Secure Learning in Adversarial Environments

Bo Li
University of Illinois at Urbana-Champaign

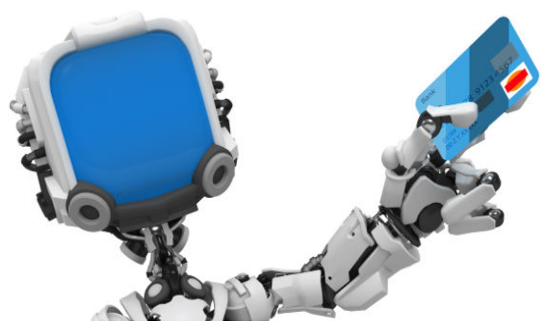
Machine Learning is Ubiquitous

Autonomous Driving

Healthcare

Smart City

Malware Classification

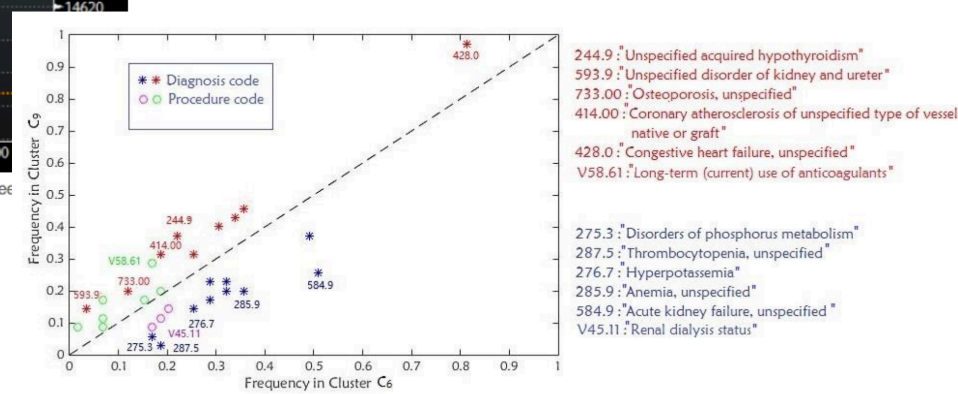


Fraud Detection

Biometrics Recognition

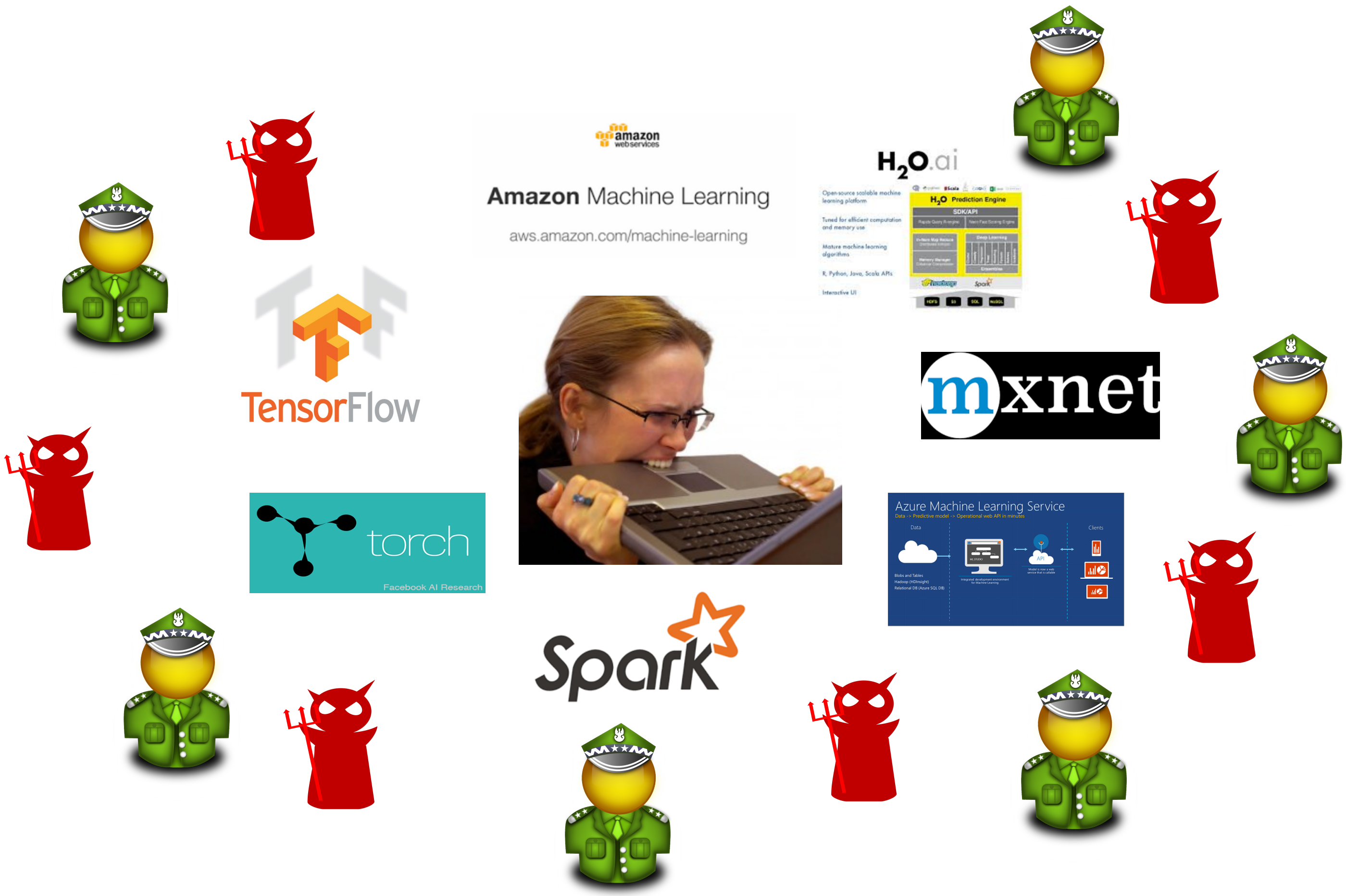
Security & Privacy Problems

Security Problems



Privacy Concerns

We Live in an Adversarial Environment



While cybersecurity R&D needs are addressed in greater detail in the NITRD Cybersecurity R&D Strategic Plan, some cybersecurity risks are specific to AI systems. ***One key research area is "adversarial machine learning"****, that explores the degree to which AI systems can be compromised by "contaminating" training data, by modifying algorithms, or by making subtle changes to an object that prevent it from being correctly identified….*

