Applications of Adversarial Attacks on Matching-based Algorithms

CS588

Speaker: Woo Jae Kim

Class Objectives

- **Up-to-date matching-based algorithms**
- **Adversarial attacks**
- **Combining adversarial attacks with matching-based algorithms**

Matching-based Algorithms

Matching-based Algorithms

- **What is matching-based algorithm?**
- **Finding correspondence or similarity between *__multiple sets of data__***
- **Can be used to estimate *__3D geometric structure__* of the data**

Stereo Matching

- **Uses two images taken from slightly different viewpoints**
- **Goal: Infer disparity or depth maps given the image pair**

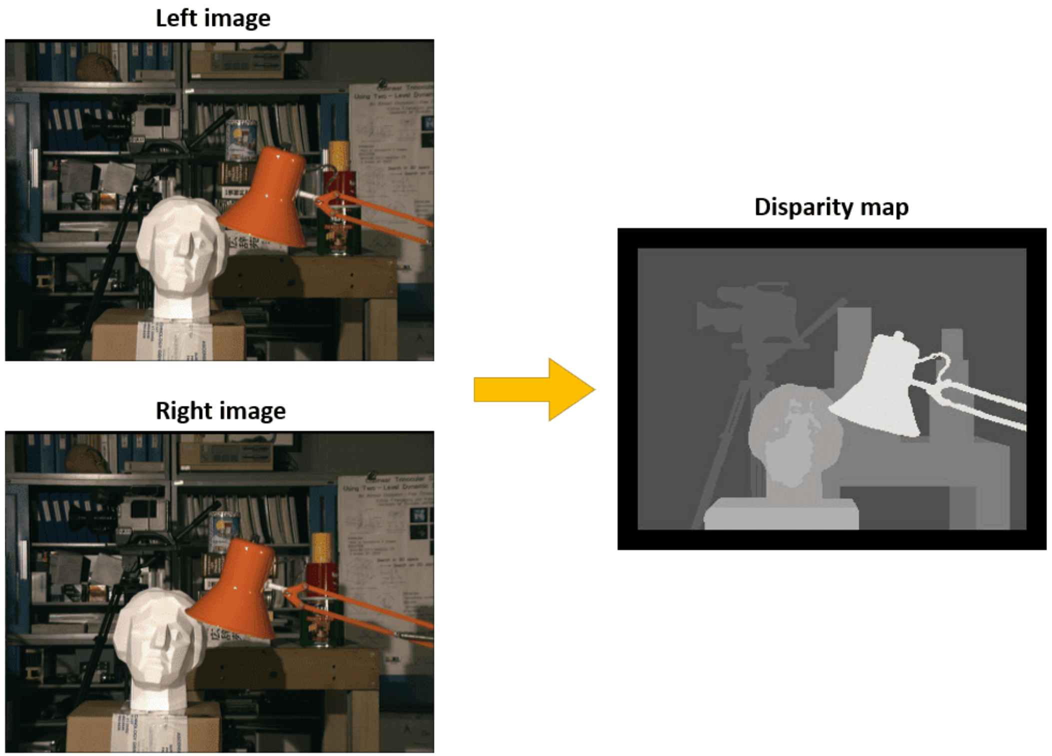
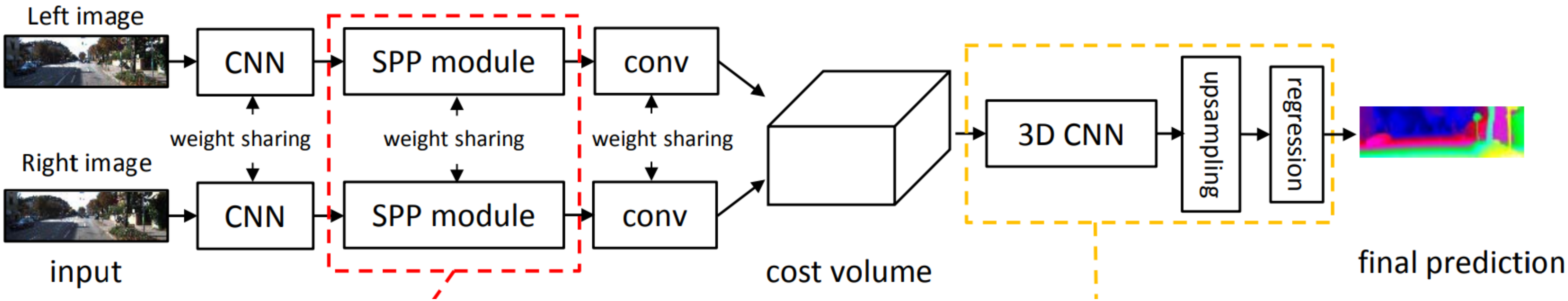


Image: https://www.baeldung.com/cs/disparity-map-stereo-vision

Stereo Matching

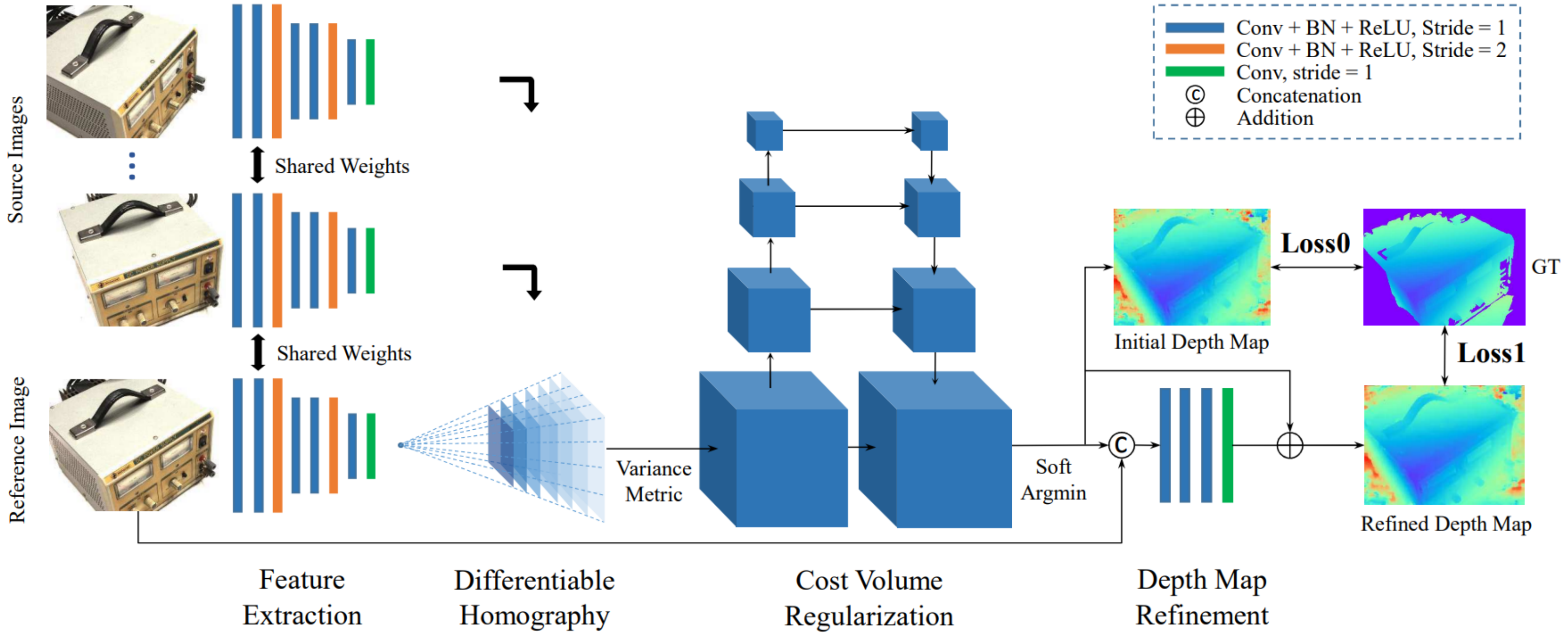
- Current methods use CNN to extract features from each image
- It then constructs a cost volume, from which the final depth is predicted



