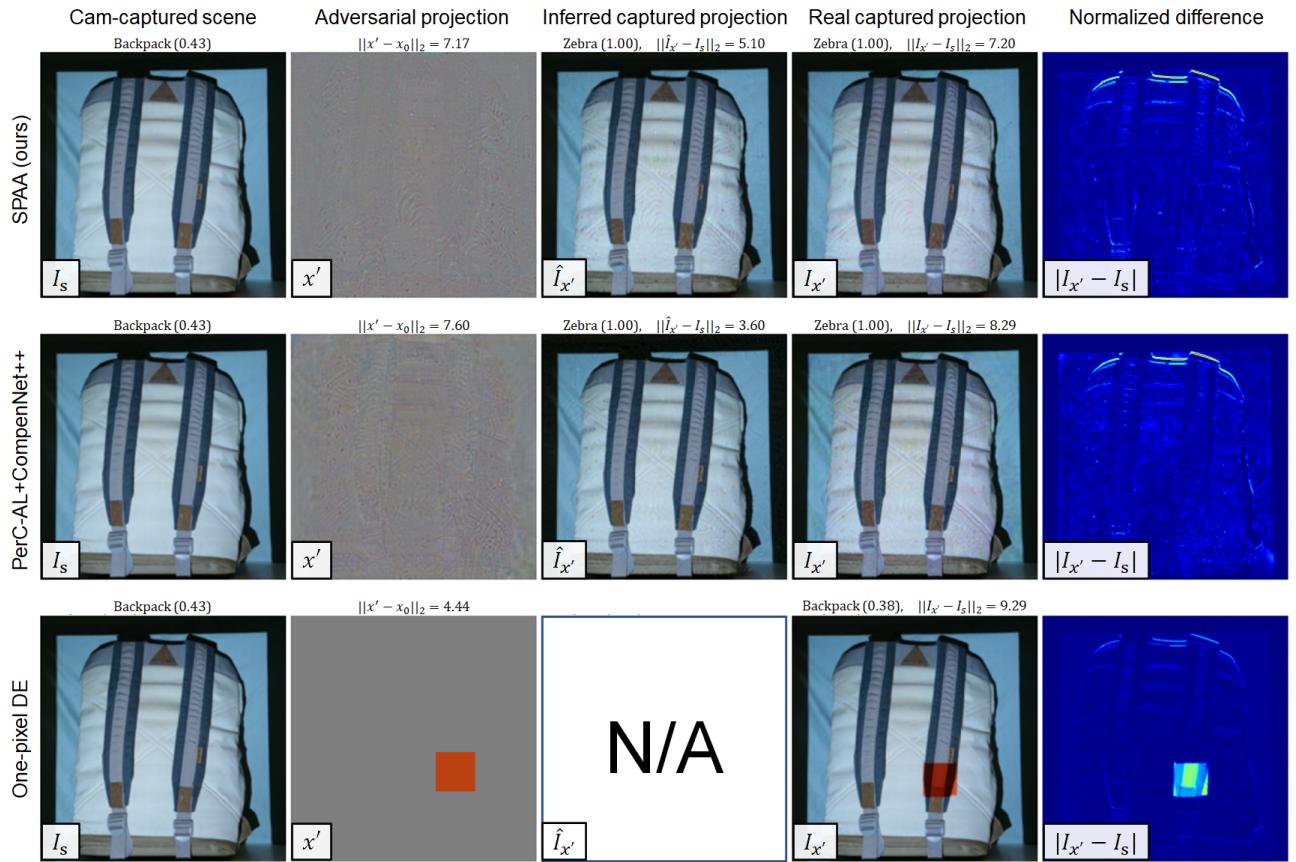


Figure 2: **Targeted** projector-based adversarial attack on ResNet-18. The goal is to cause the classifier to misclassify the captured projection as zebra.

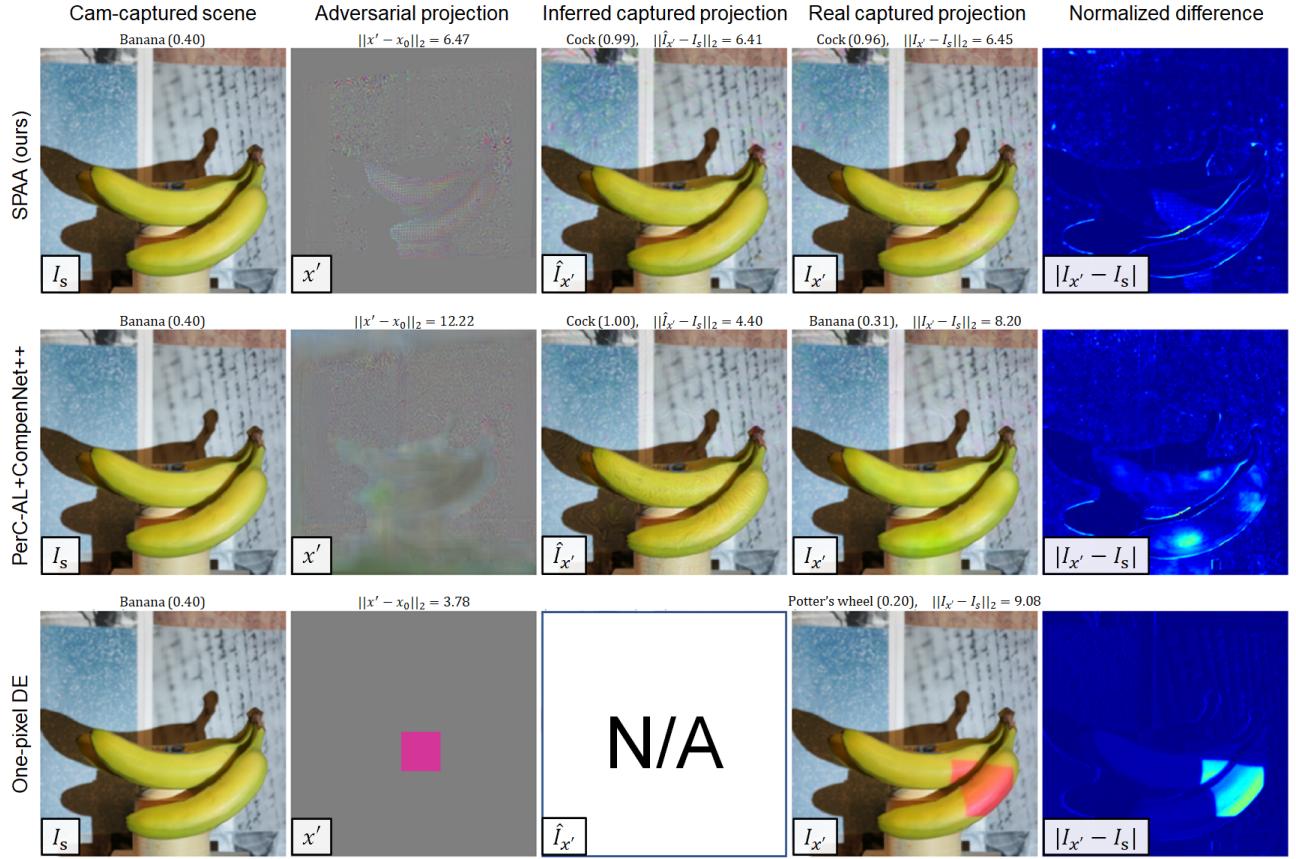


Figure 3: **Targeted** projector-based adversarial attack on VGG-16. The goal is to cause the classifier to misclassify the captured projection as **cock**.

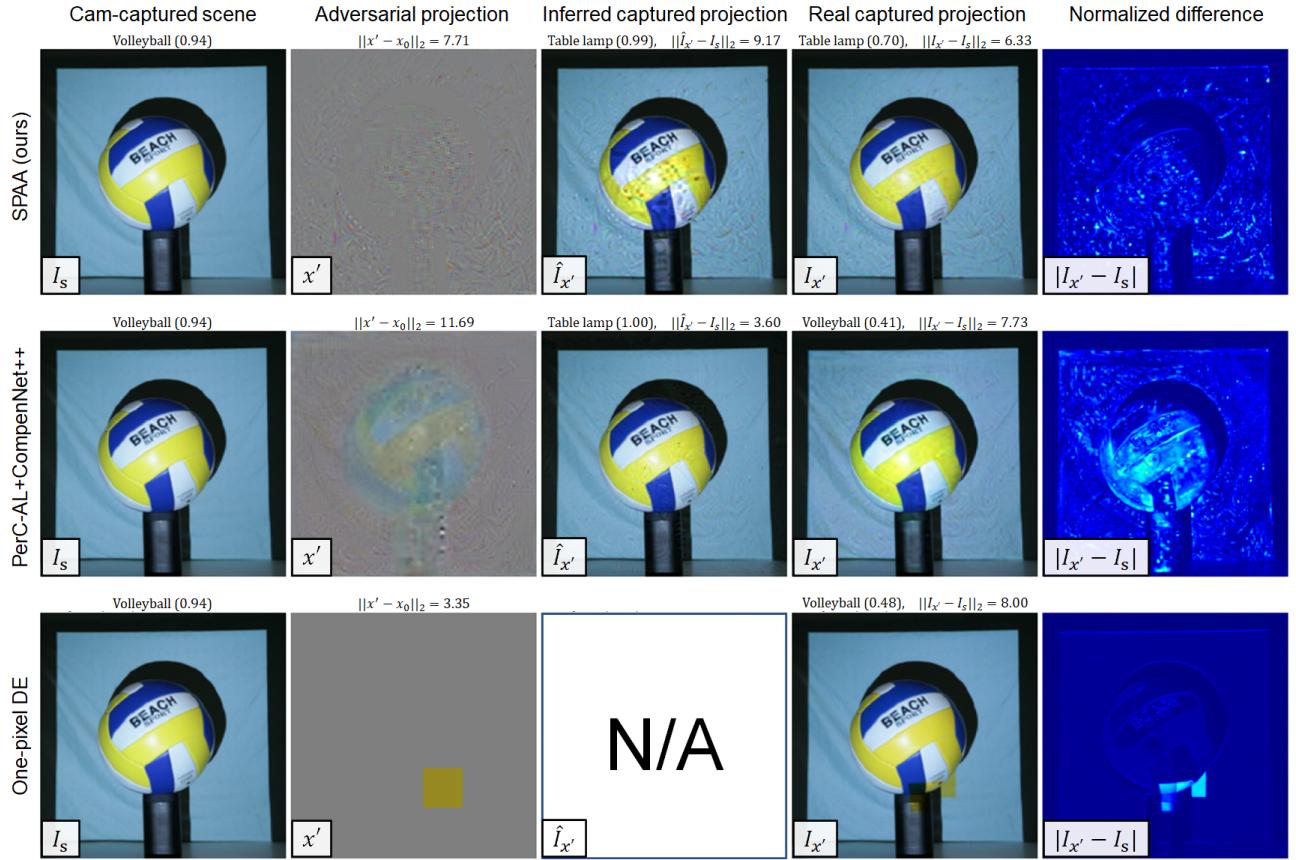


Figure 4: **Targeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection as **table lamp**.