- National Science and Technology Council
2016

DARPA

Guaranteeing AI Robustness against Deception (GARD)

Office of the Director of National Intelligence
IARPA
BE THE FUTURE

FEDBIZOPPS.GOV Federal Business Opportunities

Home | Getting Started | General Info | Opportunities | Agencies | Privacy

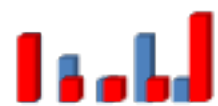
Buyers: Login | Register Vendors: Login | Register Accessibility

IARPA

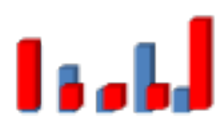
Proposers' Day Notification for Secure, Assured, Intelligent Learning Systems (SAILS) and Trojans in Artificial Intelligence (TrojAI)
Solicitation Number: IARPA-BAA-19-02_IARPA-BAA-19-03
Agency: Office of the Director of National Intelligence
Office: Intelligence Advanced Research Projects Activity
Location: IARPA1

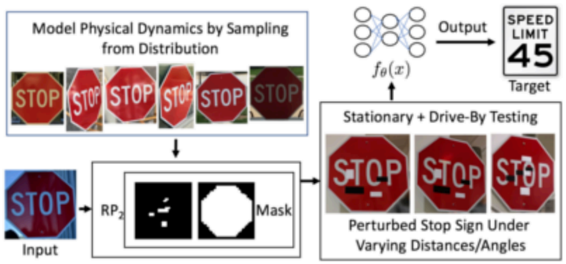
Dangers of Stationary Assumption

Traditional machine learning approaches assume

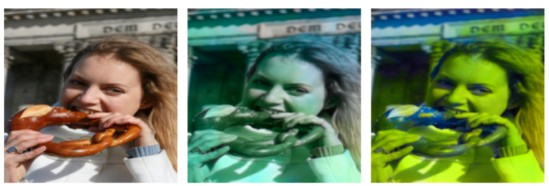
Training Data 

$\approx$

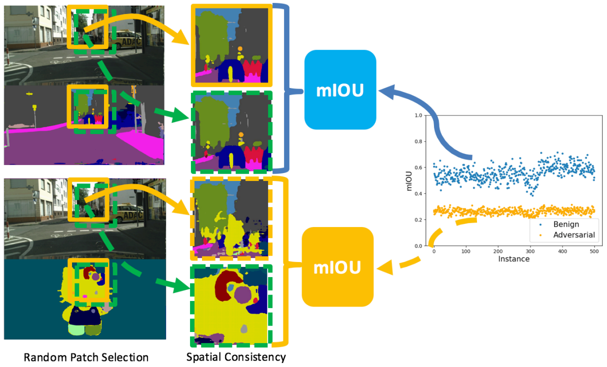
Testing Data 



Real world attacks against **different sensors**



Potential **defenses** against adversarial behaviors via game theory

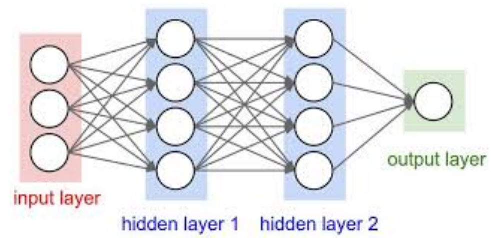


Potential **defenses** against adversarial behaviors based on learning properties

Adversarial Perturbation In ML

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

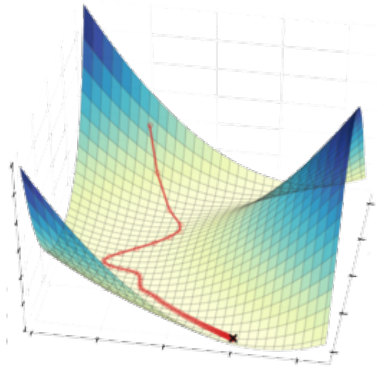
Model parameters Input feature vector label



Deep Neural Networks

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

Adversarial perturbation



Gradient Descent

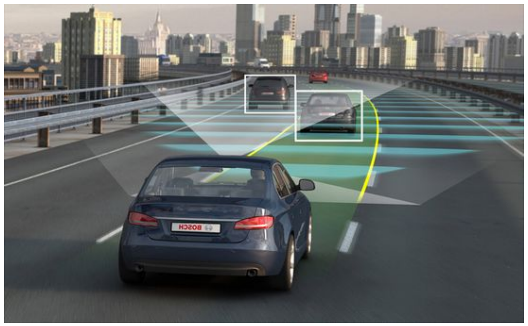
How to solve the adversary strategy

Local search

Combinatorial optimization

Convex relaxation

Autonomous Driving in Practice



However, What We Can See Everyday…

The Physical World Is... Messy

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

Physical Limits on Imperceptibility

Fabrication/Perception Error (Color Reproduction, etc.)