Chang et al., Pyramid Stereo Matching Network, CVPR 2018

Multi-view Stereo

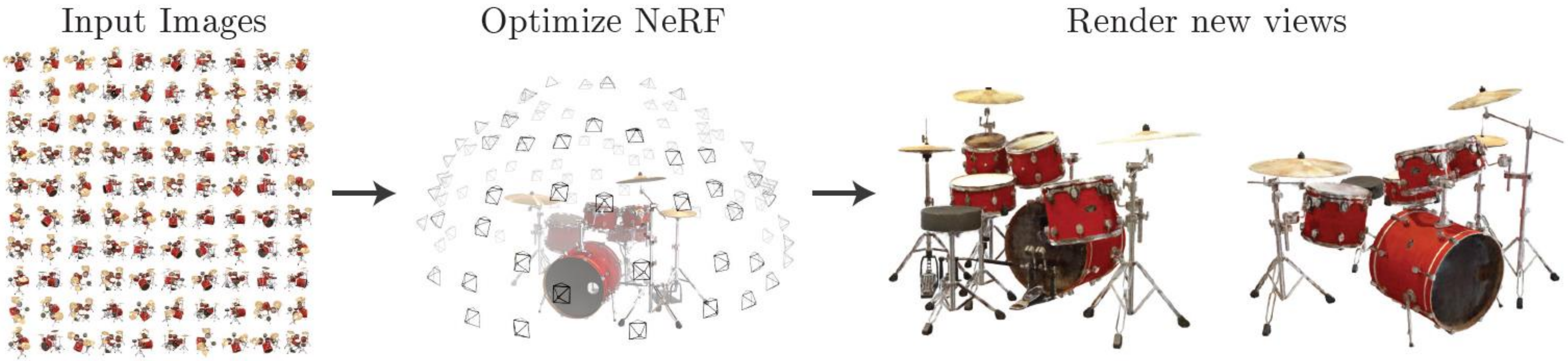
- Multi-view stereo uses ***more than two images*** captured multiple viewpoints
- Can potentially provide more accurate and detailed 3D reconstructions
- Goal: Learn ***correspondence among images*** to infer depth of objects



Yao et al., MVSNet: Depth Inference for Unstructured Multi-view Stereo, ECCV 2018

3D Matching (Novel-view Synthesis)

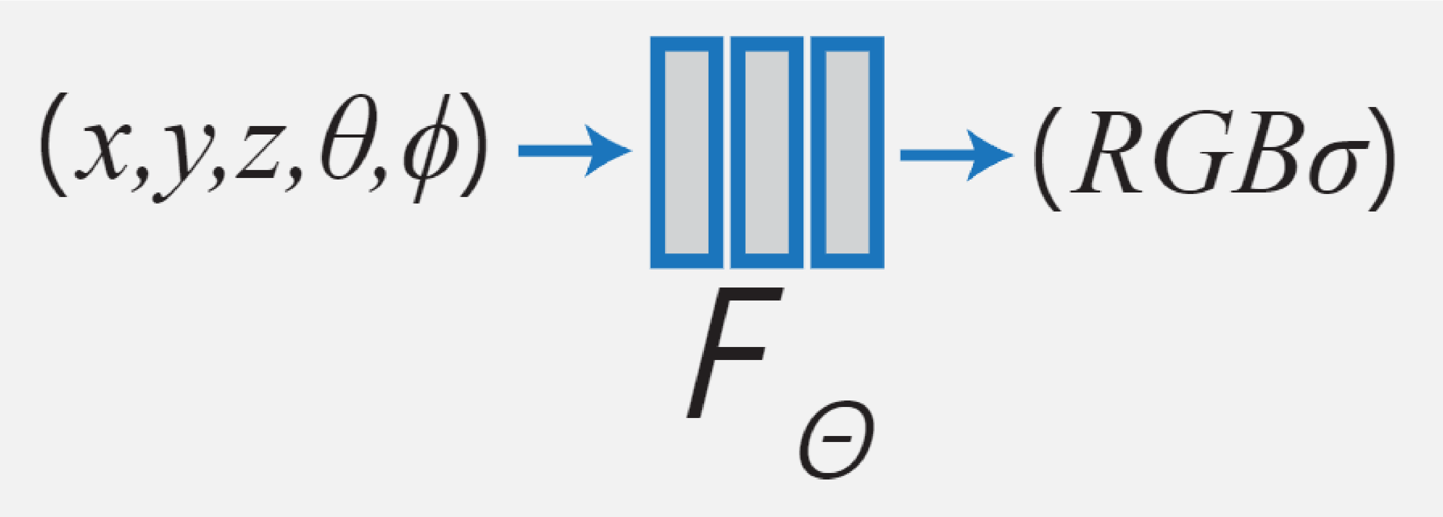
- **More current methods use correspondence among multi-view images to model a *<u>continuous 3D representation</u>* of a scene**
- **This allows a more accurate and seamless prediction of the 3D geometry**
- **Became highly popular with introduction of *<u>Neural Radiance Fields (NeRF)</u>***



