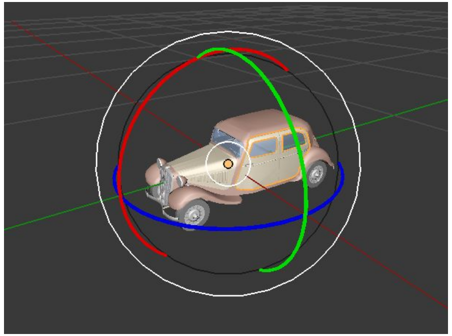
Adversarial Attacks Beyond the Image Space

Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi Keung Tang, Alan Yuille

06/19/2019

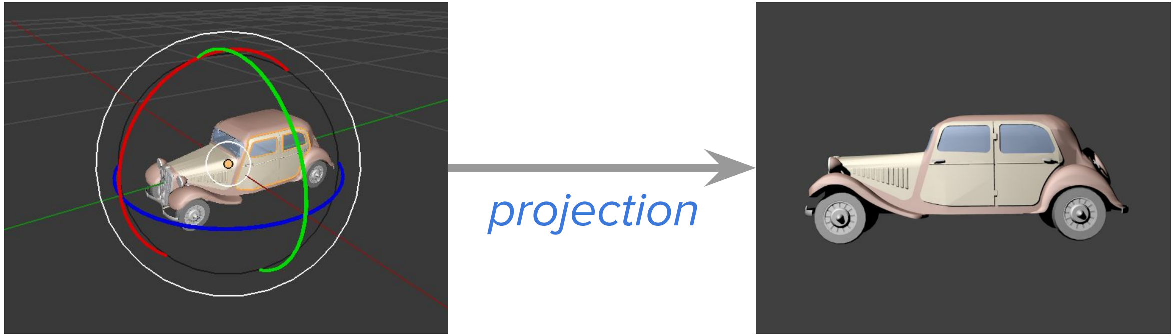
Visual Recognition Pipeline

Visual Recognition Pipeline



3D scene

Visual Recognition Pipeline

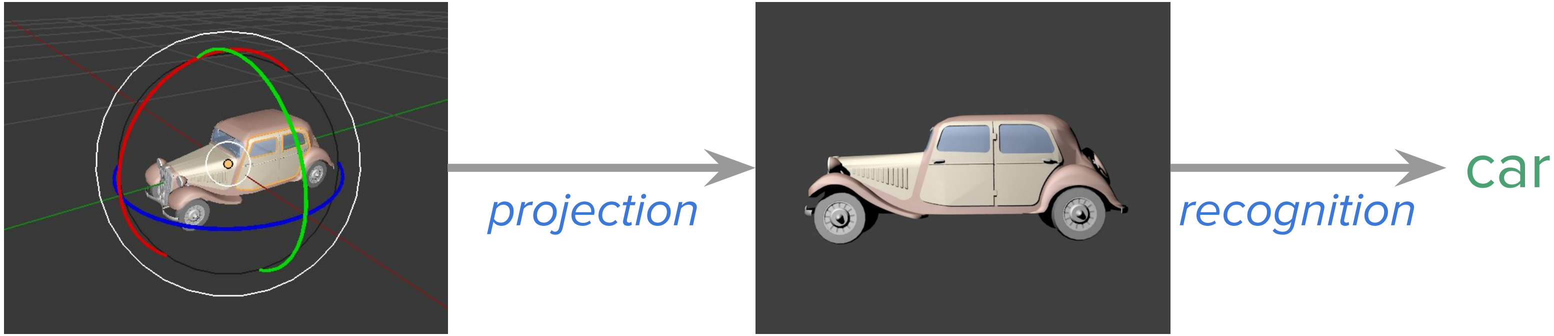