Digital Noise (What you want)

What is printed

What a camera may see

Background Modifications*

Image Courtesy, OpenAI

[CVPR, *2017*]

An Optimization Approach To Creating Robust Physical Adversarial Examples

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

Perturbation/Noise Matrix

Adversarial Target Label

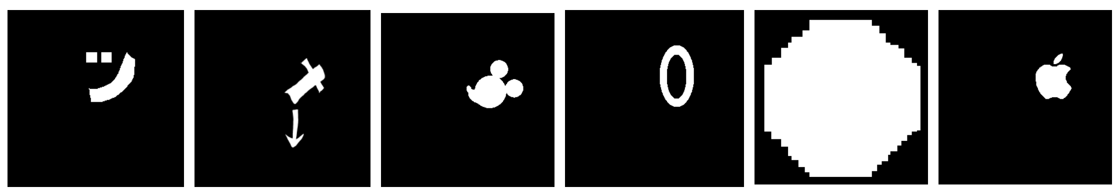
Lp norm (L-0, L-1, L-2, …)

Loss Function

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

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"

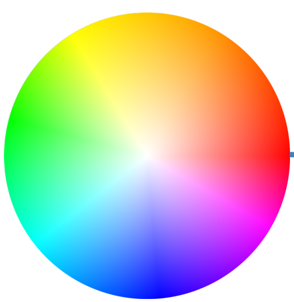
Handling Fabrication/Perception Errors

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

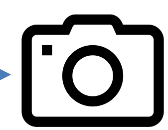
$$NPS(\delta) = \sum_{\hat{p} \in \delta} \prod_{p' \in P} |\hat{p} - p'|$$

P is a set of printable RGB triplets

Color Space



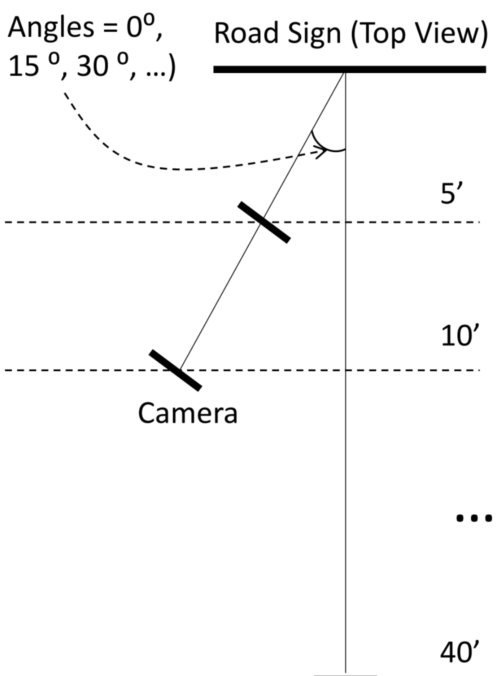
Sampled Set of RGB Triplets



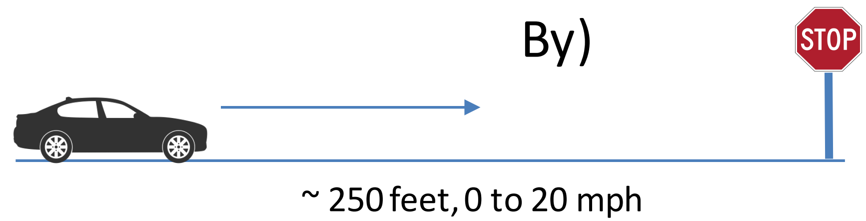
NPS based on Sharif et al., “Accessorize to a crime,” CCS 2016

How Can We Realistically Evaluate Attacks?

Lab Test (Stationary)



Field Test (Drive-By)



Record video

Sample frames every k frames

Run sampled frames through DNN

Lab Test Summary (Stationary)