Mildenhall et al., NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

Neural Radiance Fields

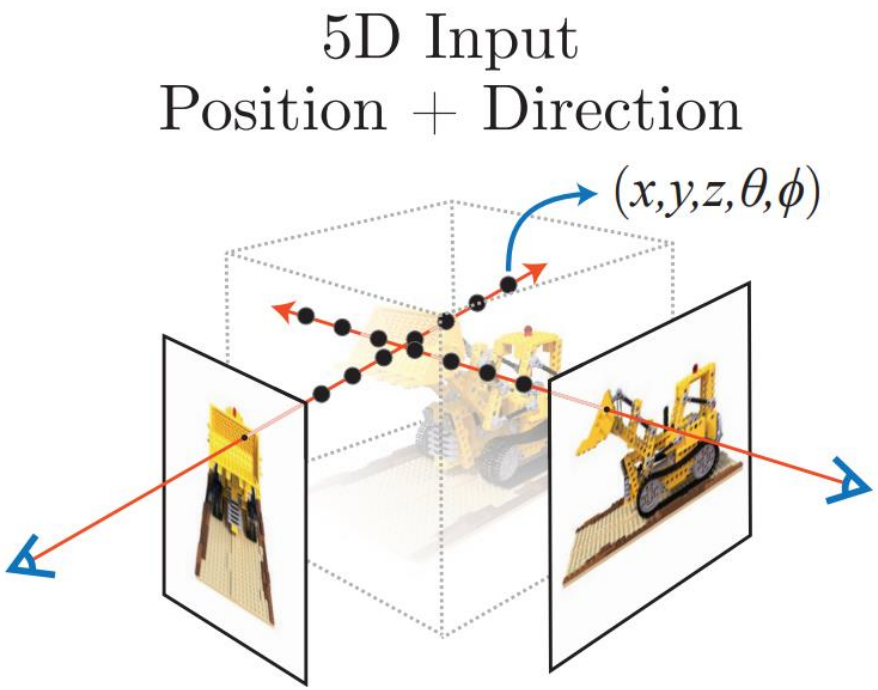
- NeRF uses an MLP network F_Θ to represent a 3D scene
- Given the position (x, y, z) and direction (θ, ϕ) of a 3D point,
- MLP F_Θ predicts the color RGB and density σ



Mildenhall et al., NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

Neural Radiance Fields: Input

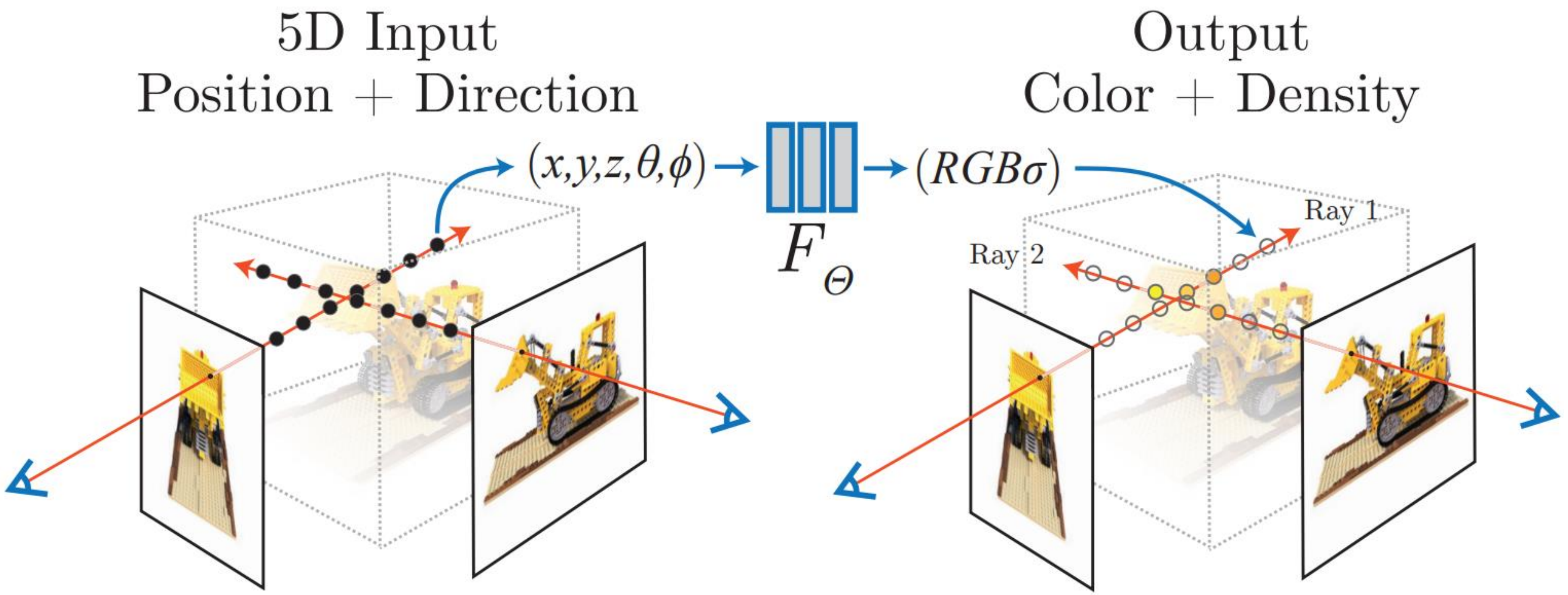
- **A set of images captured on a scene**
- **Randomly sample points along the rays from camera origin to each pixel of the image**