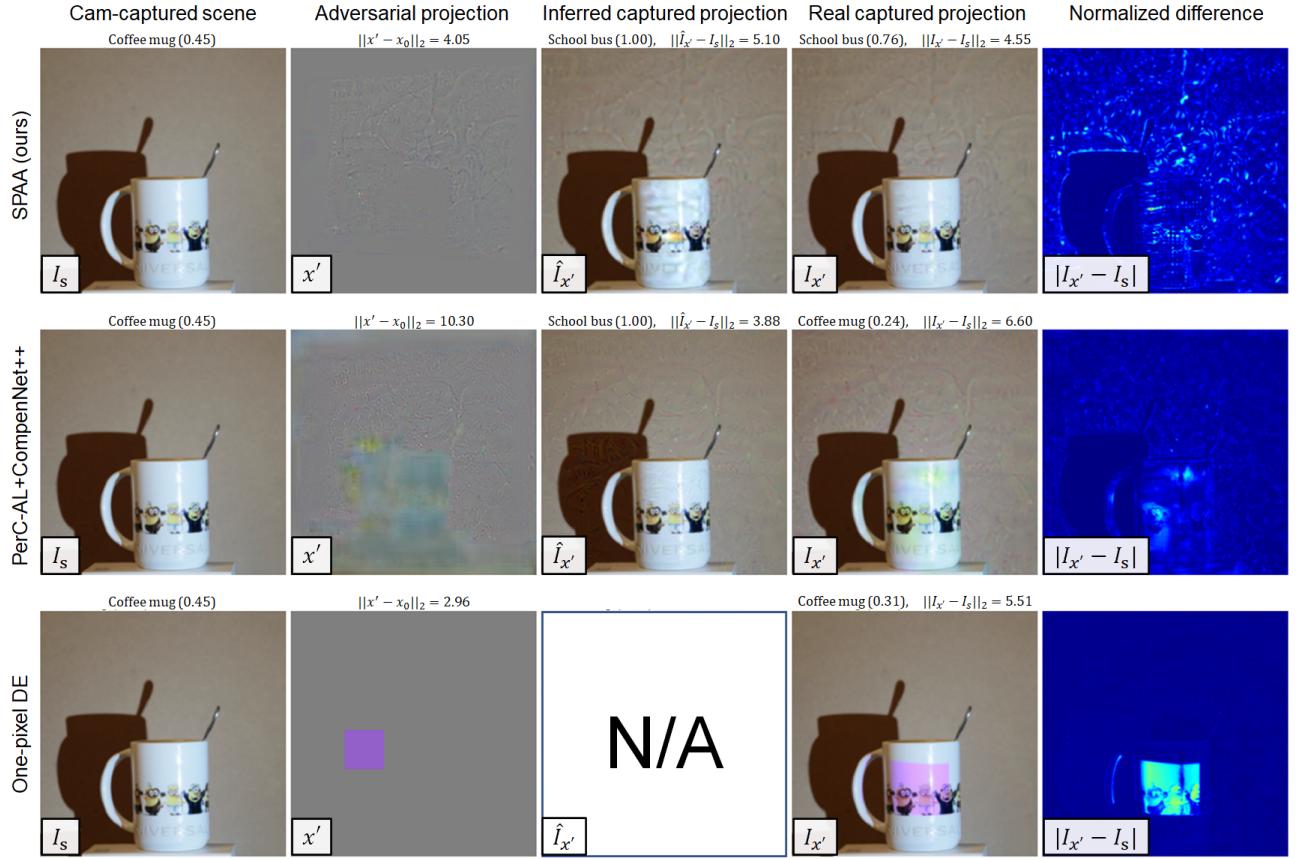


Figure 5: Targeted projector-based adversarial attack on ResNet-18. The goal is to cause the classifier to misclassify the captured projection as school bus.

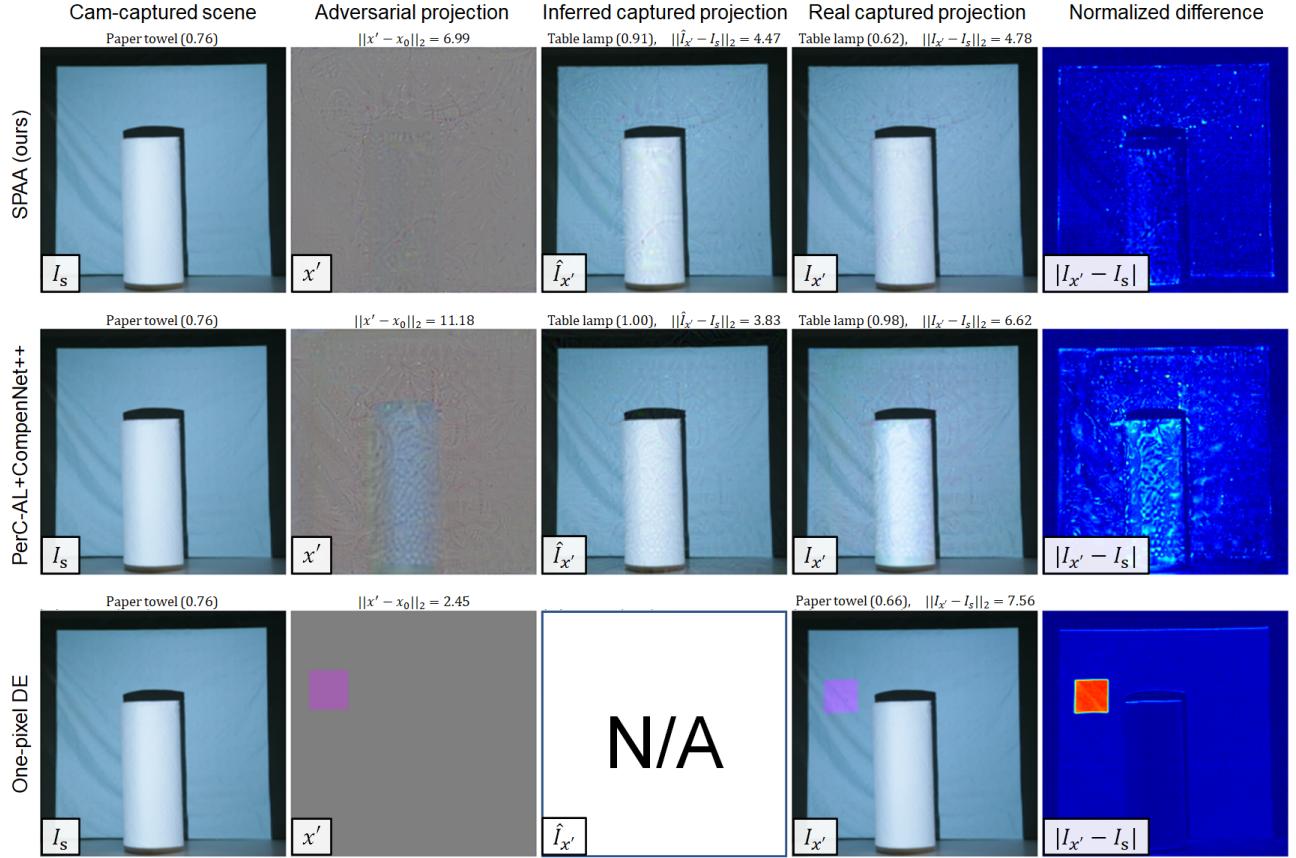


Figure 6: Targeted projector-based adversarial attack on VGG-16. The goal is to cause the classifier to misclassify the captured projection as table lamp.

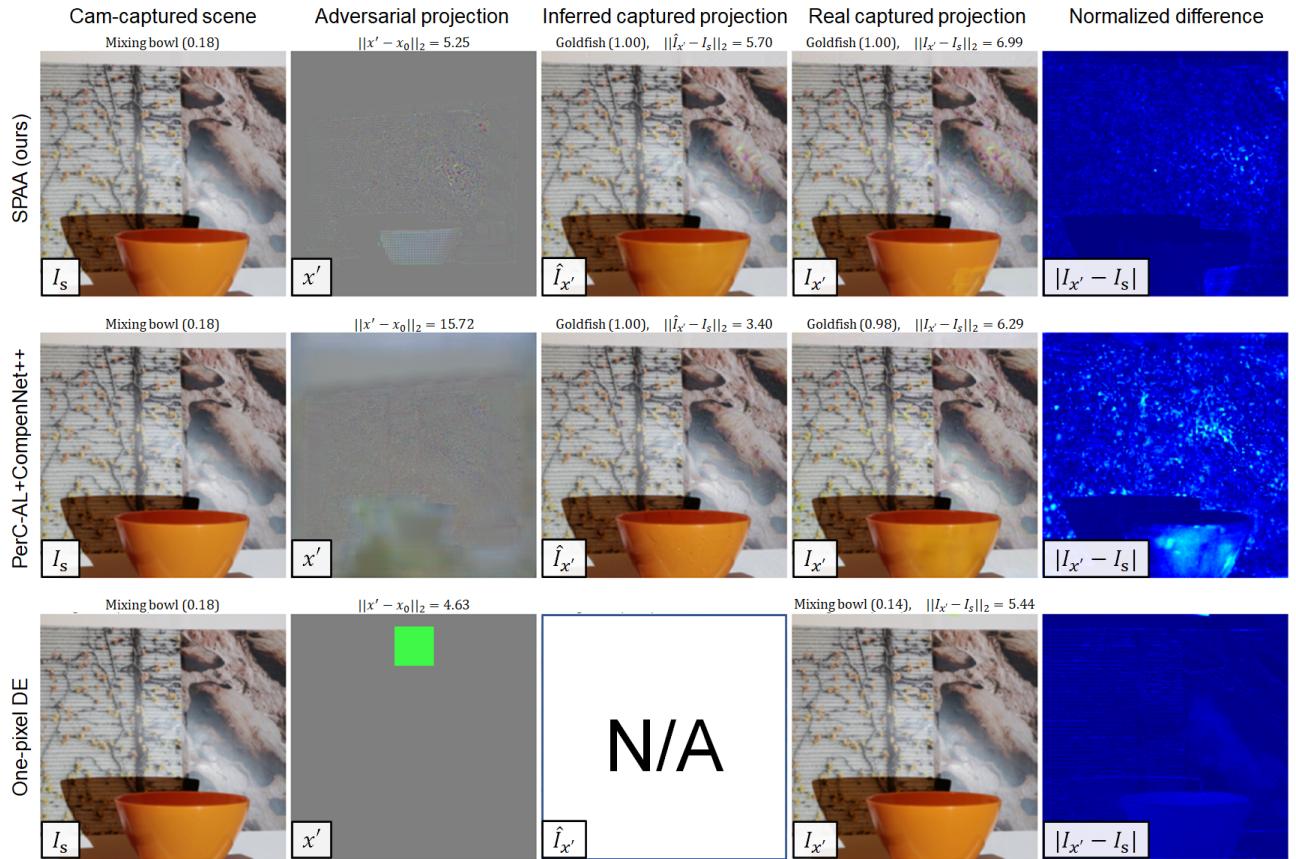


Figure 7: **Targeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection as **goldfish**.

