Secure Learning in Adversarial Environments

Bo Li

Assistant professor
University of Illinois at Urbana-Champaign

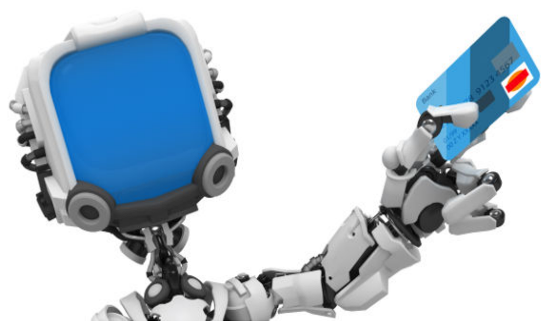
Machine Learning is Ubiquitous

Autonomous Driving

Healthcare

Smart City

Malware Classification



Fraud Detection

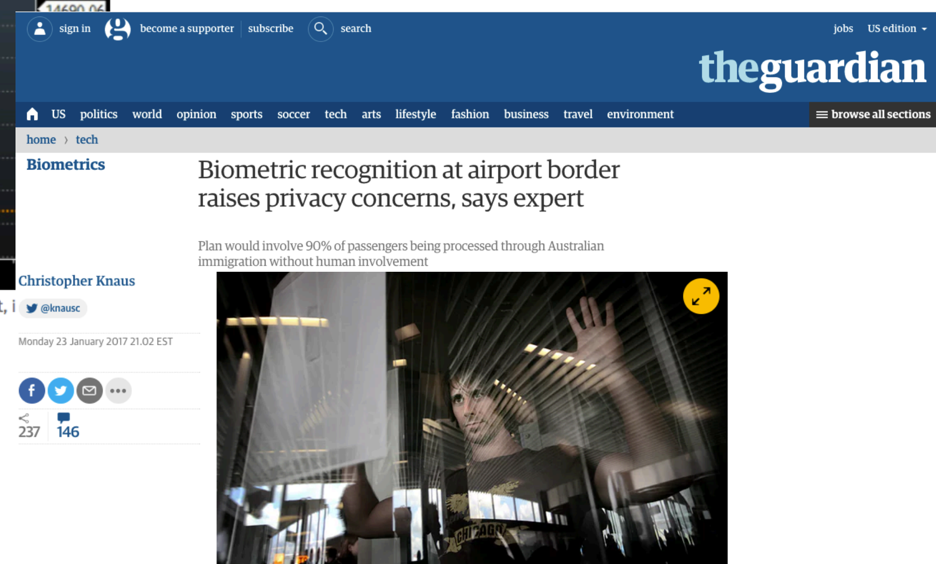
Biometrics Recognition

Security & Privacy Problems


Sections
The Washington Post
WorldViews
Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism?
By Max Fisher April 23, 2013

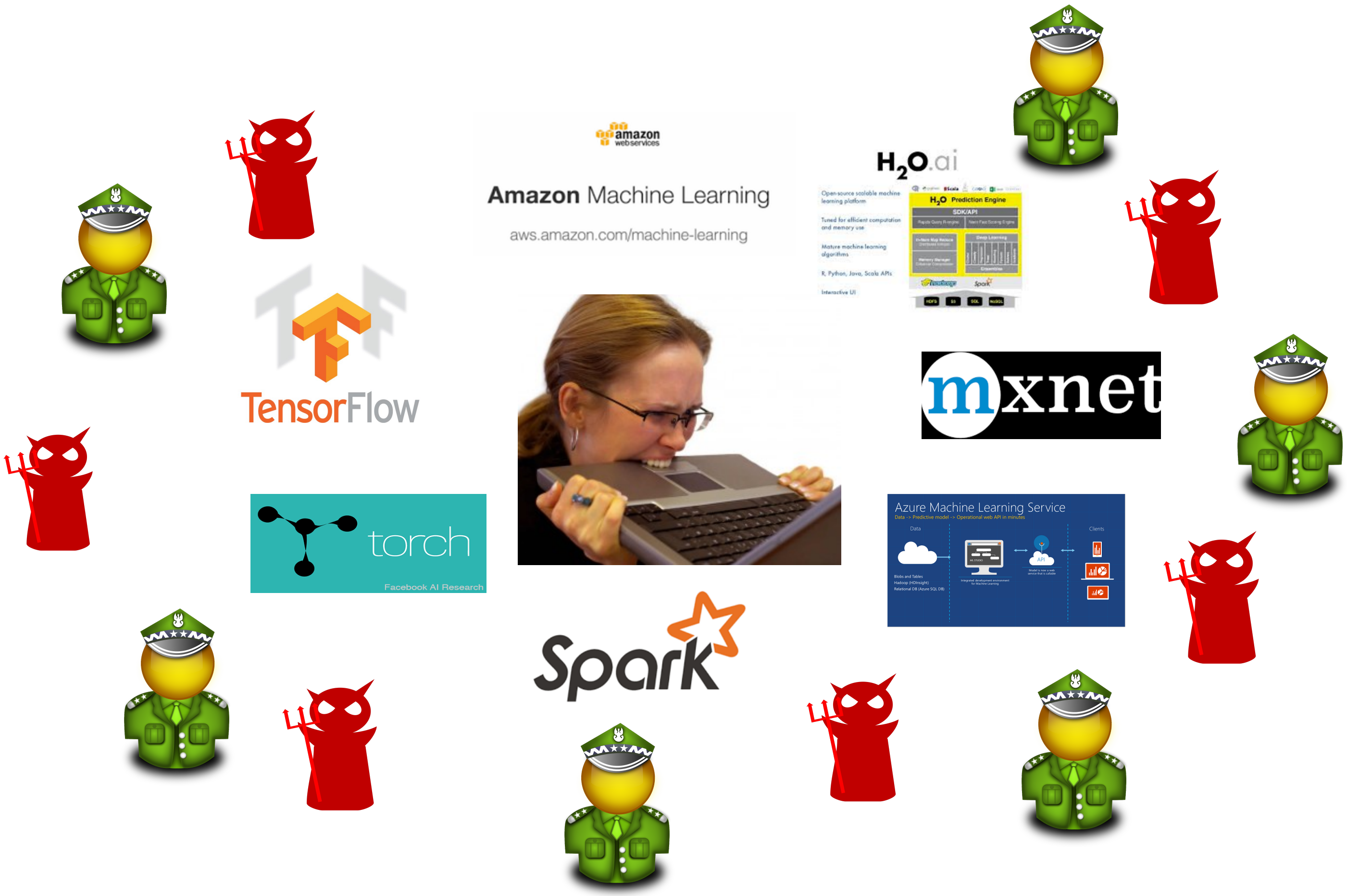
This chart shows the Dow Jones Industrial Average during Tuesday afternoon's drop, caused by a fake A.P. tweet, i

Trading Bot Crashes The Market


theguardian
home › tech
Biometrics
Biometric recognition at airport border raises privacy concerns, says expert
Plan would involve 90% of passengers being processed through Australian immigration without human involvement
Christopher Knaus
@knausc
Monday 23 January 2017 21.02 EST
237 146

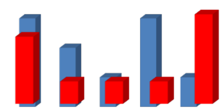
Privacy Concerns

We Live in an Adversarial Environment

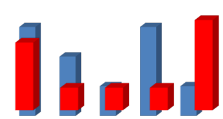


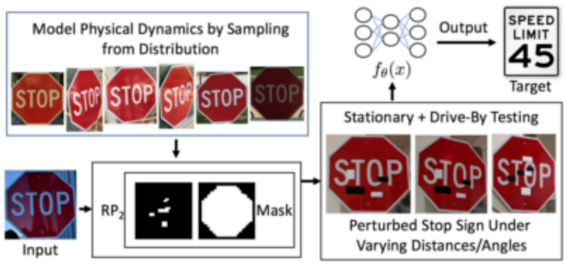
Perils of Stationary Assumption

Traditional machine learning approaches assume

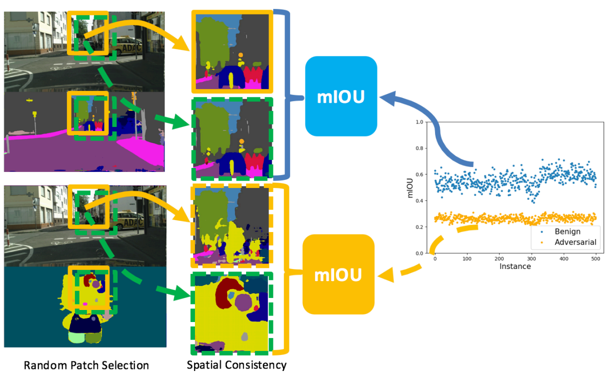
Training Data 

≈

Testing Data 



Robust physical world attacks against **different sensors**

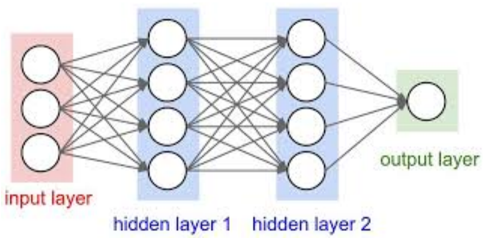


Potential **defenses** against adversarial behaviors based on intrinsic learning properties

Adversarial Perturbation In Digital World

$$\min_{\theta} J(\theta, x, y)$$

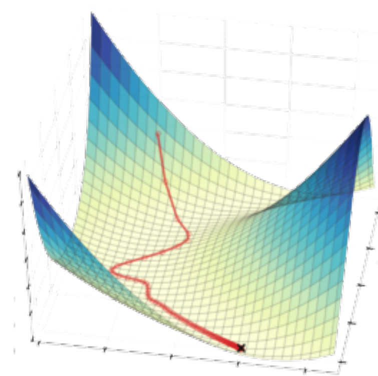
Model parameters Input feature vector label



Deep Neural Networks

$$\max_{\epsilon} J(\theta, x + \boxed{\epsilon}, y)$$

Adversarial perturbation

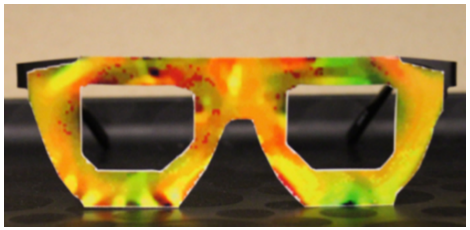


Gradient Descent

How to solve the adversary strategy

- Local search
- Combinatorial optimization
- Convex relaxation

Physical Attacks In Practice



Physical attack: Sharif et al., “Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition,” CCS 2016

However, What We Can See Everyday…

The Physical World Is… Messy

Varying Physical Conditions (Angle, Distance, Lighting, …)

Fabrication/Perception Error (Color Reproduction, etc.)