projection

3D scene

2D image

Visual Recognition Pipeline



projection

recognition

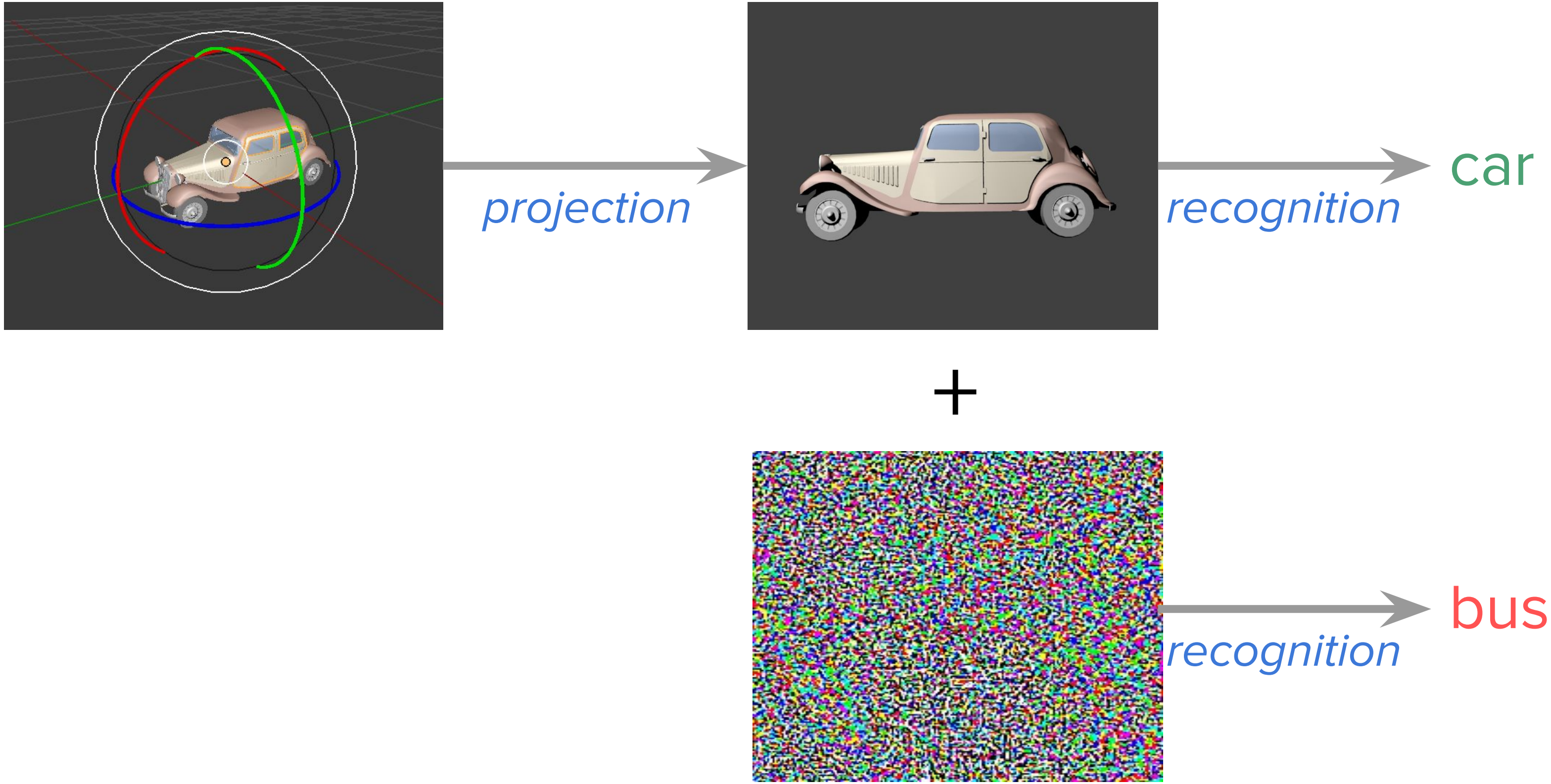
car

3D scene

2D image

label

Adversarial Attacks on 2D Image



projection

recognition

car

+

recognition

bus

Adversarial Attacks

- ***Can the network fail if we slightly modify 2D pixel values?***
  - Yes!