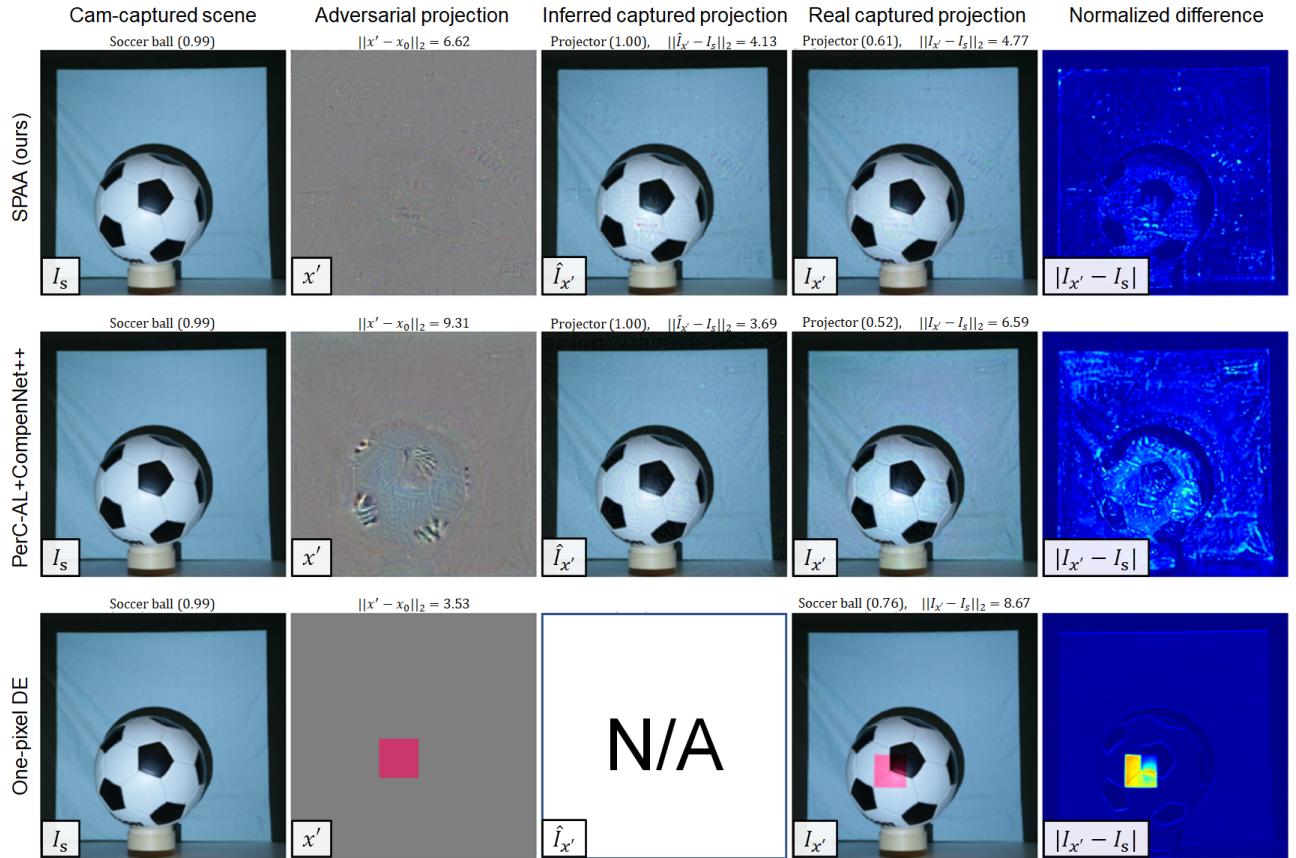


Figure 8: **Targeted** projector-based adversarial attack on ResNet-18. The goal is to cause the classifier to misclassify the captured projection as **projector**.

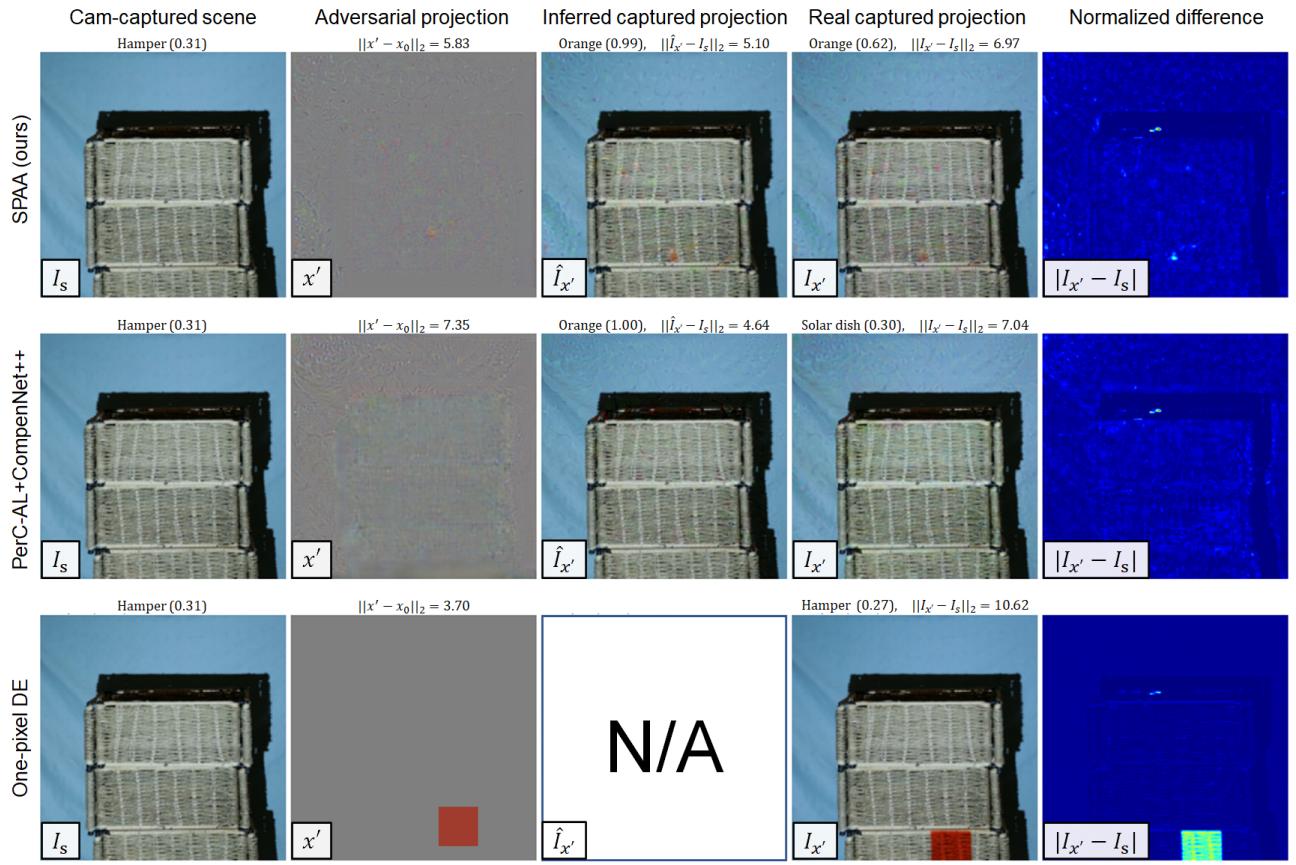


Figure 9: **Targeted** projector-based adversarial attack on VGG-16. The goal is to cause the classifier to misclassify the captured projection as **orange**.

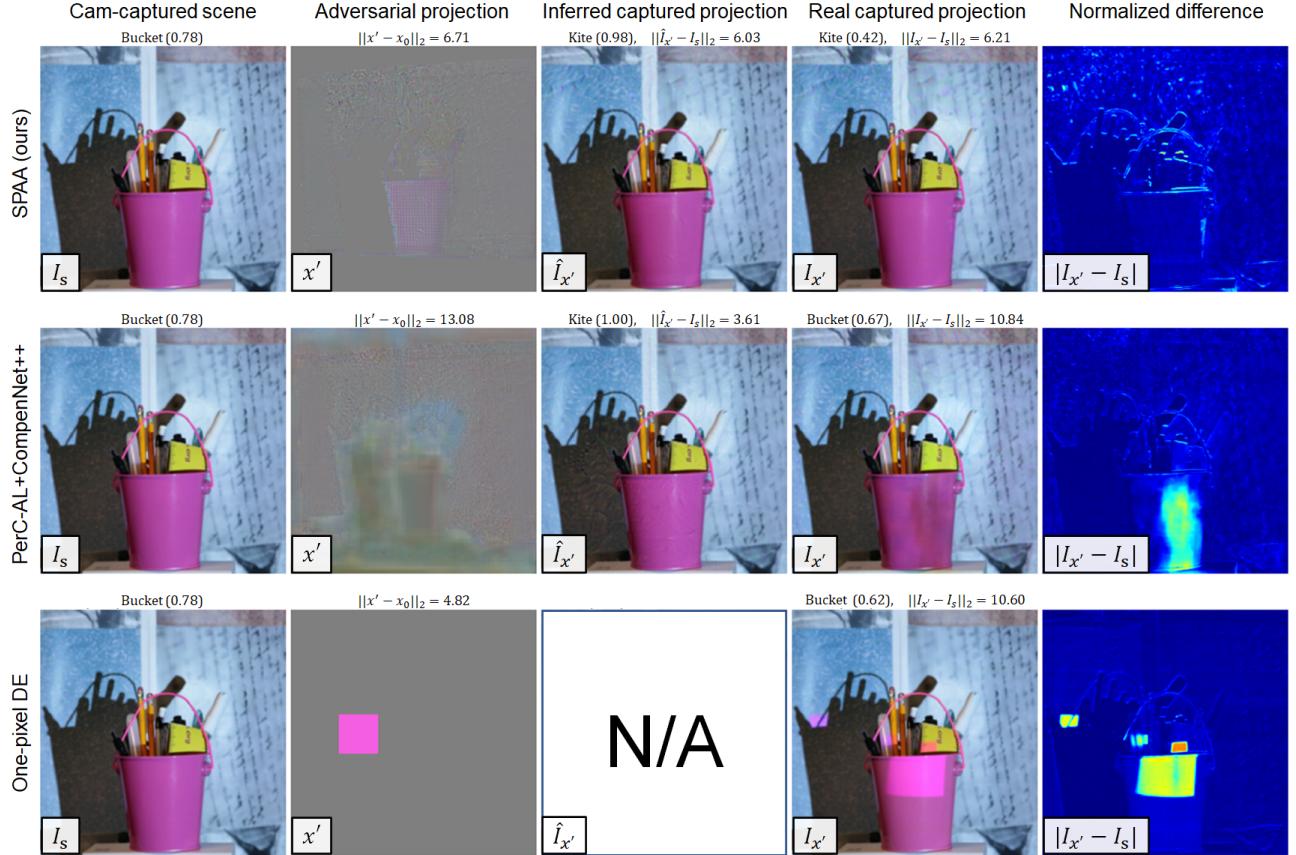


Figure 10: **Targeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection as **kite**.