Digital Noise (What you want) | What is printed | What a camera may see

Physical Limits on Imperceptibility

Background Modifications*

Image Courtesy, OpenAI

An Optimization Approach To Creating Robust Physical Adversarial Examples

$$\underset{\delta}{\operatorname{argmin}} \ \lambda ||\delta||_p + J(f_\theta(x+\delta), y^*)$$

Perturbation/Noise Matrix

Lp norm (L-0, L-1, L-2, …)

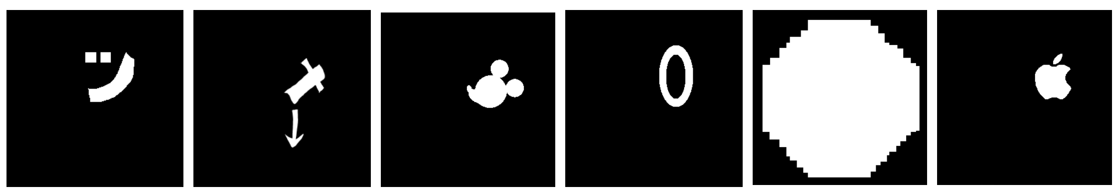
Loss Function

Adversarial Target Label

$$\underset{\delta}{\operatorname{argmin}} \ \lambda ||\delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i+\delta), y^*)$$

Optimizing Spatial Constraints (Handling Limits on Imperceptibility)

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i + M_x \cdot \delta), y^*)$$



Subtle Poster

Camouflage Sticker

Mimic vandalism

"Hide in the human psyche"

Subtle Poster

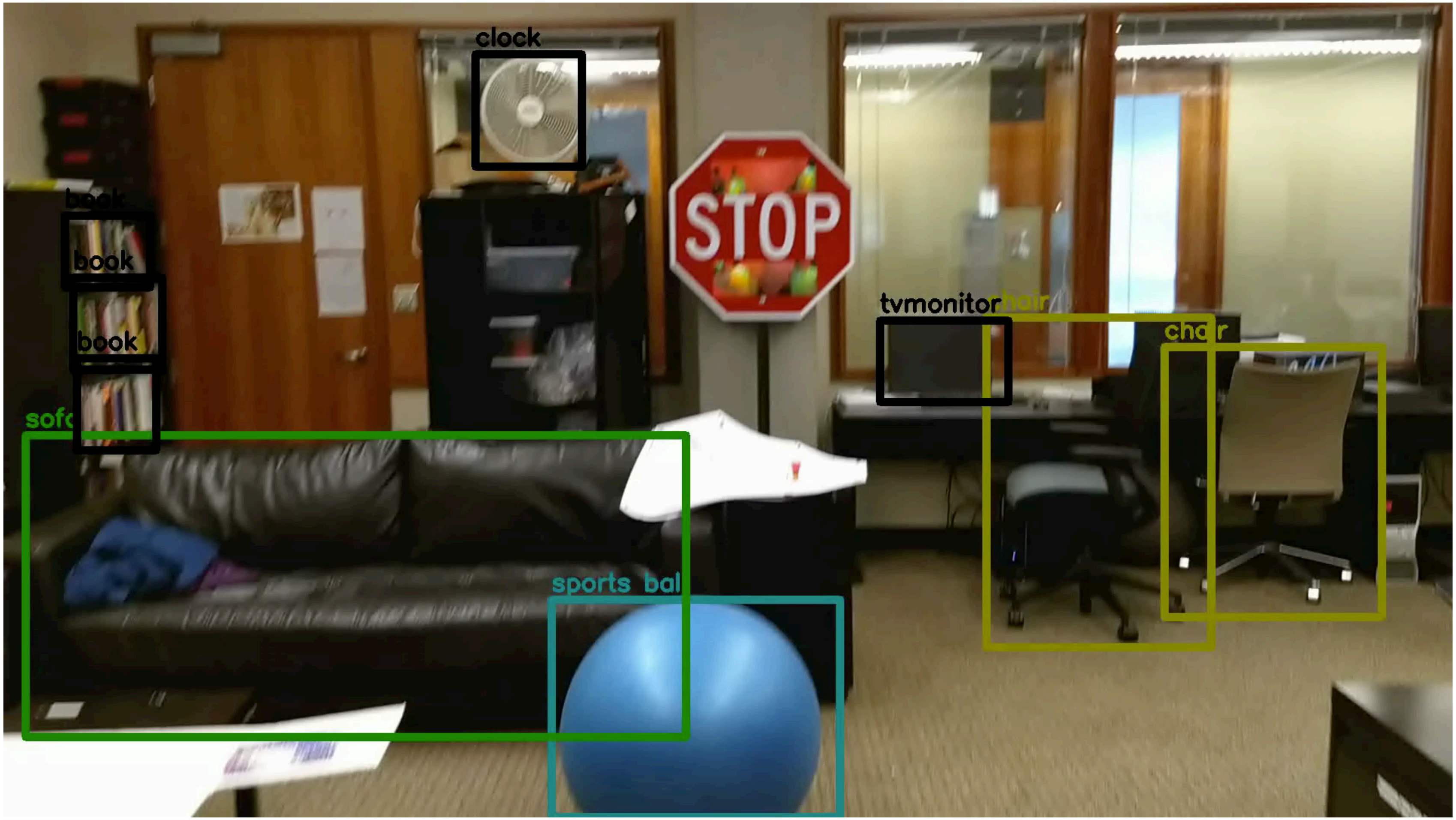
Lab Test Summary (Stationary)

Target Class: Speed Limit 45

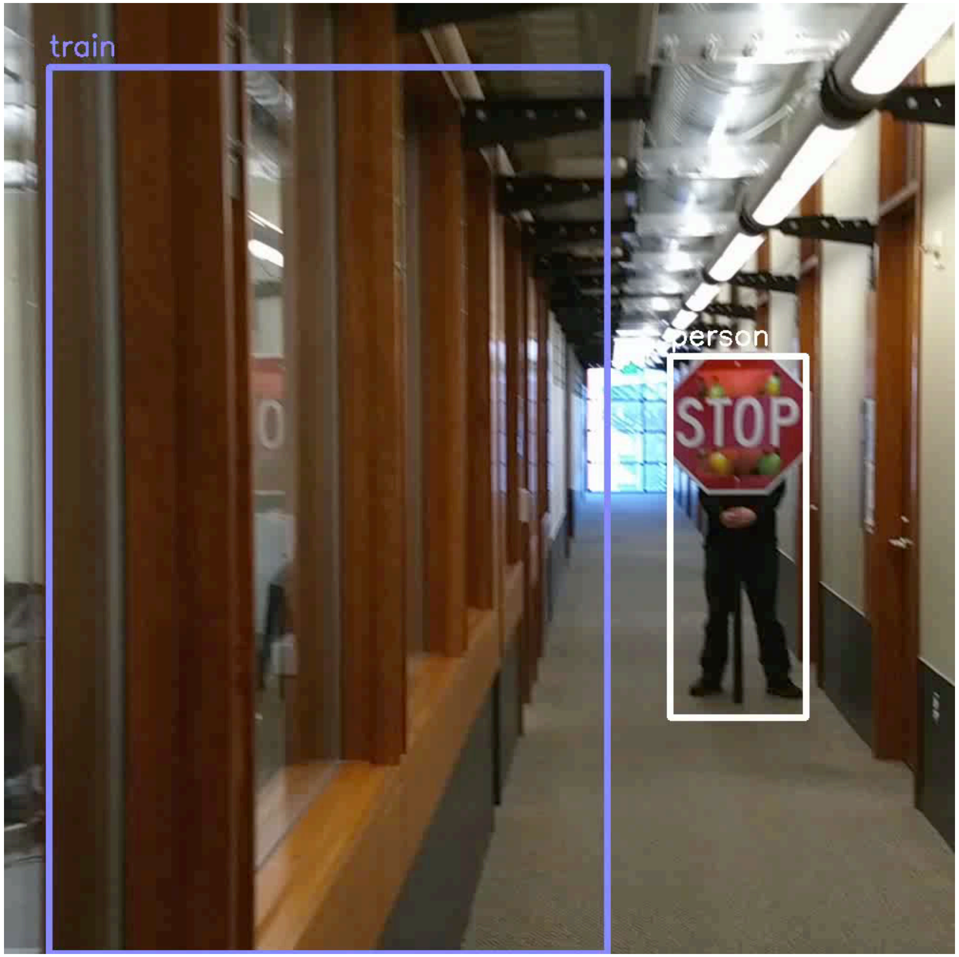
Art Perturbation

Subtle Perturbation

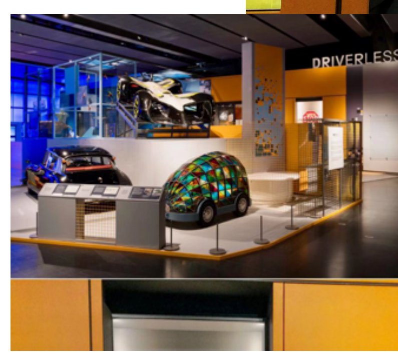
Physical Attacks Against Detectors



Physical Attacks Against Detectors



Physical Adversarial Stop Sign in the Science Museum of London



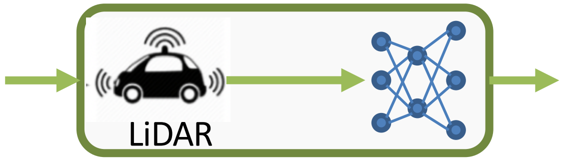
Physical Adversarial Attacks Against Sensor Fusion

Goal: we aim to generate physical **adversarial object** against **real-world LiDAR system.**