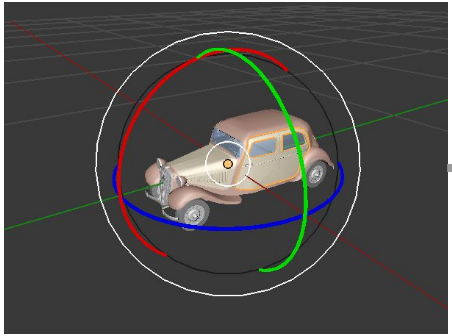
Adversarial Attacks

- ***Can the network fail if we slightly modify 2D pixel values?***
  - Yes!
- ***Should we be concerned about them in the real world?***
  - Well, sort of.

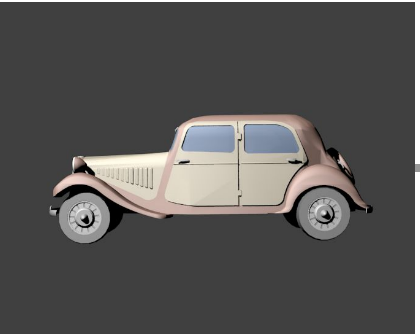
Adversarial Attacks

- ***Can the network fail if we slightly modify 2D pixel values?***
  - Yes!
- ***Should we be concerned about them in the real world?***
  - Well, sort of.
  - Potentially dangerous.
  - But require position-wise albedo change, which is unrealistic to implement.

