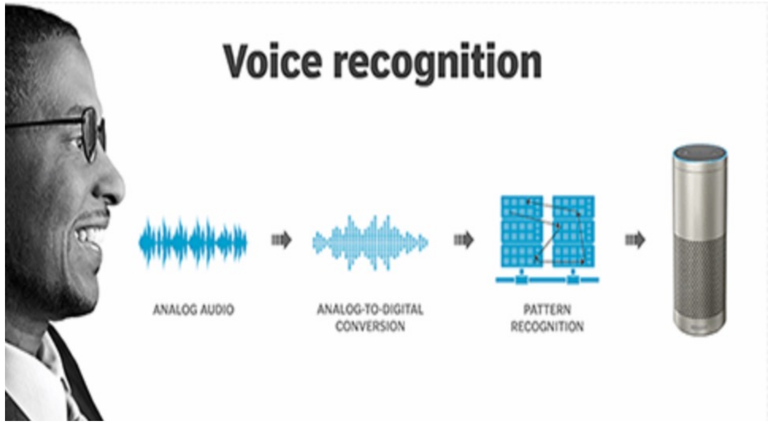
Adversarial Machine Learning in the 3D domain

Chaowei Xiao
NVIDIA & ASU

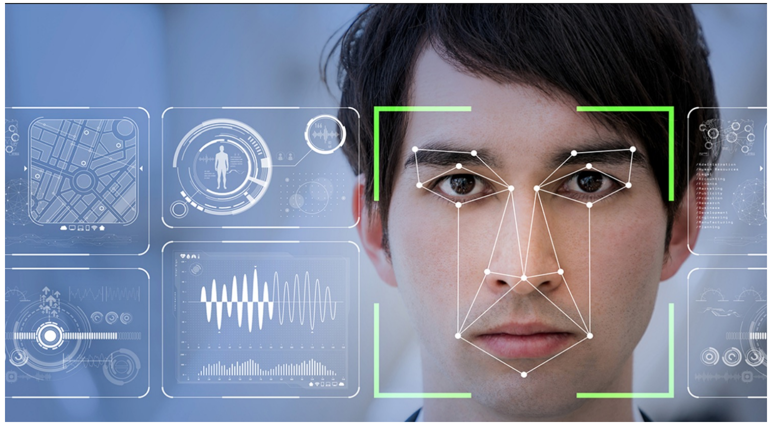
Deep Learning: Good Story

Autonomous Driving

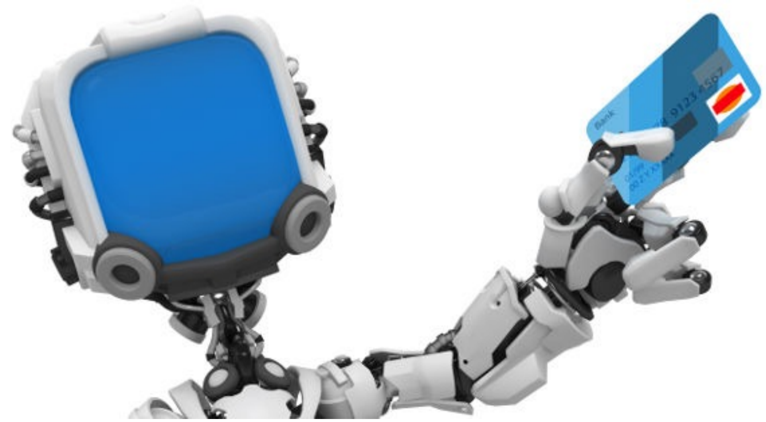


Voice recognition

Game



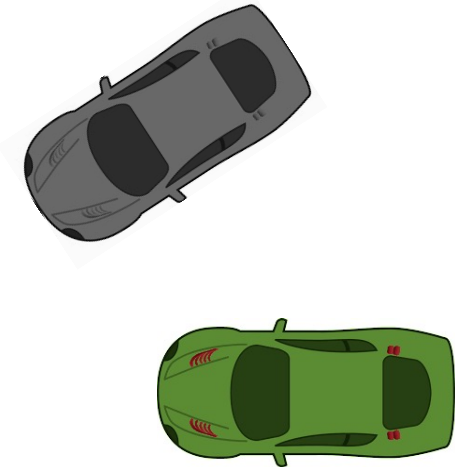
Face recognition



Fraud Detection

Malware Classification

Deep Learning: Bad Story



Not detected

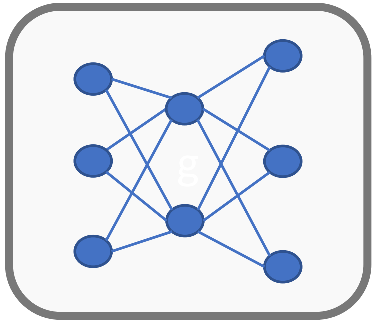
[Photo credits: nytimes]

Perils of Stationary Assumption

Input

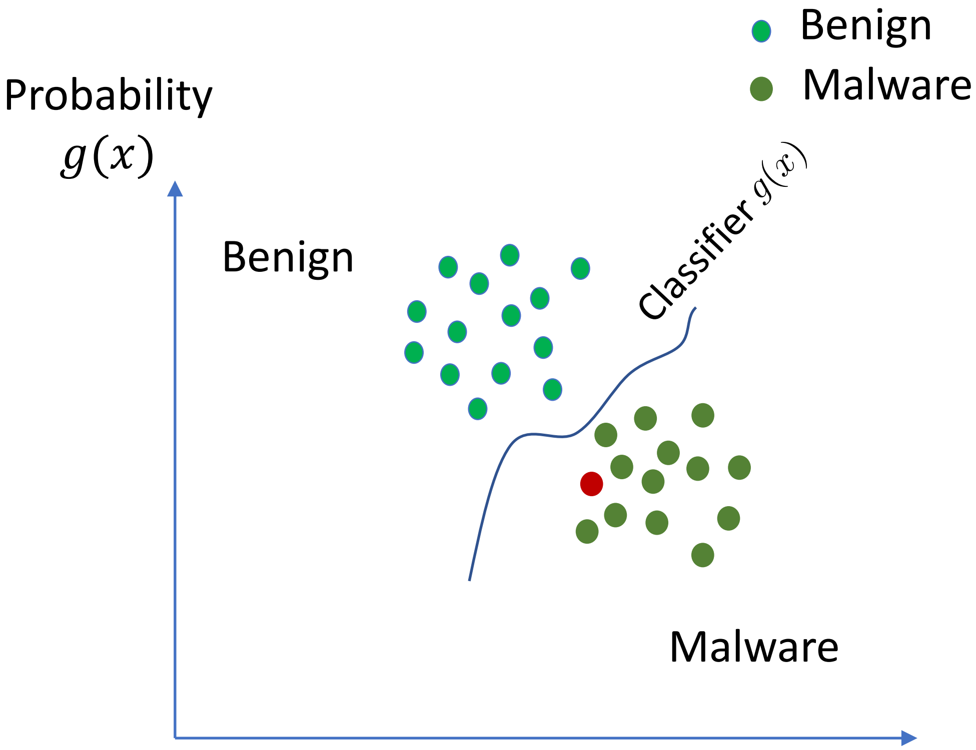
x

Machine Learning Model



g

Output



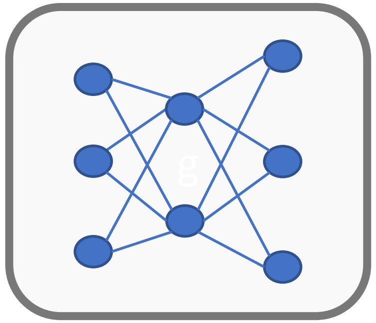
Assumption: Training Data ≈ Test Data

Perils of Stationary Assumption

Input

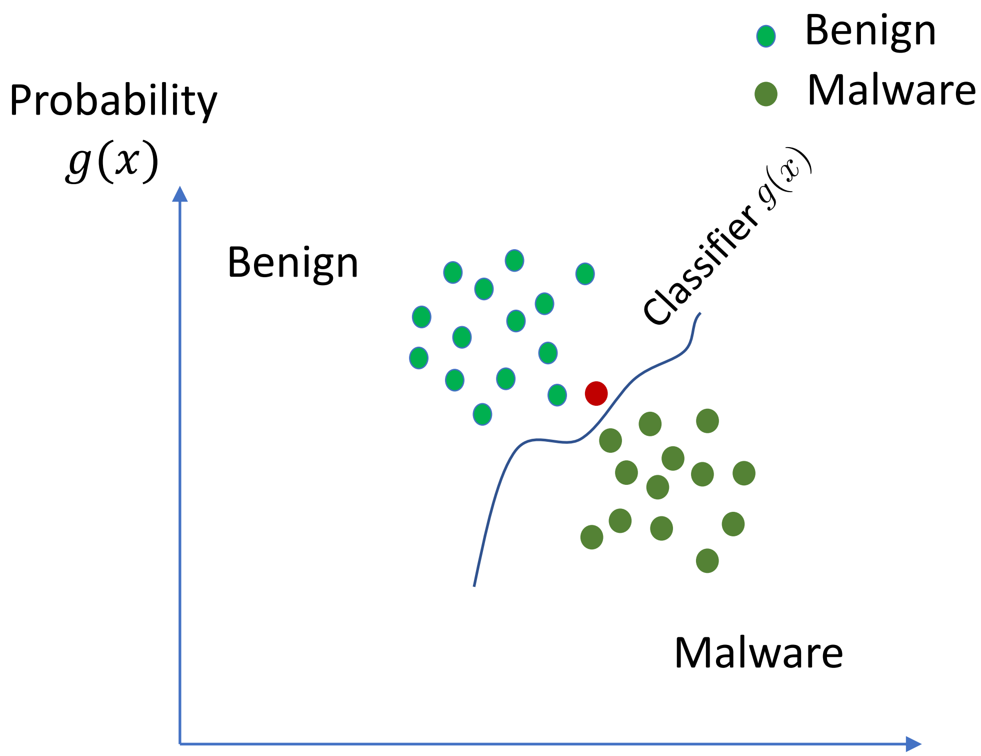
x

Machine Learning Model



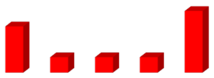
g

Output



Assumption:

Training Data

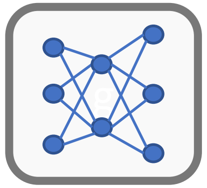


! ≈

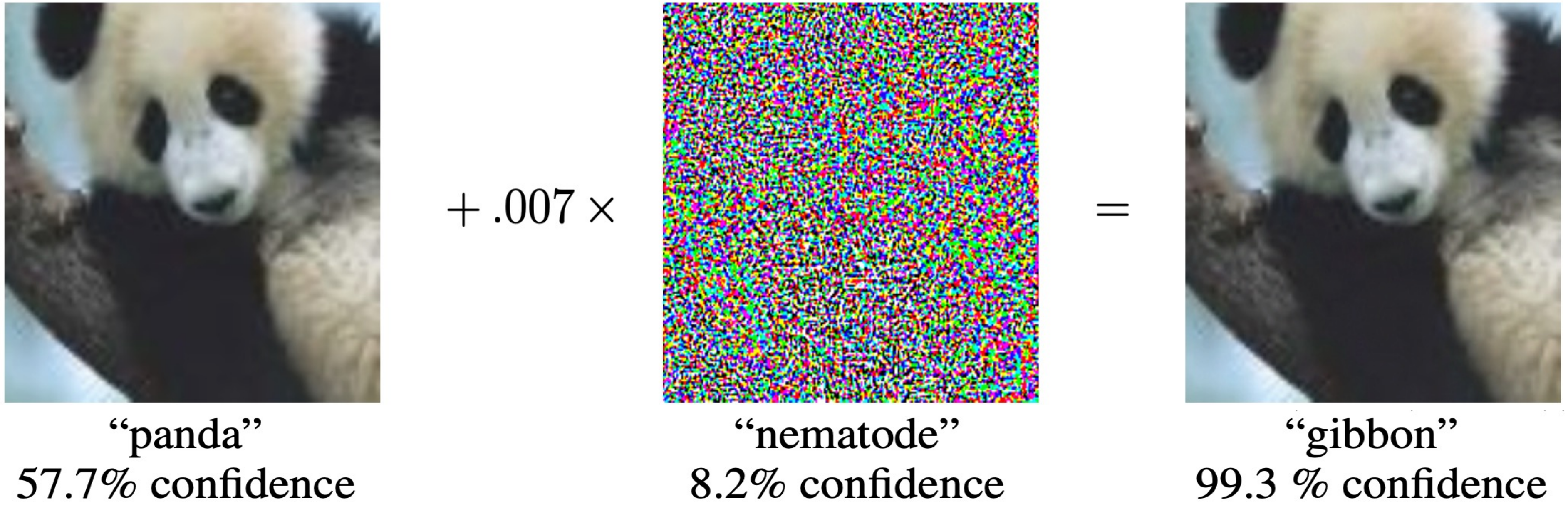
Test Data

Adversarial Environments

Could Attackers Systematically Find these Inputs?



g

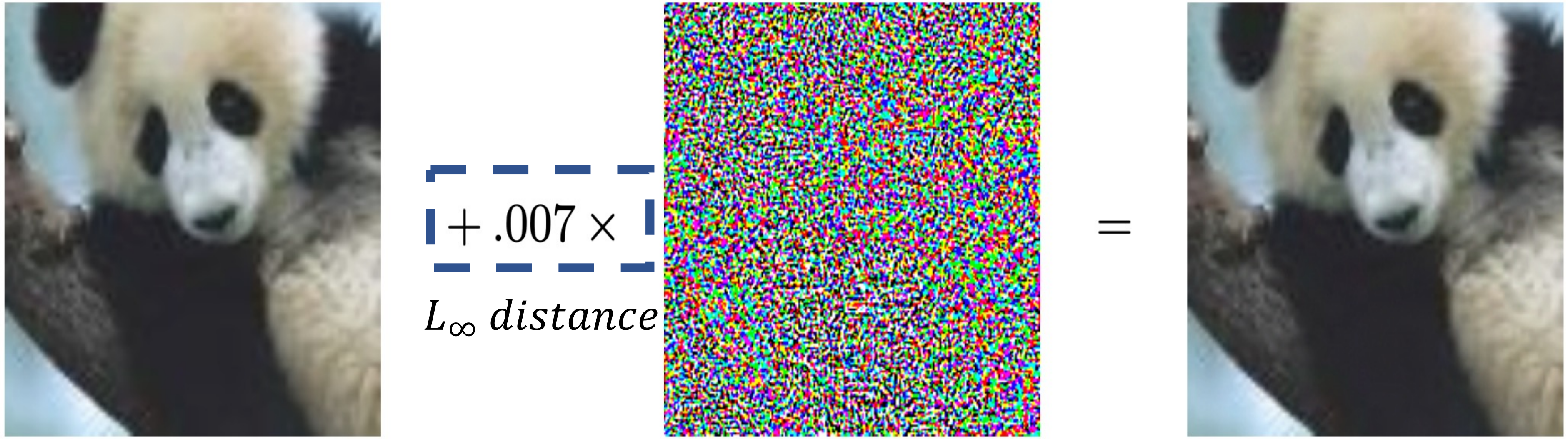


x θ x_{adv}

[Photo credit: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES]

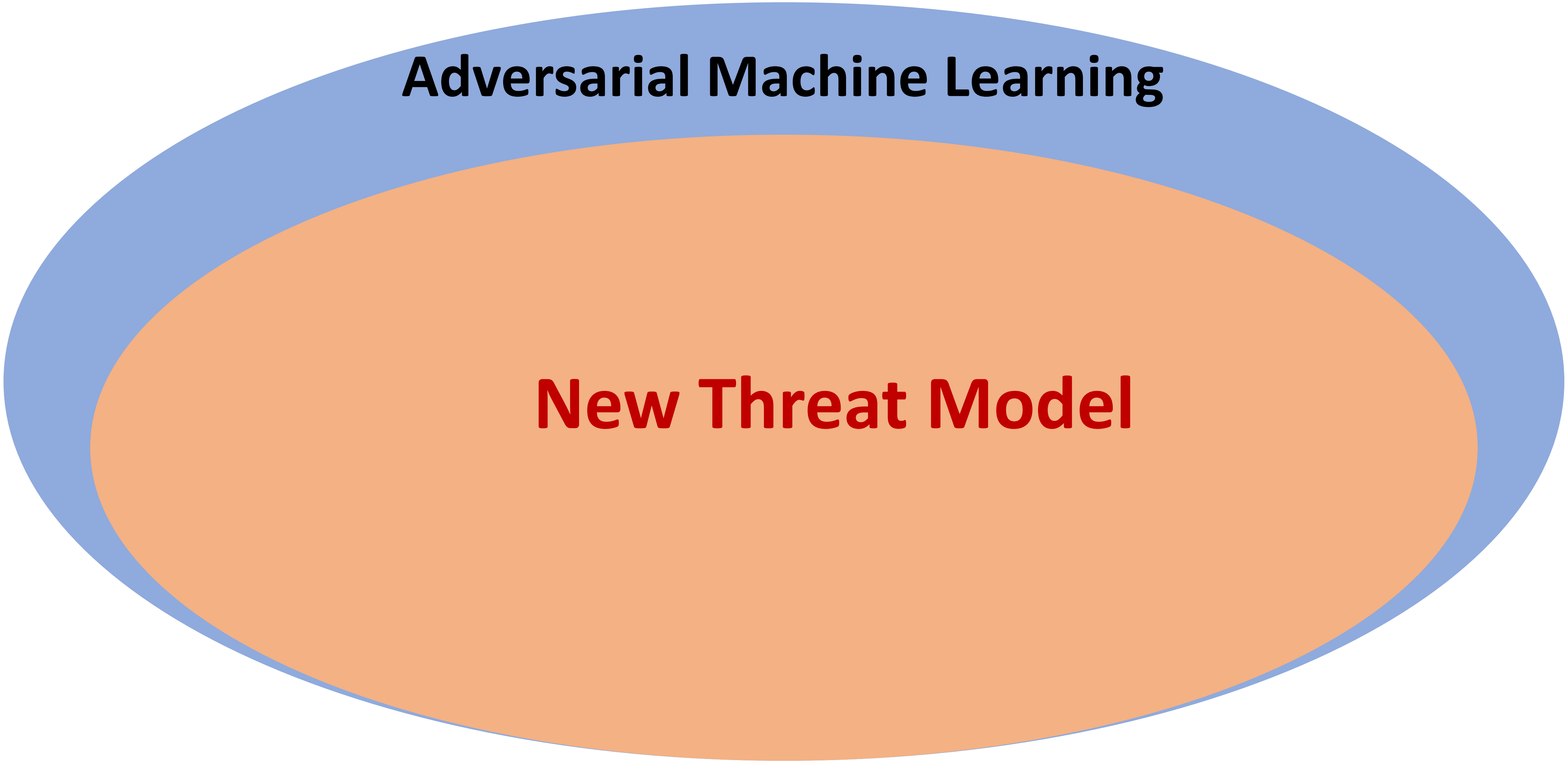
Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. ICLR' 14
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. ICLR' 15