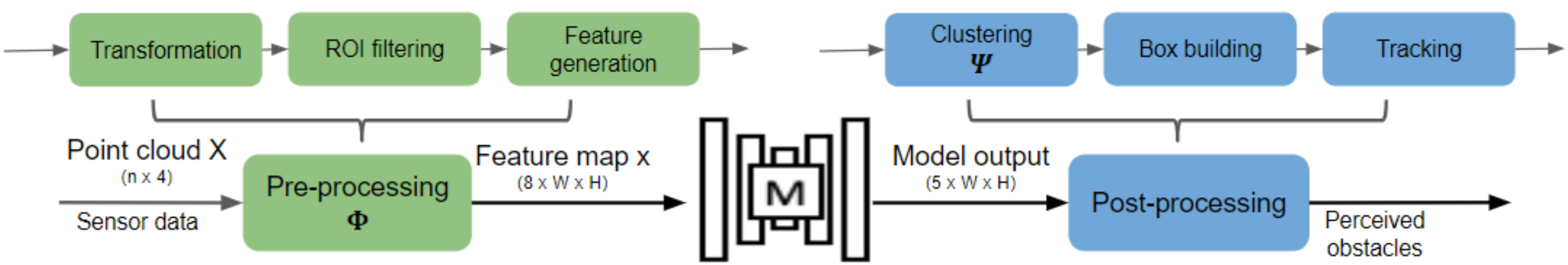
LiDAR-based perception

Challenges

- Physical LiDAR equipment
- Multiple non-differentiable pre/post-processing stages

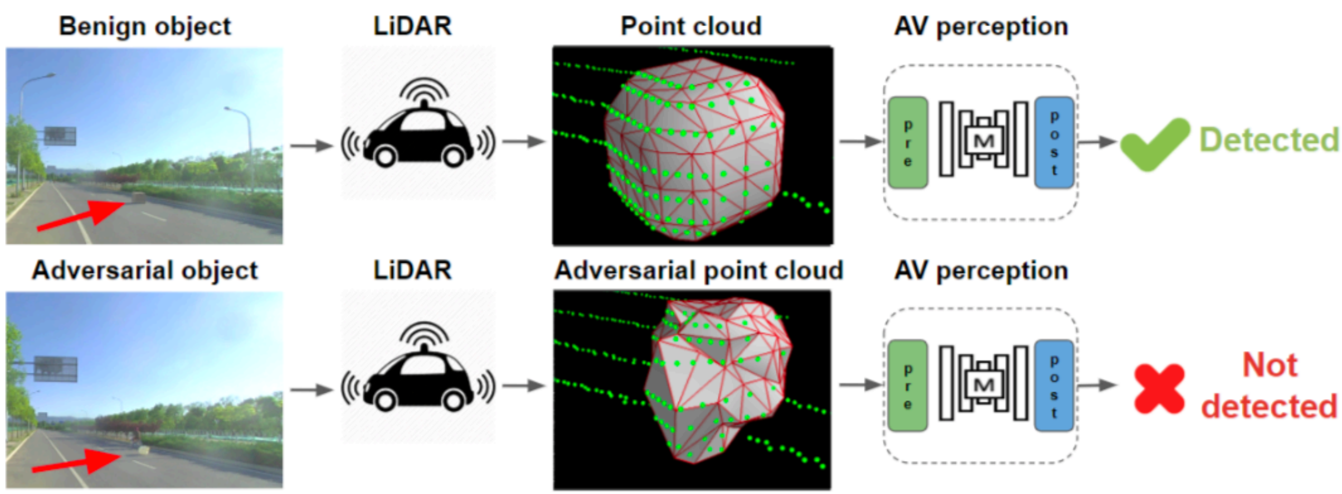


- Manipulation constraints
  - Limited by LiDAR
  - Keeping the shape plausible and smooth adds additional constraints
- Limited Manipulation Space
  - Consider the practical size of the object versus the size of the scene that is processed by LiDAR, the 3D manipulation space is rather small (< 2% in our experiments)



Pipeline of *LiDAR-adv*

- Input: a 3D mesh + shape perturbations
- Non-differentiable Pre/Post Processing
- Target: fool a machine learning model to ignore the object and keep the shape printable