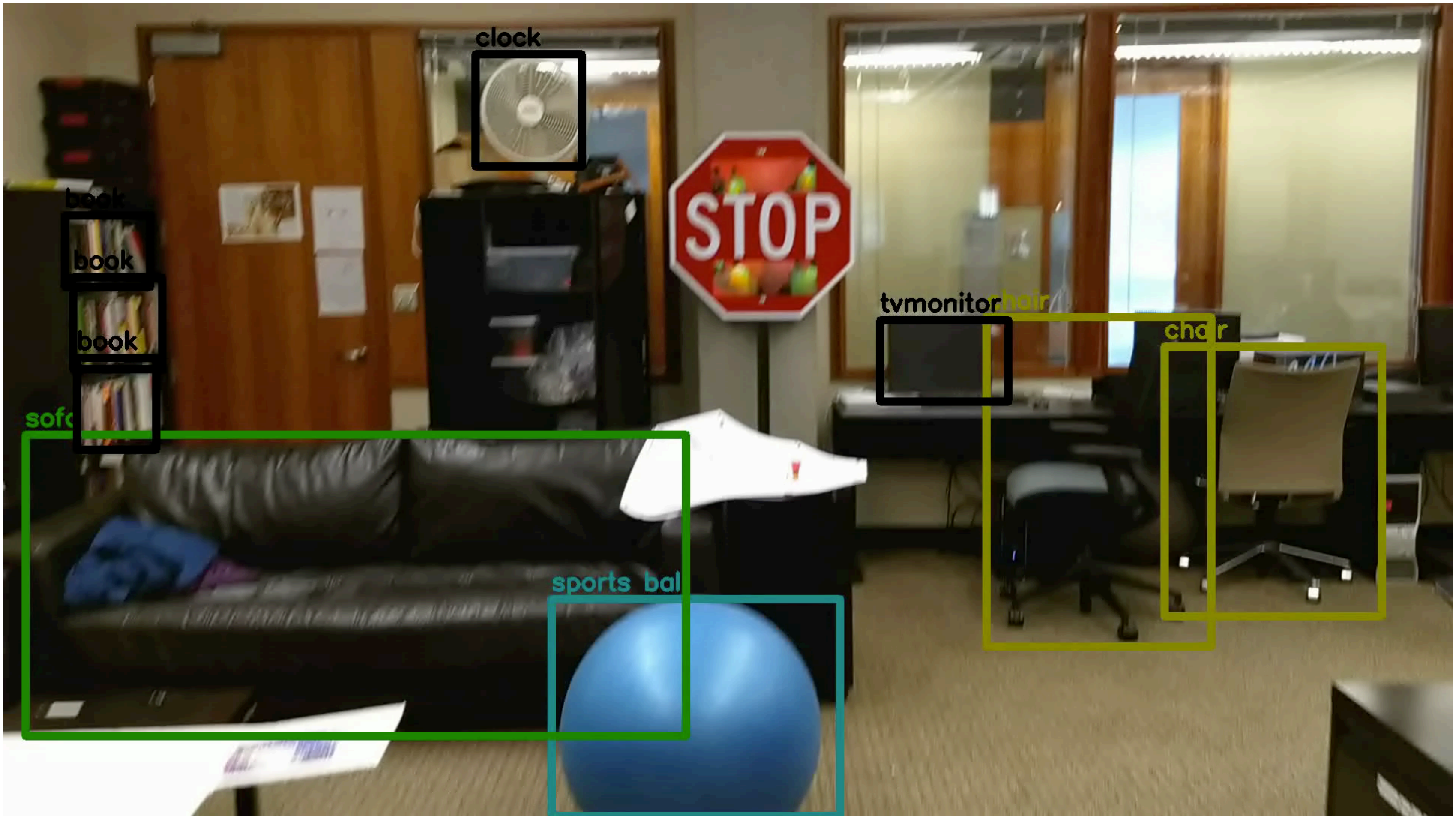
Target Class: Speed Limit 45

Subtle Poster

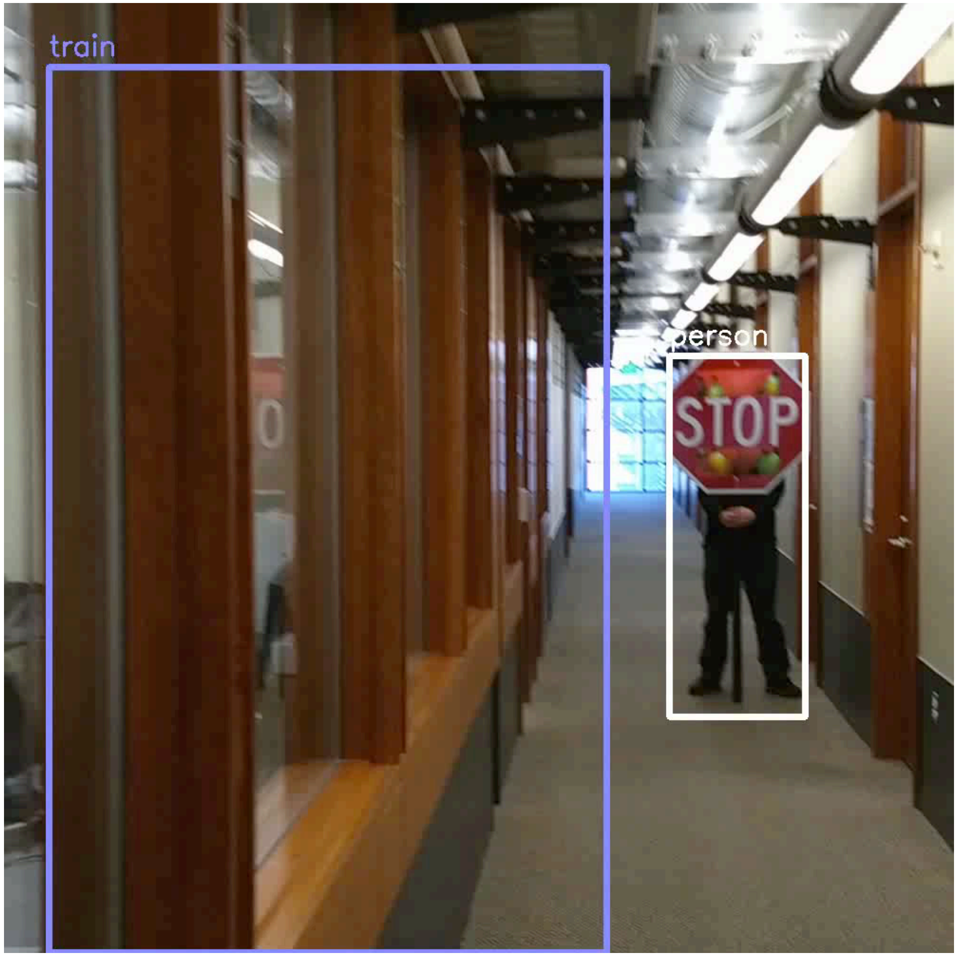
Art Perturbation

Subtle Perturbation

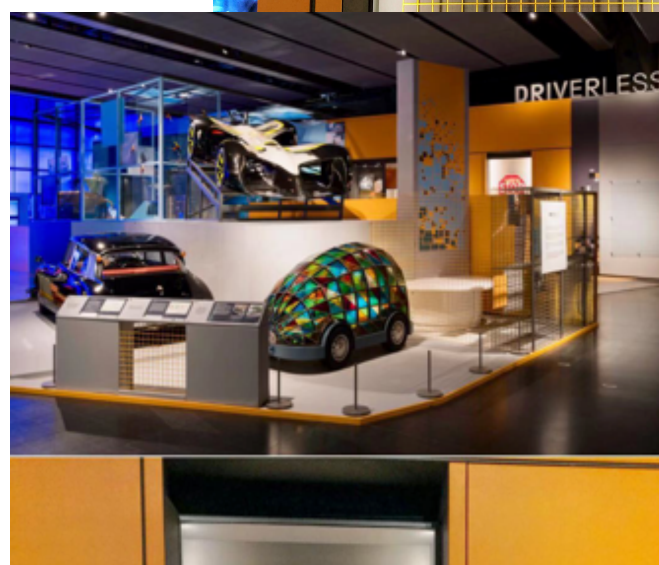
Physical Attacks Against Detectors



Physical Attacks Against Detectors



Physical Adversarial Stop Sign in the Science Museum of London



Adversarial Examples in Physical World

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