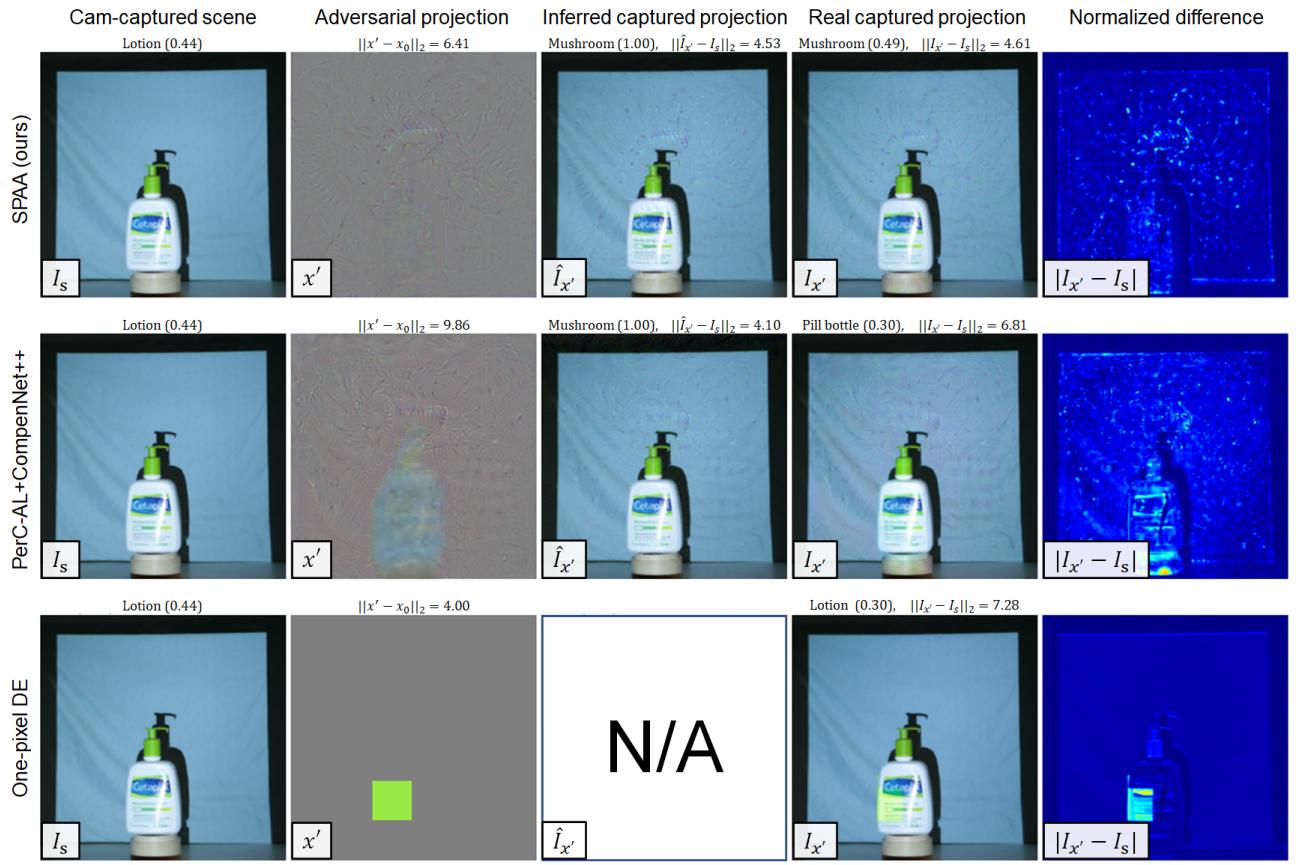


Figure 11: **Targeted projector-based adversarial attack on ResNet-18.** The goal is to cause the classifier to misclassify the captured projection as **mushroom**.

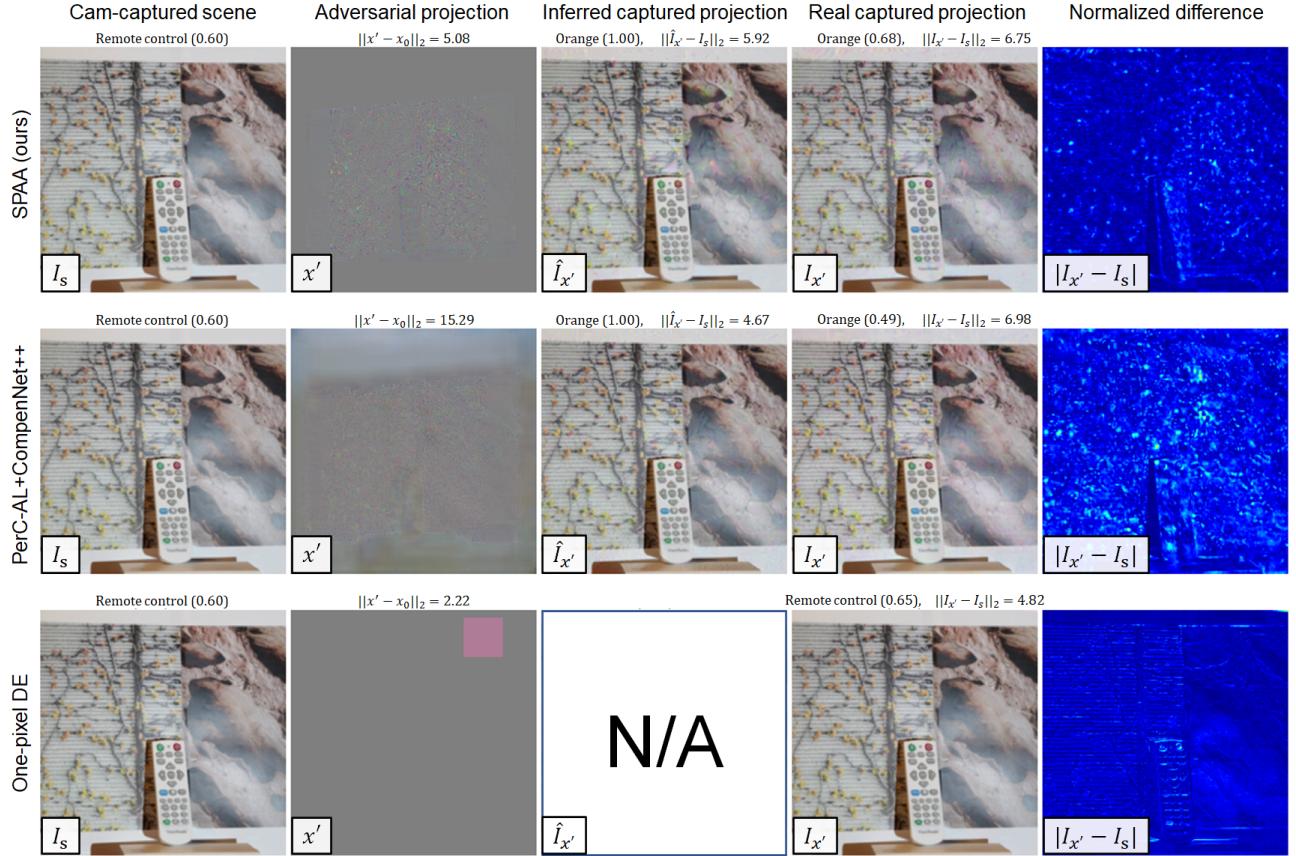


Figure 12: **Targeted projector-based adversarial attack on ResNet-18.** The goal is to cause the classifier to misclassify the captured projection as **orange**.

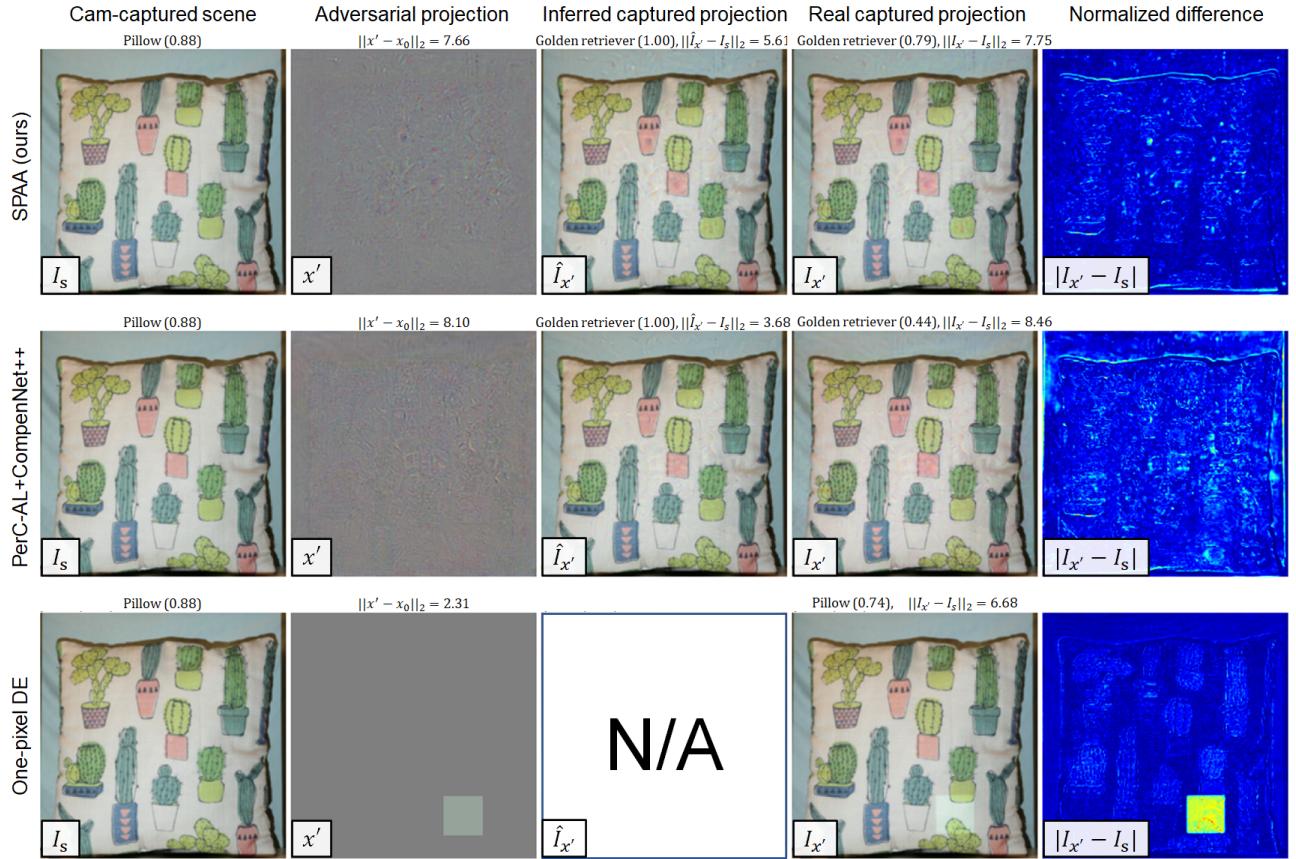


Figure 13: **Targeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection as **golden retriever**.