Mildenhall et al., NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

Neural Radiance Fields: Output

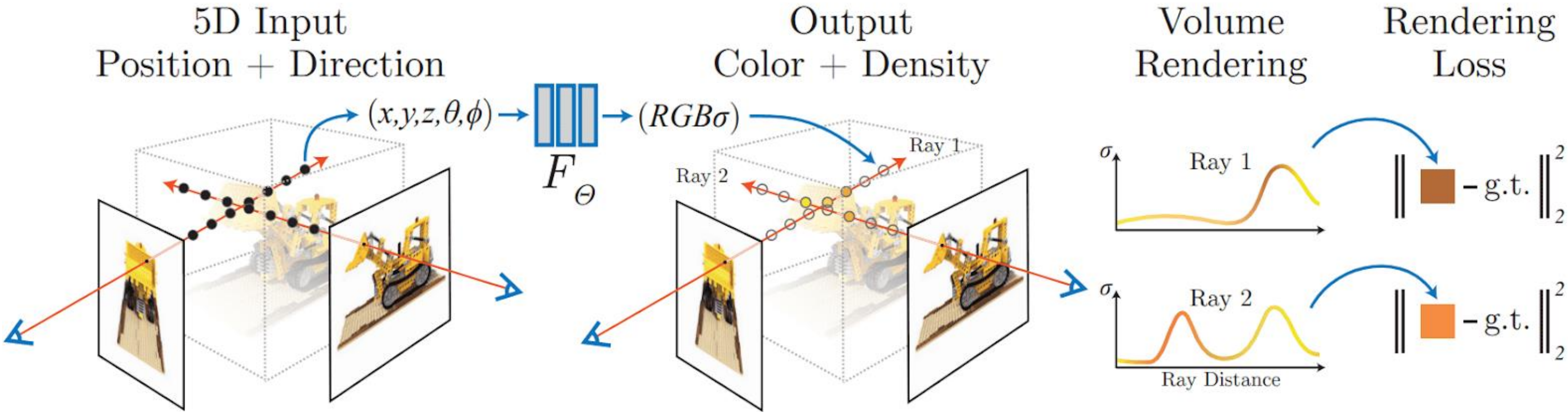
- **Pass the pos (x, y, z) and dir (θ, ϕ) of each point to the MLP network F_Θ**
- **MLP then predicts *color RGB* and *density σ* for each point**



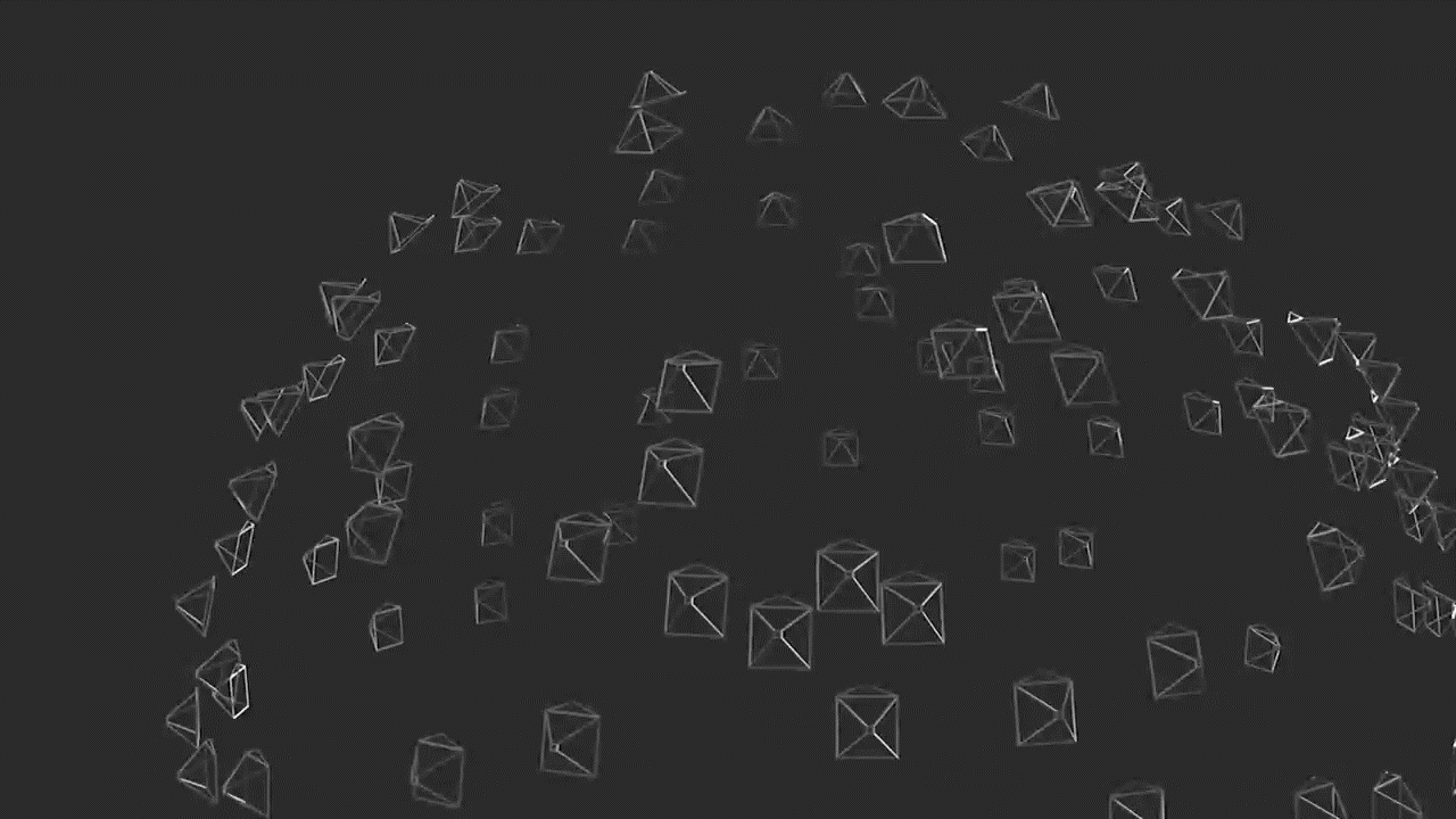
Mildenhall et al., NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

Neural Radiance Fields: Volume Rendering

- **Since NeRF models a continuous 3D volume, it can render images to an arbitrary view**
- **Volume rendering: Accumulation of colors and density of multiple points on a ray**



Mildenhall et al., NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020



NeRF for Novel-View Synthesis

- **NeRF shows excellent performance at modeling 3D scene & rendering arbitrary views**

Mildenhall et al., NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

Adversarial Attacks

Adversarial Attacks

- **Many deep learning models are known to be vulnerable to *adversarial attacks***