Adversarial Point Clouds

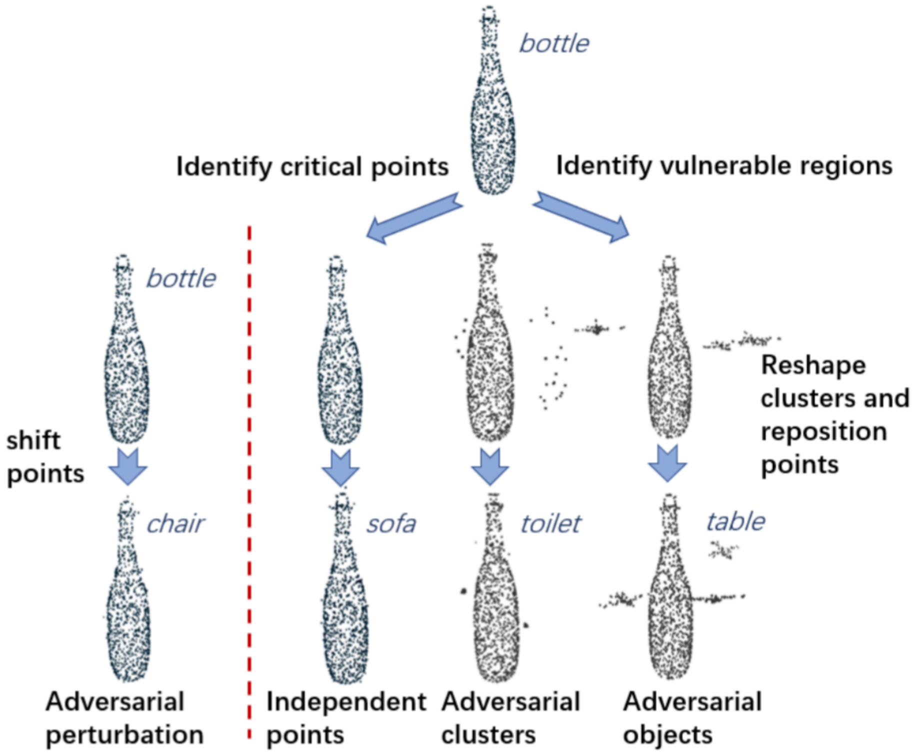
- PointNet is widely used including in autonomous driving systems to process Lidar point cloud data
- Perturbation on point cloud
  - Points shifting
  - Independent points adding
  - Adversarial clusters
  - Adversarial objects
- Adversarial objectives

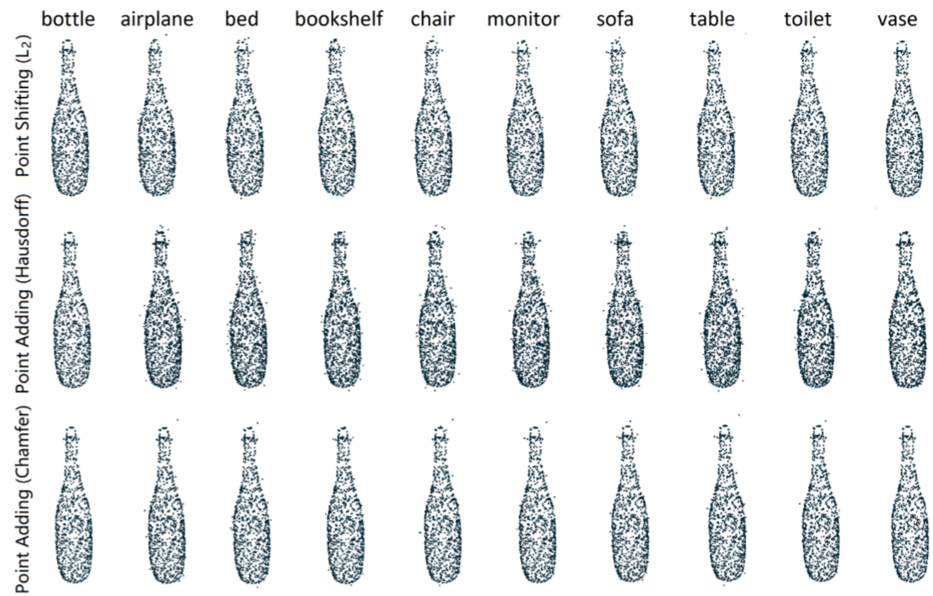
$$\min \mathcal{D}(x, x'), \qquad s.t.\ \mathcal{F}(x') = t'$$