Threat Model

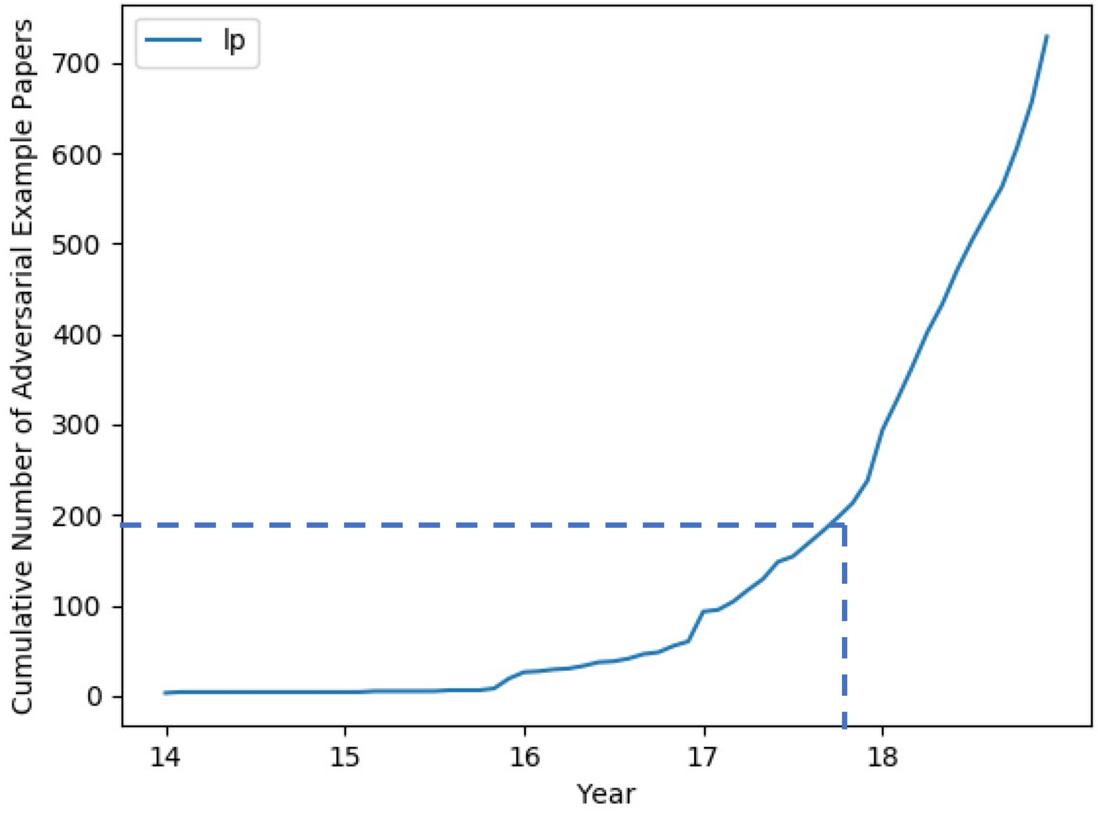


L_p has been used as threat model of adversarial examples

[Photo credit: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES]


Adversarial Machine Learning
New Threat Model

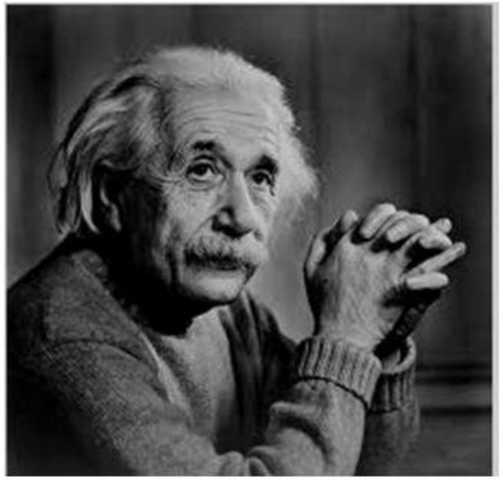
Threat Model

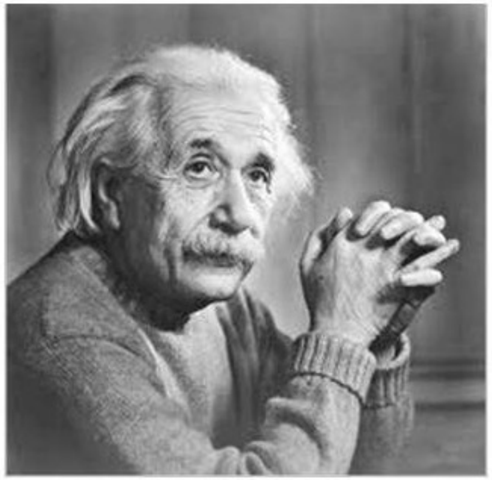


Number of Papers related to Adversarial Example in different years

[Photo credit: nicholas.carlini]

Limitation

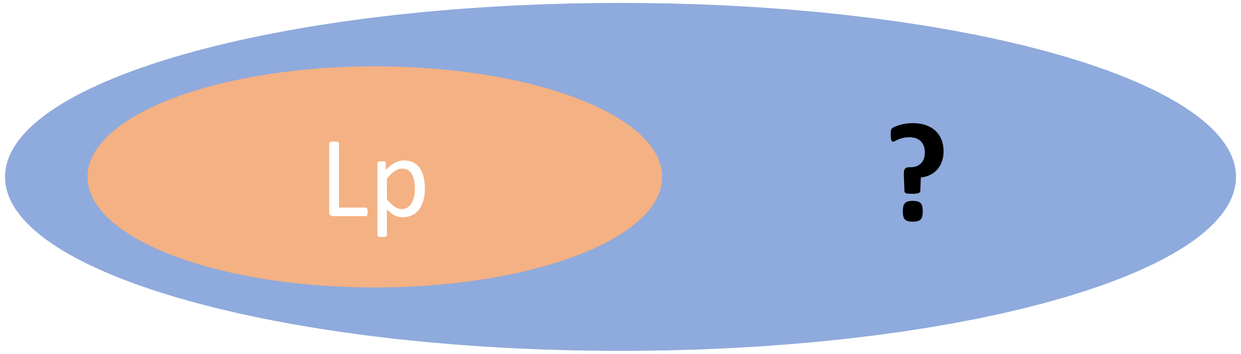




Lighting

Pixel Shift

LP is not a good metric to evaluate the "look like"

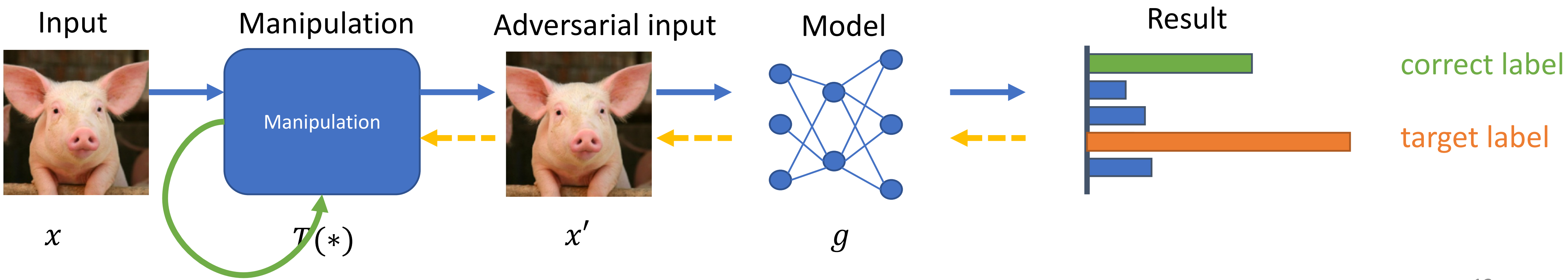


A New Threat Model

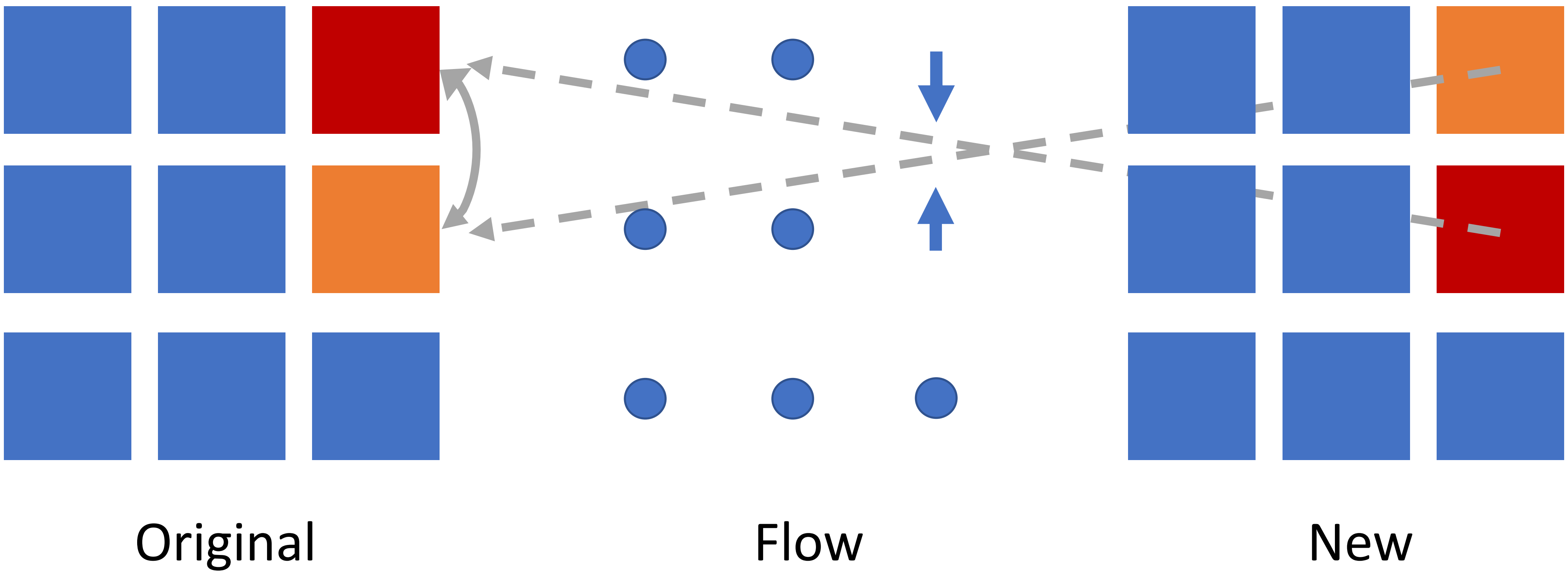
Adversarial examples should be the inputs which could be correctly recognized by humans but mislead machine learning models

$$L = L_{adv}(x; T, g) + \tau L_{perceputal}(x; T)$$

Mislead machine learning model Correctly recognized by humans



New Adversarial Examples



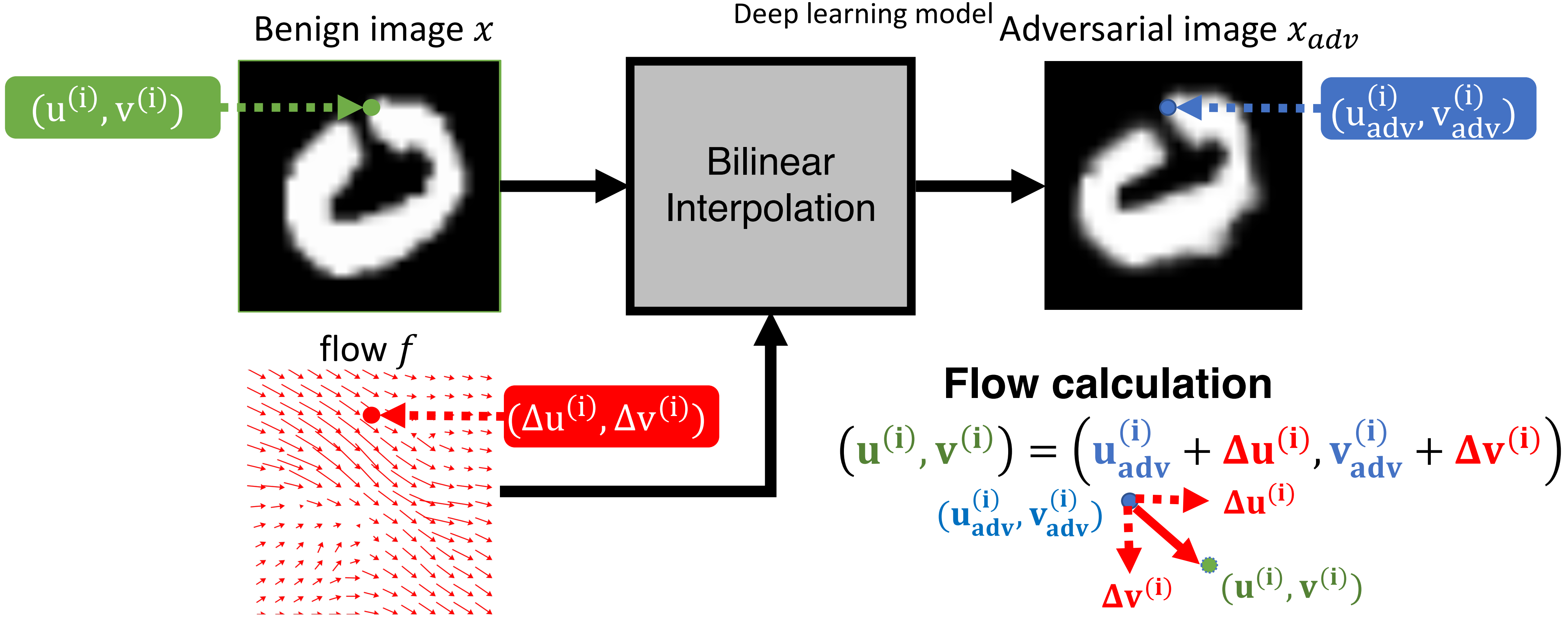
Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, Dawn Song. Spatially transformed adversarial examples. ICLR’ 18.

Spatially Transformed Adversarial Examples

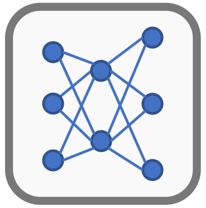
Targeted label

$$f^* = argmin_f \ L_{adv}(f; x, g, t) + \tau L_{perceptual}(f)$$

Deep learning model



Adversarial & Perceptual Loss

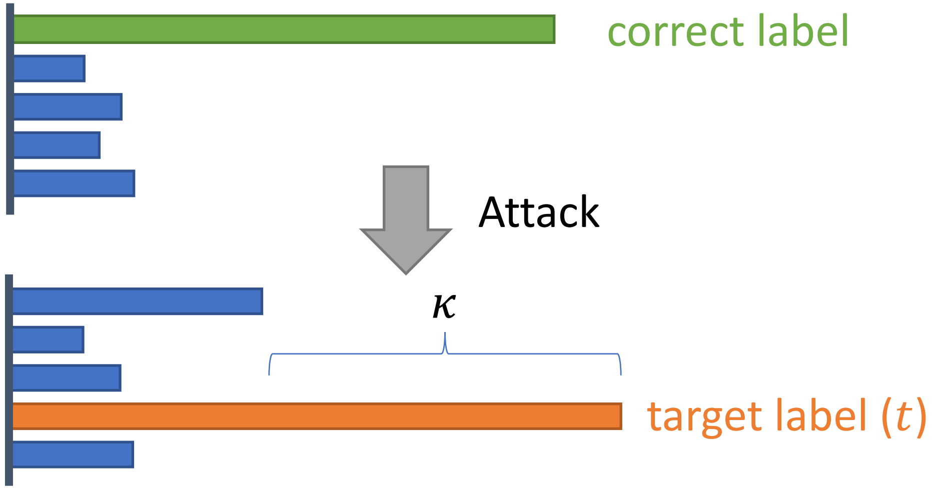


g

- Adversarial Loss L_{adv}[1]

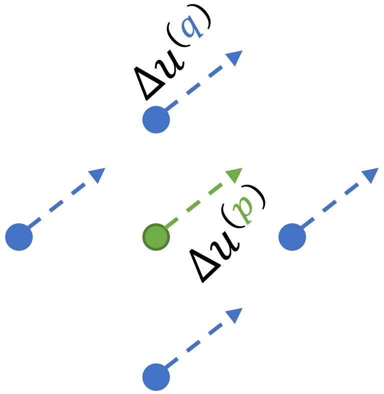
$$\max\left(\max_{i \neq t} g(x_{adv})_i - g(x_{adv})_t, -\kappa\right)$$

- Change the predicted results



- Perceptual Loss $L_{perceptual}$

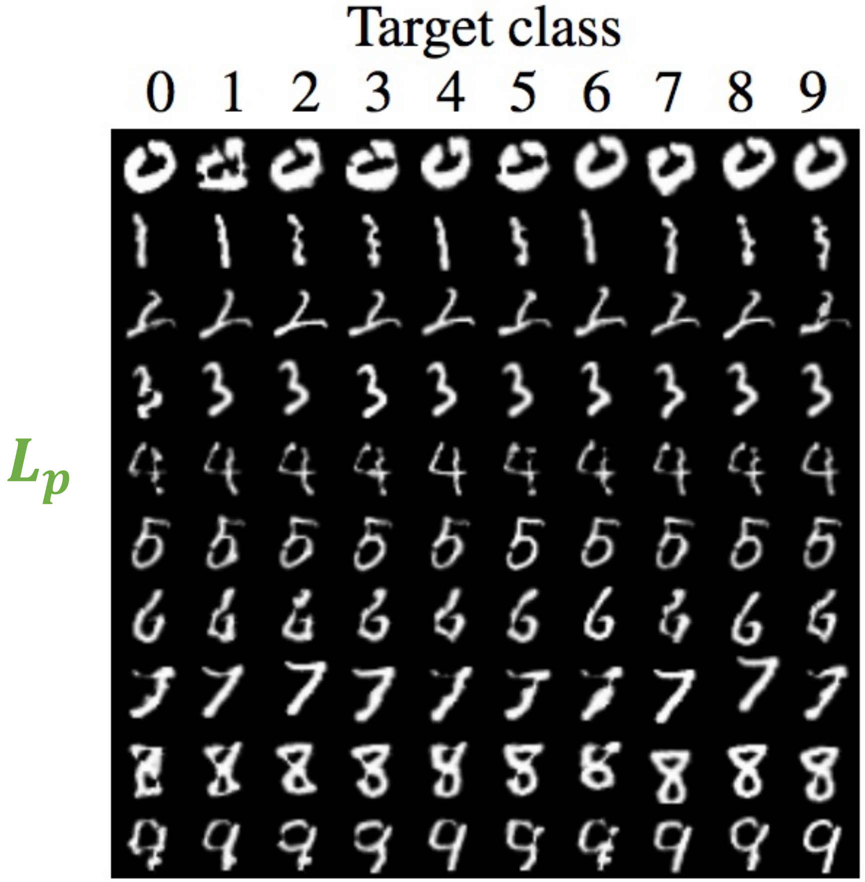
$$L_{perceptual}(f) = \Sigma_p^{all\ pixels} \Sigma_{q \in N(p)} \sqrt{\left|\Delta u^{(p)} - \Delta u^{(q)}\right|_2^2 + \left|\Delta v^{(p)} - \Delta v^{(q)}\right|_2^2}$$



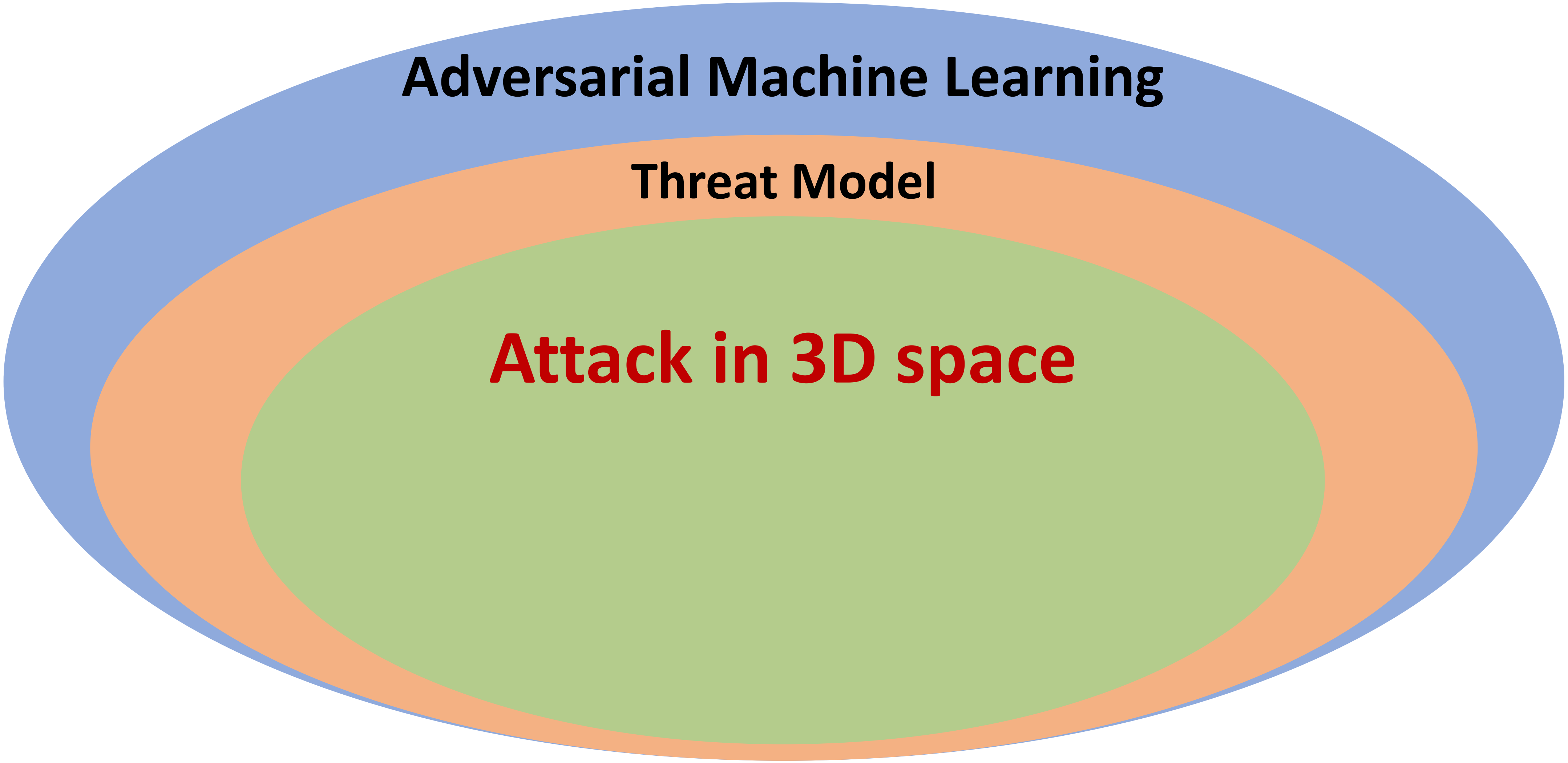
[1]Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." IEEE S&P' 17.