Adversarial Attacks on 3D Scene



projection



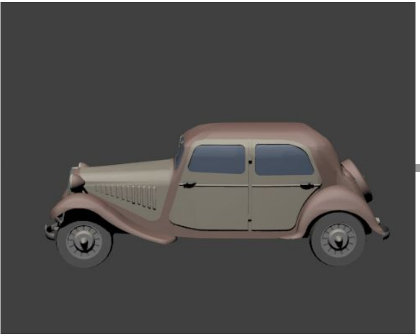
recognition

car

+

rotation,
translation

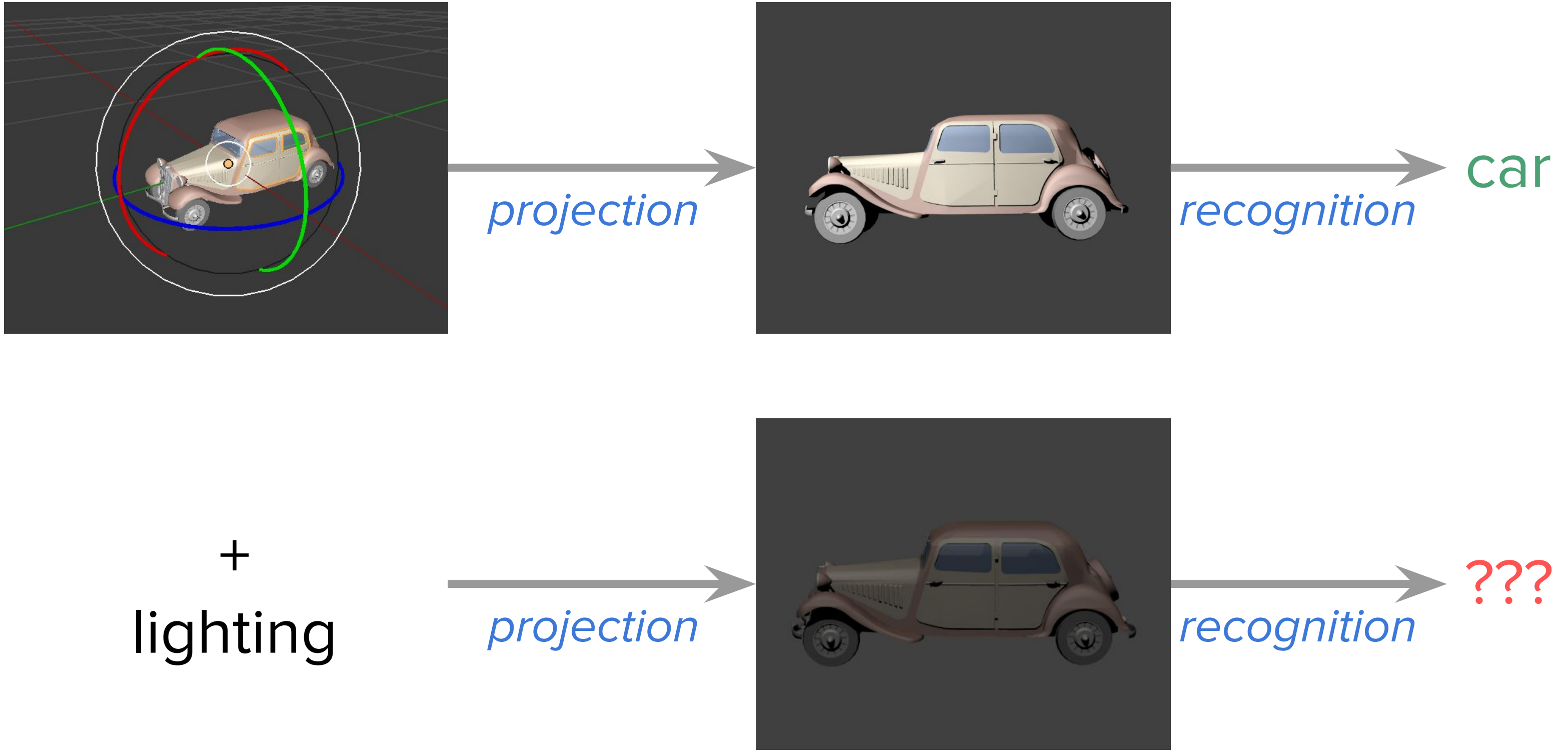
projection



recognition

???

Adversarial Attacks on 3D Scene



projection

recognition

car

+ lighting

projection

recognition

???

Adversarial Attacks

- ***Can the network fail if we slightly modify 2D pixel values?***
