Adversarial Attacks Beyond the Image Space

Xiaohui Zeng[1], Chenxi Liu[2], Yu-Siang Wang[3], Weichao Qiu[2], Lingxi Xie[2,4], Yu-Wing Tai[5], Chi-Keung Tang[6], Alan L. Yuille[2]

[1]University of Toronto [2]The Johns Hopkins University [3]National Taiwan University
[4]Huawei Noah's Ark Lab [5]Tencent YouTu [6]Hong Kong University of Science and Technology

Introduction

Adversarial examples

- trigger mis-classification by slightly perturbing the input.
- may be physically inauthentic when they remain in the image space.

Contributions

- Go beyond the image space, attack in the physical space by perturbing 3D physical parameters.
- First work to study the interpretable 3D adversarial examples that are physically authentic and plausible.

Physical Properties We Attacked

Differentiable attack:

- Surface Normal (N)
- Illumination (L)
- Material (M)

Non-differentiable attack:

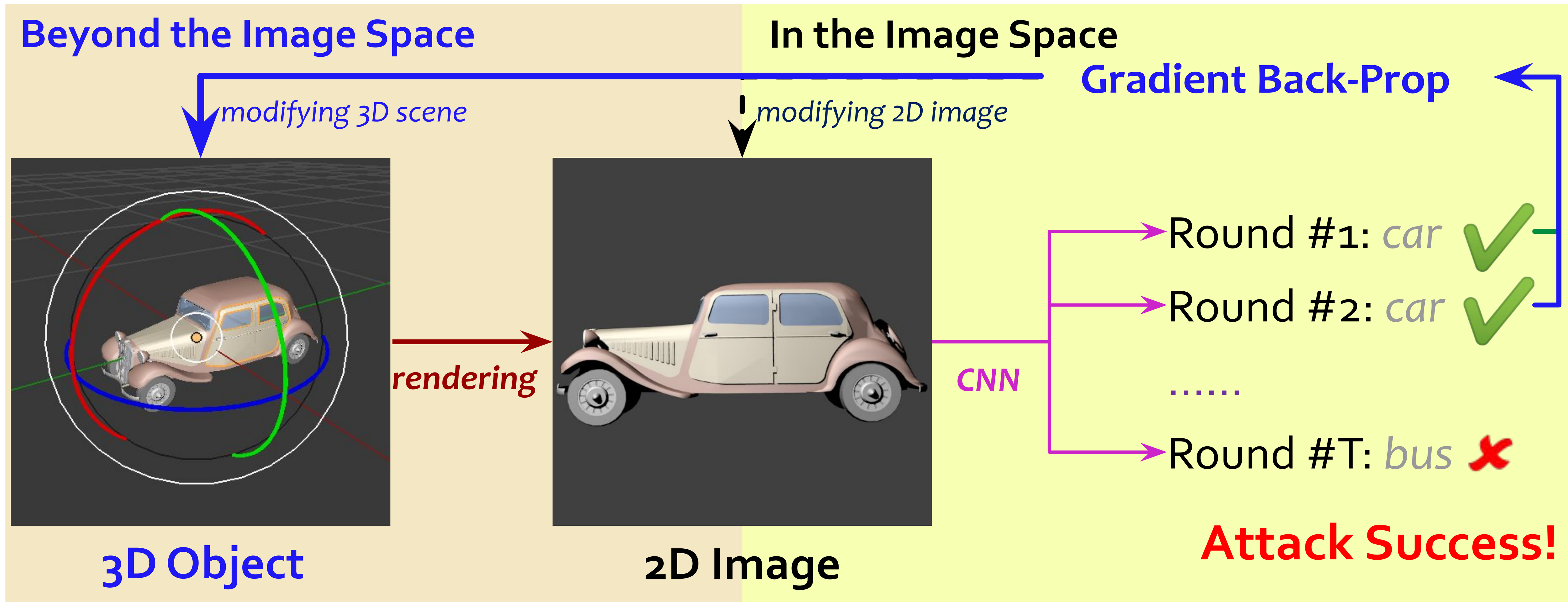
- Color (C)
- Rotation (R)
- Translation (T)
- Lighting (L)