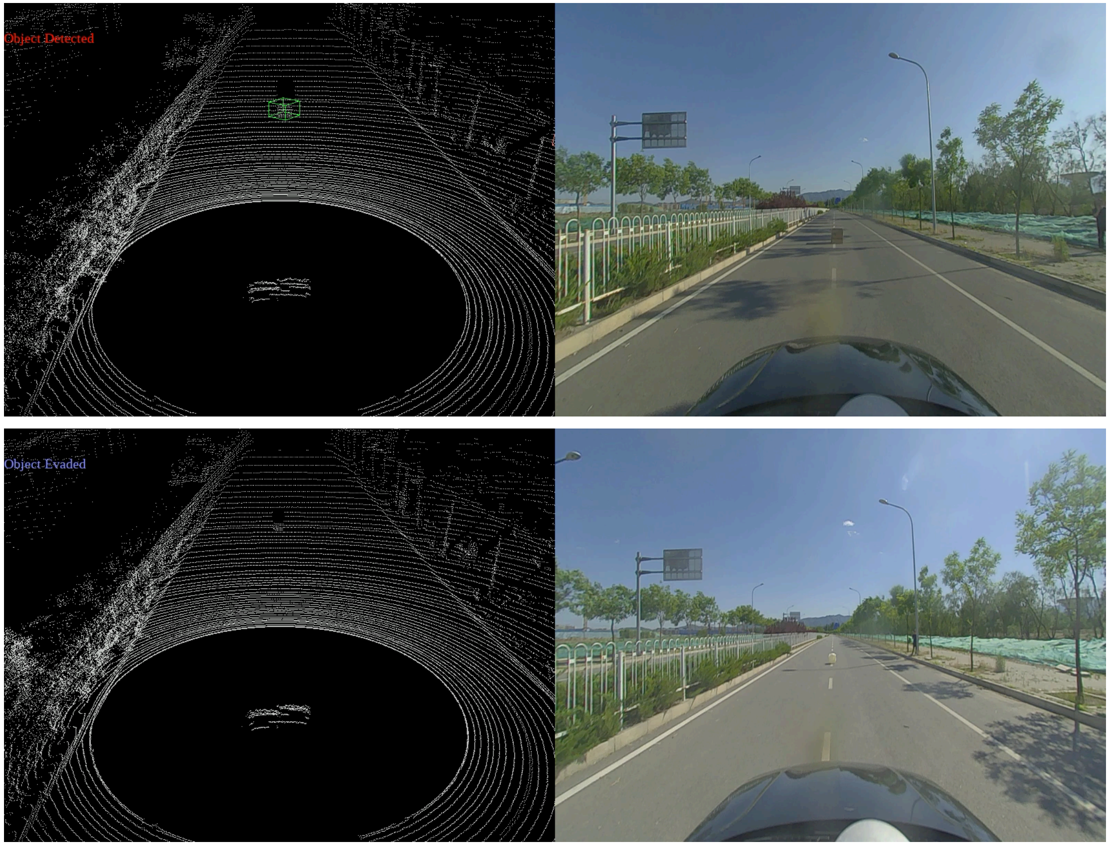


Physical Experiments

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

Benign Object

Adversarial Object

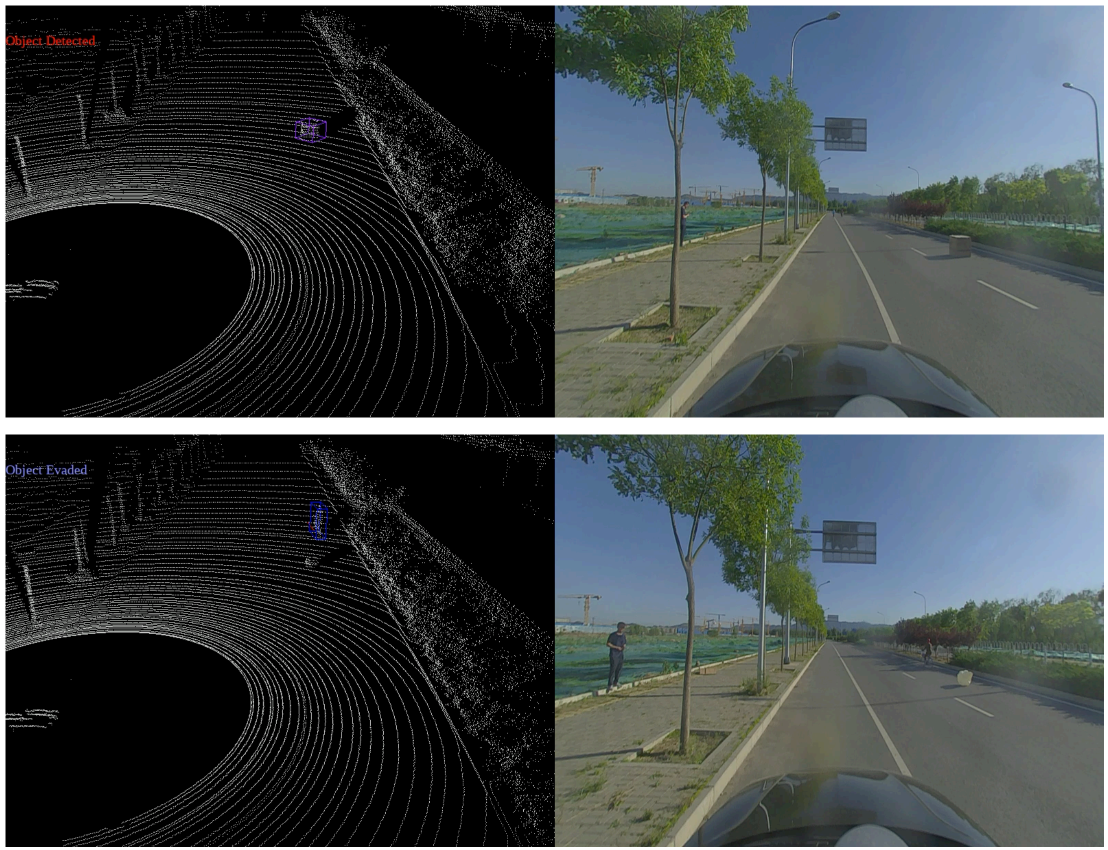


Physical Experiments

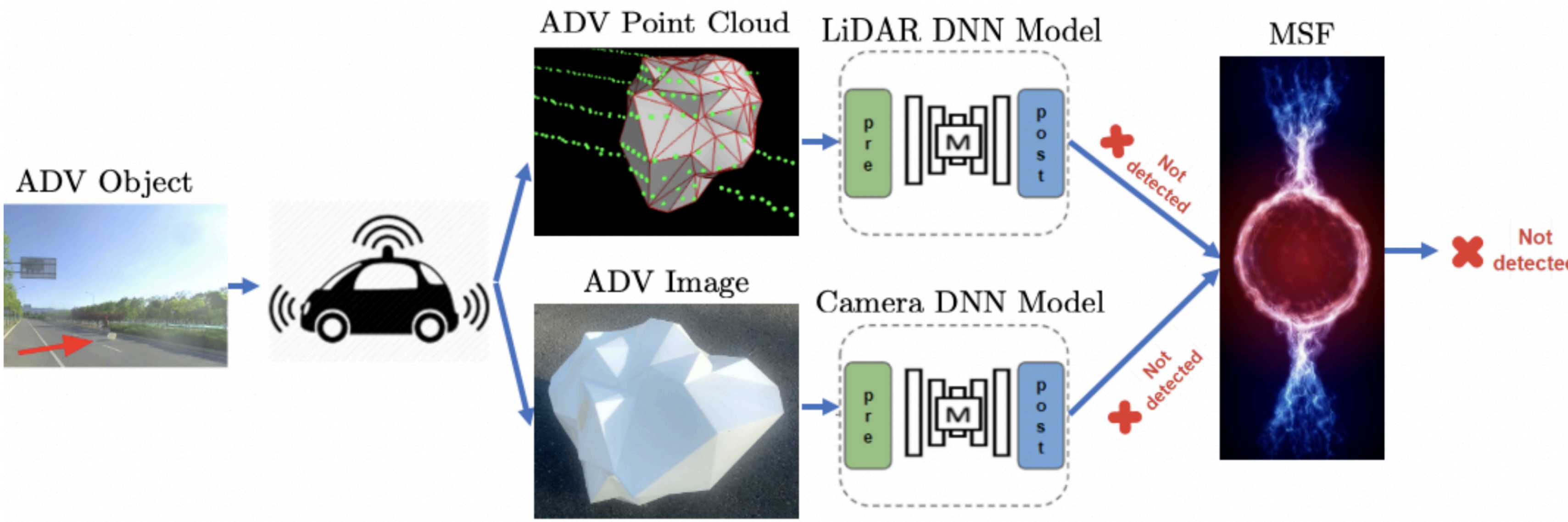
Adversarial object/benign box **on the right**

Benign Object

Adversarial Object



Physical World MSF-based Attacks

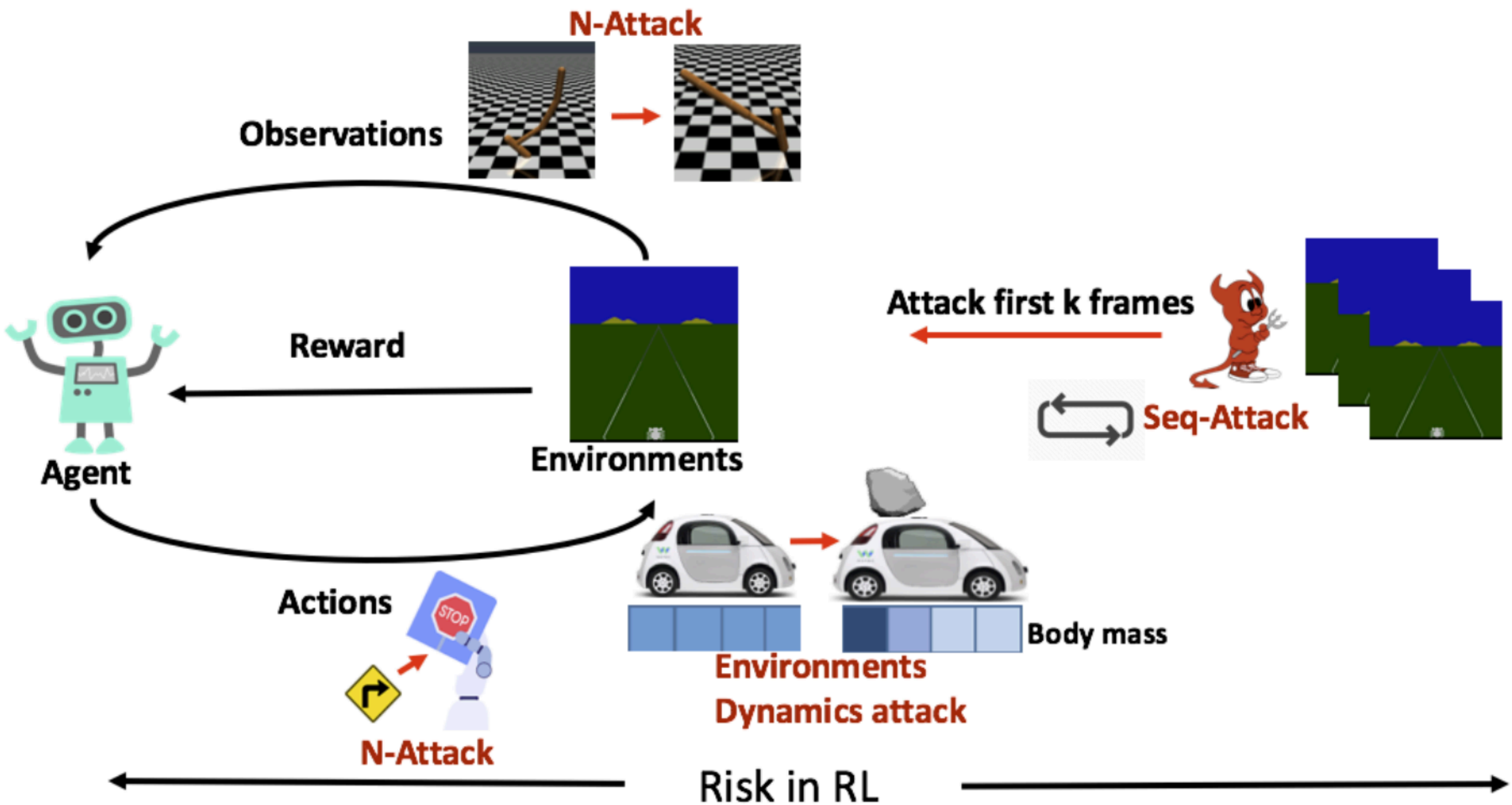


https://aisecure.github.io/BLOG/MRF/Home.html

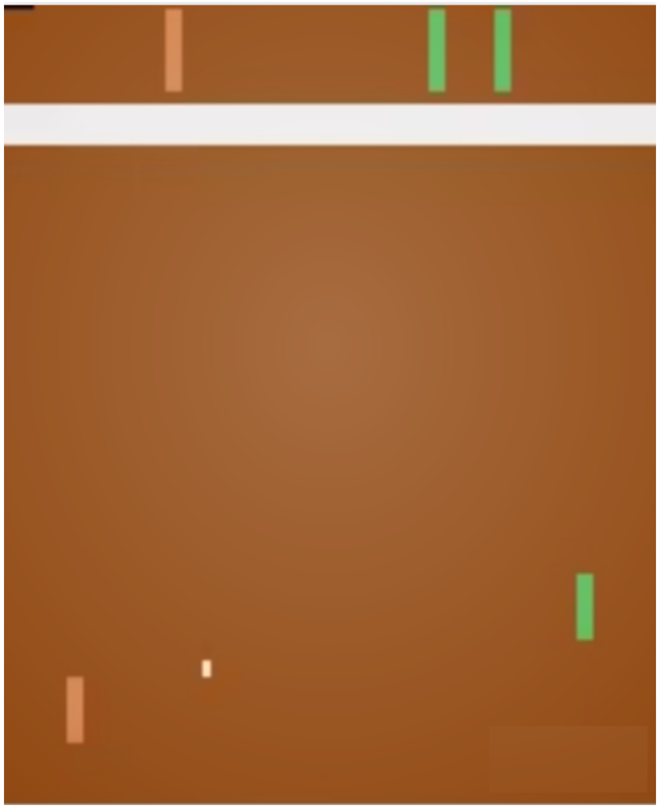
Takeaways

Adversarial perturbations are possible in physical world under different conditions and viewpoints, including the distances and angles.

Attacking Deep Reinforcement Learning



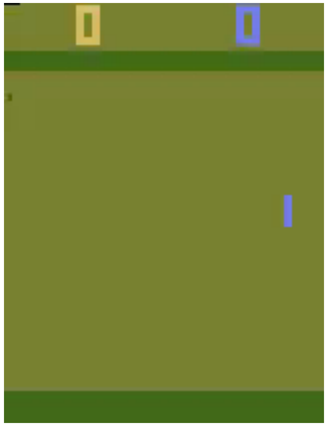
A3C: A Deep Policy on Pong



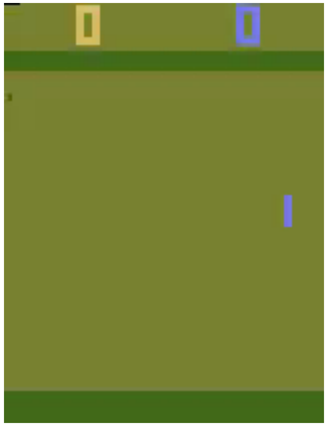
Reinforcement learning algorithms:

- Actor – **policy network** to predict the action based on each frame
- Critics – **value function** to predict the value of each frame, and the action is chosen to maximize the expected value
- Actor-critics (A3C) – combine value function into the policy network to make prediction