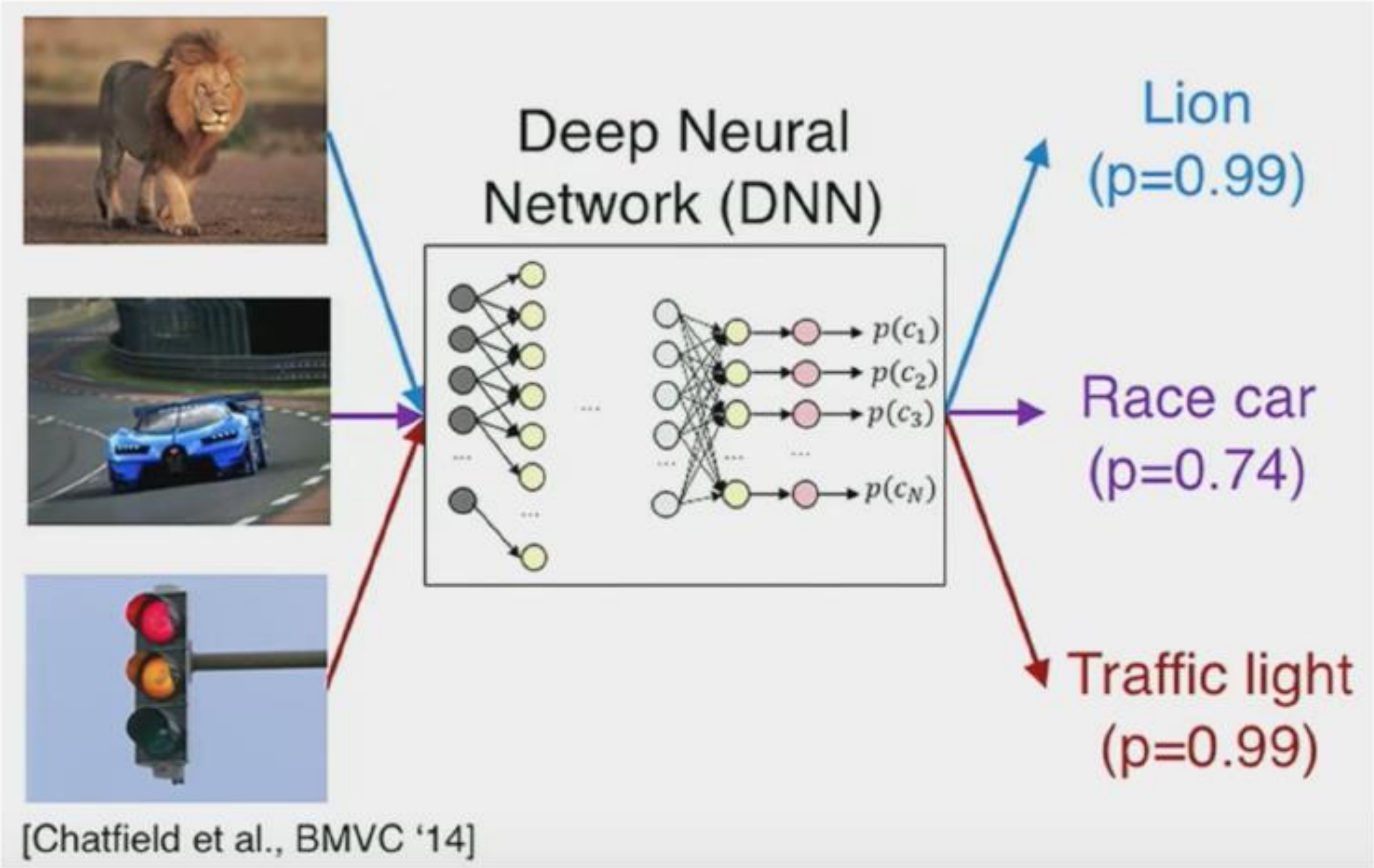
Video: https://www.youtube.com/watch?v=pc2ssNY98LA&feature=emb_logo&ab_channel=PoloClubofDataScience

Adversarial Attacks

- **Many deep learning models are known to be vulnerable to *adversarial attacks***