Spatial Transformed Adversarial Examples

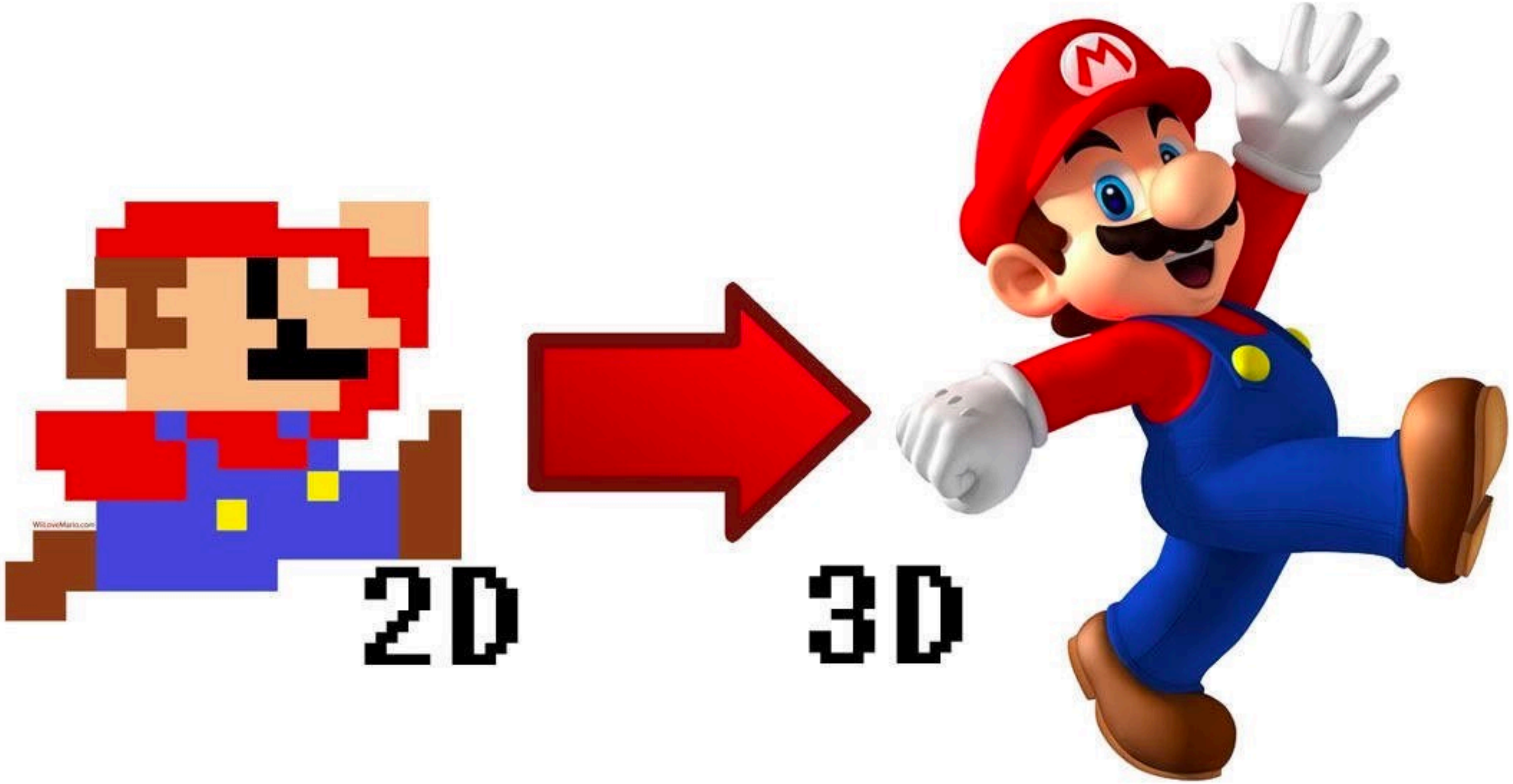
Target label: 0



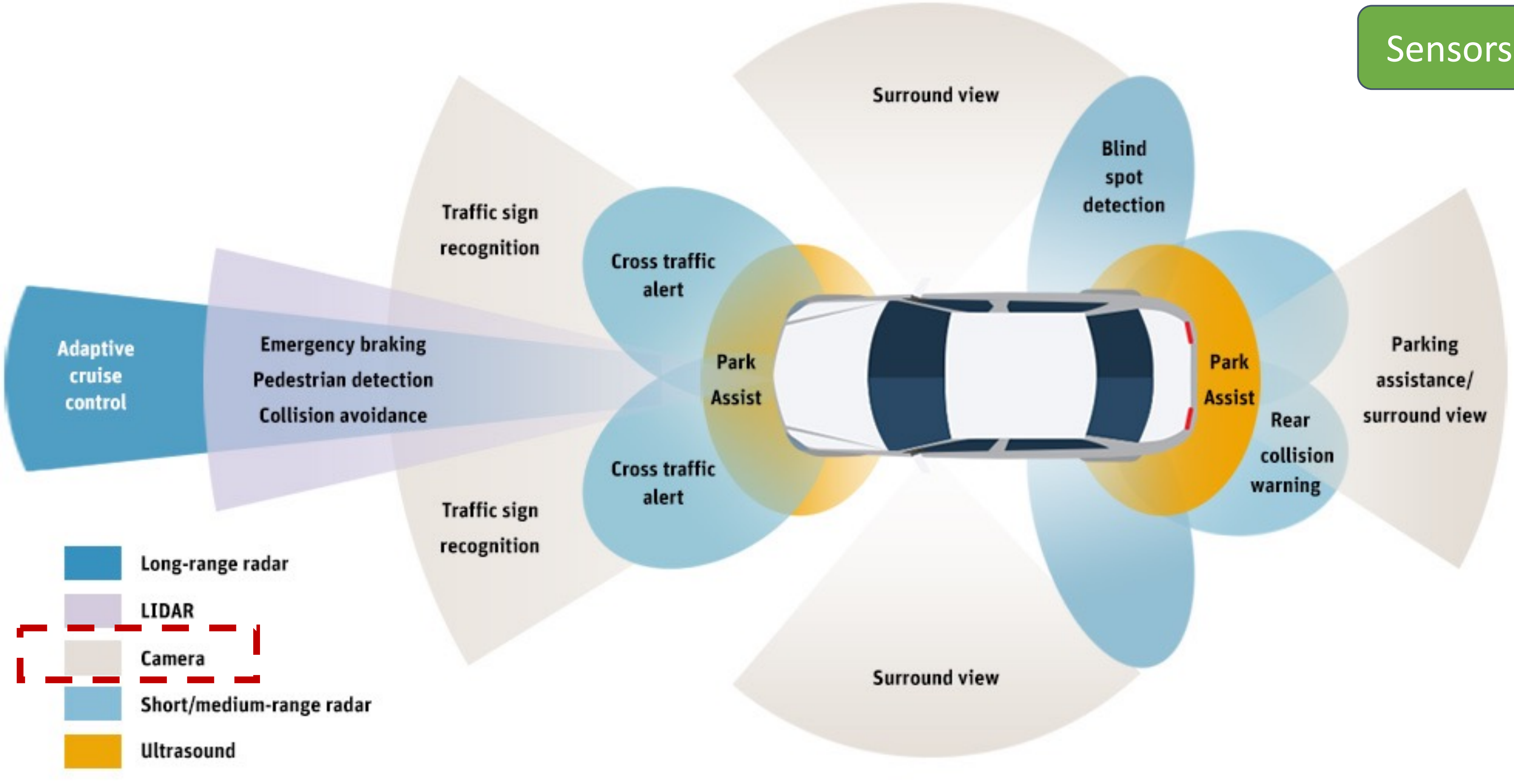
Adversarial examples generated by stAdv on MNIST.
The ground truth are shown in the diagonal.


Adversarial Machine Learning
Threat Model
Attack in 3D space

Adversarial Examples in the Physical World



Autonomous Vehicle (AV) Architecture



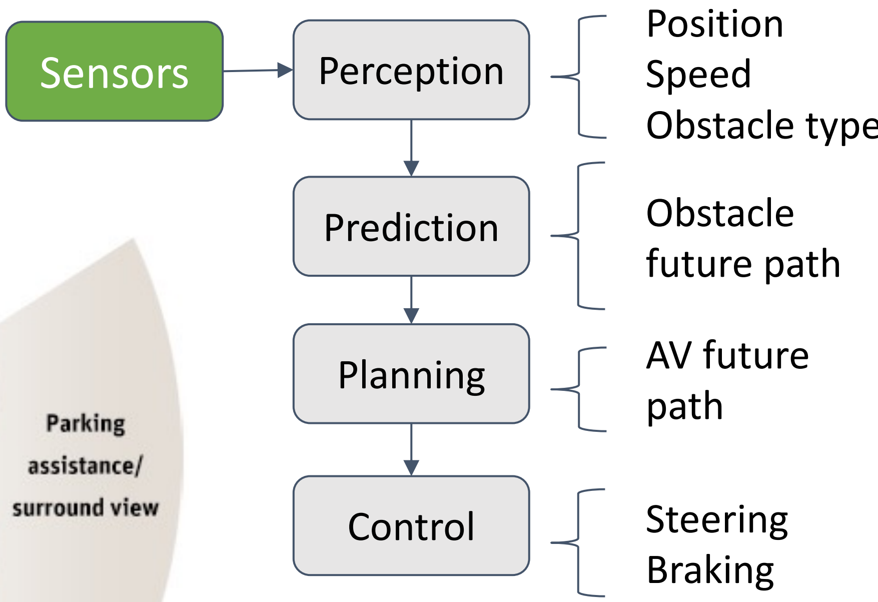
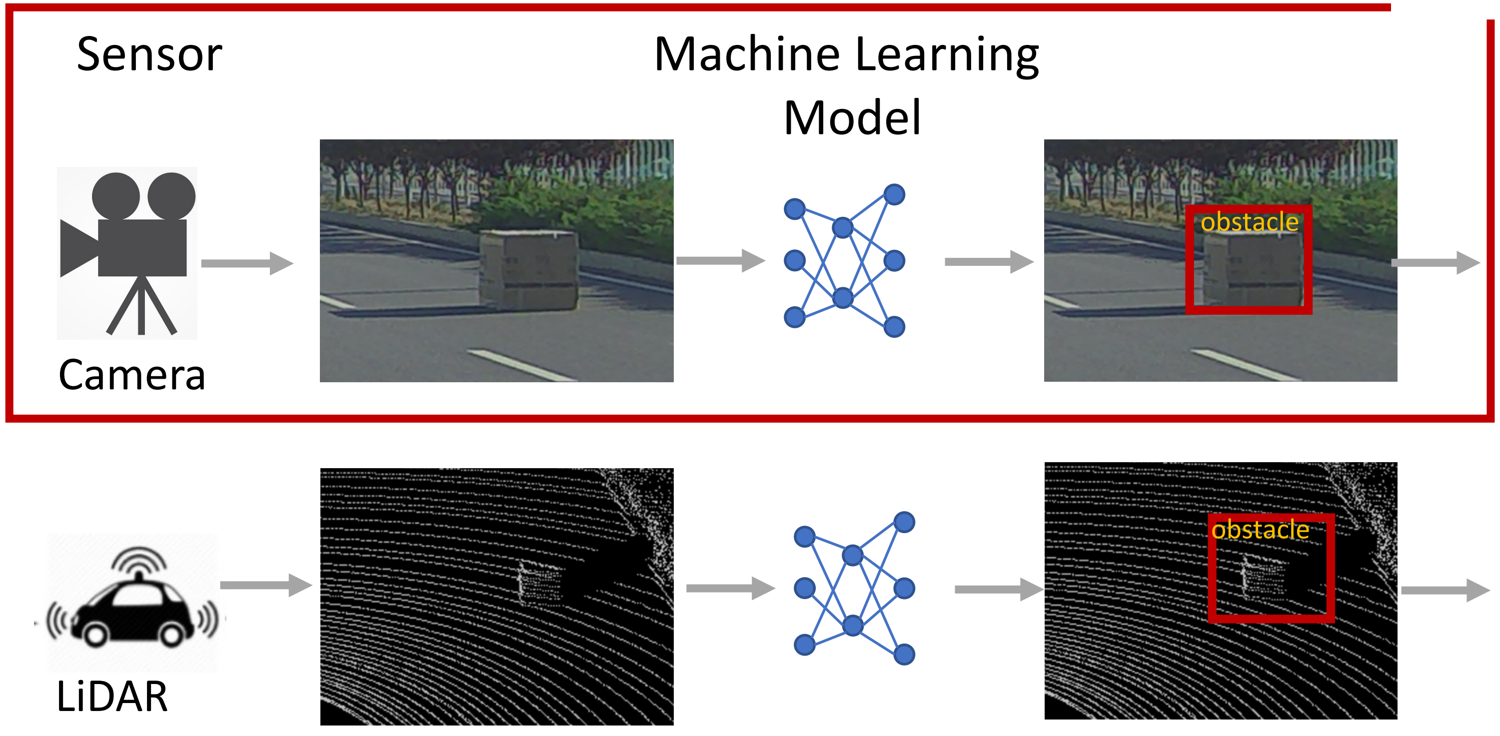


Photo credits: advantage magazine

AV Perception

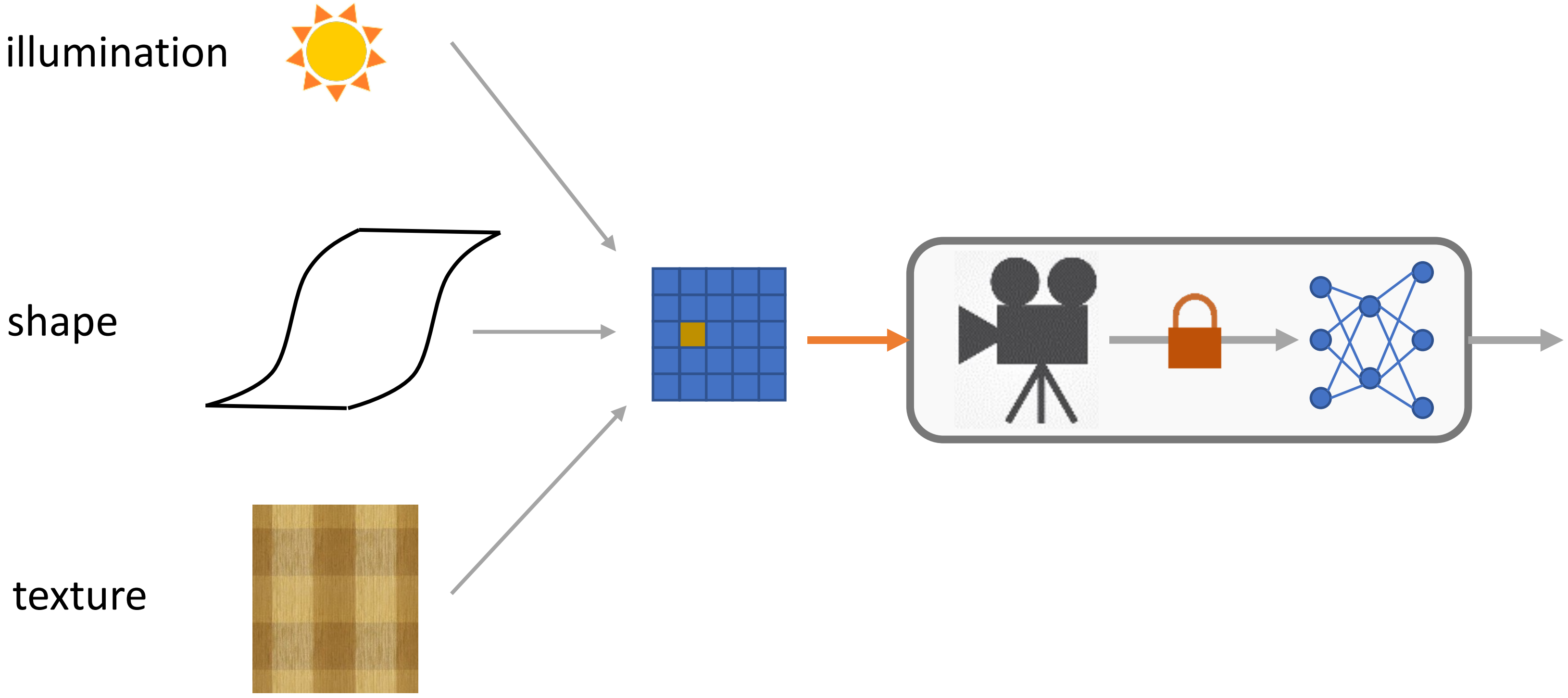


• • •

Shape information

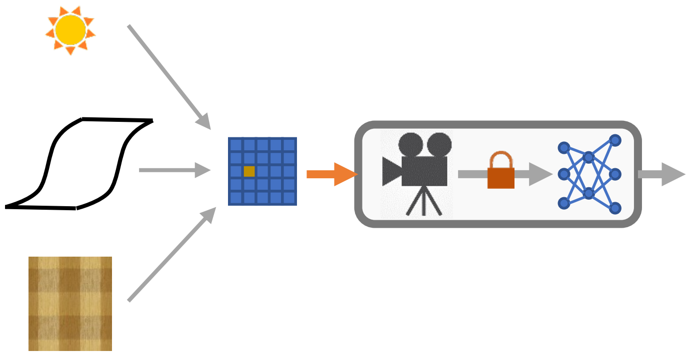
Could we generate an adversarial object to mislead the real-world LiDAR system?