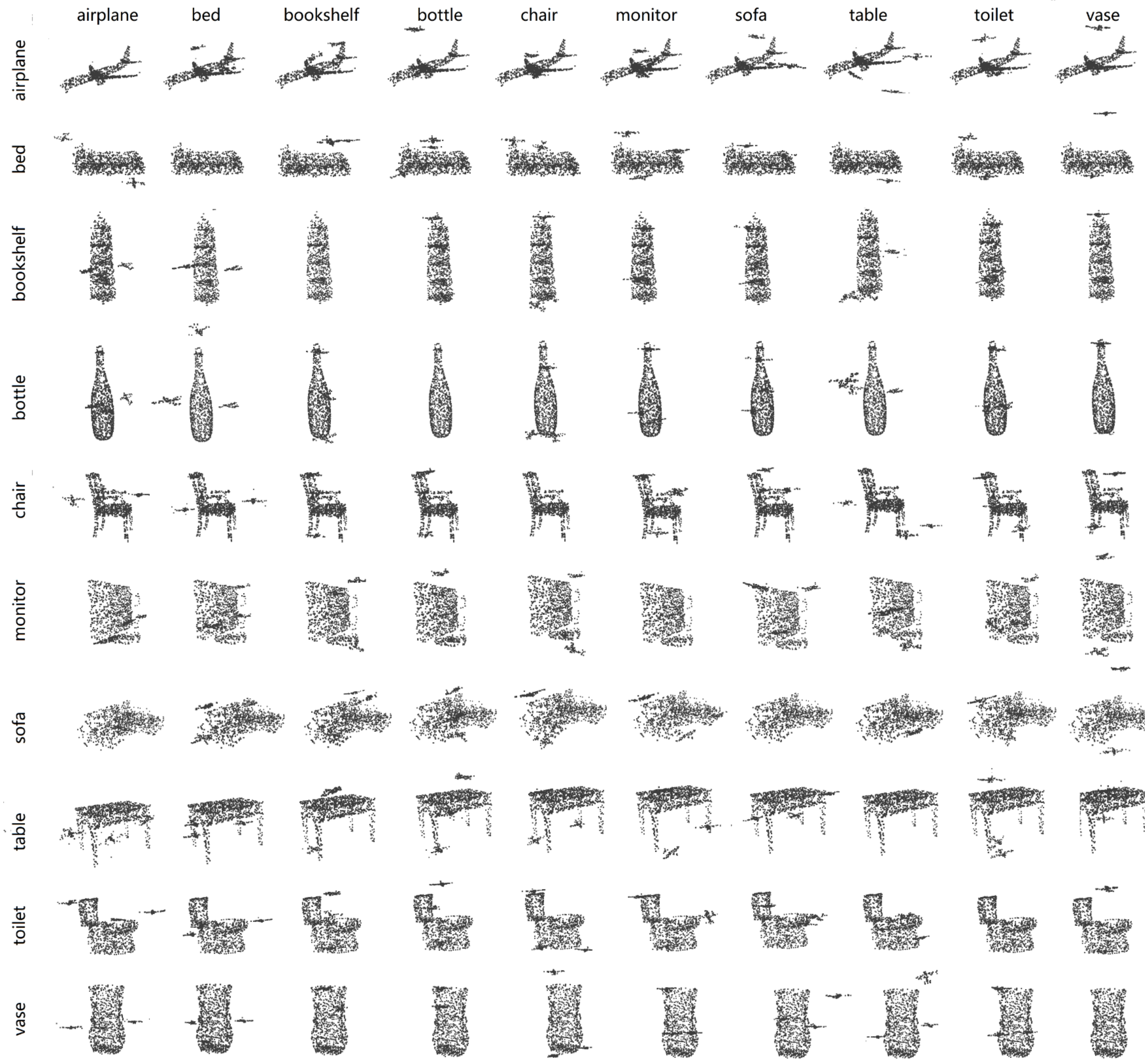
$$\mathcal{D}_C(\mathcal{S}, \mathcal{S}') = \frac{1}{\|\mathcal{S}'\|_0} \sum_{y \in \mathcal{S}'} \min_{x \in \mathcal{S}} \|x - y\|_2^2$$

$$\mathcal{D}_H(\mathcal{S}, \mathcal{S}') = \max_{y \in \mathcal{S}'} \min_{x \in \mathcal{S}} \|x - y\|_2^2$$

$$\min f(x') + \lambda \cdot \sum_i \mathcal{D}_{far}(\mathcal{S}_i) + \mu \cdot \mathcal{D}_C(\mathcal{S}_0, \mathcal{S}_i)$$


bottle
Identify critical points
Identify vulnerable regions
bottle
shift points
chair
Adversarial perturbation
sofa
toilet
table
Reshape clusters and reposition points
Independent points
Adversarial clusters
Adversarial objects


bottle
airplane
bed
bookshelf
chair
monitor
sofa
table
toilet
vase
Point Shifting (L_2)
Point Adding (Hausdorff)
Point Adding (Chamfer)