Video: https://www.youtube.com/watch?v=zQ_uMenoBCk

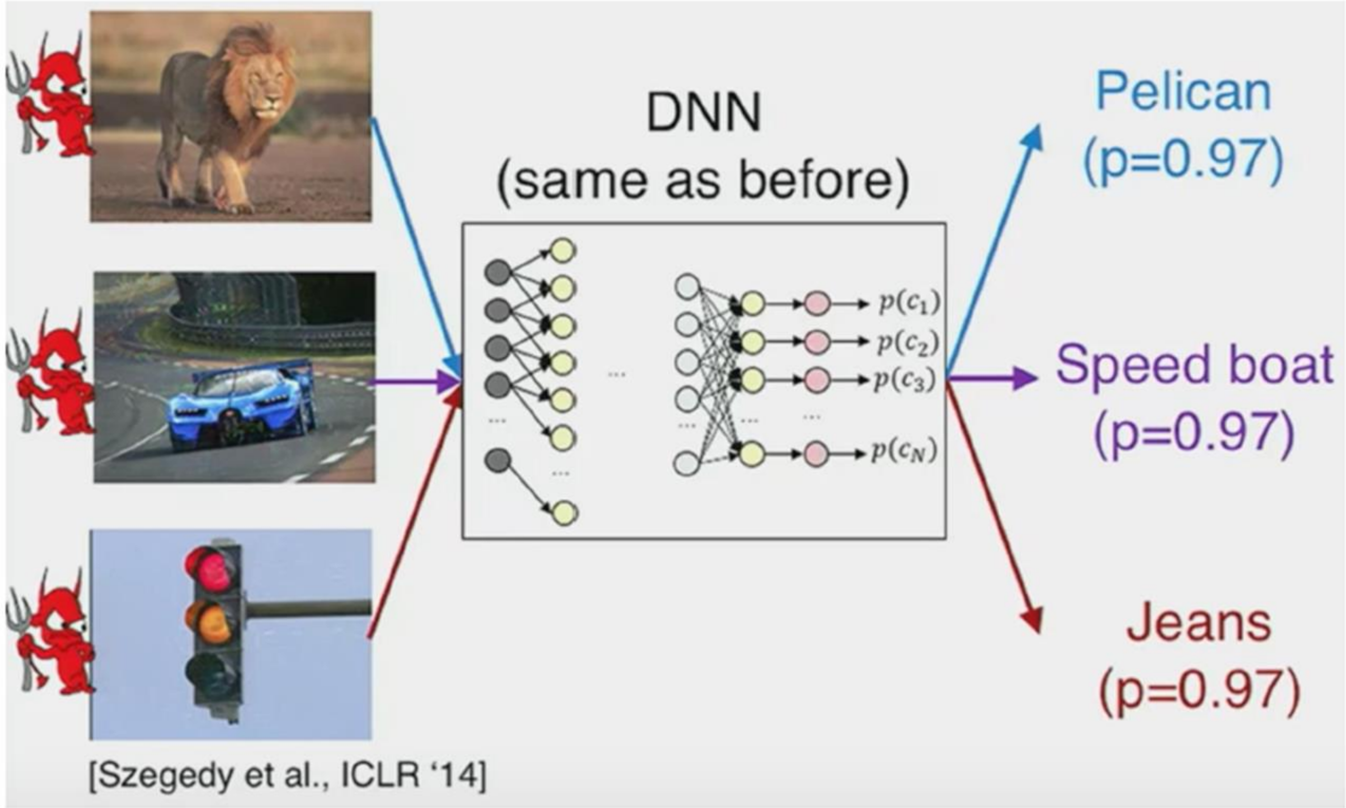
Adversarial Attacks



Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Attacks

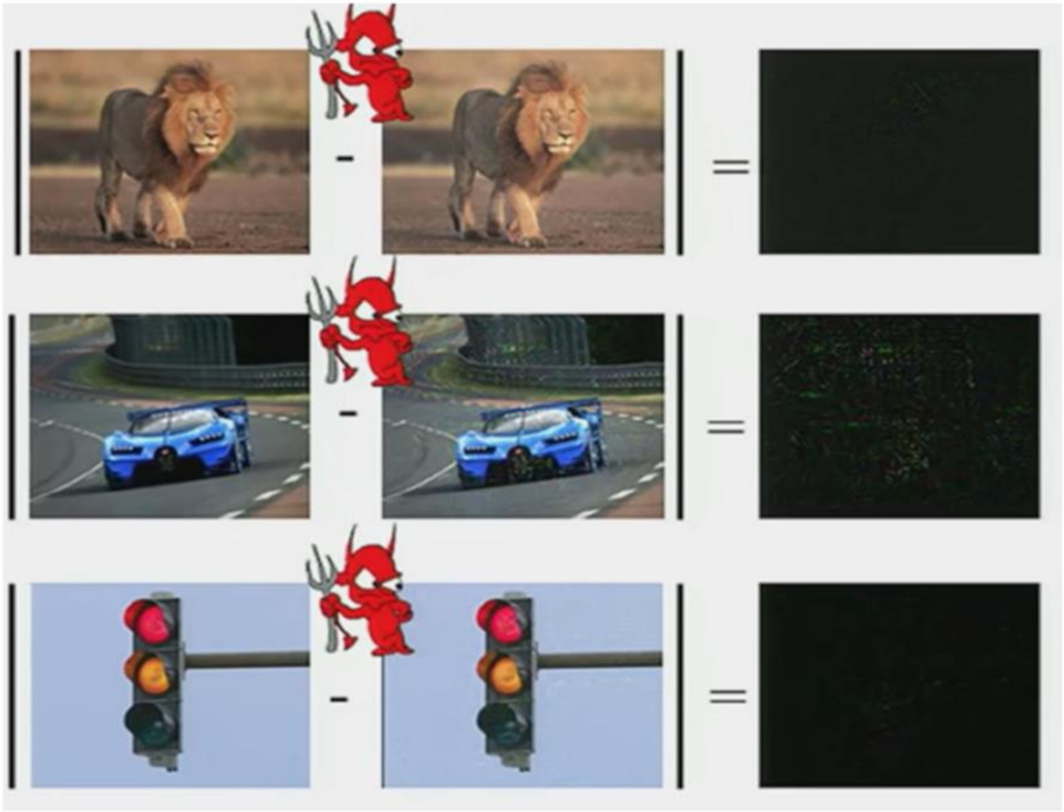
- **Adversarial example guides classifier to make wrong predictions**



Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Attacks

- **Difference between adversarial image and natural image is hardly noticeable**

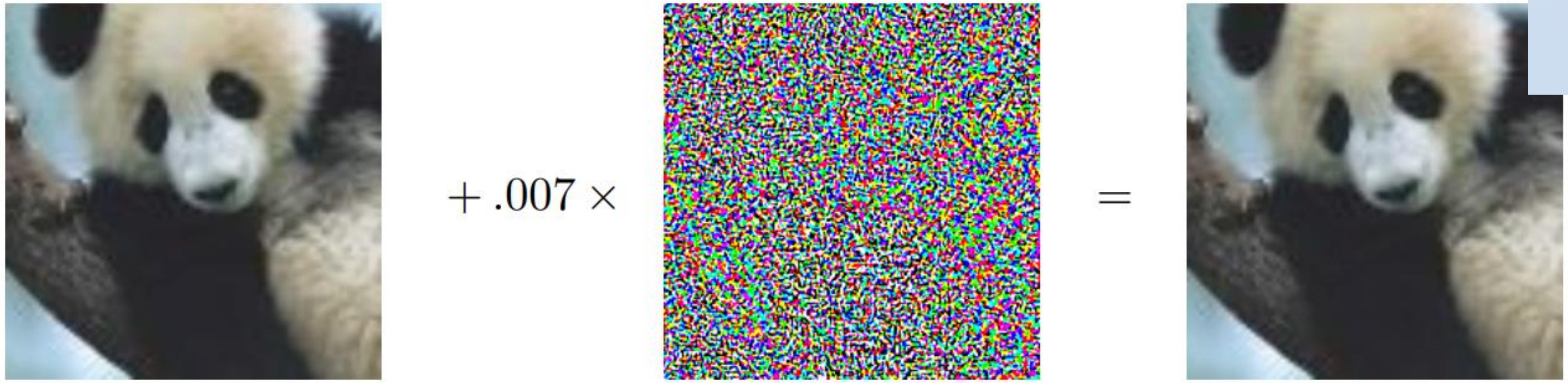


Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

How does Adversarial Attack Work?

- **Perturbation maliciously designed to fool machine learning models**
- **Formulated by *maximizing training loss***

gibbon



$\boldsymbol{x}$
"panda"
57.7% confidence

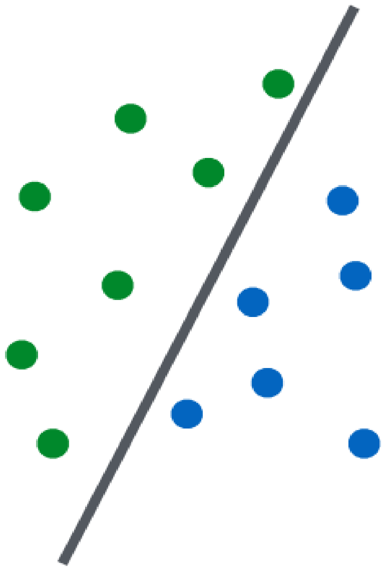
$\boxed{\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))}$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

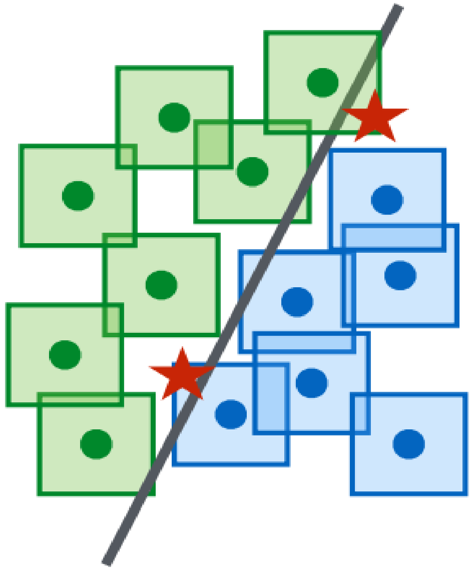
Goodfellow et al., Explaining and Harnessing Adversarial Examples. ICLR 2015

How does Adversarial Attack Work?