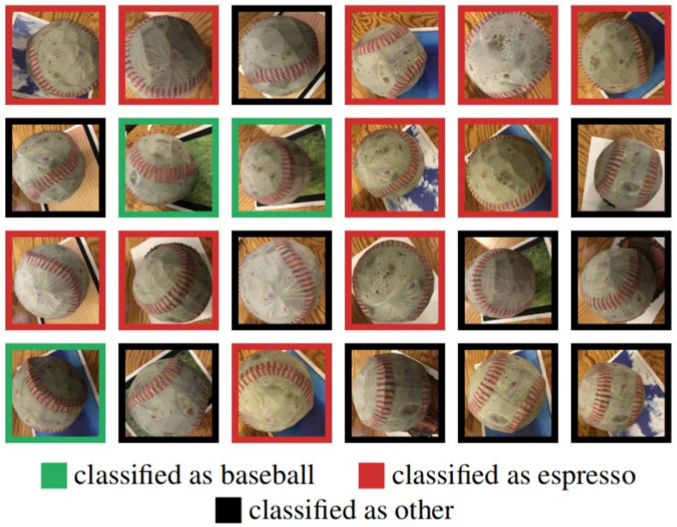
Adversarial Attacks: Physical Domain



Adversarial Attacks: Physical Domain

- Physically Possible Adversarial Examples





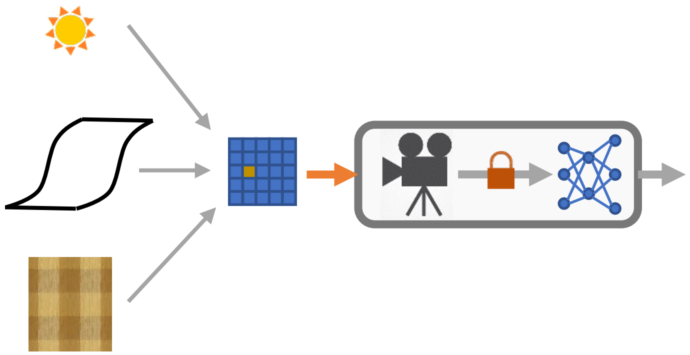
Athalye et al.

Evtimov et al.

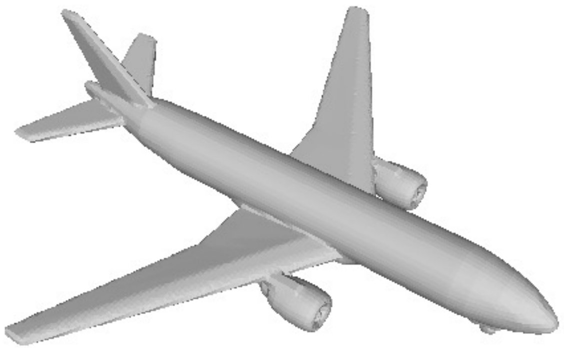
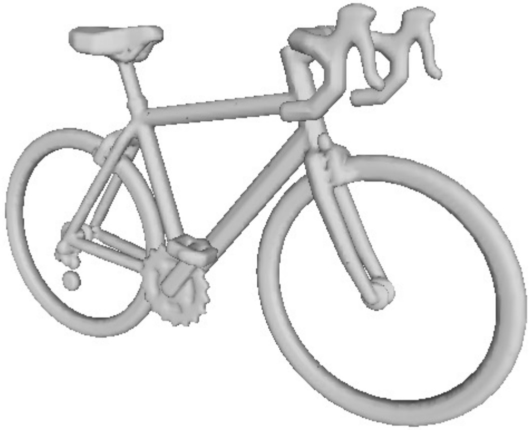
(a) Printout (b) Photo of printout (c) Cropped image

Kurakin et al.

Physical Domain: Shape and Texture



- Starting from textureless objects
- Rich geometric features but minimal texture variation

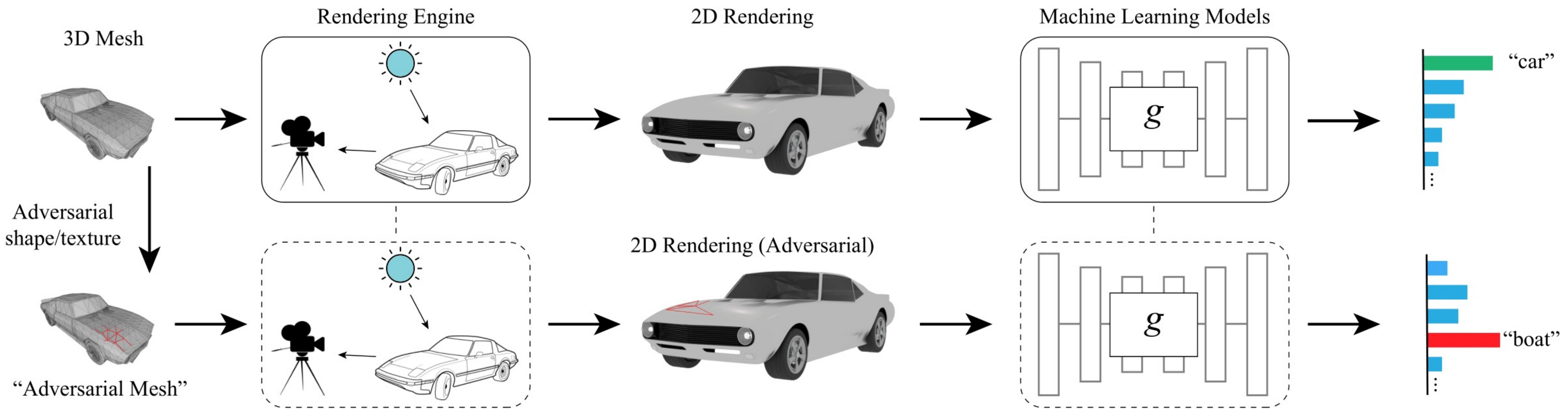




Shapes from PASCAL3D+ by Xiang et al.

Our Attacking Pipeline

- Input: a 3D mesh + shape/texture perturbations
- Render: a differentiable renderer
- Target: fool a machine learning model and keep the shape plausible

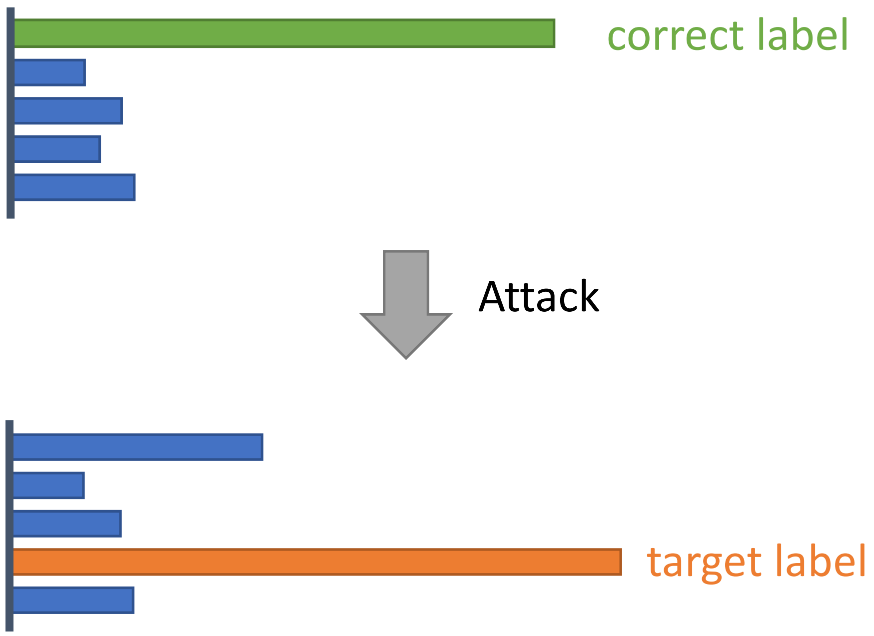
$$\mathcal{L}(S^{\text{adv}}; g, y') = \mathcal{L}_{\text{adv}}(S^{\text{adv}}; g, y') + \lambda \mathcal{L}_{\text{perception}}(S^{\text{adv}})$$



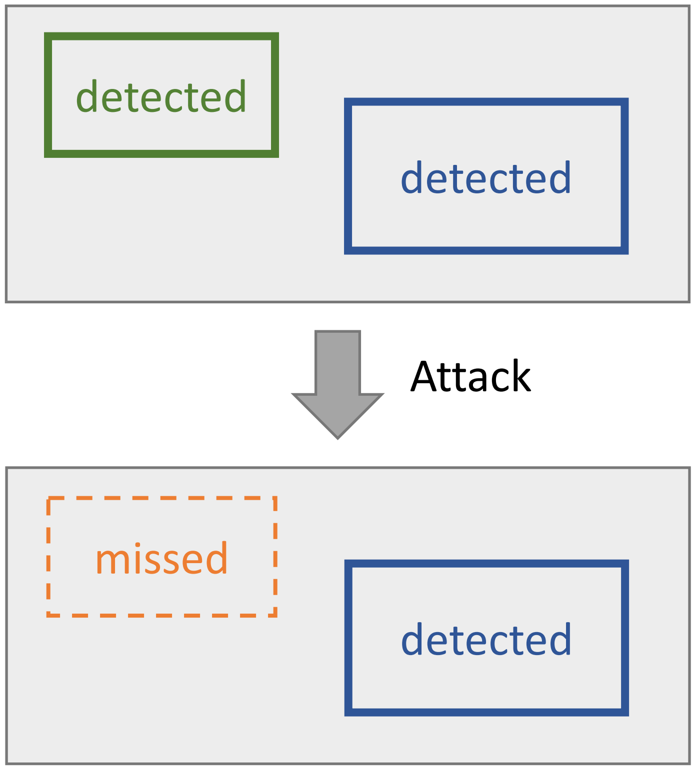
Xiao, Chaowei, Yang, Dawei et al. "Meshadv: Adversarial meshes for visual recognition." CVPR. 2019.

Adversarial Target & Loss

- Classification: cross entropy
  - Change the prediction label



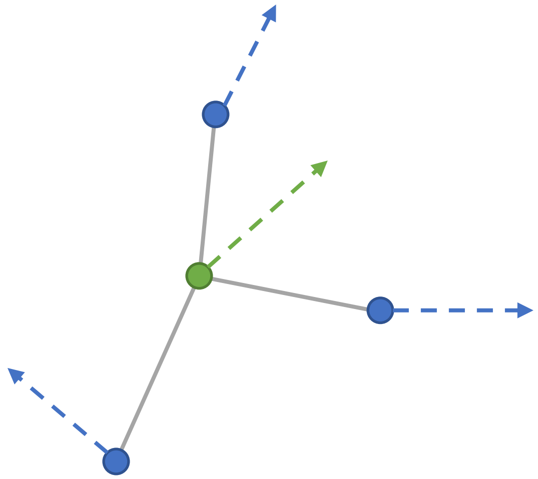
- Detection: the disappearance attack loss (Eykholt et al.)
  - Remove the targeted detection



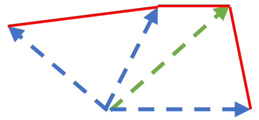
Perceptual Loss

- 3D Laplacian loss, operated on vertex displacements
  - Neighboring vertices should be perturbed along similar directions

$$\mathcal{L}_{\text{perception}}(S^{\text{adv}}) = \sum_{\vec{v}_i \in V} \sum_{\vec{v}_q \in \mathcal{N}(\vec{v}_i)} \|\Delta\vec{v}_i - \Delta\vec{v}_q\|_2^2$$



Perturbation of neighboring vertices



Perturbation differences

$\mathcal{L}_{\text{perception}}$

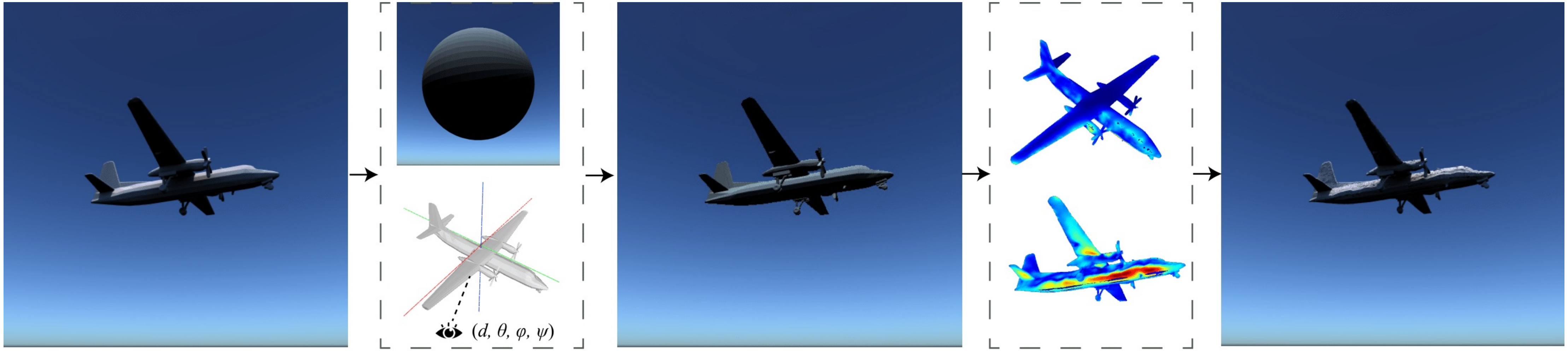
3D Laplacian Loss

Experiments: Classification

Perturb. Type	Model	Test Accuracy	Best Case	Average Case	Worst Case
Shape	DenseNet	100.0%	100.0%	100.0%	100.0%
	Inception-v3	100.0%	100.0%	99.8%	98.6%
Texture	DenseNet	100.0%	100.0%	99.8%	98.6%
	Inception-v3	100.0%	100.0%	100.0%	100.0%

Transfer to the Black-box Renderer

- Airplane + Mitsuba renderer + Skylight



Black-box rendering of an airplane mesh
Prediction: "airliner"

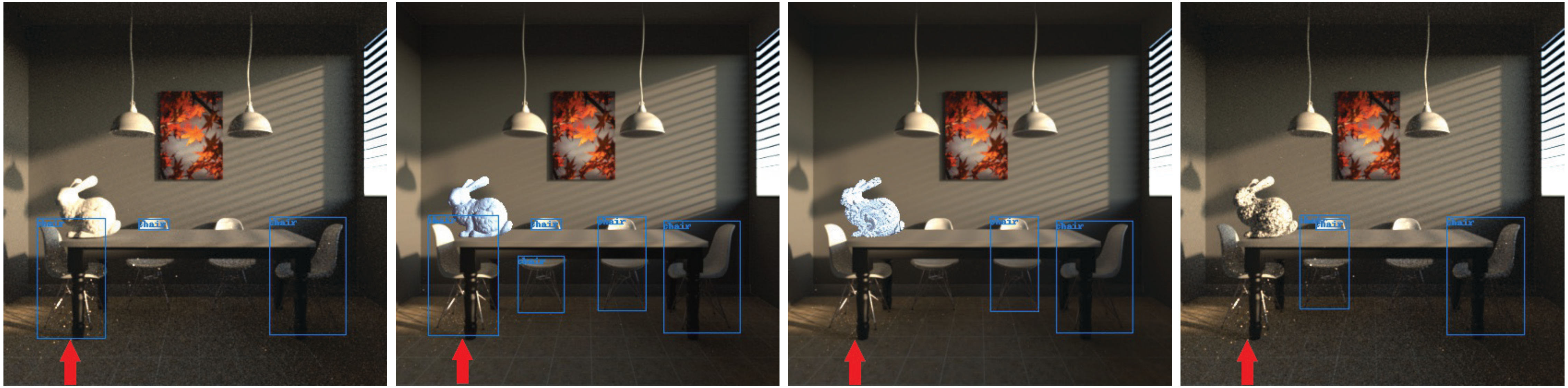
Viewpoint and lighting estimate

NMR rendering + Background

Shape perturbations by *MeshAdv*
Target: "hammerhead"

Re-render using the black-box renderer
Prediction: "hammerhead"