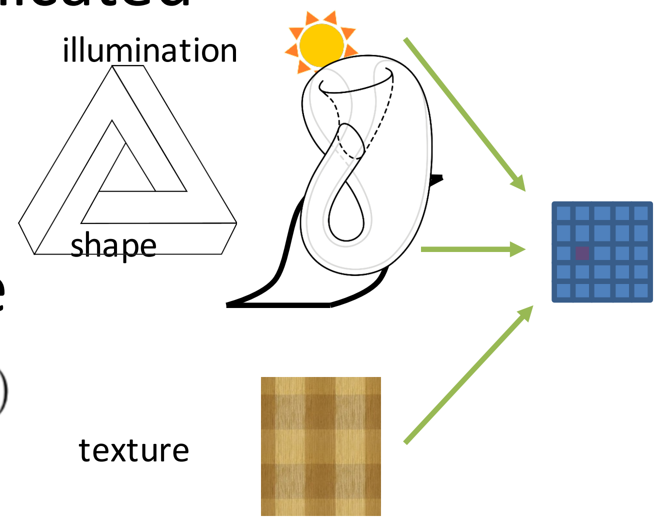

airplane
bed
bookshelf
bottle
chair
monitor
sofa
table
toilet
vase

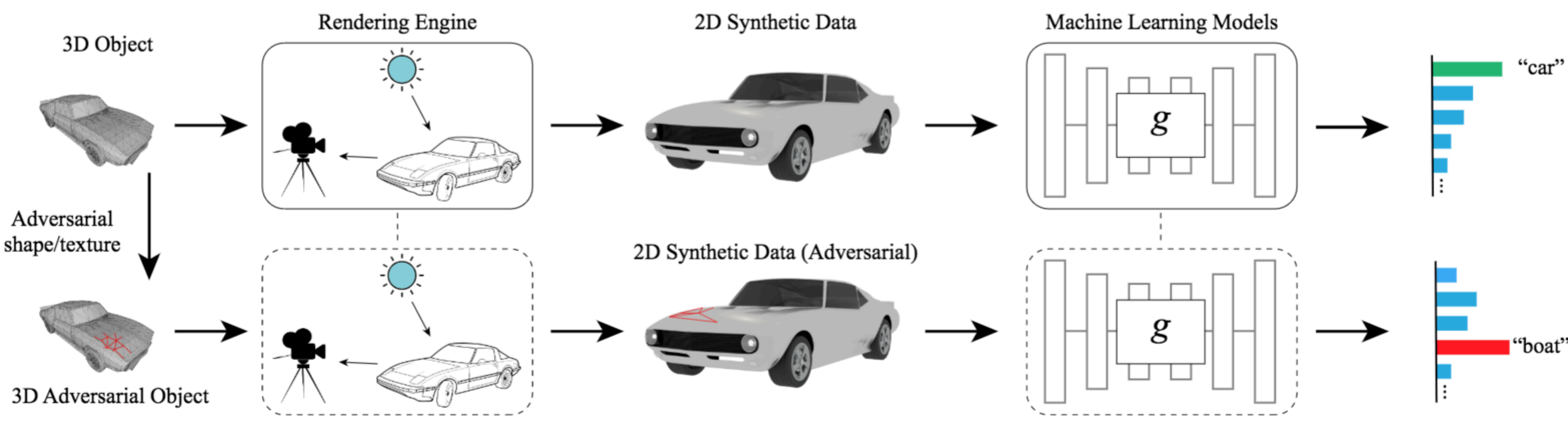
Adversarial 3D Meshes

- 3D to 2D space rendering is complicated
  - Shapes/textures/illumination
- 3D space itself is complicated
- Adversarial optimization objective