- **Adversarial examples are designed to cross the decision boundary of models**
- **The degree of perturbation on the data should be minimal**



(a) Trained binary classifier



(b) **Adversarial attack** crossing decision boundary

Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

Applicability of Adversarial Attacks

- **Adversarial examples can exist on possibly any:**
  - **Deep neural network (MLP, CNN, ViT, …)**
  - **Form of data (image, video, point cloud, mesh, …)**
  - **Task (classification, localization, generation, …)**

On the Robustness of Vision Transformers to Adversarial Examples

Kaleel Mahmood
Department of Computer Science and Engineering
University of Connecticut, CT, 06269, USA kaleel.mahmood@uconn.edu

Rigel Mahmood
Department of Computer Science and Engineering
University of Connecticut, CT, 06269, USA

Marten Van Dijk
CWI, Amsterdam
The Netherlands

On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective

Jindong Wang[1]*, Xixu Hu[1,2‡†], Wenxin Hou[3†], Hao Chen[4], Runkai Zheng[1,5‡†], Yidong Wang[6], Linyi Yang[7], Wei Ye[6], Haojun Huang[3], Xiubo Geng[3], Binxing Jiao[3], Yue Zhang[7], Xing Xie[1]

[1]Microsoft Research, [2]City University of Hong Kong, [3]Microsoft STCA, [4]Carnegie Mellon University, [5]Chinese University of Hong Kong (Shenzhen), [6]Peking University, [7]Westlake University

https://github.com/microsoft/robustlearn

Generating 3D Adversarial Point Clouds

Chong Xiang
Shanghai Jiao Tong University
Shanghai, China
xiangchong97@gmail.com

Charles R. Qi
Facebook AI Research
California, USA
charlesq34@gmail.com

Bo Li
University of Illinois at Urbana-Champaign
Illinois, USA
lxbosky@gmail.com

KAIST

Adversarial Attack on Matching-based Algorithms

- **There also have been adversarial attacks on matching-based algorithm**
- **NeRFool investigates the adversarial vulnerability of Generalizable NeRFs**

NeRFool: Uncovering the Vulnerability of Generalizable Neural Radiance Fields against Adversarial Perturbations

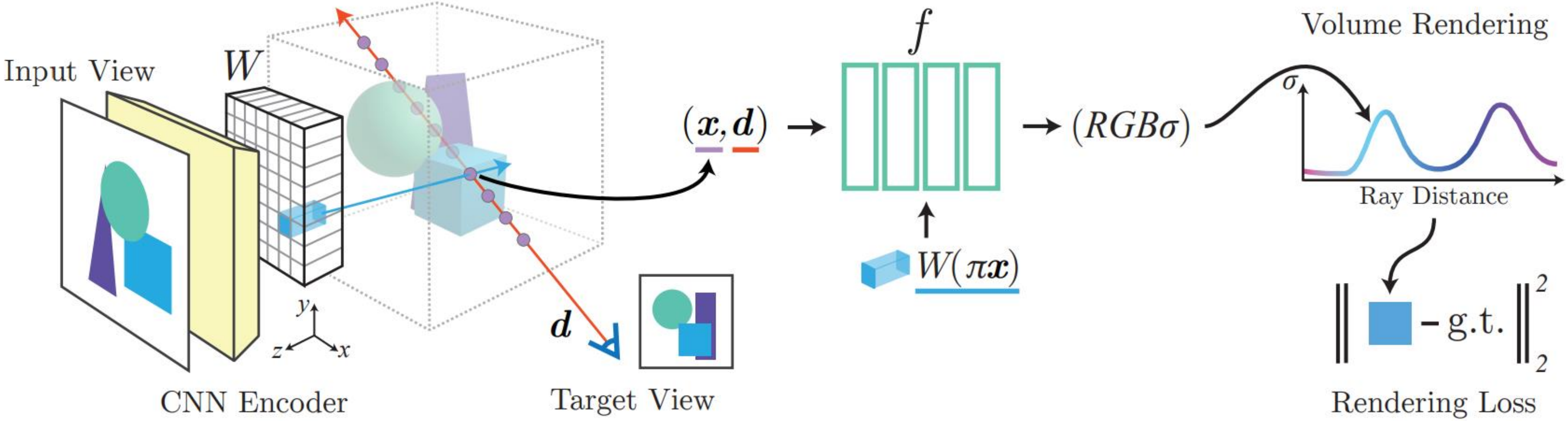
---

Yonggan Fu [1] Ye Yuan [1] Souvik Kundu [2] Shang Wu [1] Shunyao Zhang [3] Yingyan (Celine) Lin [1]

Fu et al., NeRFool: Uncovering the Vulnerability of Generalizable Neural Radiance Fields against Adversarial Perturbations, ICML 2023