Agent in Action: attack the policy network

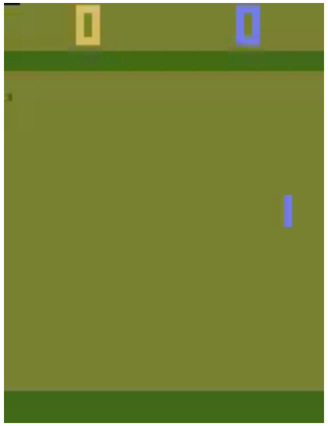


Original Frames

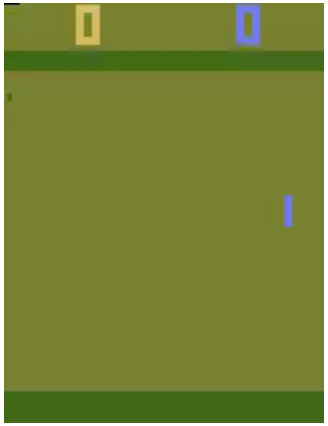


Adversarial perturbation injected into **every frame**

Agent in Action: attack the value function



Original Frames



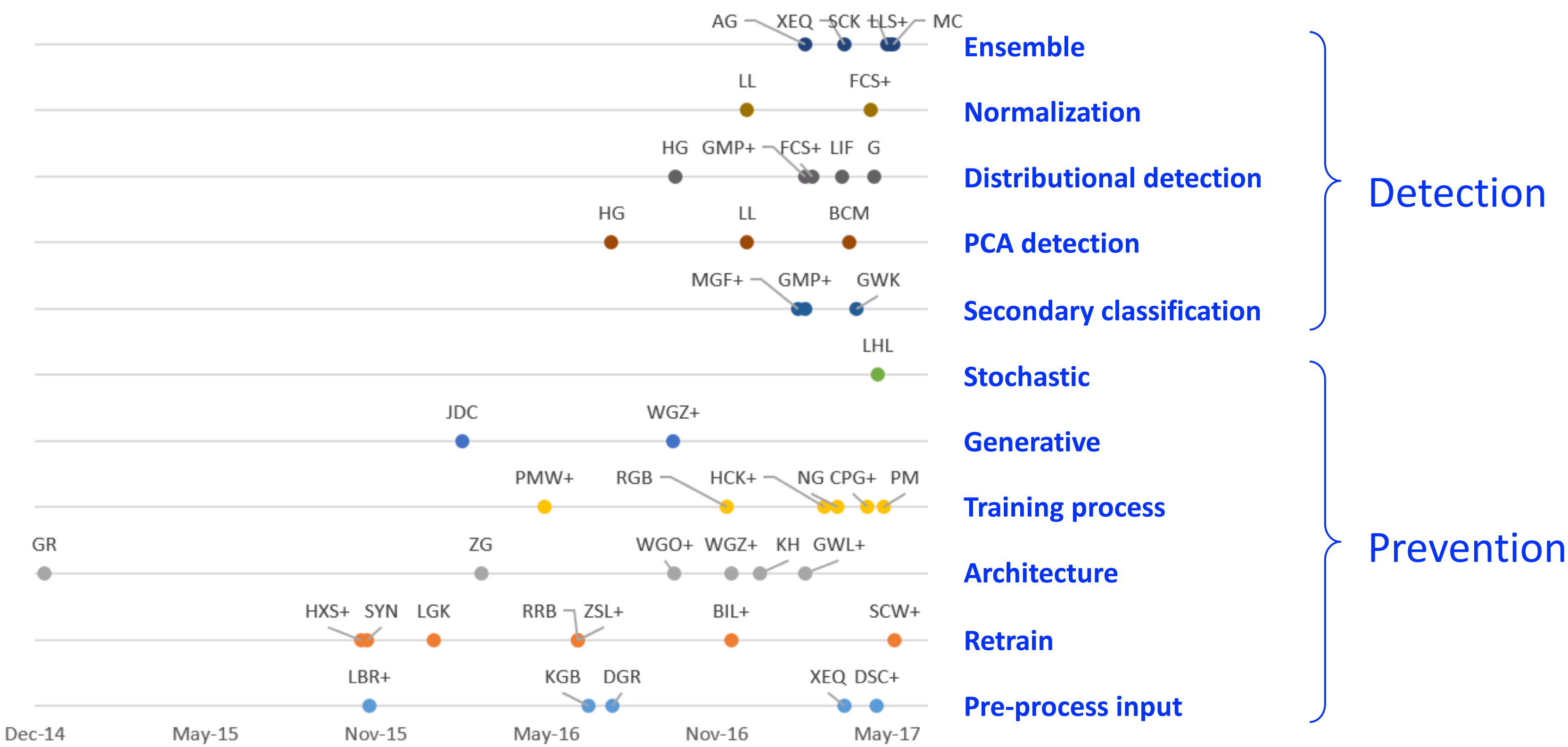
Adversarial perturbation injected into **every other 10 frames**

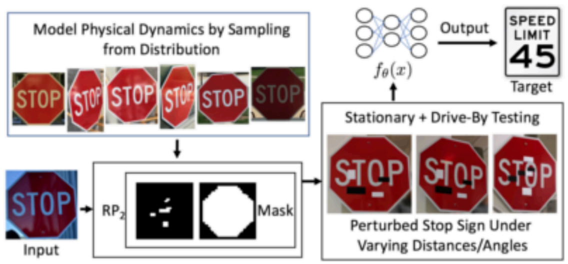

TORCS
TORCS
TORCS
TORCS
TORCS
TORCS

Takeaways

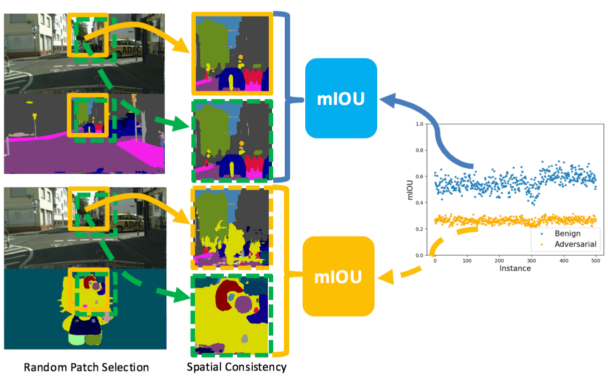
- **Reinforcement learning** systems (e.g., robotics, self-driving systems) are also **vulnerable** to adversarial examples
- To attack a reinforcement learning system, **adversarial perturbations need not be injected to every frame**.

Numerous Defenses Proposed





Robust physical world attacks against **different sensors**



Potential **defenses** against adversarial behaviors based on intrinsic learning properties

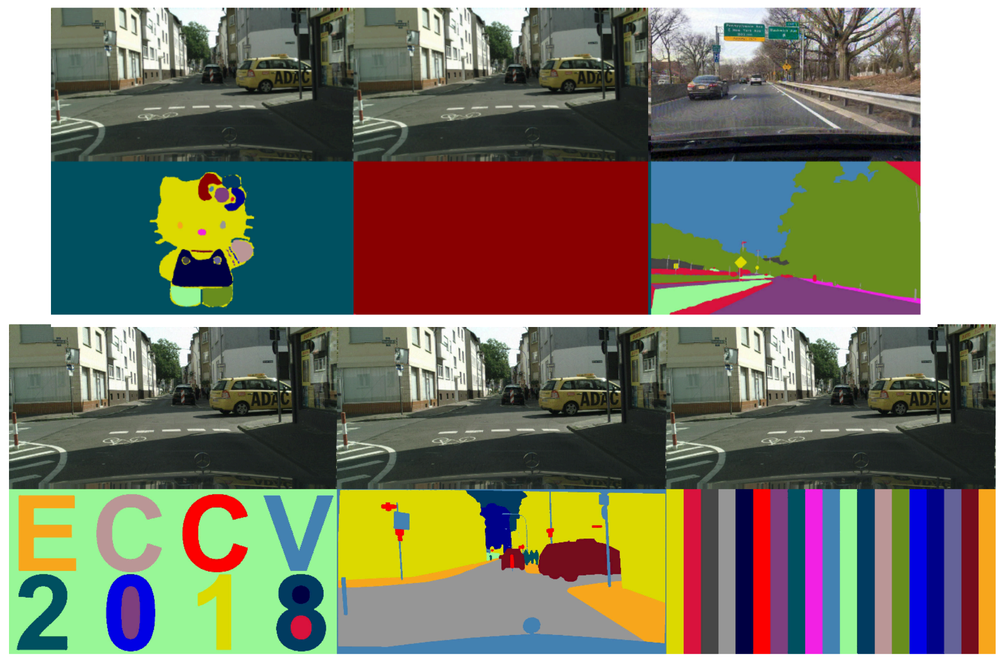
Beyond the Min-max Game

- Will it help if we have more knowledge about our learning tasks?
  - Properties of learning tasks or data
  - General understanding about ML models

Characterize Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation

- Attacks against semantic segmentation
  - State-of-the-art attacks against segmentation: Houdini [NIPS2017], DAG [ICCV 2017]
  - We design diverse adversarial targets: hello kitty, pure color, a real scene, ECCV, color shift, strips of even color of classes
  - Cityscapes and BDD datasets

Benign



Adversarial Examples

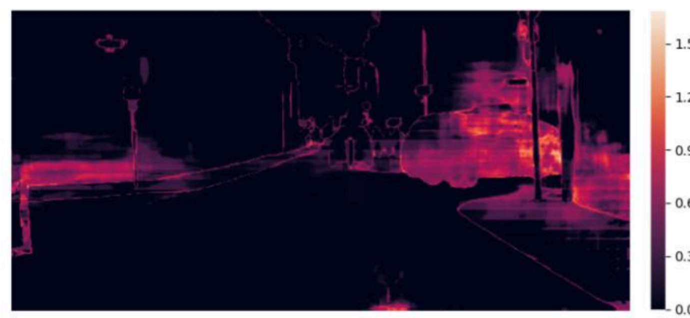
Spatial Context Information

- Spatial consistency is a distinct property of image segmentation
- Perturbation at one pixel will potentially affect the prediction of surrounding pixels

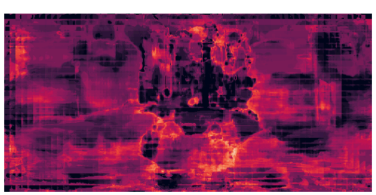
$$\mathcal{H}(m) = -\sum_{j} \mathcal{V}_m[j] \log \mathcal{V}_m[j]$$

For each pixel m, we select its neighbor pixels and calculate the entropy of their predictions for m