Transfer to the Black-box Renderer



Before Attack
Mitsuba Renderer

Before Attack
Neural Mesh Renderer

After Attack
Neural Mesh Renderer

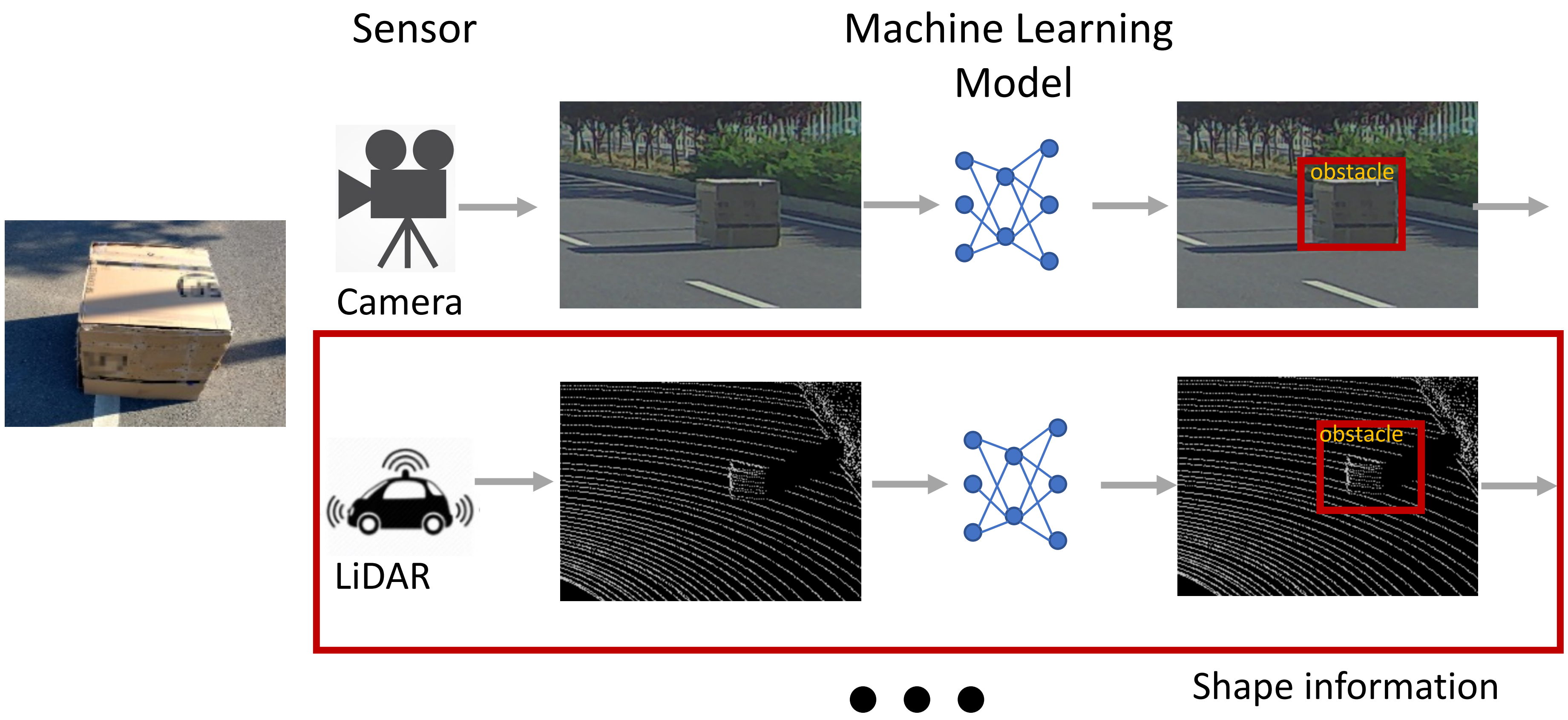
After Attack
Mitsuba Renderer

Search lighting and poses

White-box attack

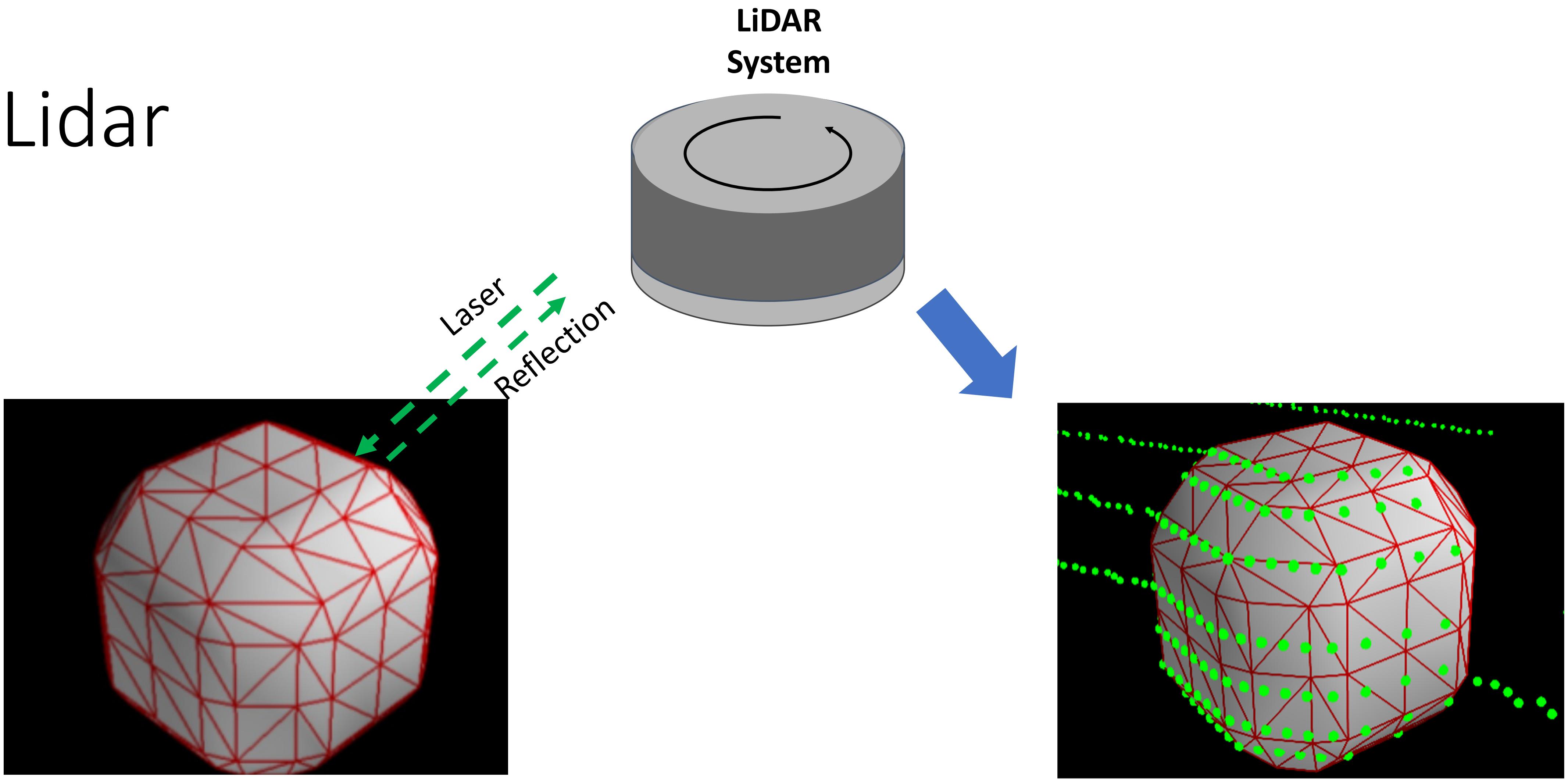
Black-box transfer

AV Perception

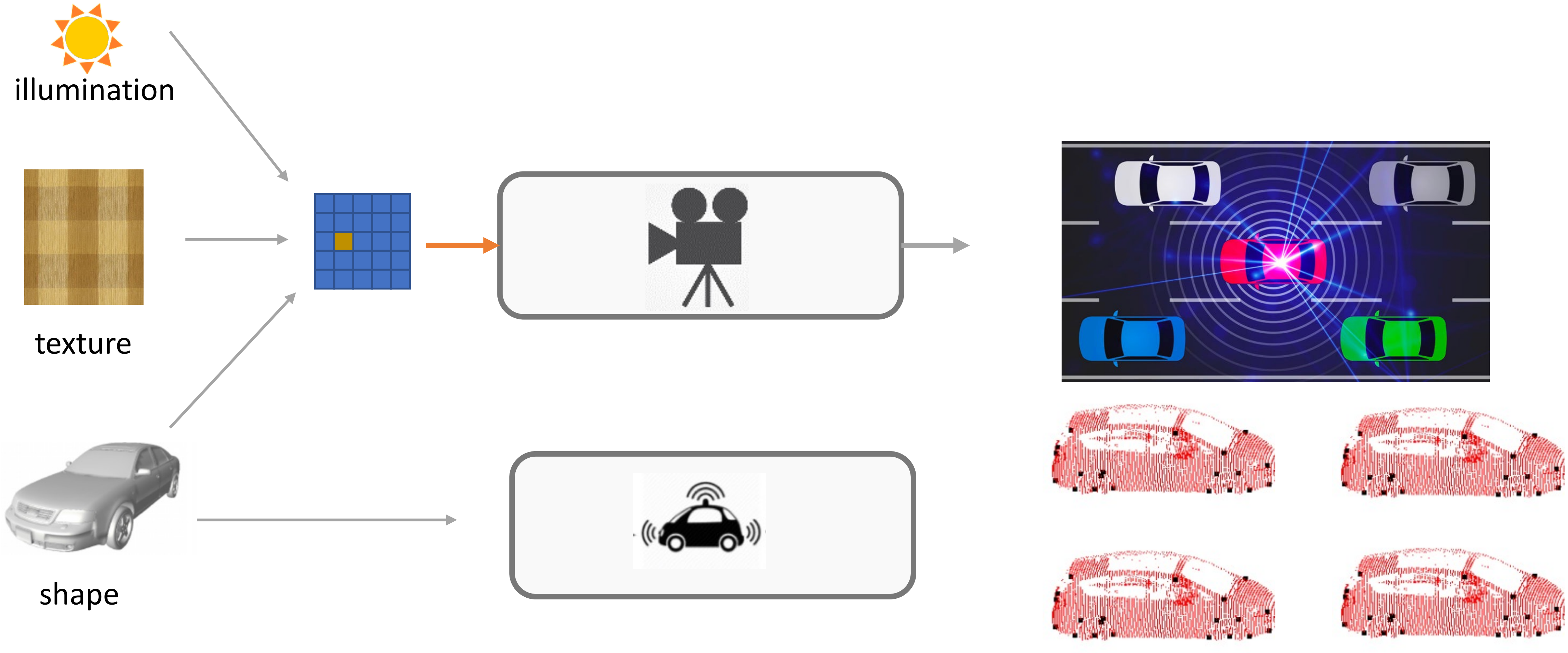


Could we generate an adversarial object to mislead the real-world LiDAR system?

Lidar



What Should We Manipulate?



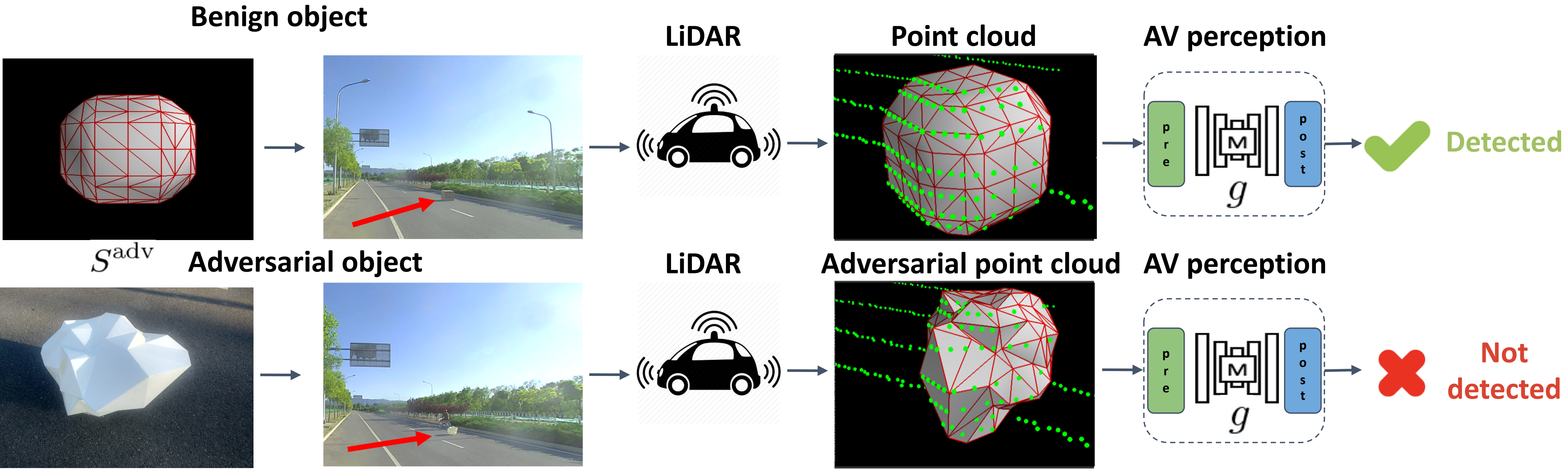
Chaowei Xiao*, Yulong Cao*, Dawei Yang*, Jin Fang, Ruigang Yang, Mingyan Liu, Bo Li. Adversarial objects against lidar-based autonomous driving systems. CVPR-AMLCV' 19

Generating Adversarial Objects

Target goal

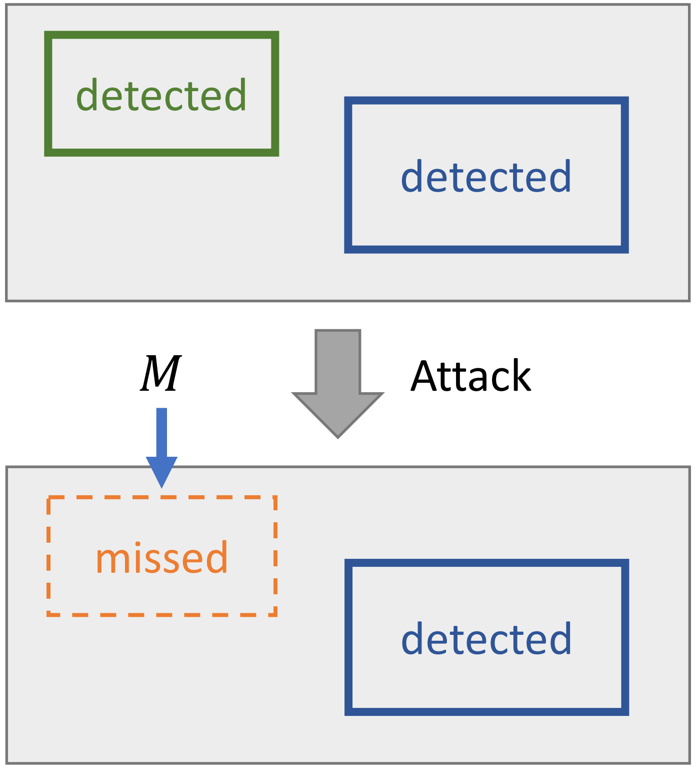
Not detected

$$S^{\mathrm{adv}} = \mathrm{argmin}_{S}\, L_{\mathrm{adv}}(S; g, t') + \tau \cdot L_{\mathrm{perceptual}}\,(S)$$



Chaowei Xiao*, Yulong Cao*, Dawei Yang*, Jin Fang, Ruigang Yang, Mingyan Liu, Bo Li. Adversarial objects against lidar-based autonomous driving systems. CVPR-AMLCV 2019

Adversarial Loss



Metric	Description
Center offset (off)	Offset to predicted center of the cluster the cell belongs to.
Objectness (obj)	The probability of a cell belonging to an obstacle.
Positiveness (pos)	The confidence score of the detection.
Object height (hei)	The predicted object height.
i**th Class Probability** (cls_i)	The probability of the cell being from class i (vehicle, pedestrian, etc.).

$$L_{adv} = H(Positiveness;\ g, S) * M$$

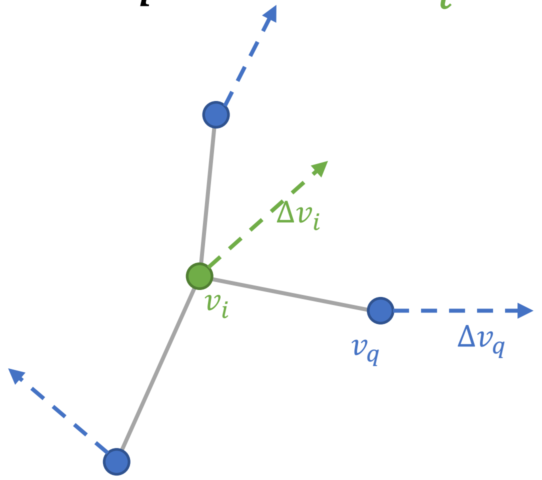
extract the *Positiveness* metric

Mask

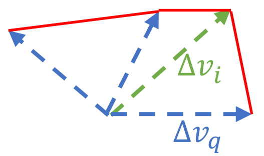
Generate Printable Shape

- 3D distance loss, operated on vertex displacements

$$L_{Perceuptual} = \Sigma_{v_i \in V} \Sigma_{q \in N(v_i)} \sqrt{\left| \Delta v_i - \Delta v_q \right|_2^2}$$



Perturbation of neighboring vertices



Perturbation differences

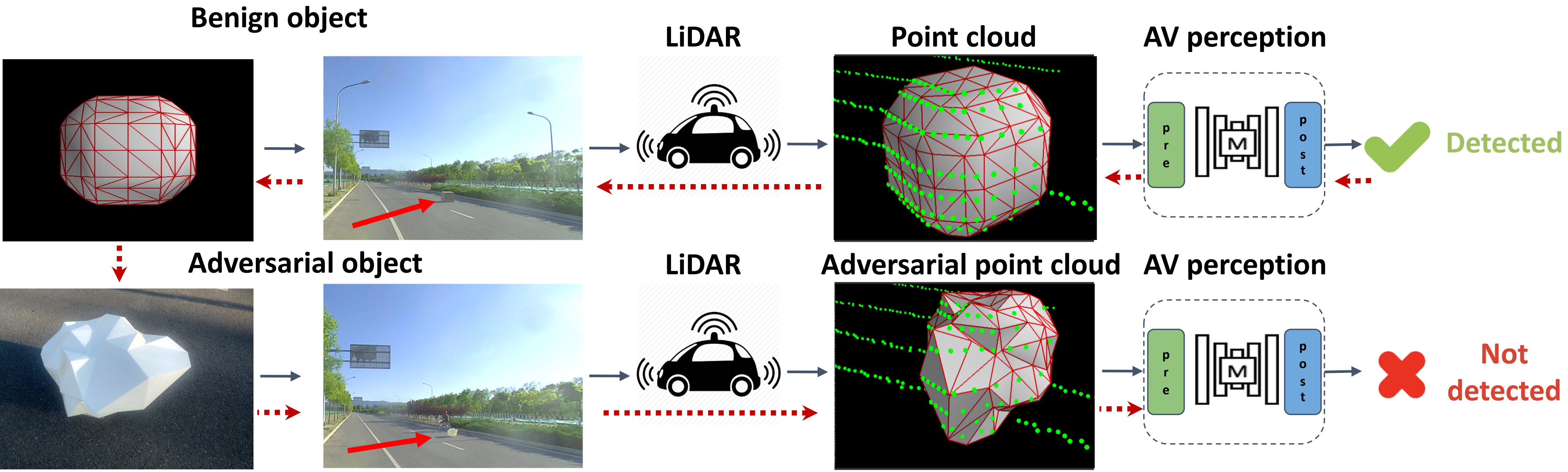
Not detected

Target goal

Pipeline of *LiDAR-adv*

- Input: a 3D mesh + shape perturbations
- **Target: fool a machine learning model and keep the shape printable**

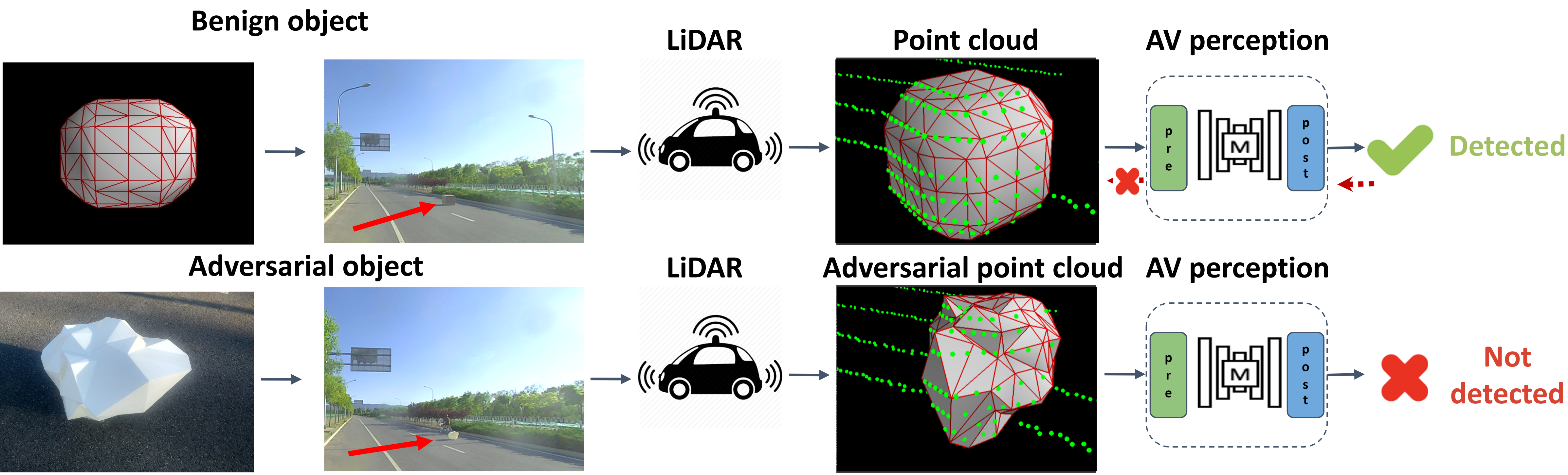
$$S^{adv} = \mathrm{argmin}_S \, L_{adv}(S; g, t') + \tau \cdot L_{perceptual}(S)$$



Pipeline of *LiDAR-adv*

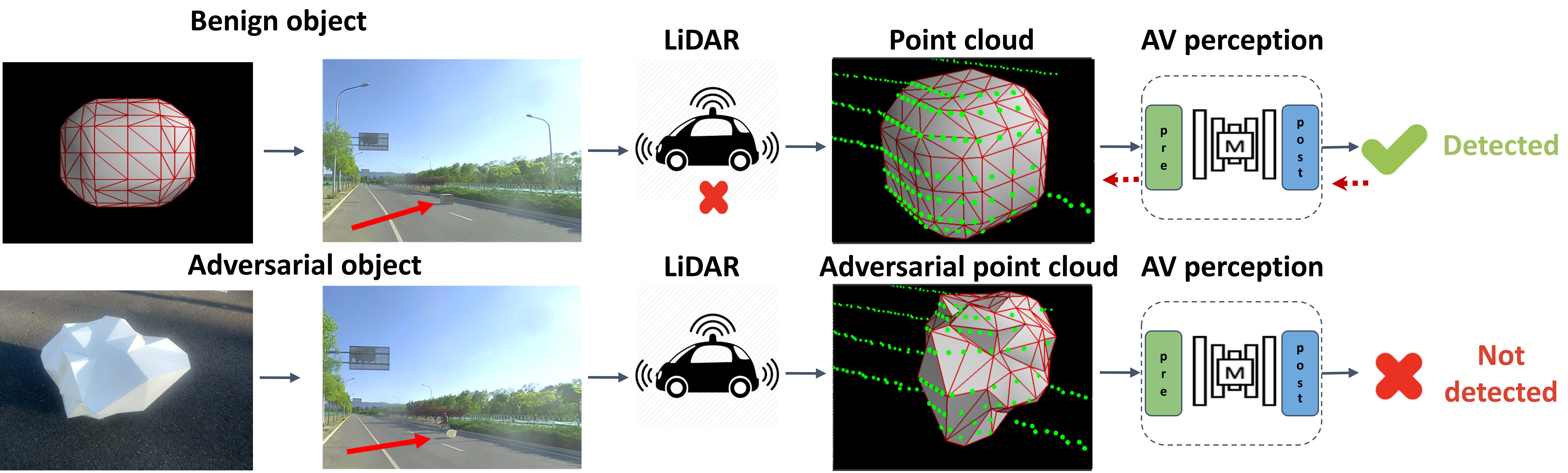
- Input: a 3D mesh + shape perturbations
- **Non-differentiable Pre/Post Processing**

$$S^{adv} = \mathrm{argmin}_S\, L_{adv}(S; g, t') + \tau \cdot L_{perceptual}\, (S)$$