Generalizable NeRFs

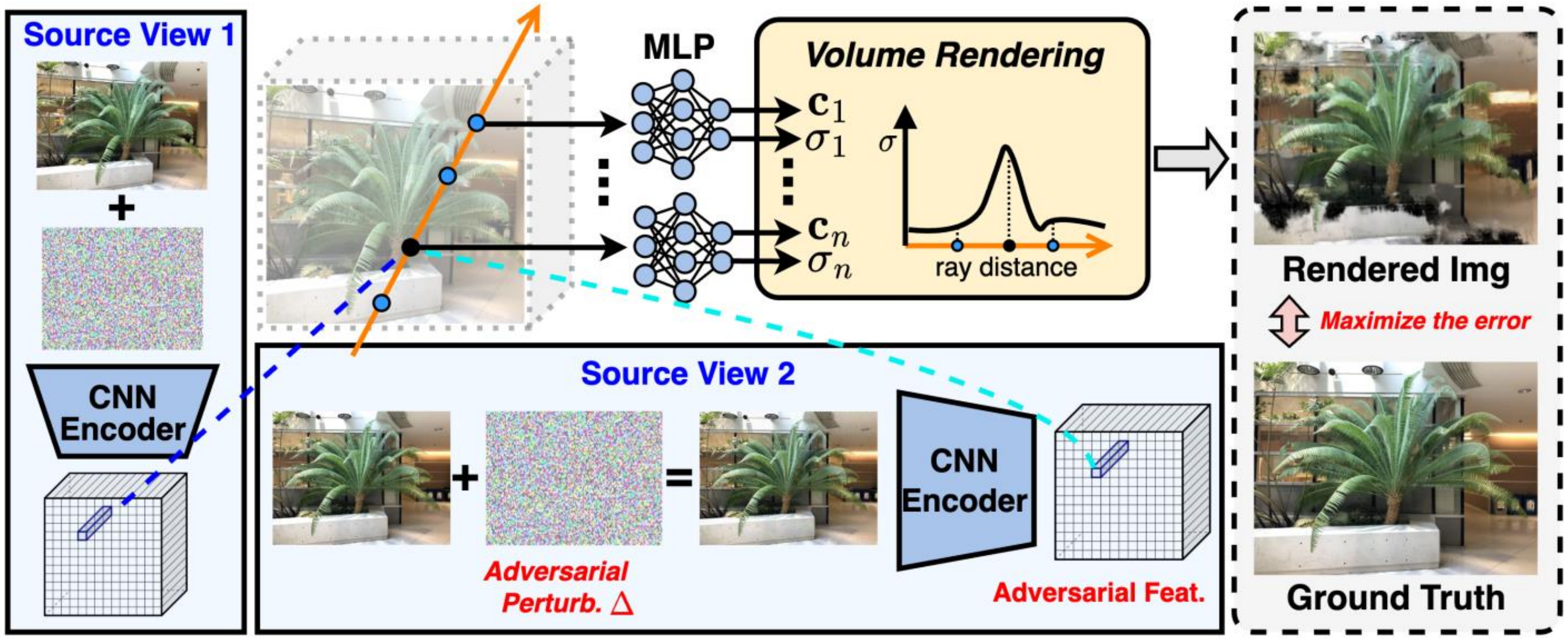
- **Originally, a single NeRF is fitted to a single 3D scene or object**
- **Generalizable NeRFs generalize a single NeRF model to *multiple scenes or objects***
- **They use a set of *"support images"* to condition the MLP network**



Yu et al., pixelNeRF: Neural Radiance Fields from One or Few Images, CVPR 2021

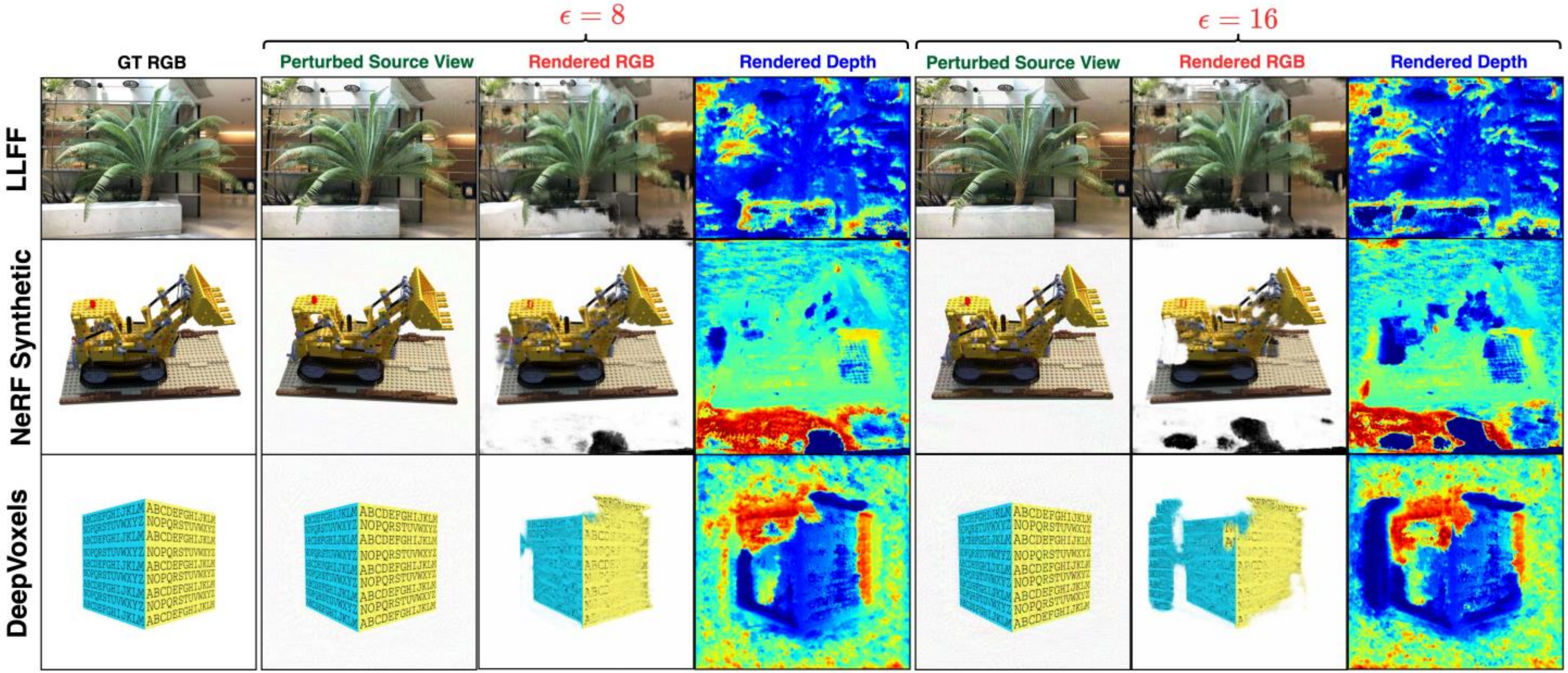
NeRFool

- **The goal of NeRFool is to fool the target G-NeRF to render disrupted images**
- **To do so, it crafts *<u>"adversarial support images"</u>***



Fu et al., NeRFool: Uncovering the Vulnerability of Generalizable Neural Radiance Fields against Adversarial Perturbations, ICML 2023

NeRFool



Fu et al., NeRFool: Uncovering the Vulnerability of Generalizable Neural Radiance Fields against Adversarial Perturbations, ICML 2023

Q&A