  - Yes!
- ***Should we be concerned about them in the real world?***
  - Well, sort of.
  - Potentially dangerous.
  - But require position-wise albedo change, which is unrealistic to implement.

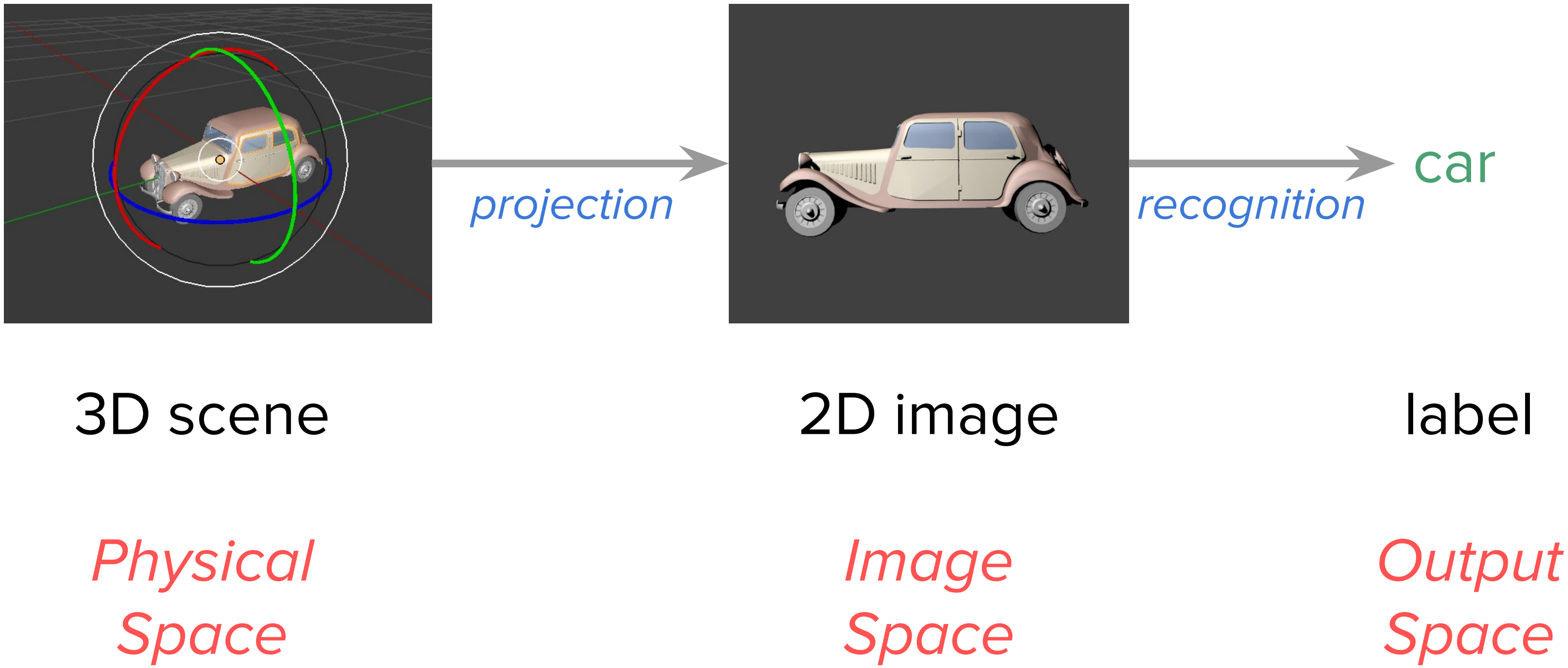
- ***Can the network fail if we slightly modify 3D physical parameters?***
  - Well, let’s find out :)