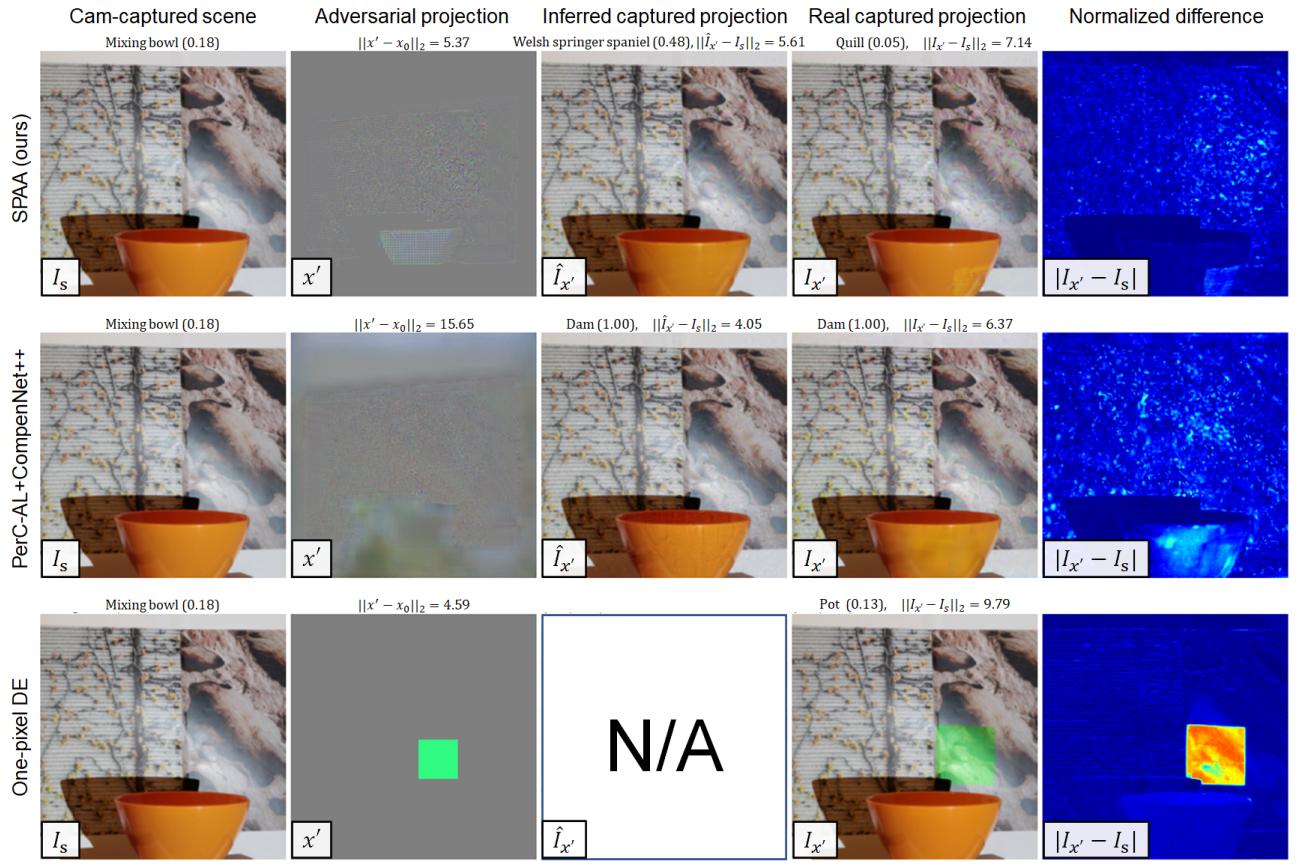


Figure 14: **Untargeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT mixing bowl**.

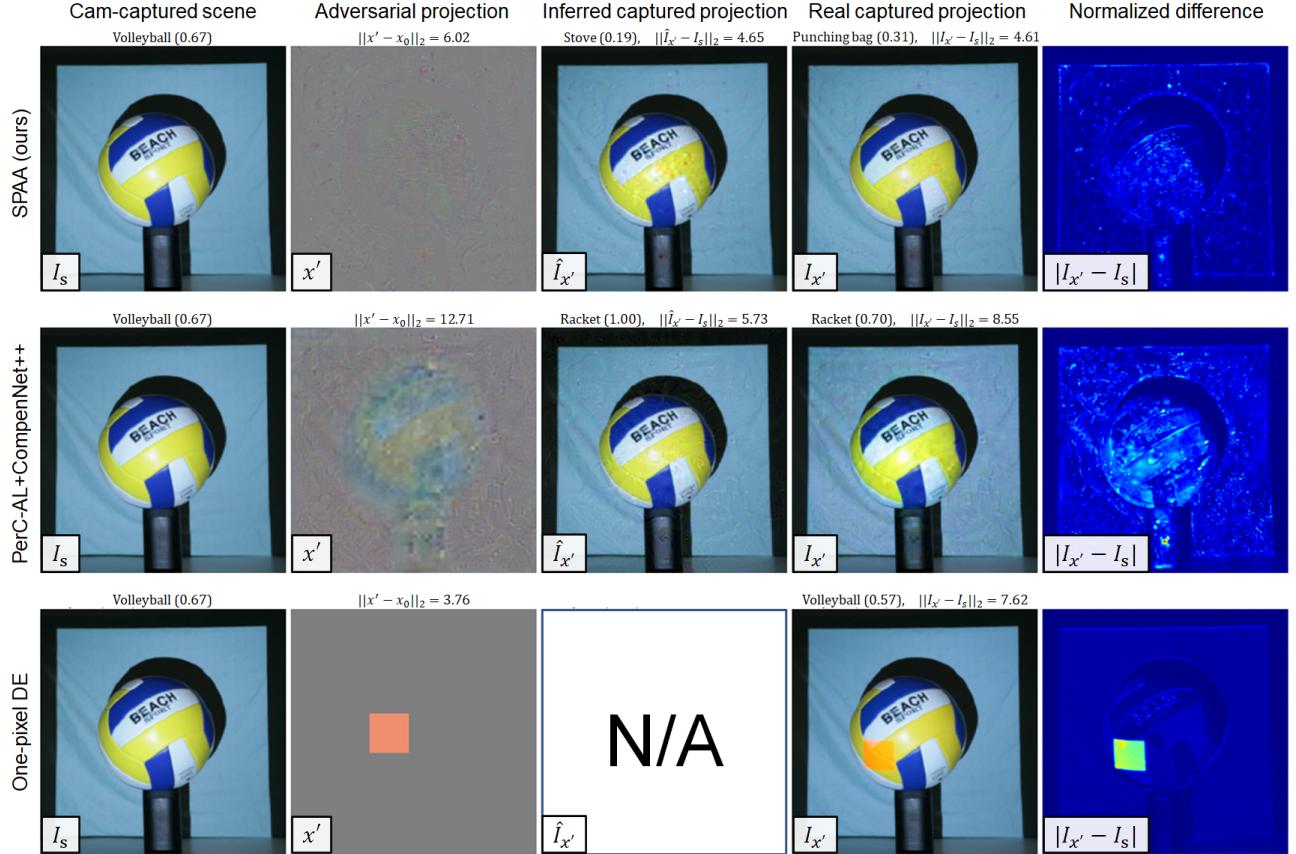


Figure 15: **Untargeted** projector-based adversarial attack on ResNet-18. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT volleyball**.

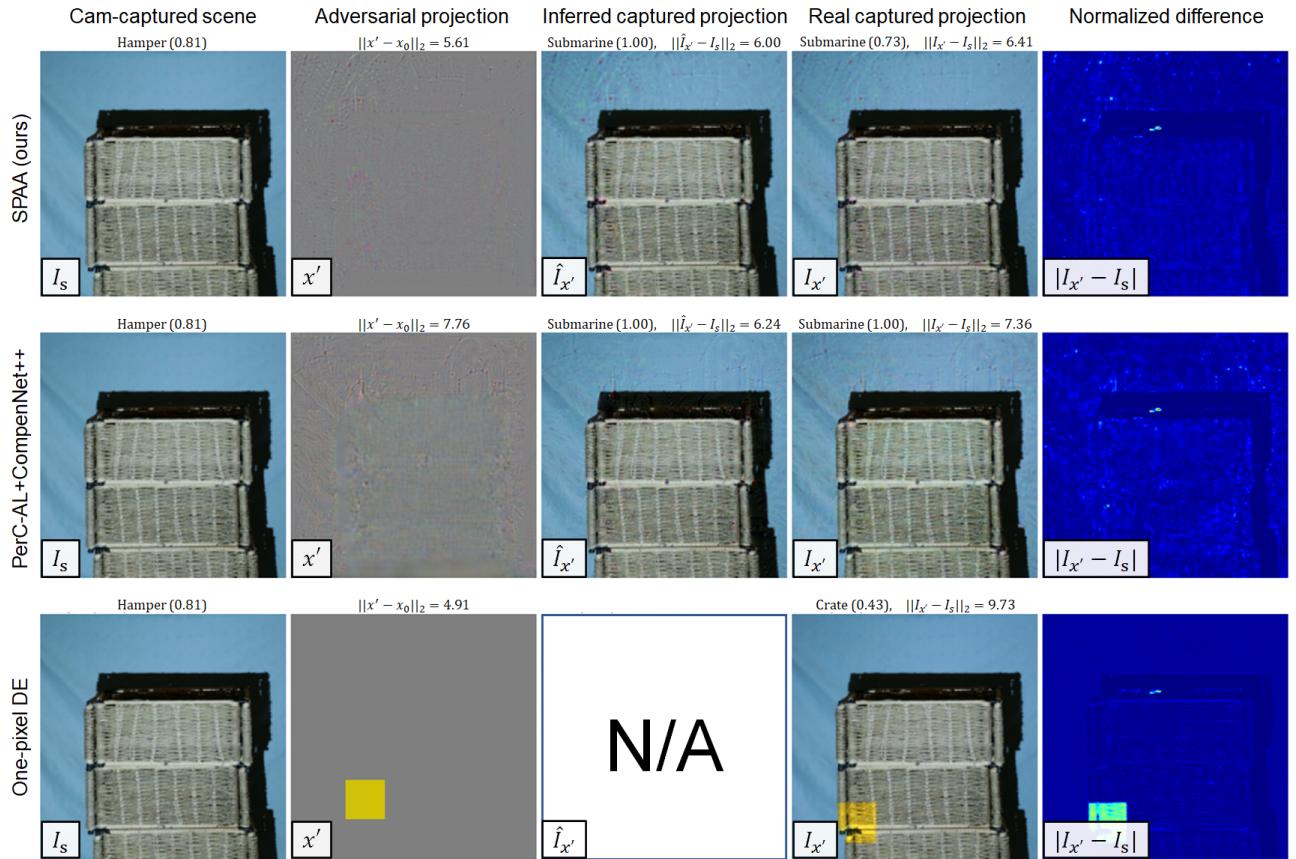


Figure 16: **Untargeted** projector-based adversarial attack on VGG-16. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT hamper**.

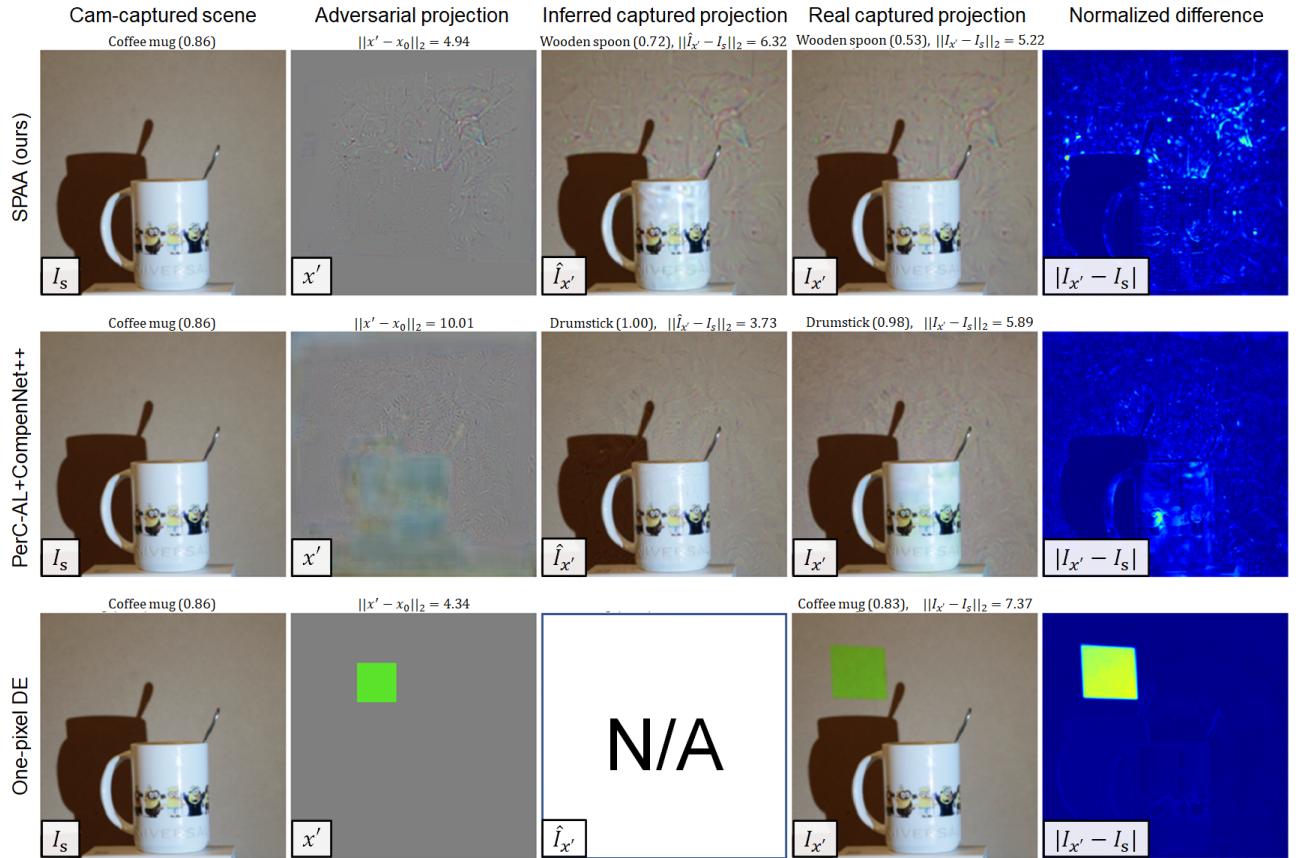


Figure 17: **Untargeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT coffee mug**.