Physical Adversarial Attack

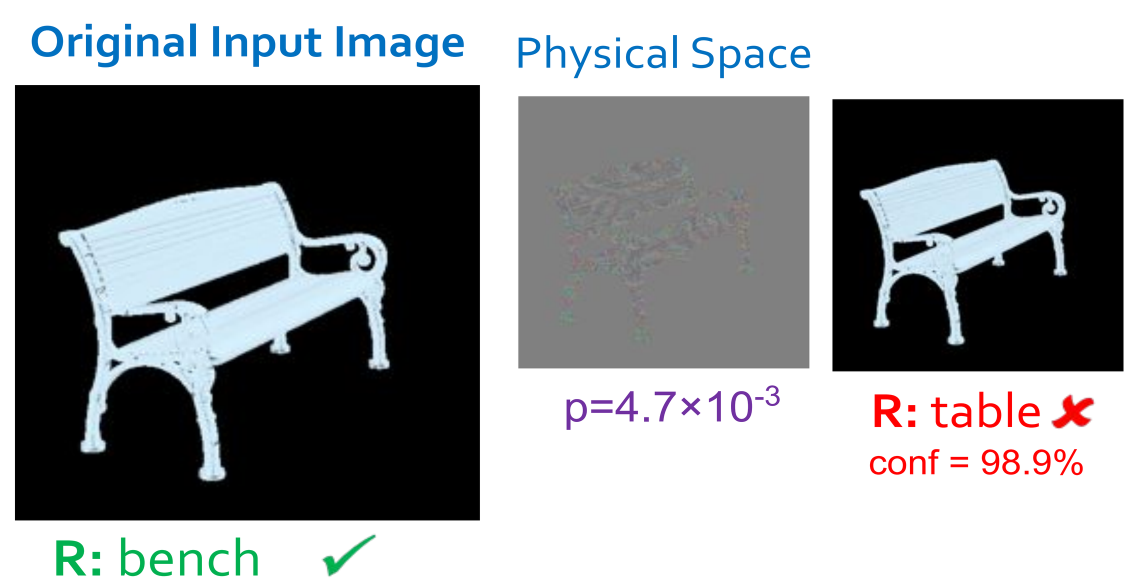
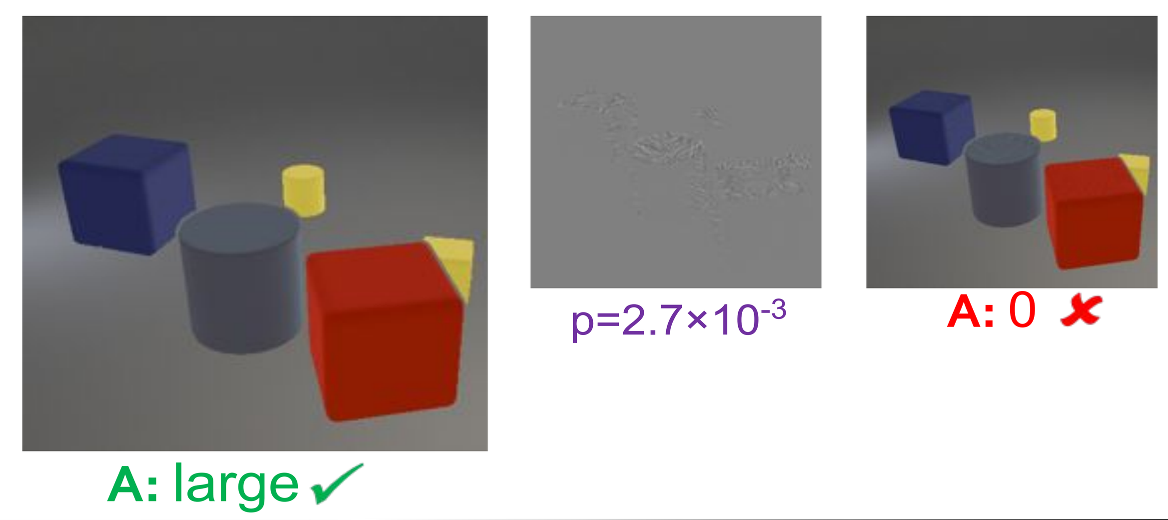
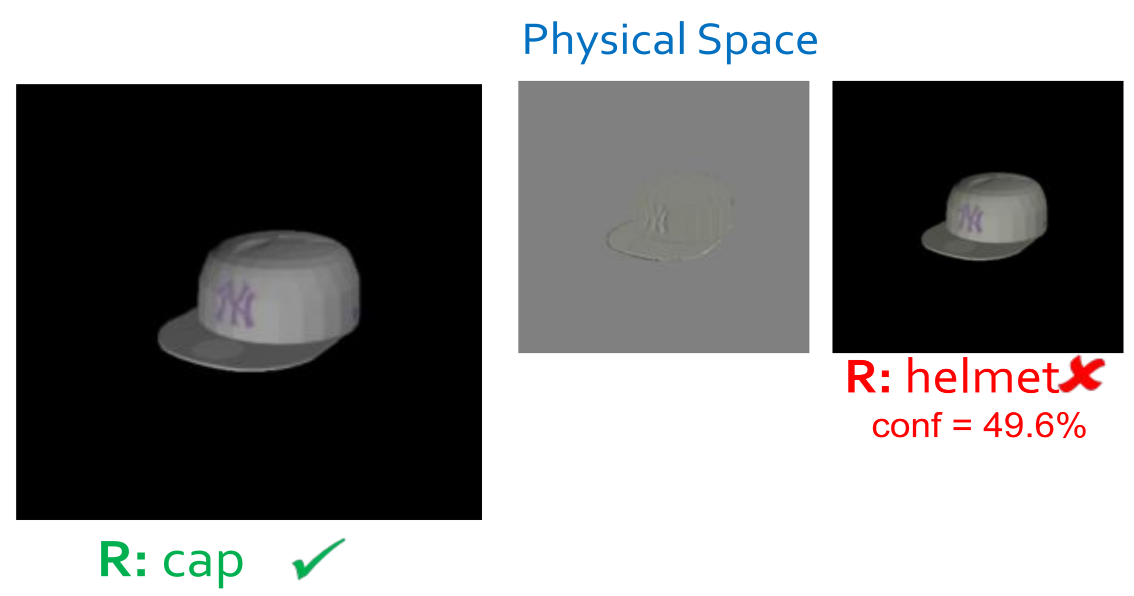
Algorithm 1