Adversarial Attacks

- ***Can the network fail if we slightly modify 2D pixel values?***
  - Yes!
- ***Should we be concerned about them in the real world?***
  - Well, sort of.
  - Potentially dangerous.
  - But require position-wise albedo change, which is unrealistic to implement.

- ***Can the network fail if we slightly modify 3D physical parameters?***
  - Well, let's find out :)
- ***Should we be concerned about them in the real world?***
  - If they exist, then we should be much more concerned than before, as they are much more easily realized.

Visual Recognition Pipeline



projection

recognition

car

3D scene

2D image

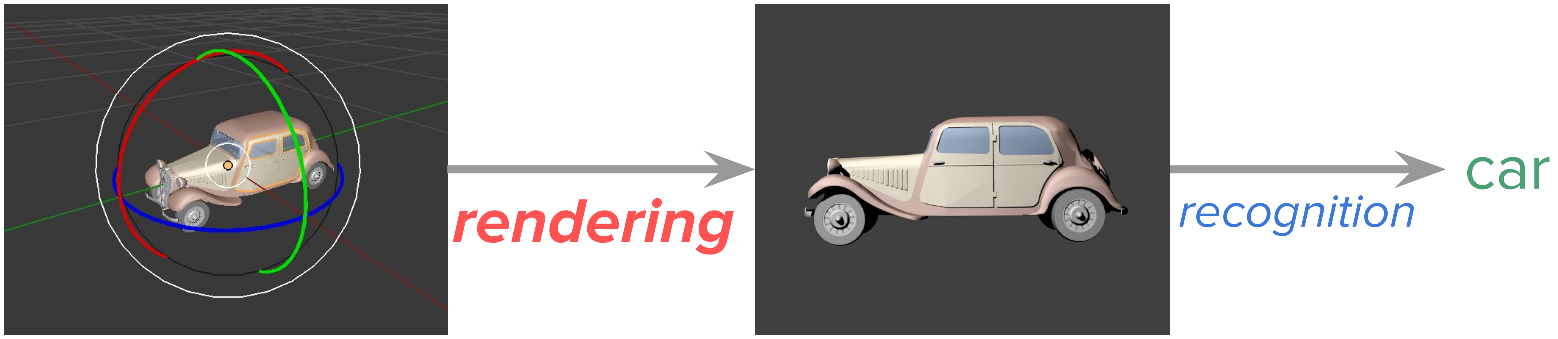
label

Physical Space

Image Space

Output Space

Visual Recognition Pipeline



rendering

recognition

car

3D scene

2D image

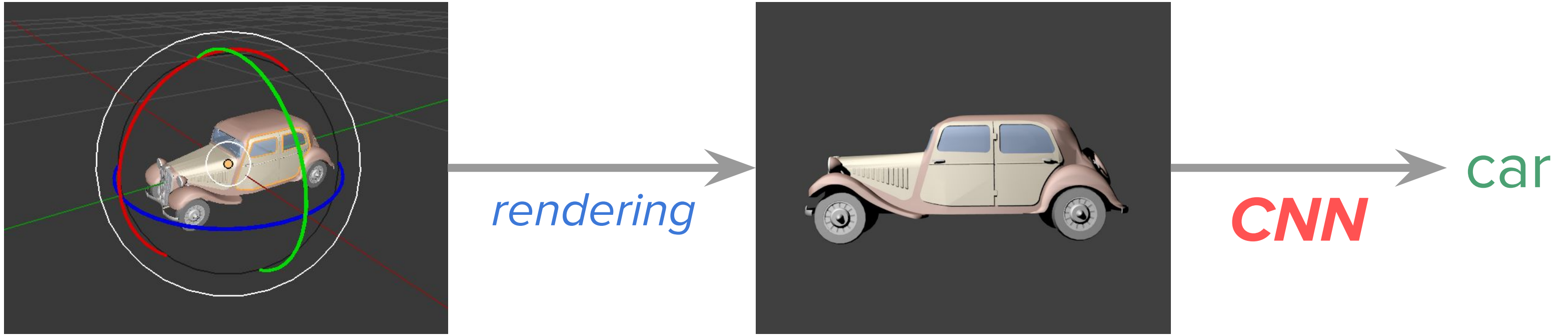
label

Physical Space

Image Space

Output Space

Visual Recognition Pipeline



rendering

CNN

car

3D scene

2D image

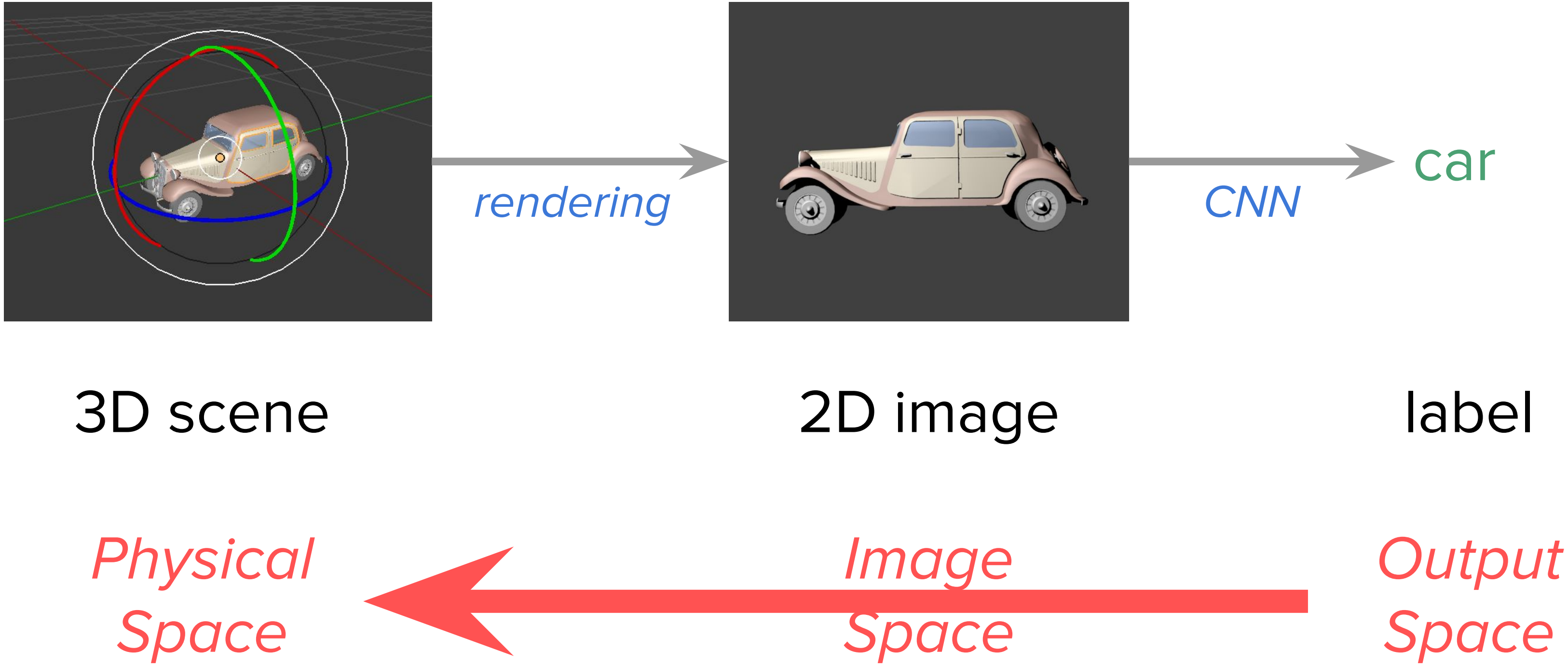
label

Physical Space

Image Space

Output Space

Visual Recognition Pipeline



rendering

CNN

car

3D scene

2D image

label

Physical Space

Image Space

Output Space

Settings & Tasks

Differentiable Renderer

- White box attack
- Use gradient descent

Settings & Tasks