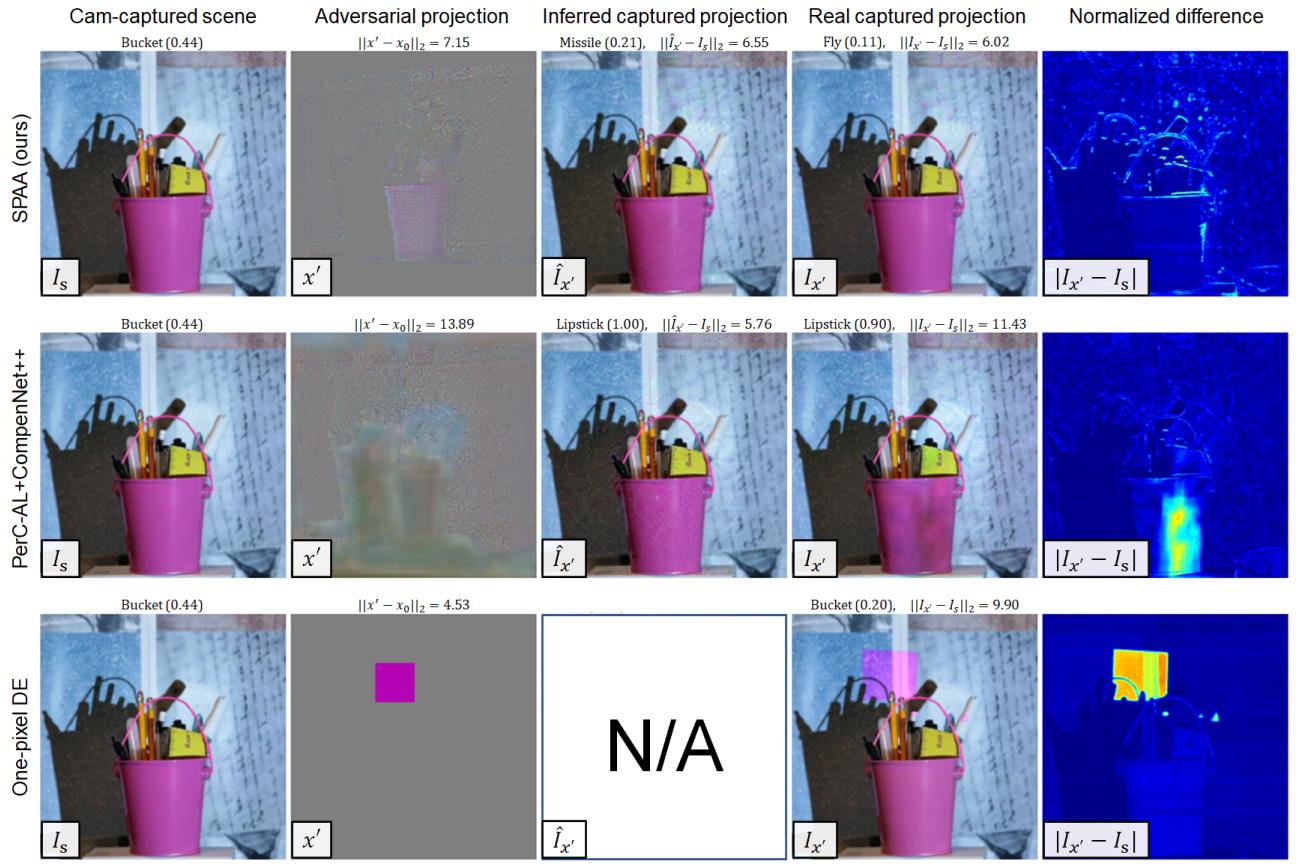


Figure 18: **Untargeted** projector-based adversarial attack on ResNet-18. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT bucket**.

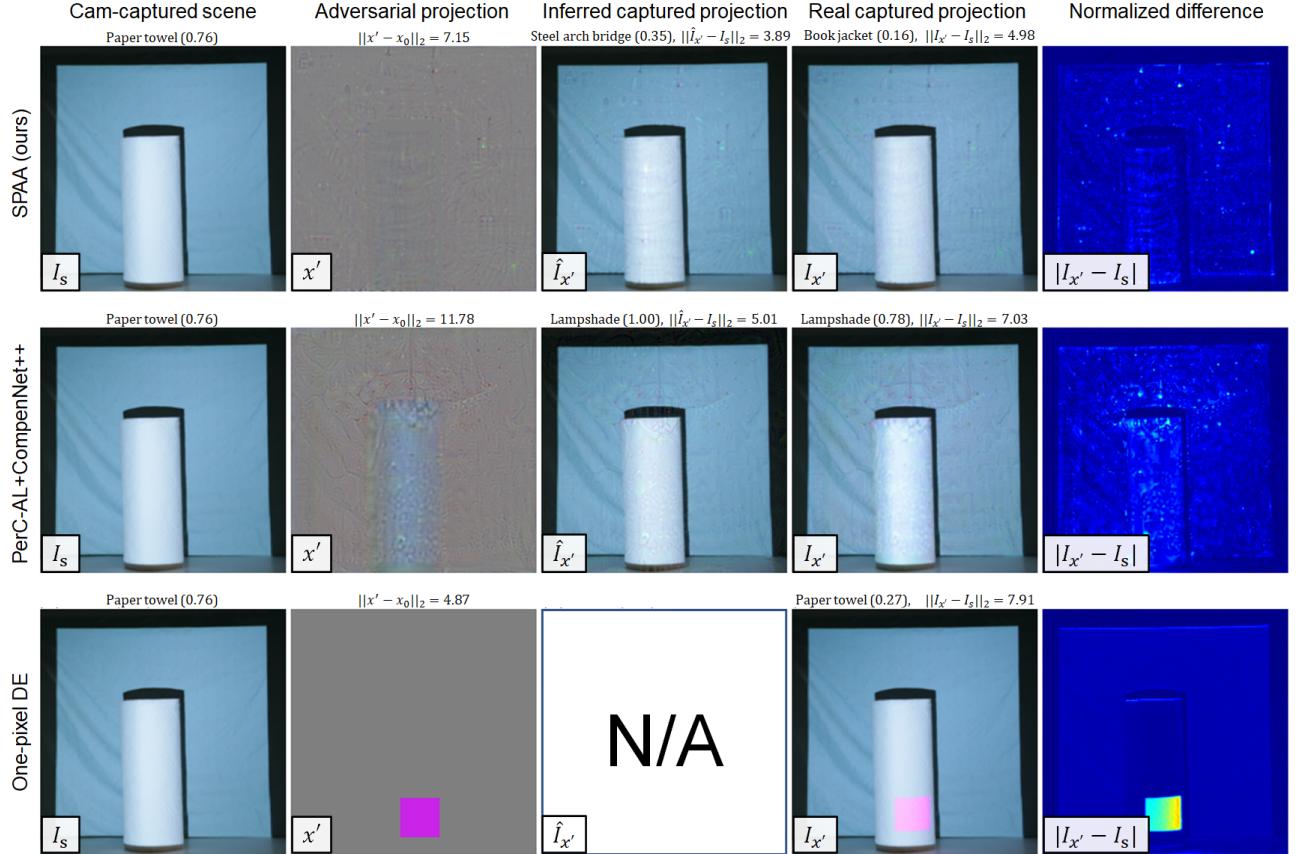


Figure 19: **Untargeted** projector-based adversarial attack on VGG-16. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT paper towel**.

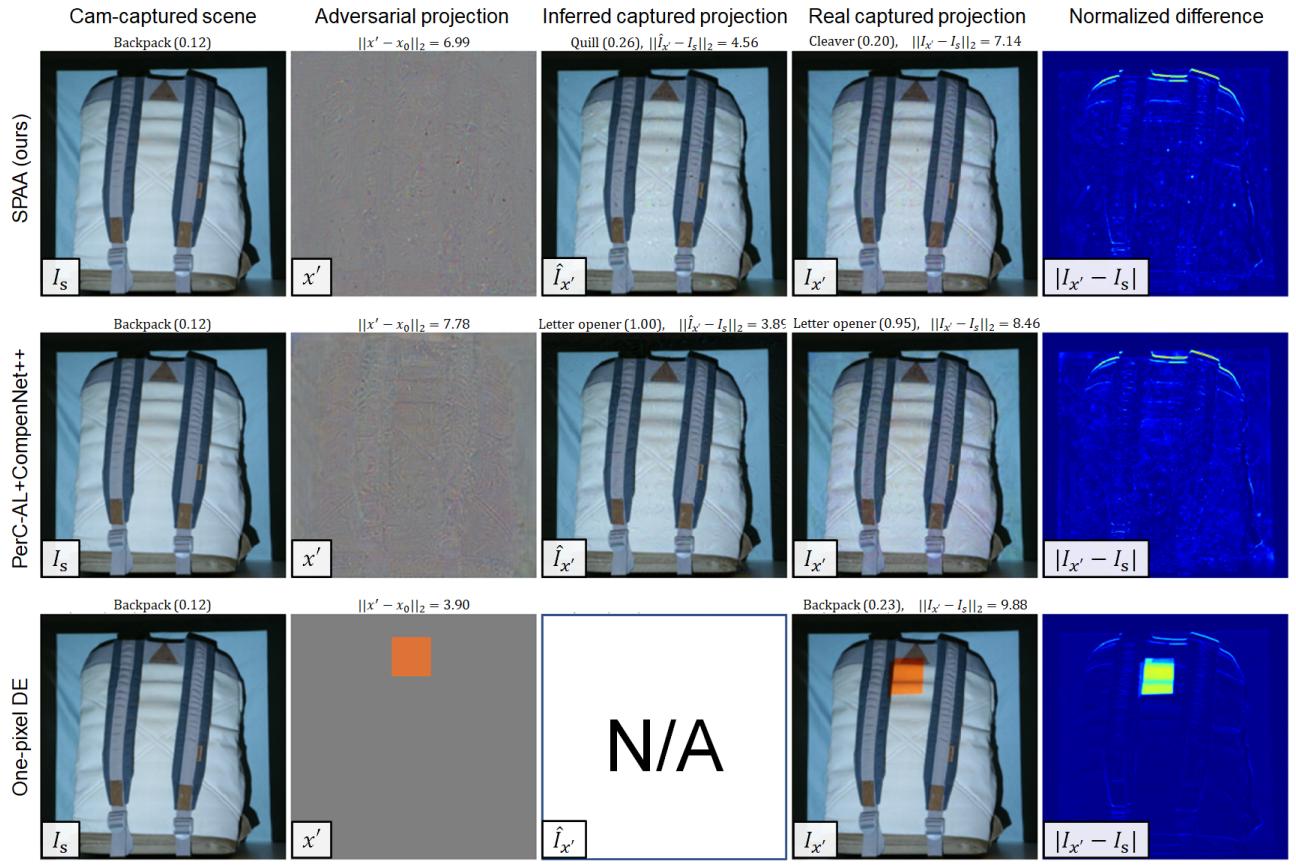


Figure 20: **Untargeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT backpack**.