Pipeline of *LiDAR-adv*

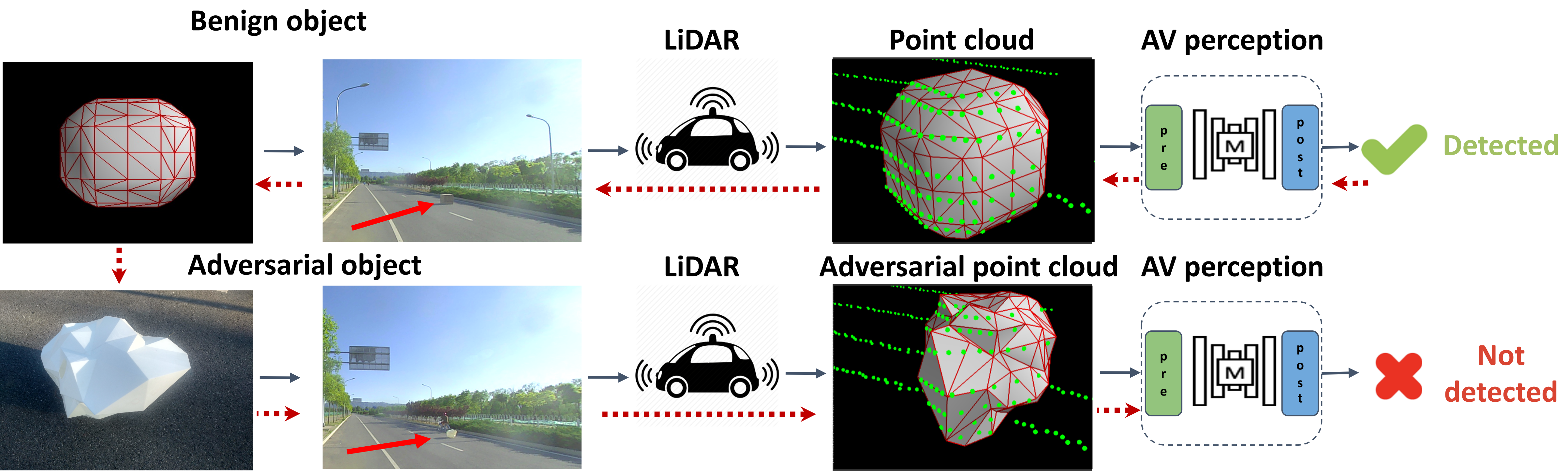
- Input: a 3D mesh + shape perturbations
- Non-differentiable pre/post processing: differentiable proxy function
- **Lidar** $S^{adv} = \mathrm{argmin}_S \, L_{adv}(S; g, t') + \tau \cdot L_{perceptual}(S)$



Pipeline of *LiDAR-adv*

- Input: a 3D mesh + shape perturbations
- LiDAR: a differentiable renderer
- Non-differentiable Pre/Post Processing: Differentiable proxy function
- Target: fool a machine learning model and keep the shape printable

$$S^{adv} = \text{argmin}_S \, L_{adv}(S; g, t') + \tau \cdot L_{perceptual}(S)$$



Physical Experiments

Adversarial object

Scene

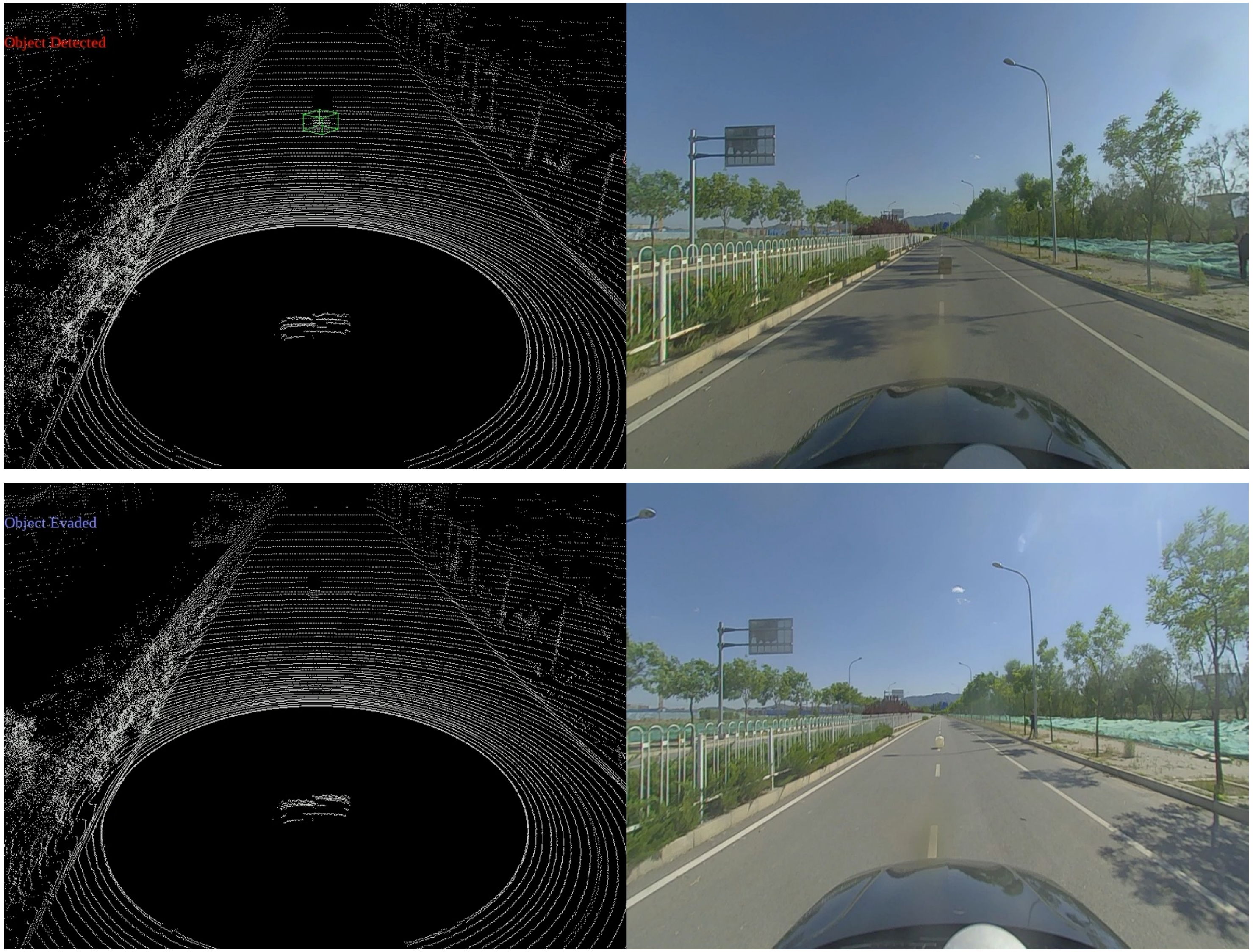
Autonomous vehicle

Physical Experiments

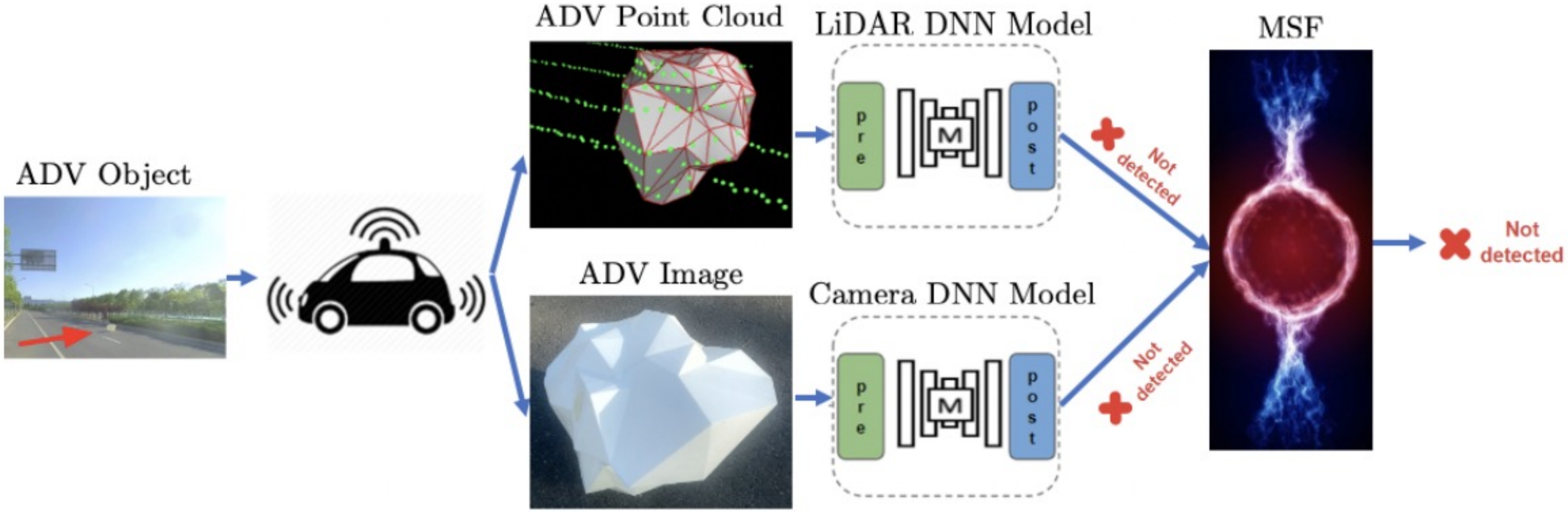
Adversarial object/benign box **in the middle lane**

Benign Object

Adversarial Object



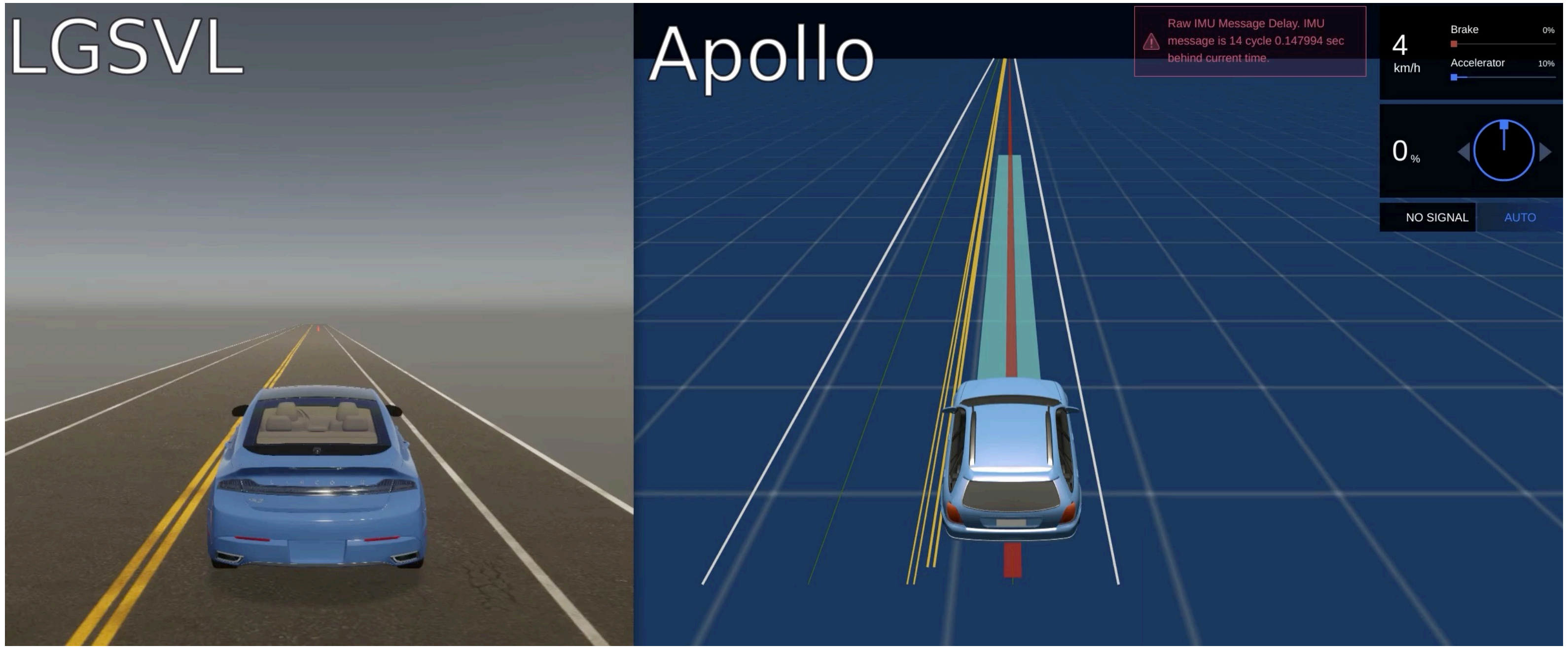
Sensor Fusion



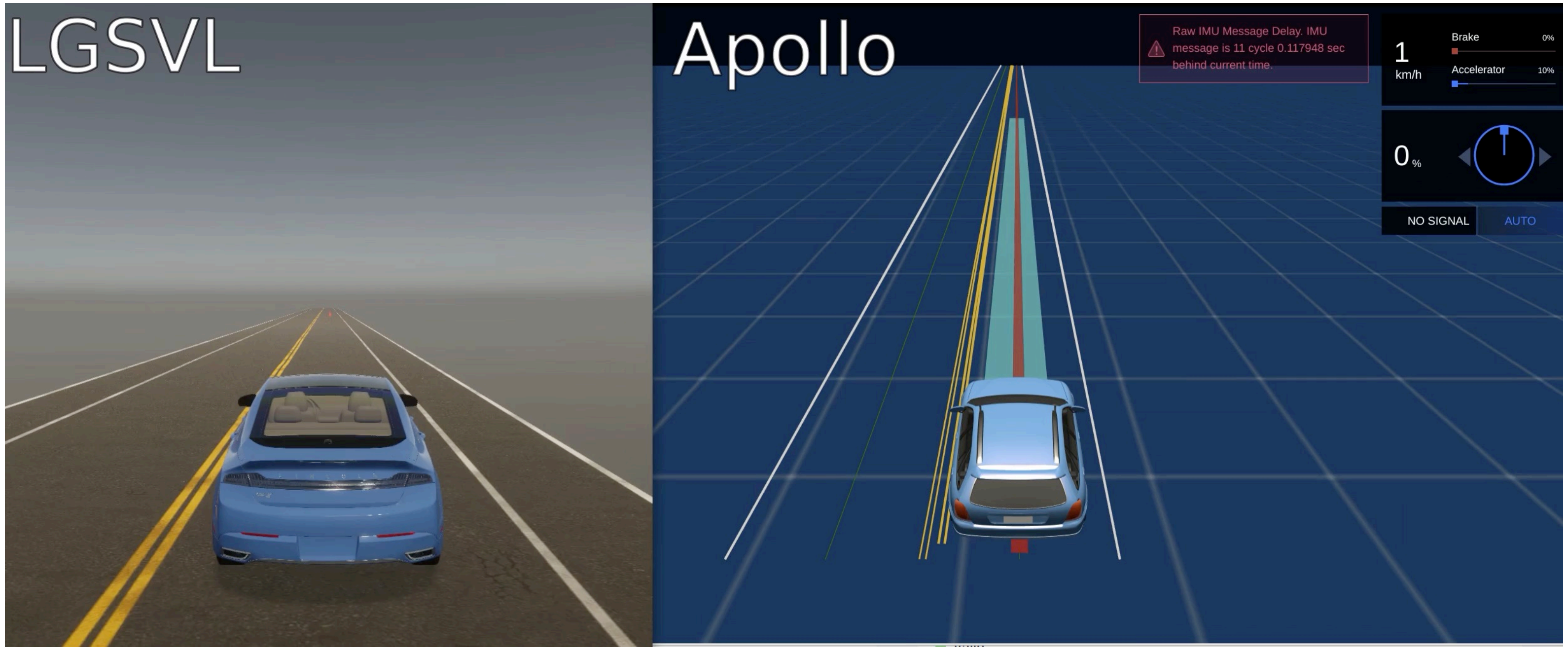
Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, et al. . Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks. S&P, 21