Differentiable Renderer	● White box attack ● Use gradient descent
Non-Differentiable Renderer	● Black box attack ● Use finite difference for the non-differentiable component

Chen, Pin-Yu, et al. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017.

Settings & Tasks

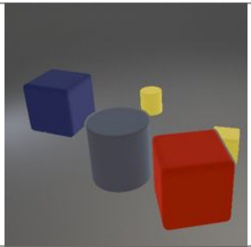
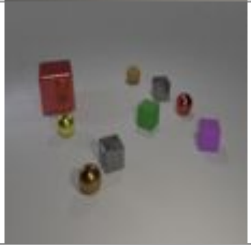
Differentiable Renderer		
Non-Differentiable Renderer		

Settings & Tasks

	Object Classification (ShapeNet)	
Differentiable Renderer		
Non-Differentiable Renderer		

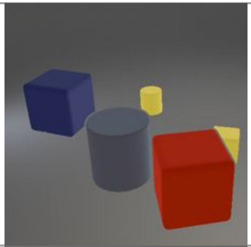
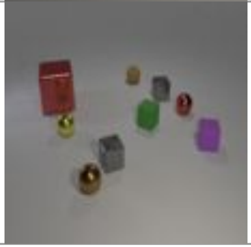
Chang, Angel X., et al. "Shapenet: An information-rich 3d model repository." arXiv preprint arXiv:1512.03012 (2015).

Settings & Tasks

	Object Classification (ShapeNet)	Visual Question Answering (CLEVR)
Differentiable Renderer		
Non-Differentiable Renderer		

Johnson, Justin, et al. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." In CVPR. 2017.

Settings & Tasks

	Object Classification (ShapeNet)	Visual Question Answering (CLEVR)
Differentiable Renderer	#1	 #2
Non-Differentiable Renderer	#3	 #4

#1: Differentiable + Object Classification

- ***Differentiable renderer: Liu et al, 2017***
  - surface normal
  - illumination
  - material

Liu, Guilin, et al. "Material editing using a physically based rendering network." In ICCV. 2017.