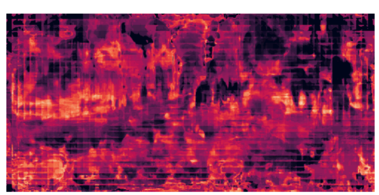
(a) Benign example



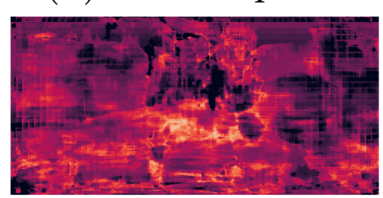
(b) Heatmap of benign image



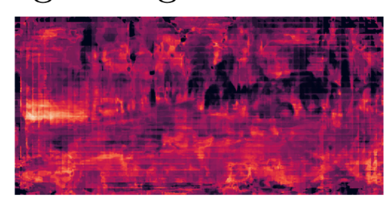
(c) DAG | Kitty



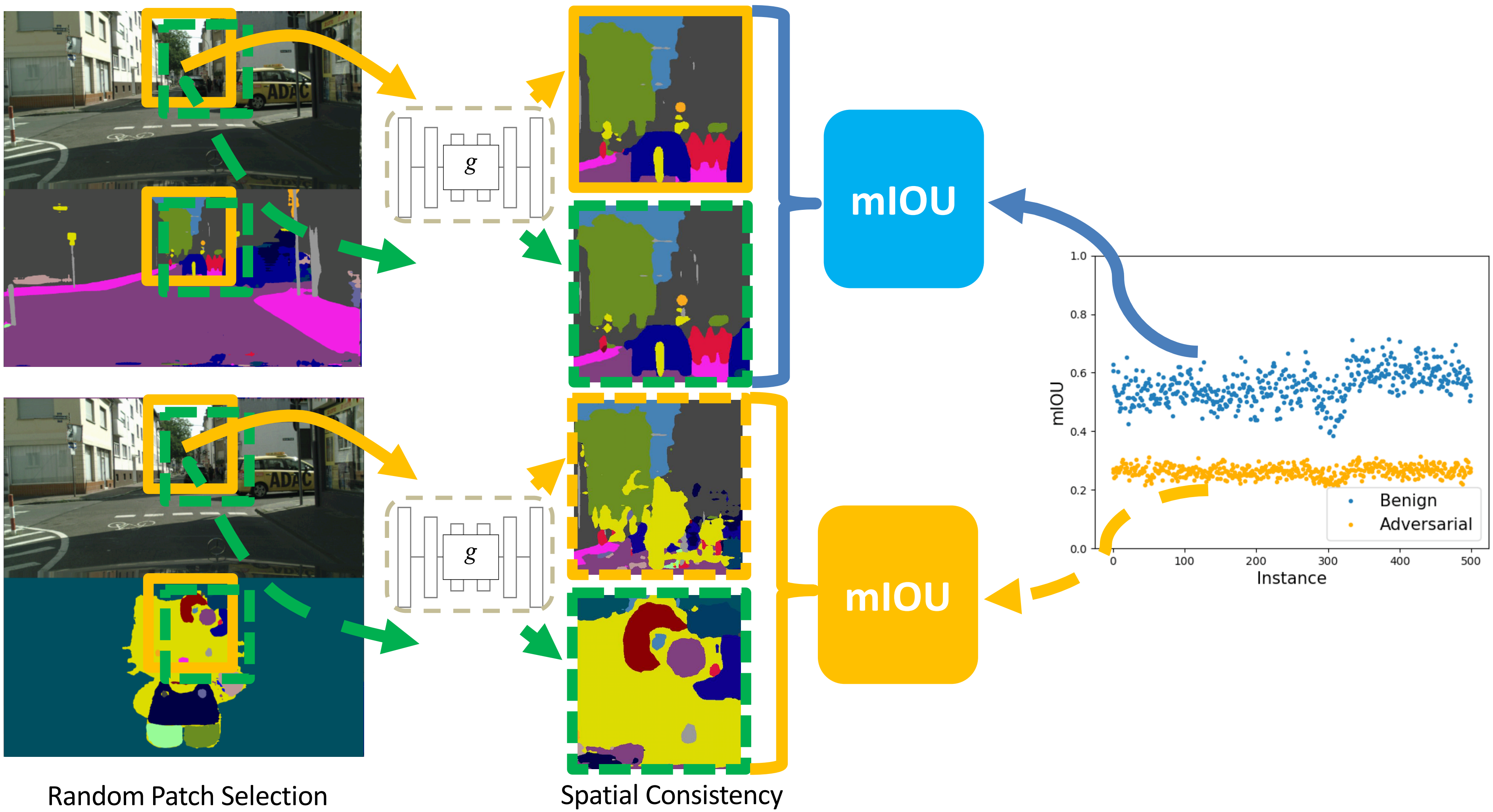
(d) DAG | Pure



(e) Houdini | Kitty



(f) Houdini | Pure



Pipeline of spatial consistency based detection for adversarial examples on semantic segmentation

Detecting adversarial instances based on spatial consistency information

- Both the spatial consistency based detection and the scaling based baseline achieve promising detection rate on different attacks
- The scaling based baseline fails to detect strong adaptive attacks while the spatial based method can

Method		Model	mIOU	Detection				Detection Adap			
				DAG		Houdini		DAG		Houdini	
				Pure	Kitty	Pure	Kitty	Pure	Kitty	Pure	Kitty
Scale (std)	0.5	DRN (16.4M)	66.7	100%	95%	100%	99%	100%	67%	100%	78%
	3.0			100%	100%	100%	100%	100%	0%	97%	0%
	5.0			100%	100%	100%	100%	100%	0%	71%	0%
Spatial (K)	1	DRN (16.4M)	66.7	91%	91%	94%	92%	98%	94%	92%	94%
	5			100%	100%	100%	100%	100%	100%	100%	100%
	10			100%	100%	100%	100%	100%	100%	100%	100%
	50			100%	100%	100%	100%	100%	100%	100%	100%

Takeaways

Spatial consistency information can be potentially applied to help distinguish benign and adversarial instances against segmentation models.

Temporal consistency?