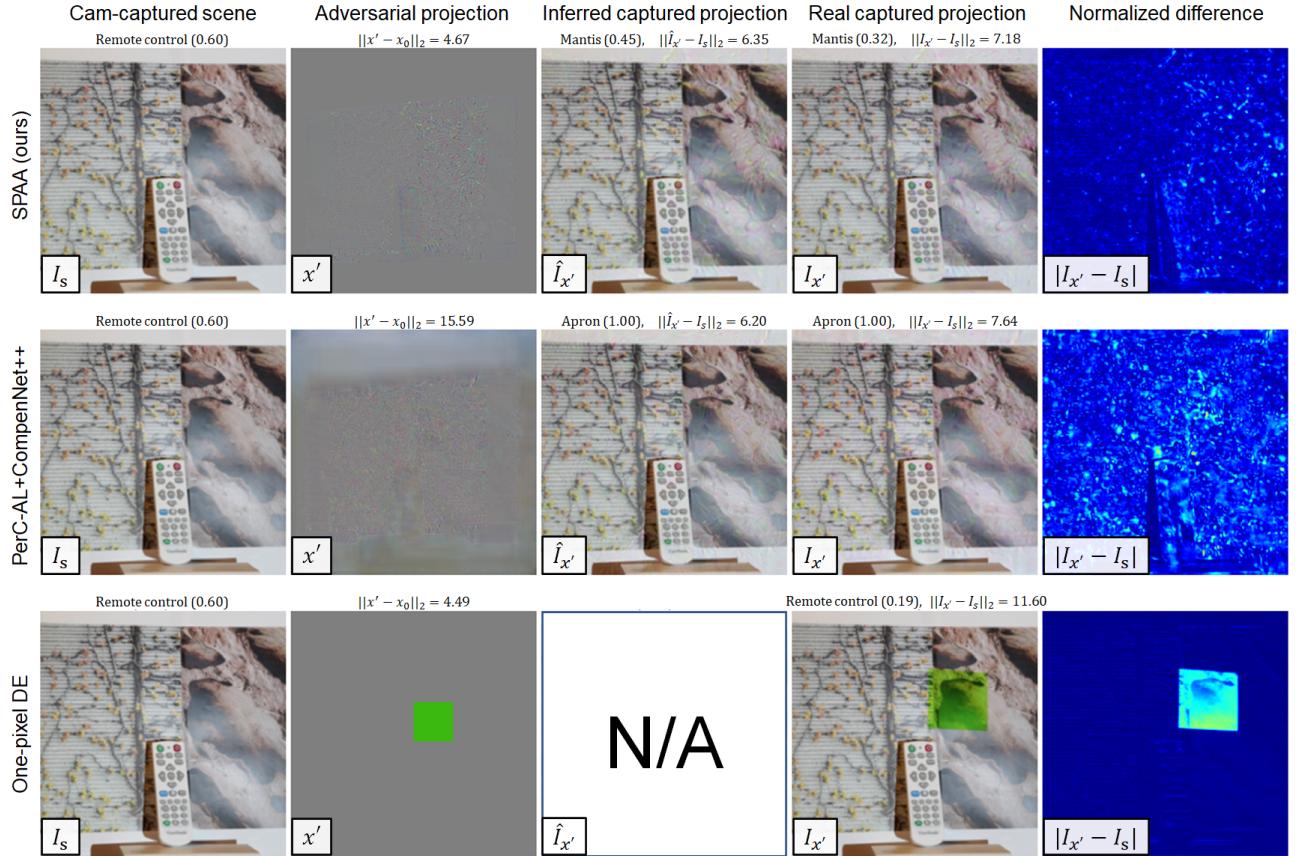


Figure 21: **Untargeted** projector-based adversarial attack on ResNet-18. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT remote control**.

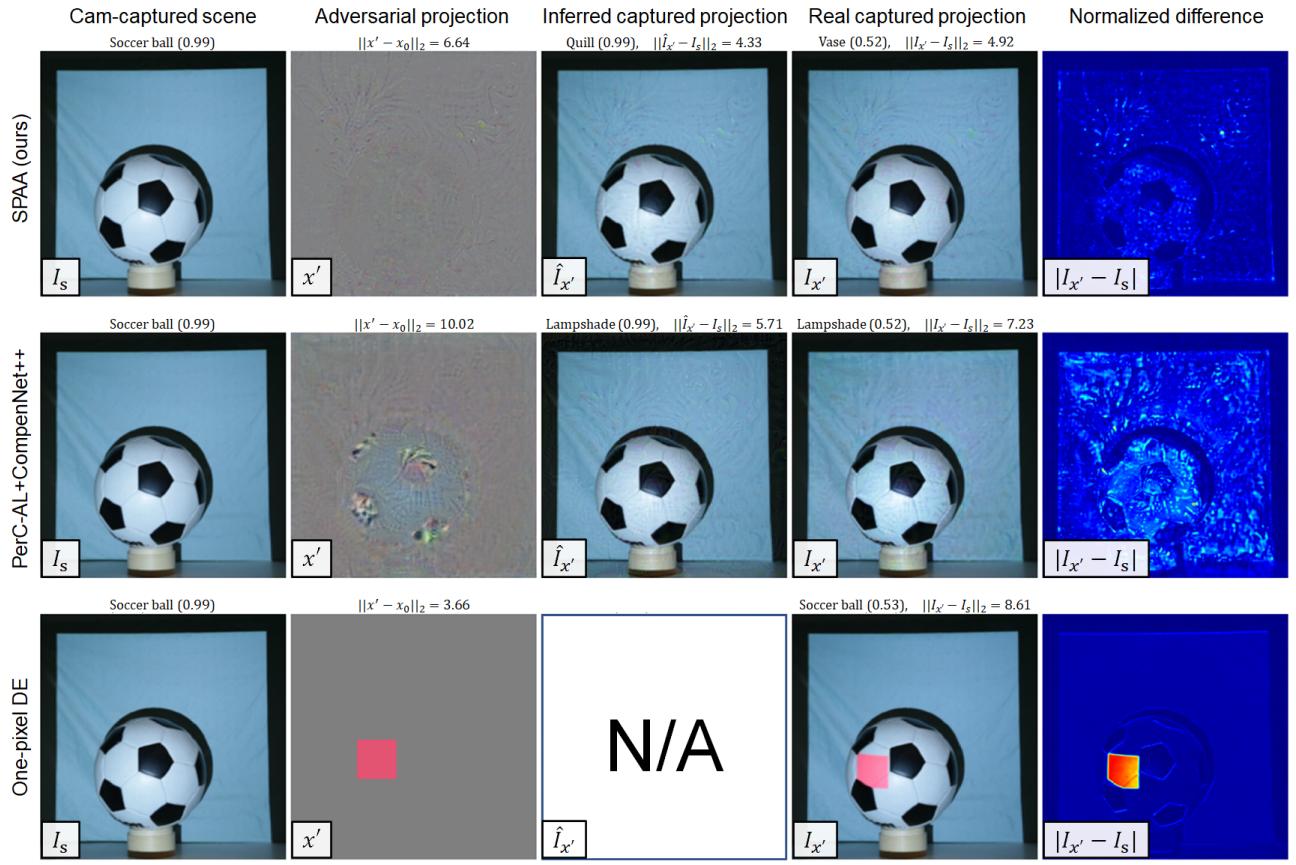


Figure 22: **Untargeted** projector-based adversarial attack on VGG-16. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT soccer ball**.

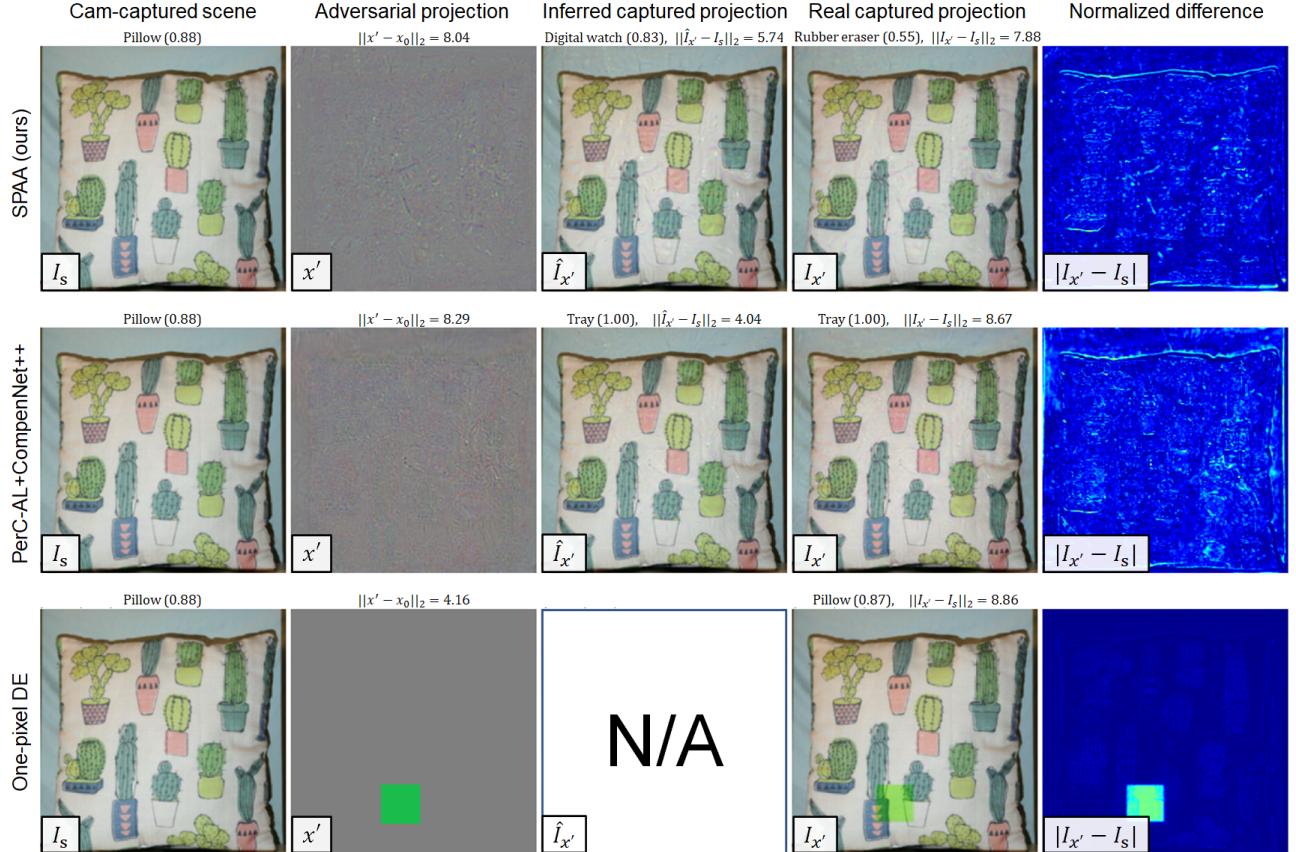


Figure 23: **Untargeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT pillow**.