#1: Differentiable + Object Classification

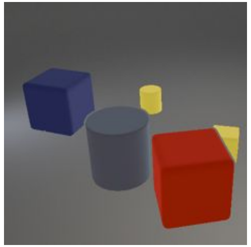
- ***Differentiable renderer: Liu et al, 2017***
  - surface normal
  - illumination
  - material
- ***Can image space adversarial noise be explained by physical space?***
  - No for 97% of the case

#1: Differentiable + Object Classification

- ***Differentiable renderer: Liu et al, 2017***
  - surface normal
  - illumination
  - material
- ***Can image space adversarial noise be explained by physical space?***
  - No for 97% of the case
- ***Attacking image space vs physical space:***

	Image	Surface N.	Illumination	Material	Combined
Attack success %	**100.00**	89.27	29.61	18.88	**94.42**

#2: Differentiable + VQA



- ***Attacking image space vs physical space:***

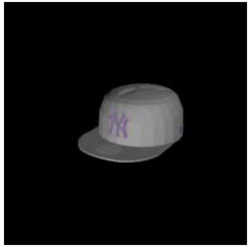
	Image	Surface N.	Illumination	Material	Combined
Attack success %	**96.33**	83.67	48.67	8.33	**90.67**

#3: Non-Differentiable + Object Classification

- ***Non-Differentiable renderer: Blender***
  - color
  - rotation
  - translation
  - lighting

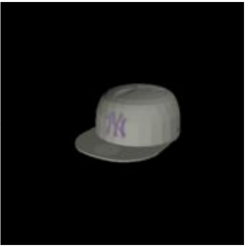
#3: Non-Differentiable + Object Classification

- ***Non-Differentiable renderer: Blender***
  - color
  - rotation
  - translation
  - lighting
- ***How often does physical space attack succeed?***
  - ~10% of the time
  - But highly interpretable:



cap

Rotate (-2.9, 9.4, 2.5) x 10^{-3} rad along x, y, z
Move (2.0, 0.0, 0.2) x 10^{-3} along x, y, z
Change RGB color by (9.1, 5.4, -4.8) x 10^{-2}
Adjust light source by -0.3
Change the light angle by (9.5, 5.4, 0.6) x 10^{-2}

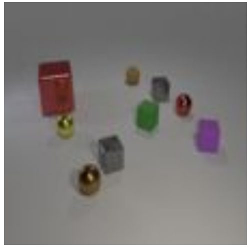


helmet

#4: Non-Differentiable + VQA

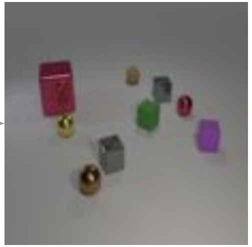
- ***How often does physical space attack succeed?***
  - ~20% of the time
  - But highly interpretable:

Q:*How many other purple objects have the same shape as the purple matte object?*



A: 0

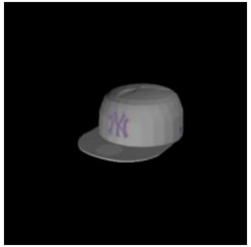
Move light source by (0.0, 3.0, -1.0, -1.7) x 10^{-2}
Rotate object 2 by (-1.6, 4.1) x 10^{-2}
Move object 3 by (-3.1, 6.2) x 10^{-2}
Change RGB of object 9 by (-3.7, -1.1, -4.5) x 10^{-2}
......



A: 1

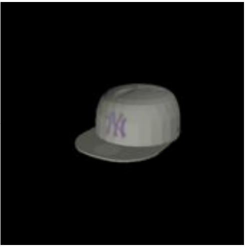
Conclusion

- We study adversarial attacks beyond the image space on the physical space
- Such attacks (via rotation, translation, color, lighting etc) can still succeed
- They pose more serious threat



cap

Rotate (-2.9, 9.4, 2.5) x 10^{-3} rad along x, y, z
Move (2.0, 0.0, 0.2) x 10^{-3} along x, y, z
Change RGB color by (9.1, 5.4, -4.8) x 10^{-2}
Adjust light source by -0.3
Change the light angle by (9.5, 5.4, 0.6) x 10^{-2}



helmet

Thank you!