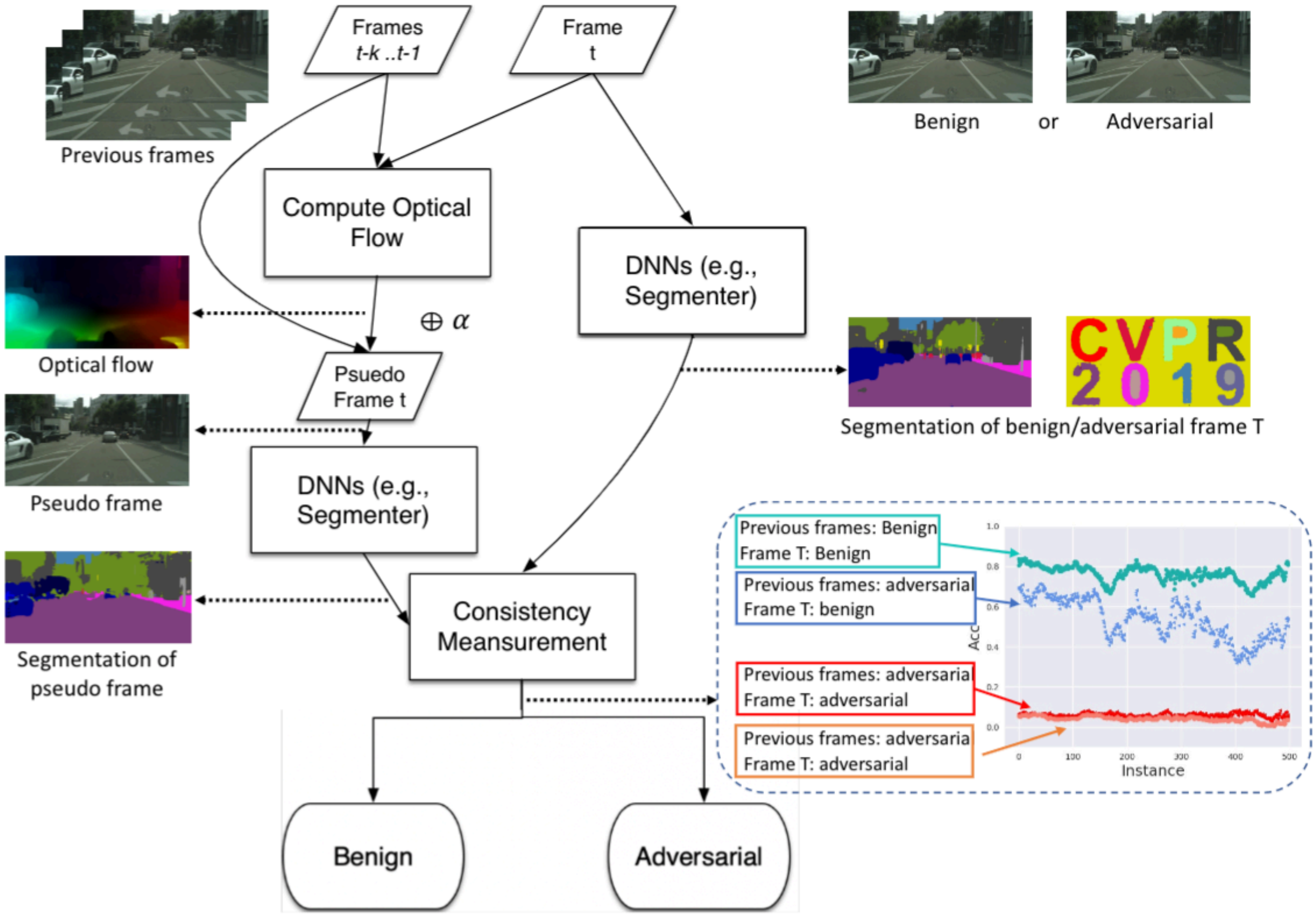
Adversarial Frames In Videos

Attacks on segmentation

Attacks on pose estimation

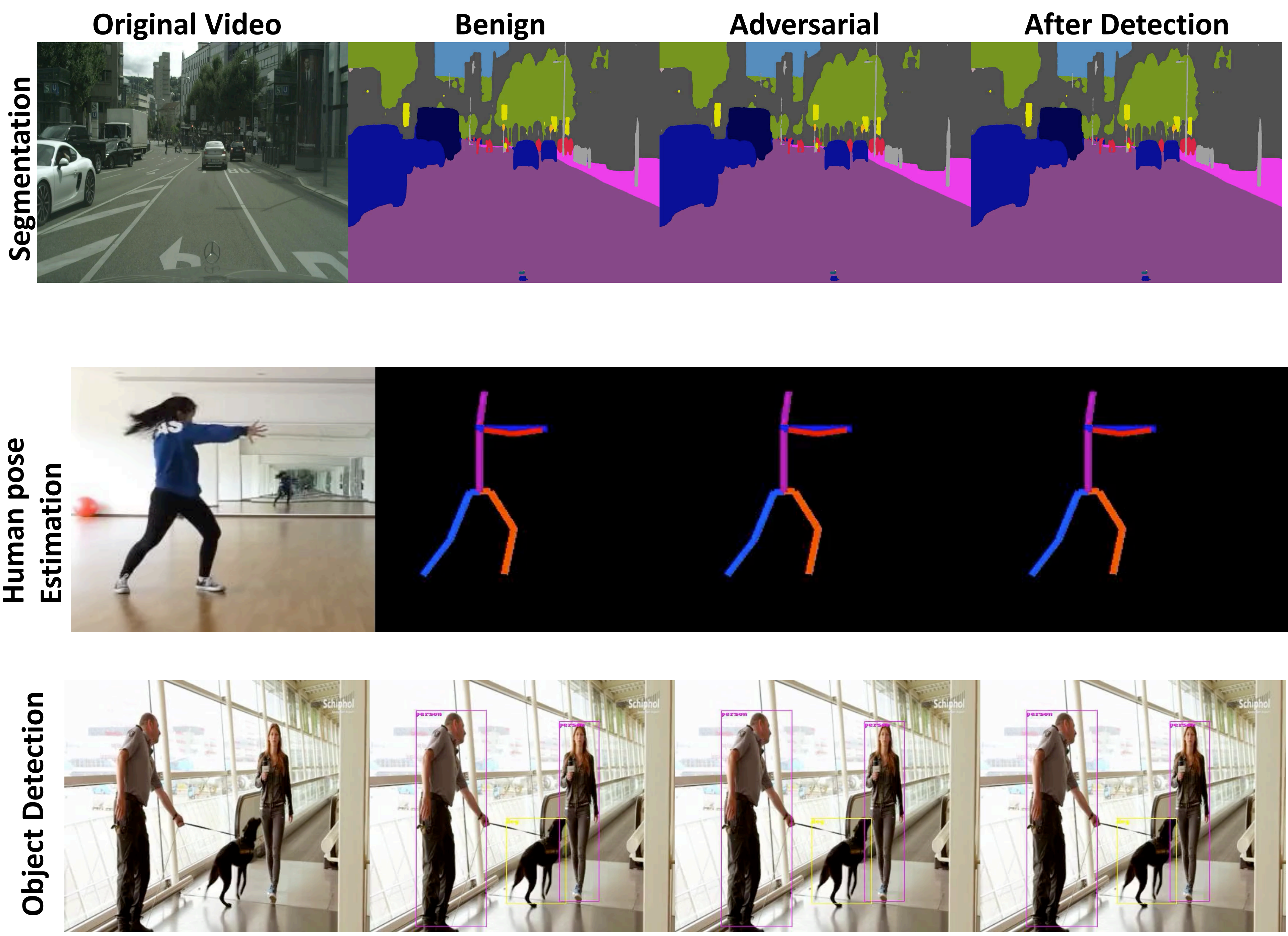
Attacks on object detection

Defensing Adversarial behaviors in Videos – Temporal Dependency



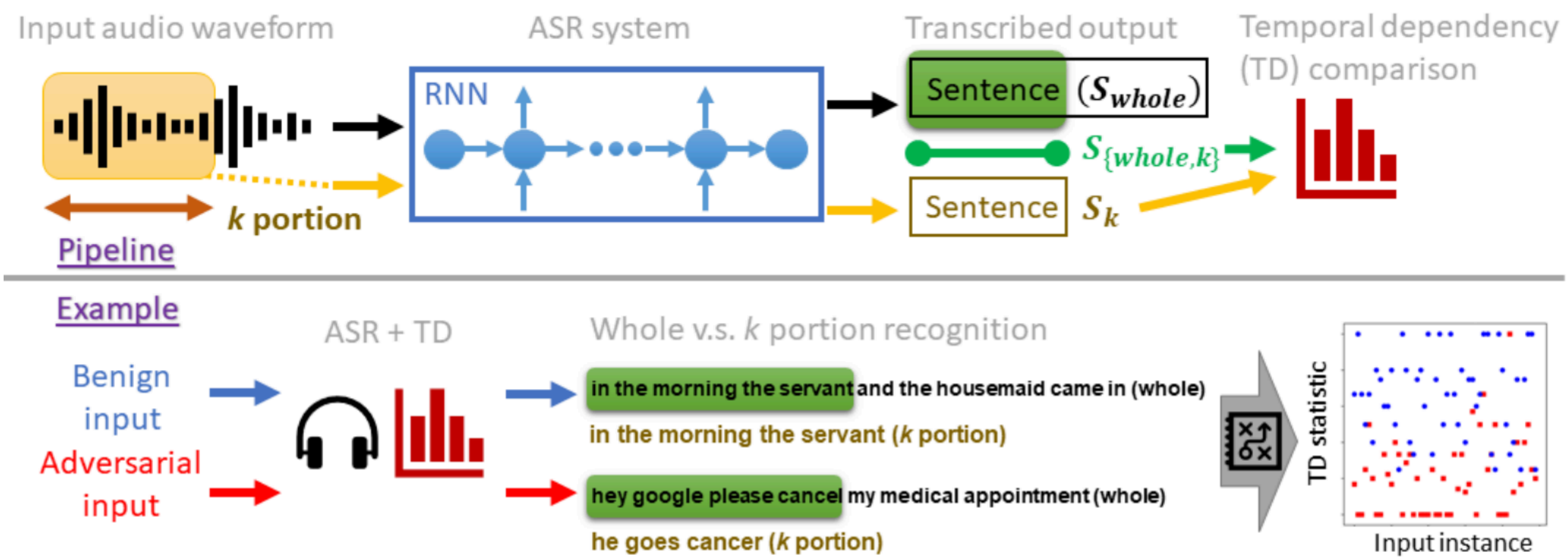
Task	Attack Method	Target	Previous Frames	Detection 1	Detection 3	Detection 5	Detection Adap 1	Detection Adap 3	Detection Adap 5
Semantic Segmentation	Houdini	CVPR	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Remapping	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Stripe	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	99%	100%	100%
	DAG	CVPR	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Remapping	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
		Stripe	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	100%	100%	100%
Human Pose Estimation	Houdini	shuffle	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	99%	100%	100%
		Transpose	Benign	100%	100%	100%	98%	100%	100%
			Adversarial	98%	99%	100%	98 %	99%	100%
Object Detection	DAG	all	Benign	100%	100%	100%	100%	100%	100%
			Adversarial	100%	100%	100%	98%	100%	100%
		person	Benign	99%	100%	100 %	100%	100%	100%
			Adversarial	97%	98%	100%	96 %	97%	100%

- The results show that choosing more random patches can improve detection rate while k=5 is enough to achieve AUC 100%
- The spatial consistency based detection is robust against strong adaptive attackers due to the randomness in patch selection


Original Video
Benign
Adversarial
After Detection
Segmentation
Human pose
Estimation
Object Detection
Schiphol
person
Dog

Temporal Consistency Based Analysis

- “Yanny” or “Laurel”? – adversarial audio



[ICLR 2019]

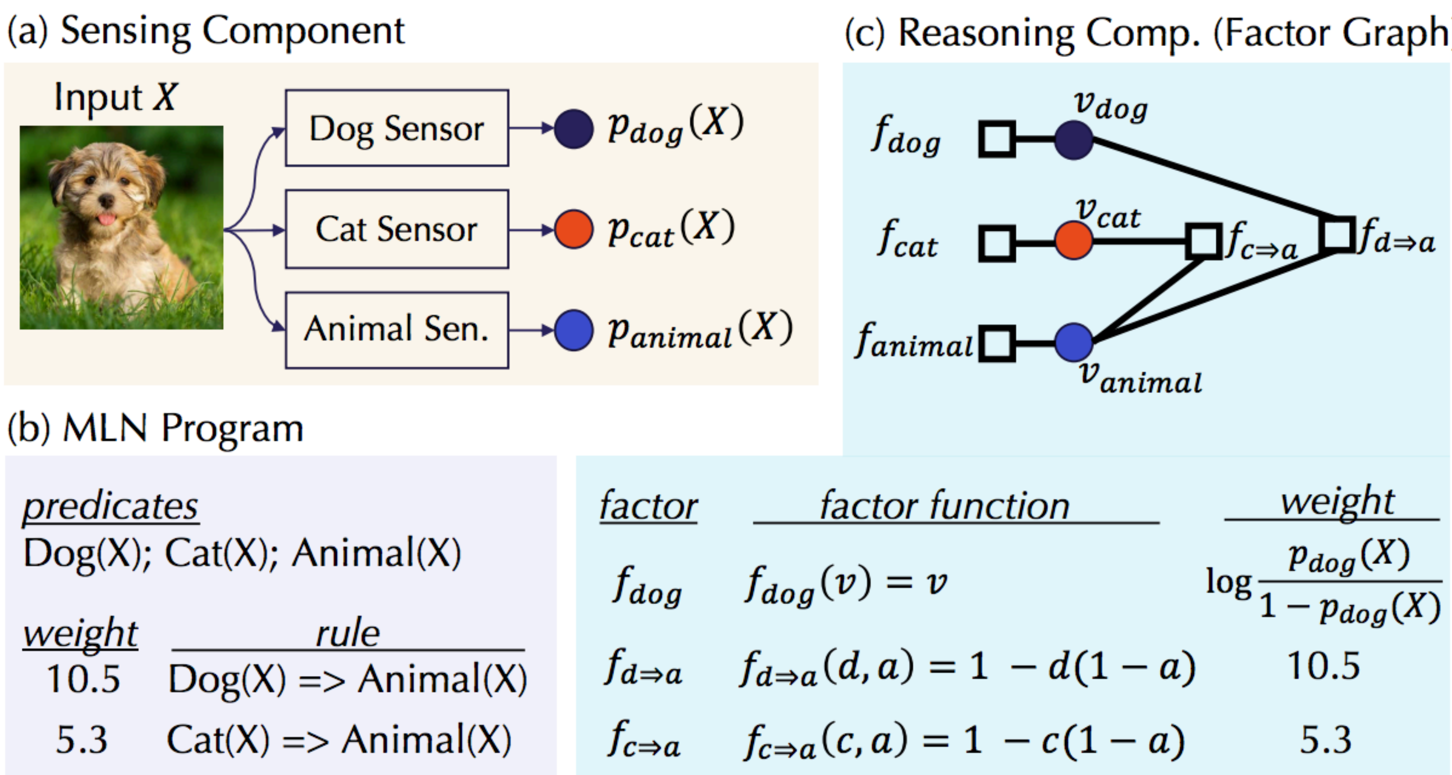
Temporal Consistency (TD) Based Detection

Type	Transcribed results
Original	then good bye said the rats and they went home
the first half of Original	then good bye said the raps
Adversarial (short)	hey google
First half of Adversarial	he is
Adversarial (medium)	this is an adversarial example
First half of Adversarial	thes on adequate
Adversarial (long)	hey google please cancel my medical appointment
First half of Adversarial	he goes cancer

Dataset	LSTM	TD (WER)	TD (CER)	TD (LCP ratio)
Common Voice	0.712	**0.936**	0.916	0.859
LIBRIS	0.645	0.930	**0.933**	0.806