Adversarial object/benign box **in the middle lane**

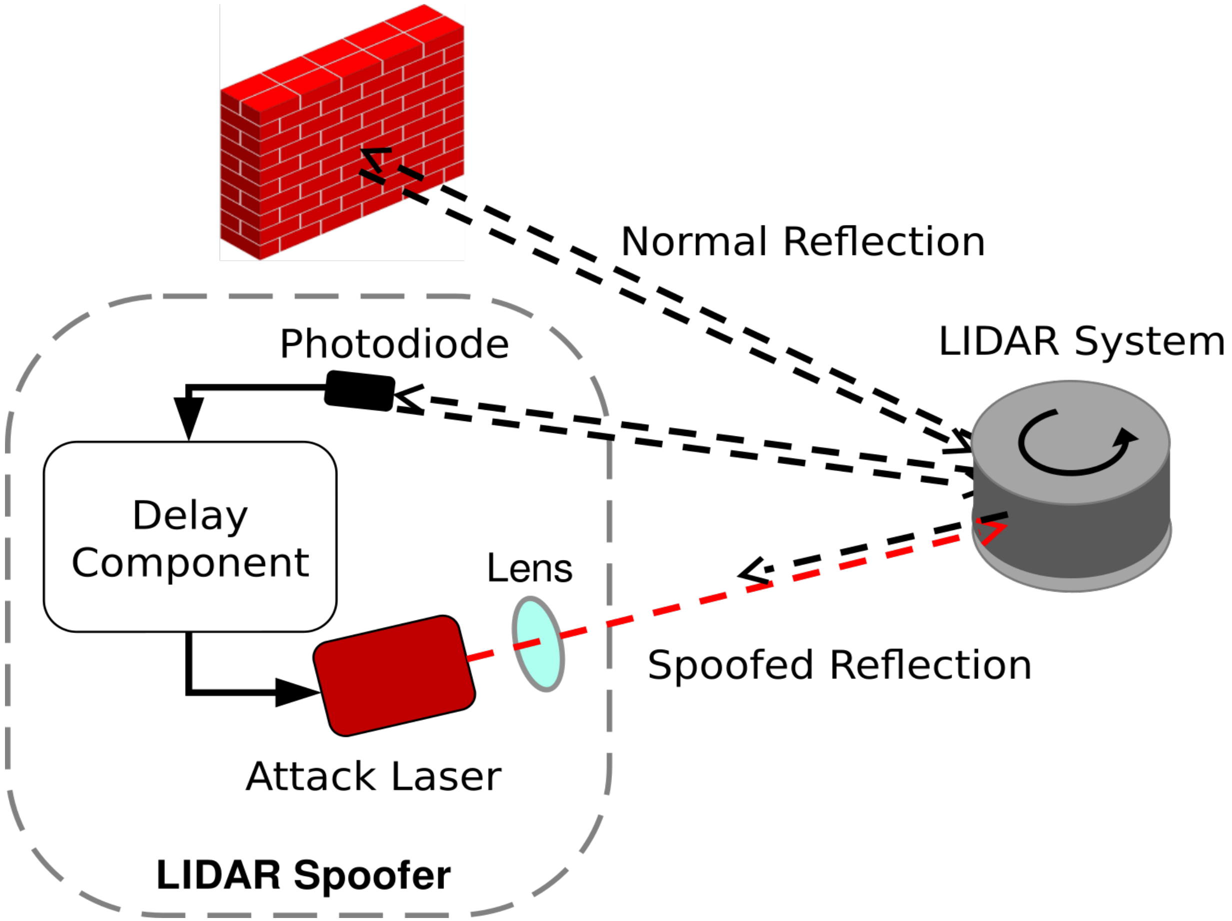
Benign Cone

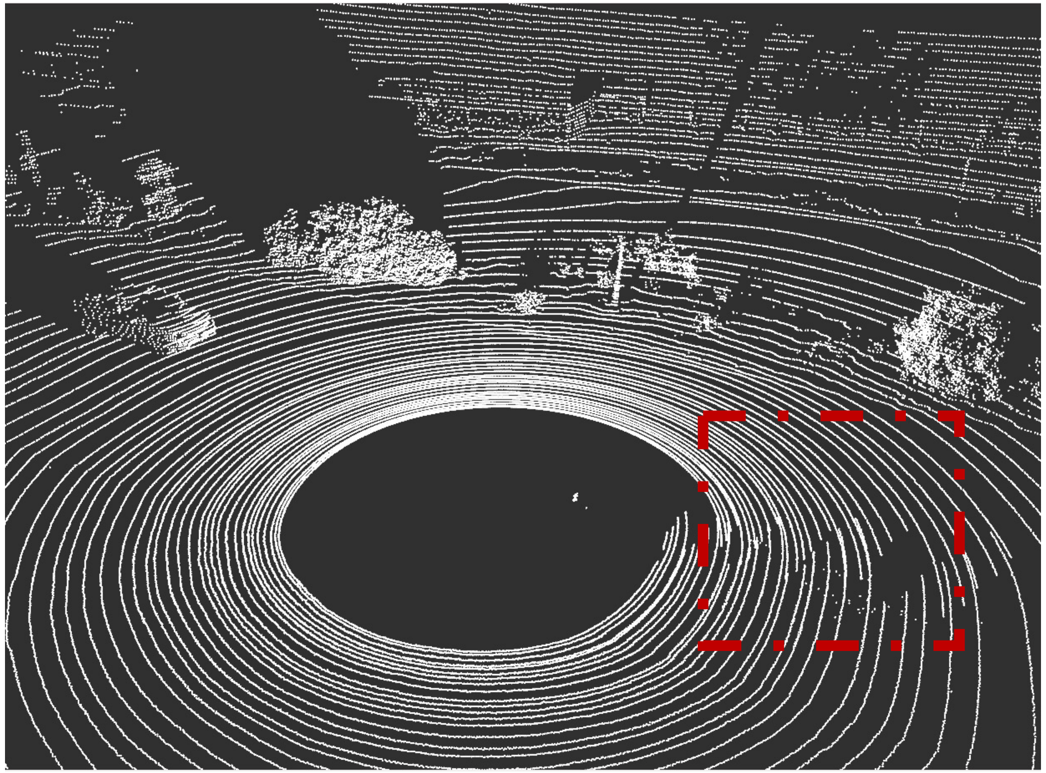


Adversarial Cone



LiDAR Spoofing Attack





Yulong Cao, **Chaowei Xiao**, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, Z. Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. CCS' 19

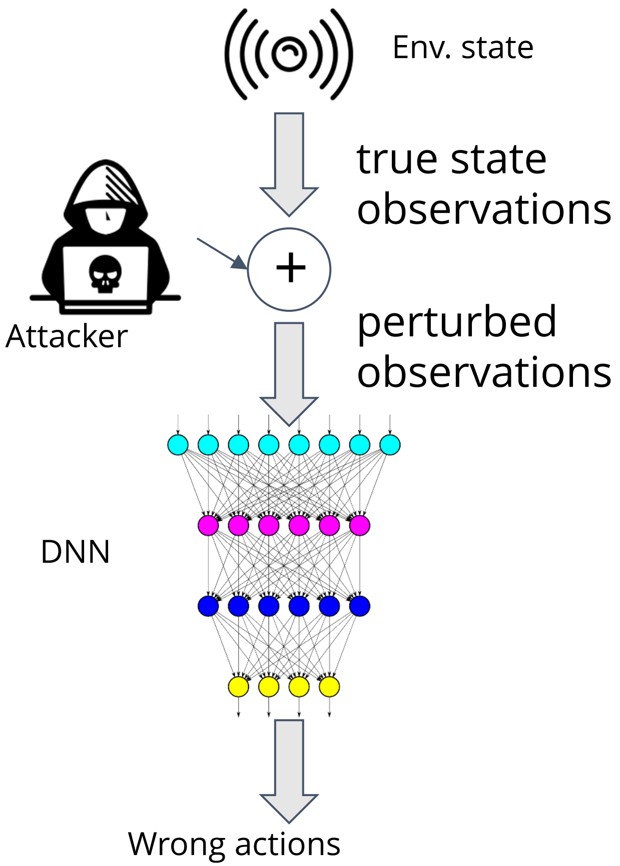
LiDAR Spoofing Attacks

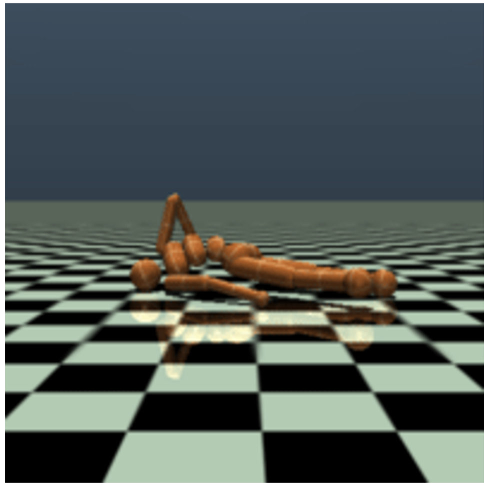
Yulong Cao, **Chaowei Xiao**, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, Z. Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. CCS 2019

Deep reinforcement learning can be vulnerable

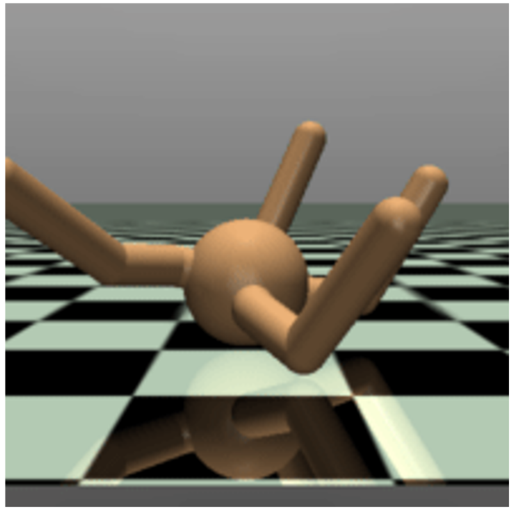
Successful attacks by adding small perturbations to state observations

(Huang et al., Kos & Song et al., Lin et al., Behzadan & Munir, Pattanaik et al., Xiao et al. ...)

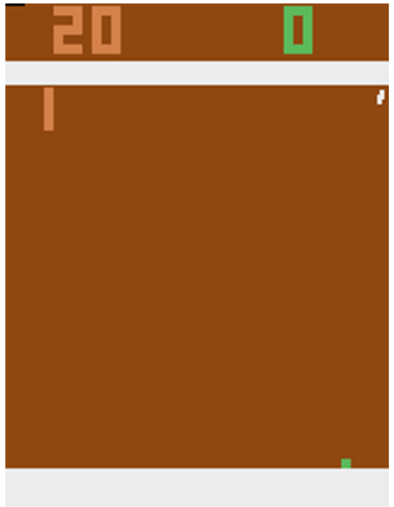




PPO **Humanoid**
Robust Sarsa Attack
Reward: 719
(original 4386)



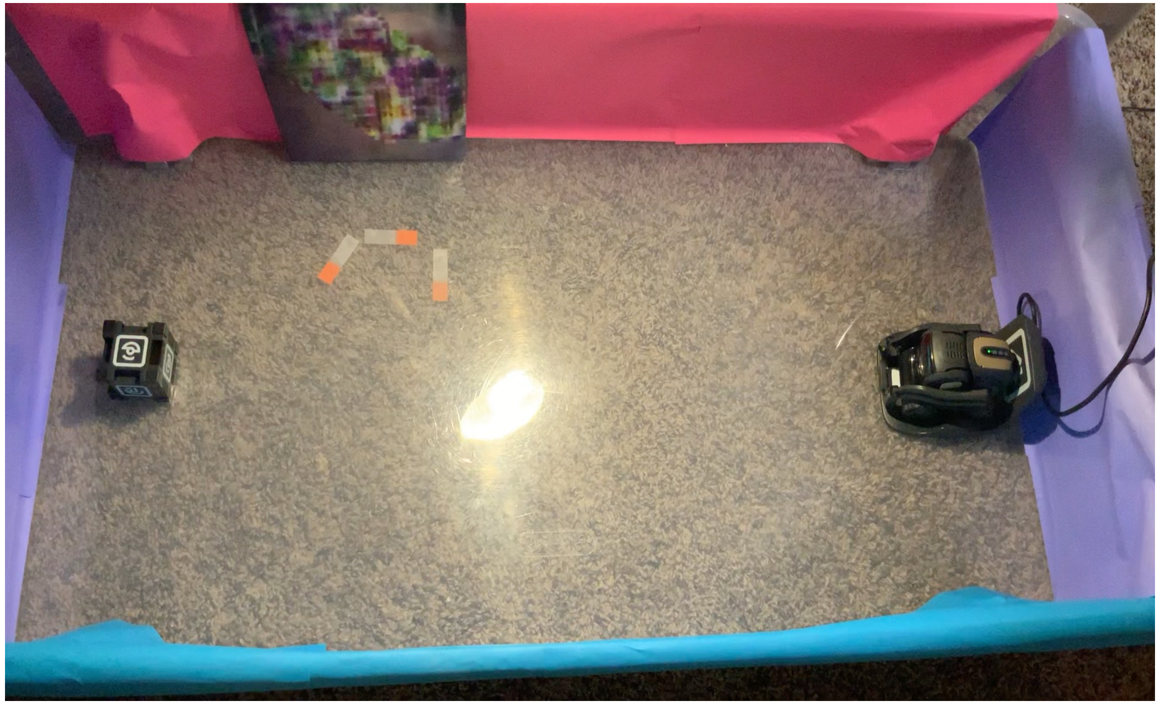
DDPG **Ant**
Robust Sarsa Attack
Reward: 258
(original 2462)



DQN **Pong**
PGD attack
Reward: -21
(lowest)

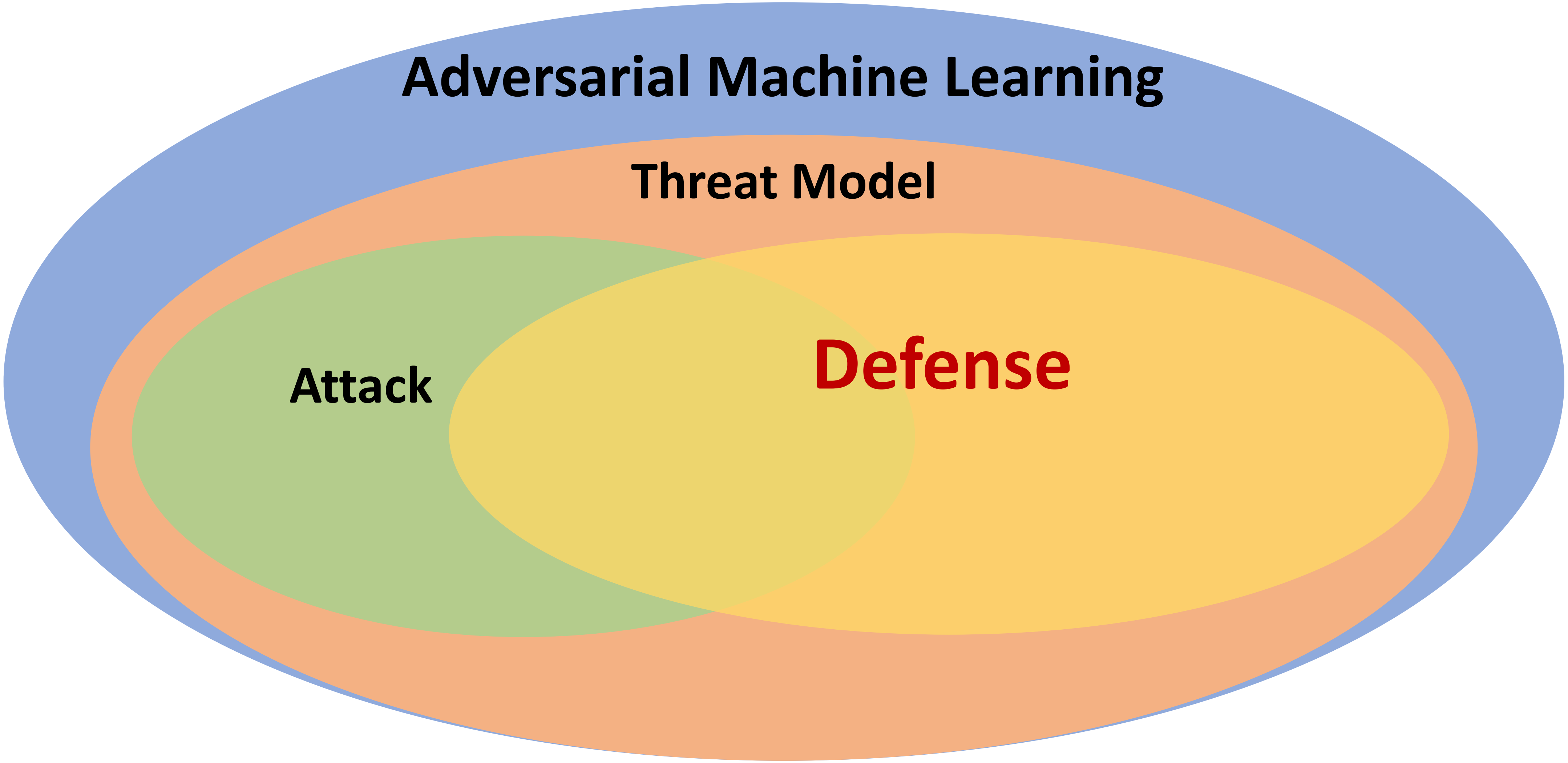
Huan Zhang, Hongge Chen, Chaowei Xiao et.al. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. NeurIPS 2020

Deep reinforcement learning can be vulnerable



Reinforcement Learning

Xinlei Pan, Chaowei Xiao, Warren He, Bo Li, Jian Peng, Mingjie Sun, Jinfeng Yi, Mingyan Liu, Dawn Song. Characterizing attacks on deep reinforcement learning. ICML SPML' 19

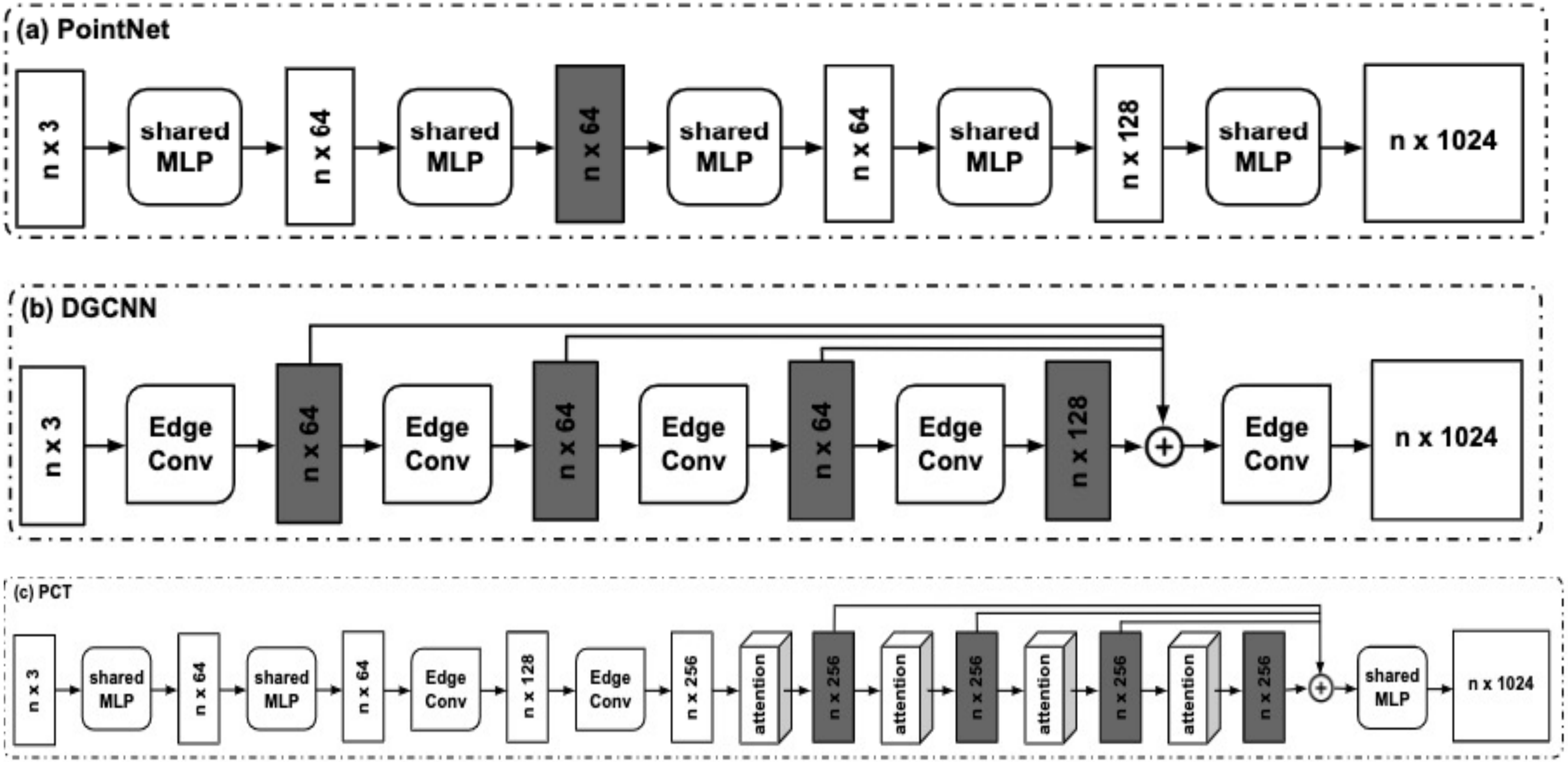

Adversarial Machine Learning
Threat Model
Attack
Defense

Defending against Adversarial Examples is Hard

- A Brief History of defense[1]
  - Oakland’ 16- broken
  - ICLR’ 17- broken
  - CCS’ 17- broken
  - ICLR’ 18 - broken (mostly)
  - CVPR’ 18 – broken
  - NeurIPS’ 18 –broken (some)
- Dup-net (broken), gather-vector guidance (broken).
- Error spaces containing adversarial are large[2]