Input: physical params $\mathbf{X} \in \mathbb{R}^D$;
black-box render $\mathbf{r}(\cdot)$ and model $\mathbf{f}(\cdot;\boldsymbol{\theta})$;
loss function $\mathcal{L}(\cdot)$ with parameter λ;
learning rate η; max steps T;
Output: adversarial perturbation $\Delta\mathbf{X}$;
Init: $\mathbf{I} = \mathbf{r}(\mathbf{X})$, $\mathbf{Z} = \mathbf{f}(\mathbf{I};\boldsymbol{\theta})$, $c = \arg\max_{c'} Z_{c'}$;
    $t \leftarrow 0$, $\mathbf{X}^{(0)} \leftarrow \mathbf{X}$, $\mathbf{I}^{(0)} \leftarrow \mathbf{I}$, $\mathbf{Z}^{(0)} \leftarrow \mathbf{Z}$, $\Delta\mathbf{X} \leftarrow \mathbf{0}$;
repeat
    $t \leftarrow t+1$
    if FGSM:
        $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)} + \eta \cdot \mathit{sign}(\nabla\mathbf{X}^{(t)})$
    else: # use ZOO
        sample: $\mathcal{D}^{(t)} \subseteq \{1, 2, \ldots, D\}$;
        $R_d^{(t)} \leftarrow \mathbb{I}\left[d \in \mathcal{D}^{(t)}\right] \cdot \frac{\partial\mathcal{L}(\mathbf{X}^{(t-1)})}{\partial X_d^{(t-1)}}$, $d = 1, 2, \ldots, D$;
        $\mathbf{X}^{(t)} = \mathbf{X}^{(t-1)} + \eta \cdot \mathbf{R}^{(t)}$;
    $\mathbf{I}^{(t)} = \mathbf{r}(\mathbf{X}^{(t)})$,
    $\mathbf{Z}^{(t)} = \mathbf{f}(\mathbf{I}^{(t)};\boldsymbol{\theta})$;
until $t = T$ **or** $Z_c^{(t)} < \max_{c'}\left\{Z_{c'}^{(t)}\right\}$;
Return: $\Delta\mathbf{X} = \mathbf{X}^{(t)} - \mathbf{X}$.

The Proposed Approach



Examples

	Object Classification	Visual Question Answering
White Box Attack		
Black Box Attack		

Object Classification:

- Illumination: $\Delta\mathbf{L}_{key} = (9.5, 5.4, 0.6)/100$
- Object rotation: $(-2.9, -9.4, -2.5)/1000$ rad
- Object translation: $(\Delta x, \Delta z) = (2.0, 0.2)/1000$,
- Object color: $(\Delta R, \Delta G, \Delta B) = (9.1, 5.4, -4.8)/100, \ldots$

Visual Question Answering:

- Illumination: $\Delta\mathbf{L}_{key} = (0.0, 3.0, -1.0, -1.7)/100$
- Object 2: $(\Delta r, \Delta\theta) = (-1.6, 4.1)/100$,
- Object 3: $(\Delta x, \Delta y) = (-3.1, 6.2)/100$,
- Object 9: $\Delta c = (-3.7, -1.1, -4.5)/100, \ldots$

Results

White-box adversarial attacks on classification model for ShapeNet object:

Perturbing	AlexNet		ResNet-34	
	Succ.	p^*	Succ.	p^*
Image	100.00	5.7	99.57	5.1
Surface N.	89.27	10.8	88.41	9.3
Illumination L	29.61	25.8	14.16	29.3
Material M	18.88	25.8	3.43	55.2
Combined	94.42	18.1	94.85	16.4

Visual question answering model for CLEVR Dataset;

Perturbing	IEP	
	Succ.	p^*
Image	96.33	2.1
Surface N.	83.67	6.8
Illumination L	48.67	9.5
Material M	8.33	12.3
Combined	90.67	8.8

p^* stands for perceptibility x 10^{-3}

Conclusion

- Image space adversaries can not be explained by simple physical space changes with current optimization algorithms.
- Directly constructing physical space adversaries can still succeed, which poses more serious threats.

Code will be released soon on github:
`https://github.com/ZENGXH/adversarial_attack_beyond_the_img_space`