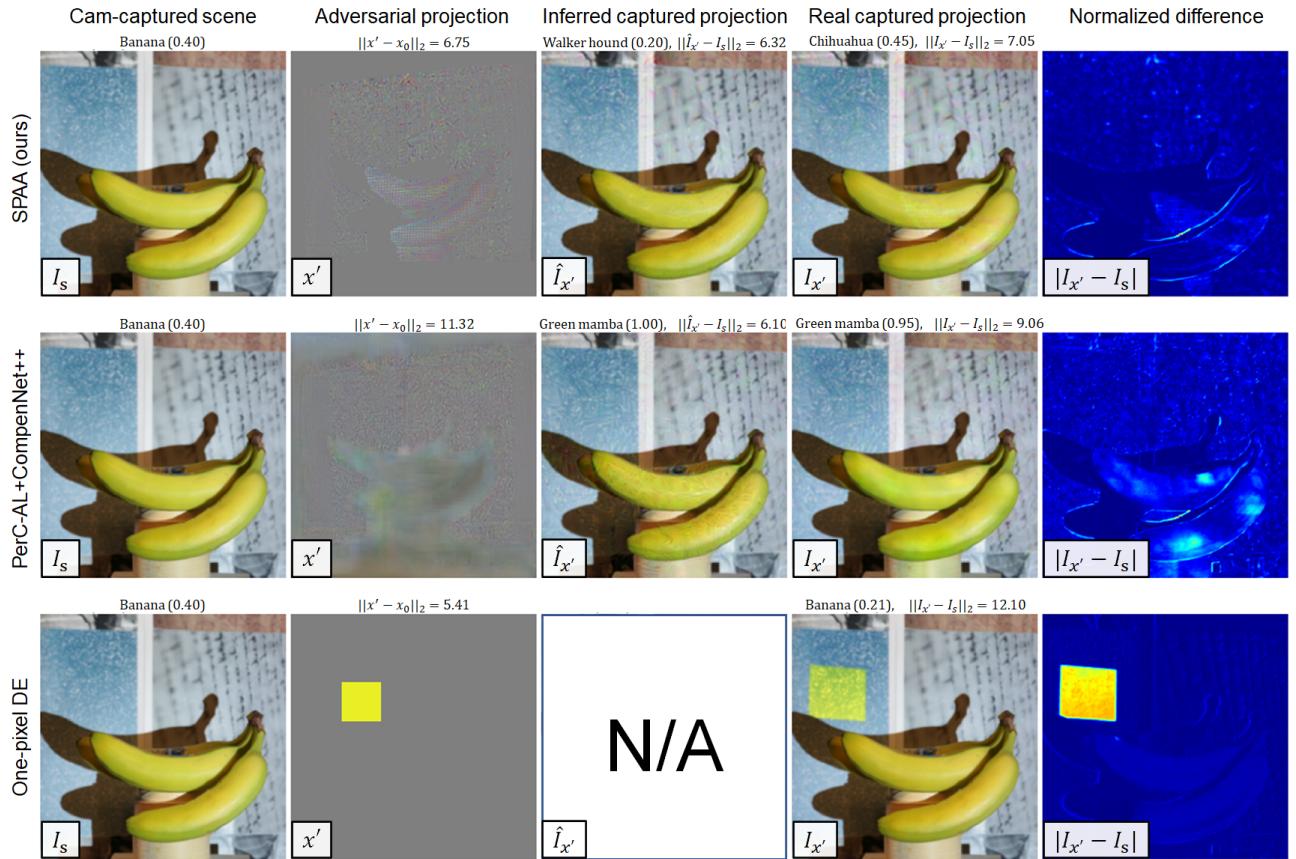


Figure 24: **Untargeted** projector-based adversarial attack on ResNet-18. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT banana**.

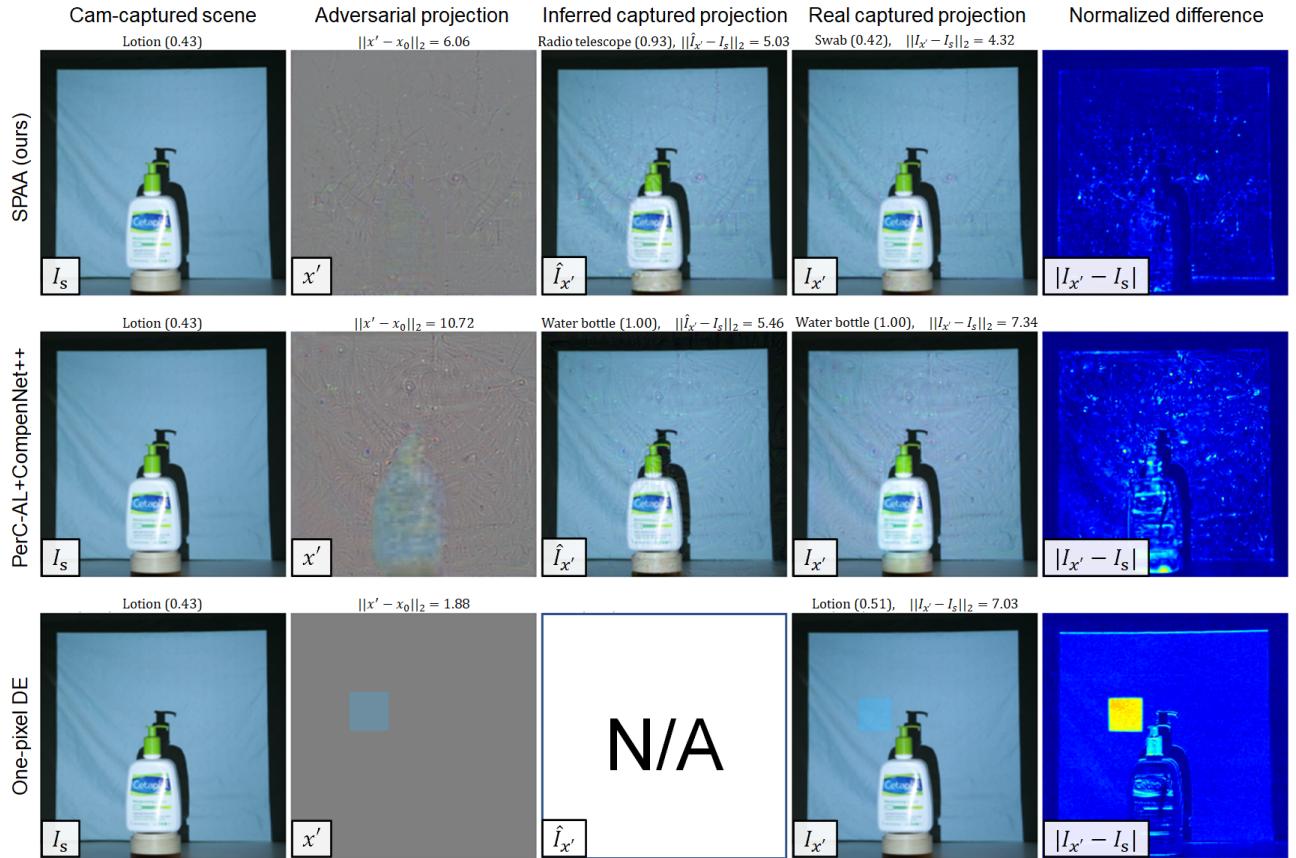


Figure 25: **Untargeted** projector-based adversarial attack on VGG-16. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT lotion**.

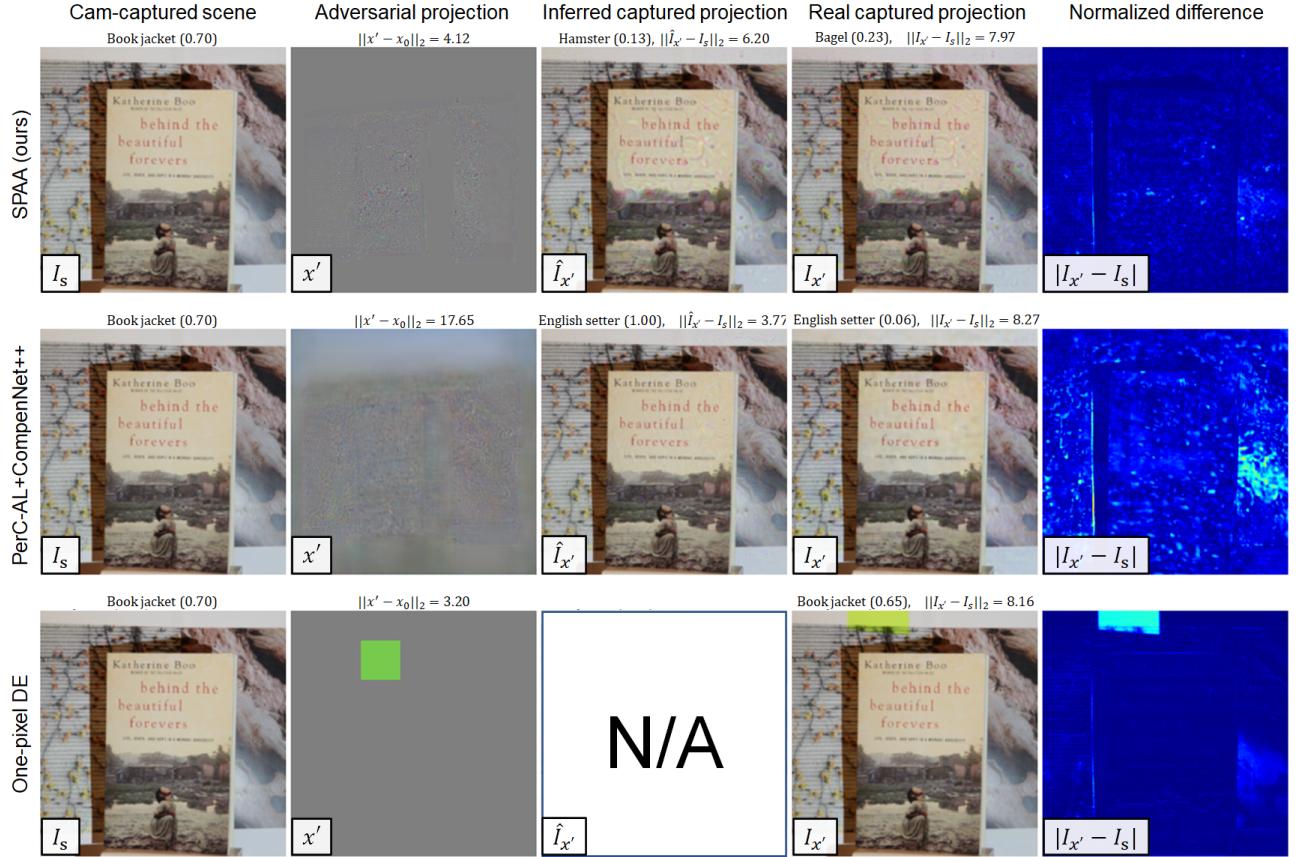


Figure 26: **Untargeted** projector-based adversarial attack on Inception v3. The goal is to cause the classifier to misclassify the captured projection, such that the output is **NOT book jacket**.