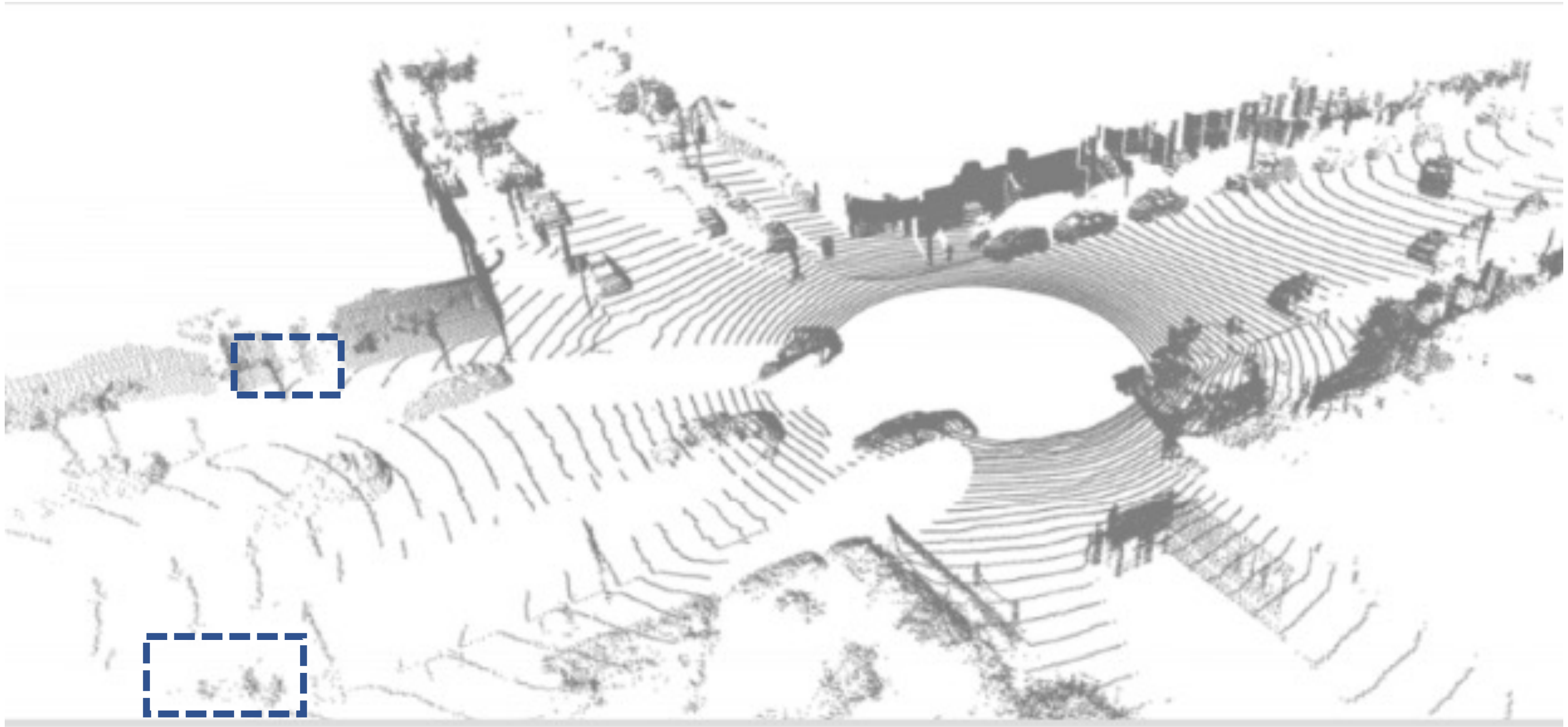
[1]Nicholas Carlini: Making and Measuring Progress in Adversarial Machine Learning
[2]Ian Goodfellow and Nicolas Papernot. Is attacking machine learning easier than defending it ? Blog

Defense in 3D domain

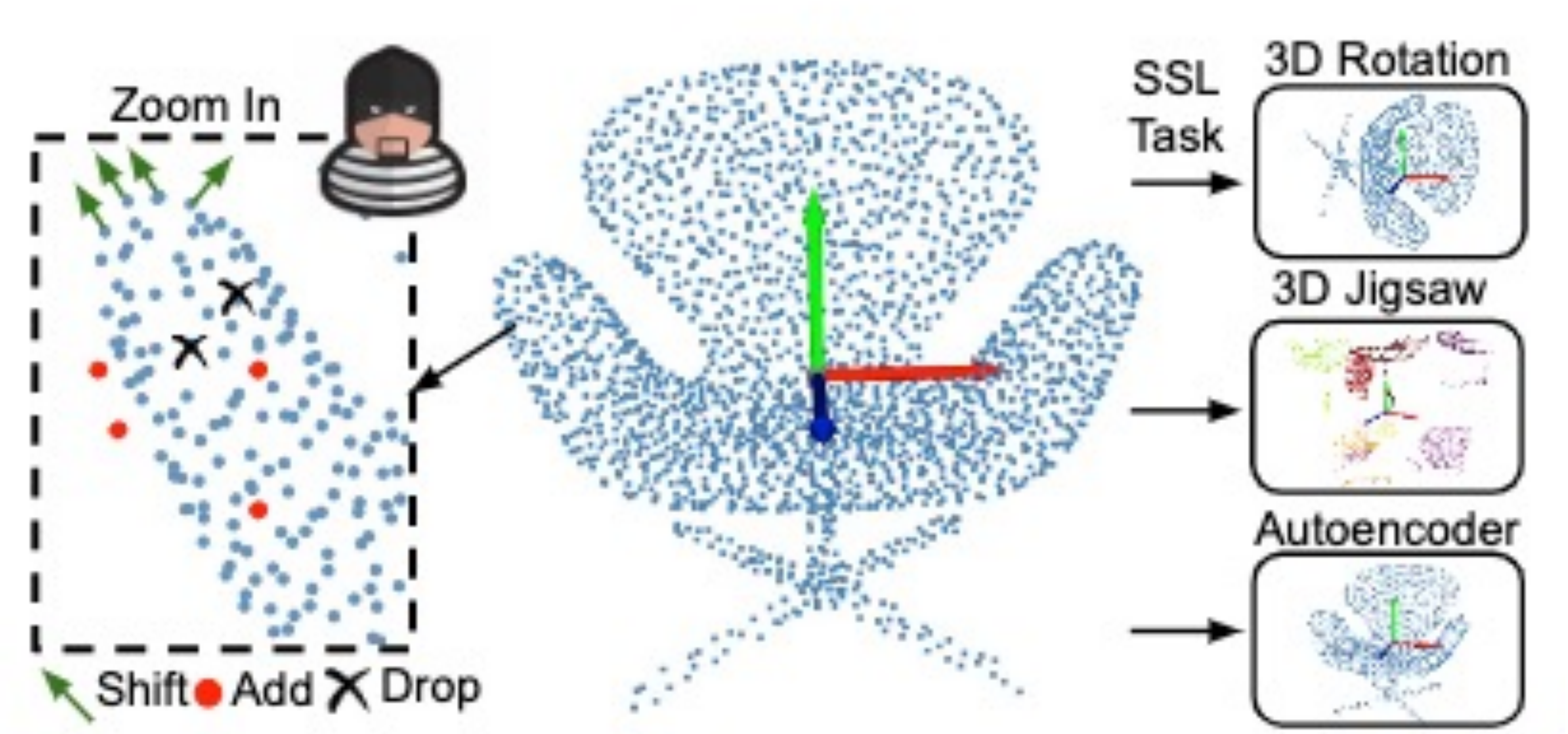


Defense in 3D domain

- Annotation is expensive

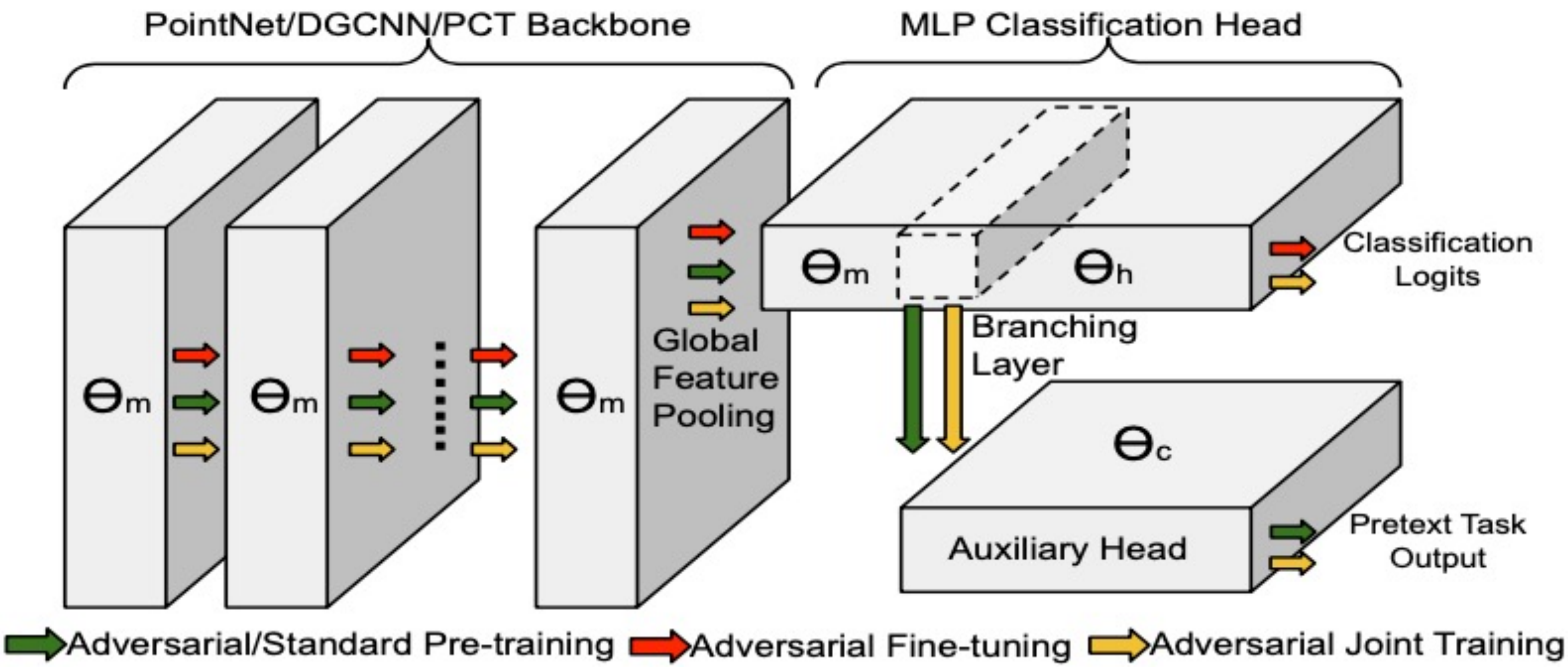


Adversarially Robust 3D Point Cloud Recognition Using Self-Supervisions



Jiachen Sun, Yulong Cao, Christopher Choy, Zhiding Yu, Chaowei Xiao, Anima Anandkumar, Zhuoqing Mao, Adversarially Robust 3D Point Cloud Recognition Using Self-Supervisions

Adversarially Robust 3D Point Cloud Recognition Using Self-Supervisions



Adversarial Pre-training for Fine-tuning

		ModelNet40						ScanObjectNN						ModelNet10					
		PointNet		DGCNN		PCT		PointNet		DGCNN		PCT		PointNet		DGCNN		PCT	
Pretext Task	Parameters	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA
AT Baseline	N/A	87.7	37.9	90.6	62.0	89.7	49.1	69.9	23.7	74.4	30.9	72.4	20.5	96.6	79.7	98.1	86.3	97.4	80.0
3D Rotation	$\eta = 6$	87.2	48.0	91.4	63.6	90.2	50.7	69.1	24.5	75.7	32.9	72.6	20.6	96.8	79.0	97.7	84.9	97.2	80.4
	$\eta = 18$	87.2	48.3	91.1	64.1	90.2	49.5	69.5	25.0	73.8	32.2	72.5	20.1	97.1	79.3	98.5	85.3	97.8	80.3
Adversarial 3D Rotation	$\eta = 6$	87.6	42.1	90.8	61.8	90.4	50.8	69.6	25.3	75.0	36.8	71.6	28.7	97.0	79.9	97.7	87.5	98.0	82.2
	$\eta = 18$	87.4	45.7	90.9	62.9	90.4	50.1	69.3	24.5	75.0	36.3	73.1	26.9	97.0	79.7	98.0	88.2	97.4	83.7
3D Jigsaw	$k = 3$	87.6	50.1	90.0	67.4	90.4	51.1	70.8	25.5	79.0	33.8	73.4	23.2	96.8	80.0	98.0	89.6	97.8	81.5
	$k = 4$	87.6	50.9	90.1	65.3	90.3	50.2	70.2	25.4	76.2	35.3	73.8	24.6	96.7	80.2	98.0	89.0	97.7	81.9
Adversarial 3D Jigsaw	$k = 3$	88.2	52.1	89.6	65.8	89.8	51.3	69.0	24.8	77.5	41.3	72.5	26.3	97.0	80.6	98.5	90.5	97.4	83.5
	$k = 4$	87.8	50.5	89.9	65.3	89.6	51.0	69.9	25.5	76.1	40.6	73.1	27.4	97.0	80.5	98.0	89.1	97.3	83.9
Autoencoder	sphere	87.4	50.0	89.9	62.8	90.2	50.7	69.9	25.1	76.1	36.0	71.3	24.1	97.0	80.5	98.2	86.8	97.1	80.1
	plane	87.1	48.8	90.1	62.2	90.2	50.2	69.4	25.5	76.2	35.6	71.1	22.6	96.8	80.8	97.8	87.6	97.0	80.1
	gaussian	87.4	48.9	90.8	63.3	89.7	50.3	69.7	23.8	75.6	35.8	71.3	24.8	96.8	80.5	97.8	86.4	97.1	80.1
Adversarial Autoencoder	sphere	87.1	49.7	90.0	62.2	90.3	50.0	70.4	25.2	75.2	36.2	72.6	22.2	96.7	80.4	97.5	87.3	97.5	82.1
	plane	86.9	46.6	89.7	61.8	89.7	50.0	69.2	24.0	75.6	38.0	73.3	21.6	97.0	80.6	98.0	86.1	97.7	82.5
	gaussian	87.1	48.5	90.7	62.7	90.2	50.5	68.8	25.0	74.7	36.3	72.6	23.4	97.0	80.2	97.8	88.4	97.4	83.2

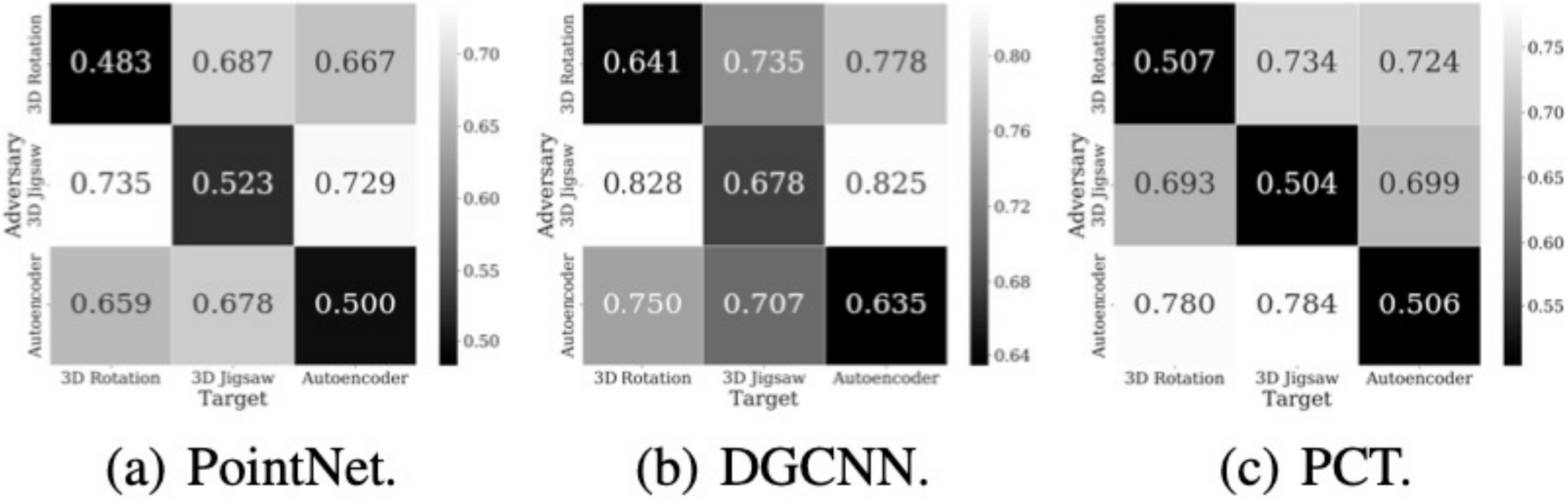
Table 2: Evaluation Results (%) of Adversarial Pre-training for Fine-tuning

Adversarial Joint Training.

		ModelNet40						ScanObjectNN						ModelNet10					
		PointNet		DGCNN		PCT		PointNet		DGCNN		PCT		PointNet		DGCNN		PCT	
Pretext Task	Parameters	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA	CA	RA
AT Baseline	N/A	87.7	37.9	90.6	62.0	89.7	49.1	69.9	23.7	74.4	30.9	72.4	20.5	96.6	79.7	98.1	86.3	97.4	80.0
3D Rotation	$\eta = 6$	86.8	45.0	91.2	60.7	89.5	44.3	67.8	24.3	74.2	37.8	72.3	20.3	96.6	79.0	98.1	86.3	97.8	73.8
	$\eta = 18$	86.5	46.4	91.3	62.0	88.9	42.9	68.7	25.1	76.2	37.2	72.1	19.8	97.0	79.9	97.9	85.7	98.1	75.6
3D Jigsaw	$k = 3$	87.6	42.5	91.0	62.3	90.2	43.1	69.4	25.5	77.1	38.9	72.1	20.7	96.8	79.8	98.4	87.9	97.7	76.8
	$k = 4$	87.2	46.7	91.1	61.7	89.8	40.9	70.0	24.6	75.9	38.4	73.7	20.8	96.8	77.9	98.0	88.6	97.1	78.0

Table 3: Evaluation Results (%) of Adversarial Joint Training.

Transferability Analysis



Robust Accuracy on Transfer Attacks among Fine-tuned Models
from Different SSL Tasks on ModelNet40.

Collaborators

Anima Anandkumar (Caltech&Nvidia)
Alfred Chen (UCI)
Hongge Chen (Waymo)
Yulong Cao (Umich)
Jia Deng (Princeton U)
Cho-Jui Hsieh (UCLA)
Warren He (UC Berkeley)
Jean Kossaifi (Nvidia)
Bo Li (UICU)
Mingyan Liu (Umich)
Morely Mao (Umich)
Xinlei Pan (UC Berkeley)
Haonan Qiu (CUHK)
Dawn Song (UC Berkeley)
Jiachen Sun (Umich)
Ningfei Wang (UCI)
Zhiding Yu (Nvidia)
Dawei Yang (Google)
Ruigang Yang (Inceptio)
Xinchen Yan (Uber)
Junyan Zhu (CMU)
Huan Zhang (CMU)

frontiers

2020 PROGRESS REPORT

The year of Open Science

- **5th most-cited** publisher
- **1 billion** article views and downloads
- **16,000** new editors on our boards
- **27** new journals
- **77** new institutional members

See Progress Report

frontiers

Research Topic

Trustworthy Machine Learning

Manage topic | Submit your abstract | Submit your manuscript | Participate

Overview | Articles | Authors | Impact

About this Research Topic

Recent studies have shown that machine learning (ML) models could be deliberately fooled, evaded, misled, and stolen. These studies result in profound security and privacy implications, especially when employing ML to critical applications such as autonomous driving, surveillance systems, and disease diagnosis. Additionally, recent studies have revealed potential societal biases in ML models, where the models learn inappropriate correlations between th final predictions and sensitive attributes such as gender and race. Without properly quantifying and reducing the reliance on such correlations, the broad adoption of ML models can have the inadvertent effect of magnifying stereotypes. To allow wide deployment of ML and enable pro-social outcomes, we desire trustworthy ML systems that are able to resist attacks from strong adversaries, protect user privacy, and produce fair decisions.

Thanks

- Q&A
- xiaocw@umich.edu