TD achieves high detection rate for adversarial audio

Certified Robustness for Sensing-Reasoning ML Pipelines

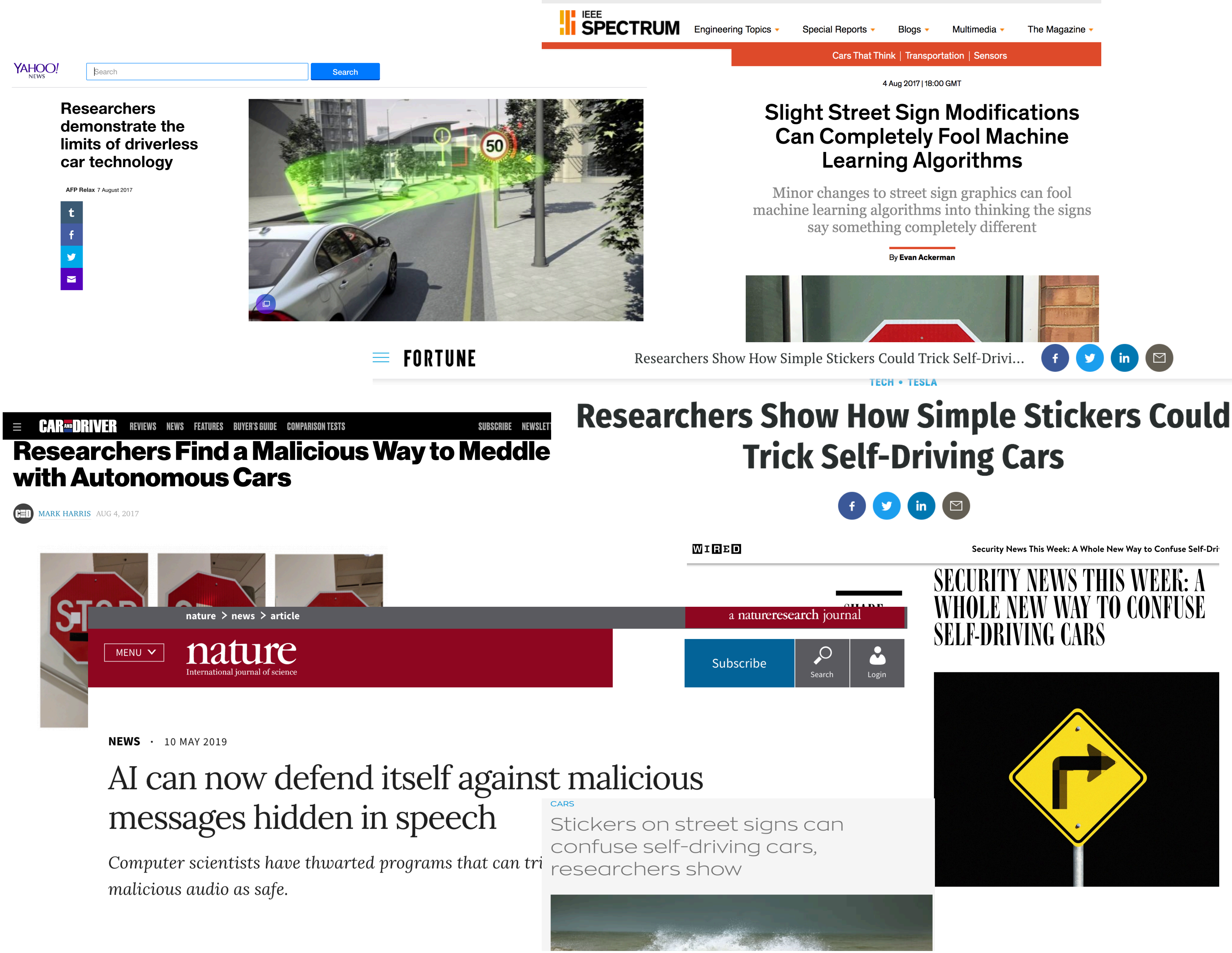


Definition 3 (ROBUSTNESS). Given input polynomial-time computable weight function $w(\cdot)$ and query function $Q(\cdot)$, parameters α, two real numbers $\epsilon > 0$ and $\delta > 0$, a ROBUSTNESS oracle decides, for any $\alpha' \in P^{[m]}$ such that $\|\alpha - \alpha'\|_\infty \leq \epsilon$, whether the following is true:

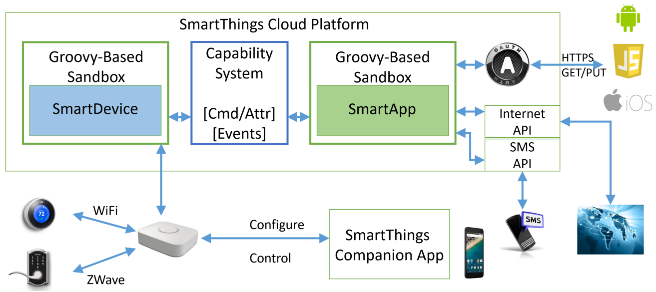
$$\left|\mathbf{E}_{\sigma\sim\pi_\alpha}[Q(\sigma)] - \mathbf{E}_{\sigma\sim\pi_{\alpha'}}[Q(\sigma)]\right| < \delta.$$

Conclusions

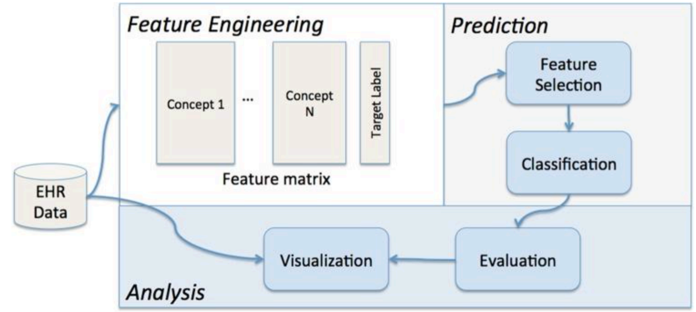
- ML models are vulnerable to sophisticated adversarial attacks (e.g. evasion, poisoning)
- Any ML models can be adversarially attacked
- Lead board of the certified robustness: https://github.com/AI-secure/Provable-Training-and-Verification-Approaches-Towards-Robust-Neural-Networks
- First certified robustness against backdoor attacks: https://arxiv.org/abs/2002.11750



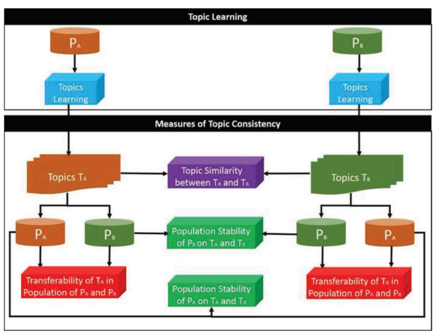
YAHOO! NEWS
Search
Researchers demonstrate the limits of driverless car technology
AFP Relax 7 August 2017
50
IEEE SPECTRUM
Engineering Topics
Special Reports
Blogs
Multimedia
The Magazine
Cars That Think | Transportation | Sensors
4 Aug 2017 | 18:00 GMT
Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms
Minor changes to street sign graphics can fool machine learning algorithms into thinking the signs say something completely different
By Evan Ackerman
FORTUNE
Researchers Show How Simple Stickers Could Trick Self-Drivi...
TECH • TESLA
Researchers Show How Simple Stickers Could Trick Self-Driving Cars
CAR AND DRIVER
REVIEWS
NEWS
FEATURES
BUYER'S GUIDE
COMPARISON TESTS
SUBSCRIBE
NEWSLET
Researchers Find a Malicious Way to Meddle with Autonomous Cars
MARK HARRIS AUG 4, 2017
WIRED
Security News This Week: A Whole New Way to Confuse Self-Dri
SECURITY NEWS THIS WEEK: A WHOLE NEW WAY TO CONFUSE SELF-DRIVING CARS
nature > news > article
a natureresearch journal
MENU
nature
International journal of science
Subscribe
Search
Login
NEWS · 10 MAY 2019
AI can now defend itself against malicious messages hidden in speech
Computer scientists have thwarted programs that can tri
malicious audio as safe.
CARS
Stickers on street signs can confuse self-driving cars, researchers show



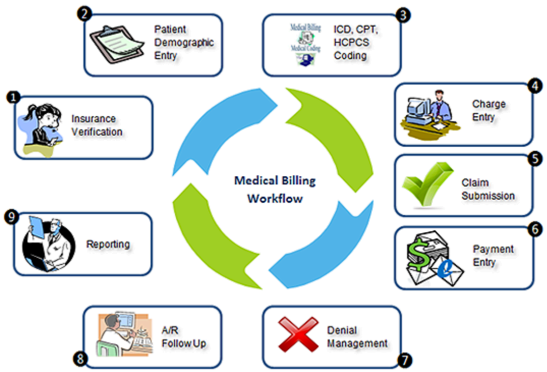
Robust Smart Home



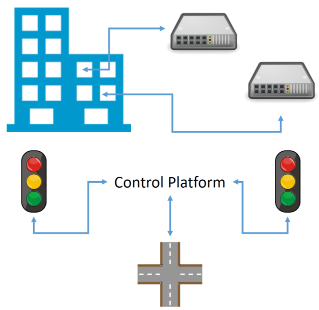
Privacy-Preserving Data Analysis



Topic of Workflow Analysis

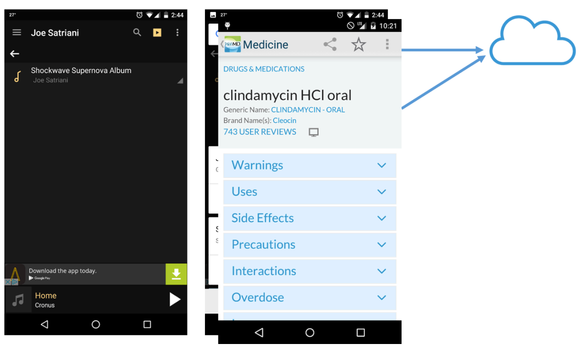


Game Theoretic Auditing System for EMR

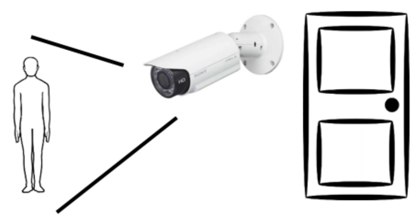


Large-Scale Auditing Game With Human In the Loop

Robust Learning



Privacy Protected Mobile Healthcare



Robust Face Recognition Against Poisoning Attack

Thank You!

Bo Li
lbo@illinois.edu

http://boli.illinois.edu/