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

$$I^{\text{adv}} = R(S^{\text{adv}}; P, L)$$

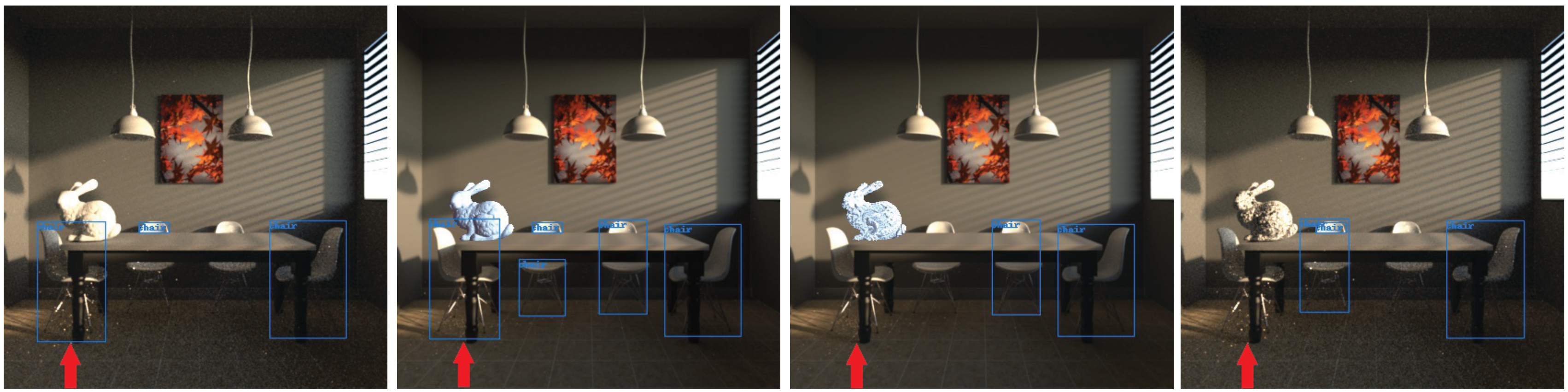




Adversarial Goal: Misclassification

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: Misdetection



Before Attack
Mitsuba Renderer

Before Attack
Neural Mesh Renderer

After Attack
Neural Mesh Renderer

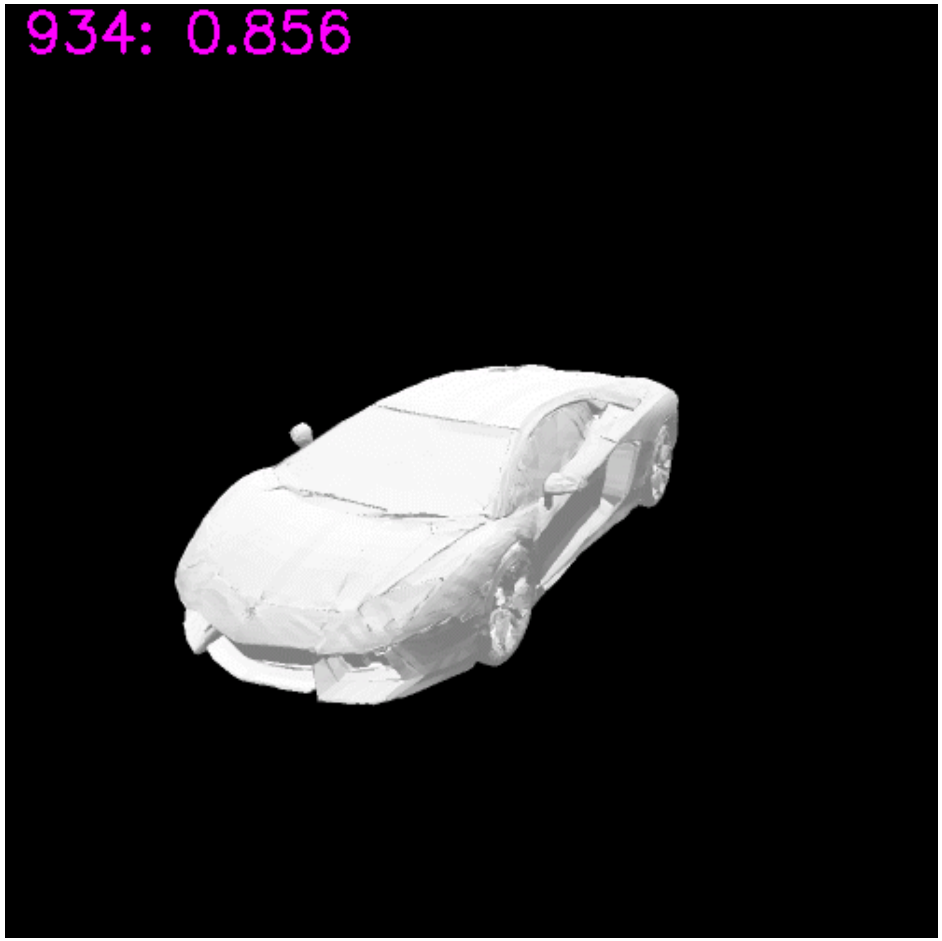
After Attack
Mitsuba Renderer

Search lighting and poses

White-box attack

Black-box transfer

Adversarial 3D Meshes



- 934 : hot dog

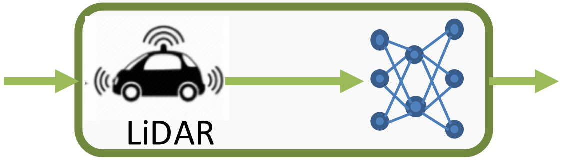
[CVPR 2019]

LiDAR-based perception

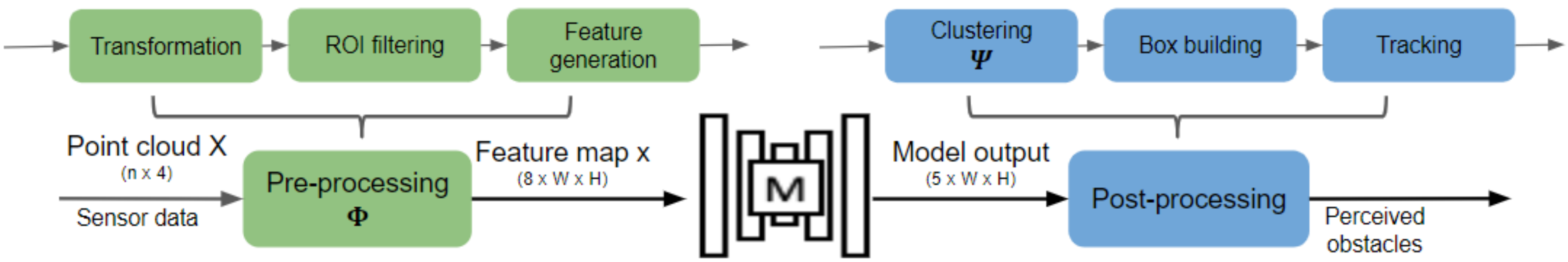
Goal: we aim to generate physical **adversarial object** against **real-world LiDAR system**.

Challenges

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

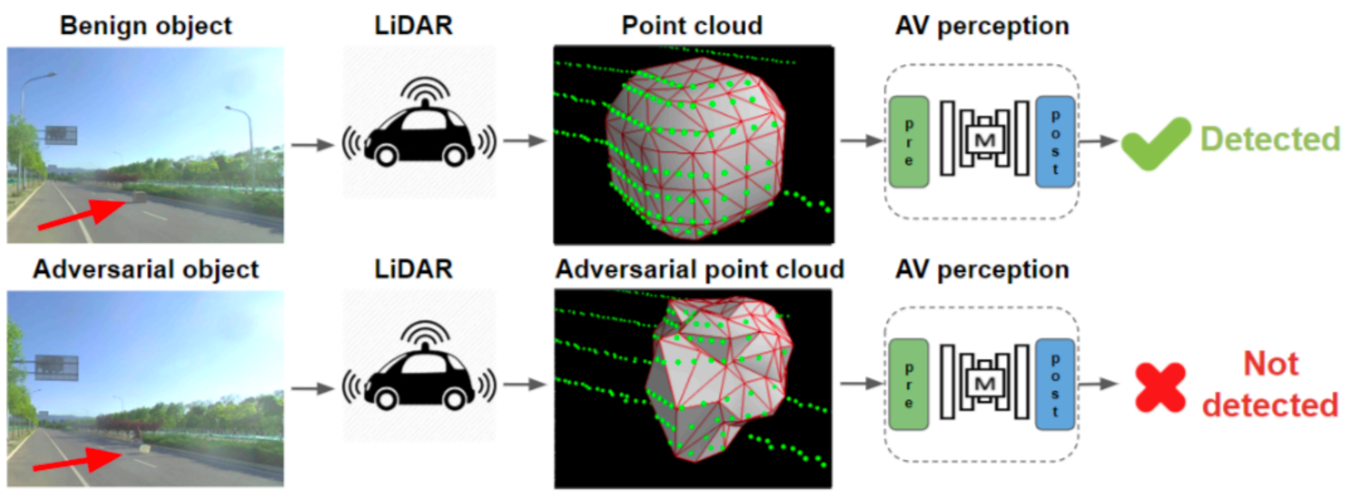


- Manipulation constraints
  - Limited by LiDAR
  - Keeping the shape plausible and smooth adds additional constraints
- Limited Manipulation Space
  - Consider the practical size of the object versus the size of the scene that is processed by LiDAR, the 3D manipulation space is rather small (< 2% in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

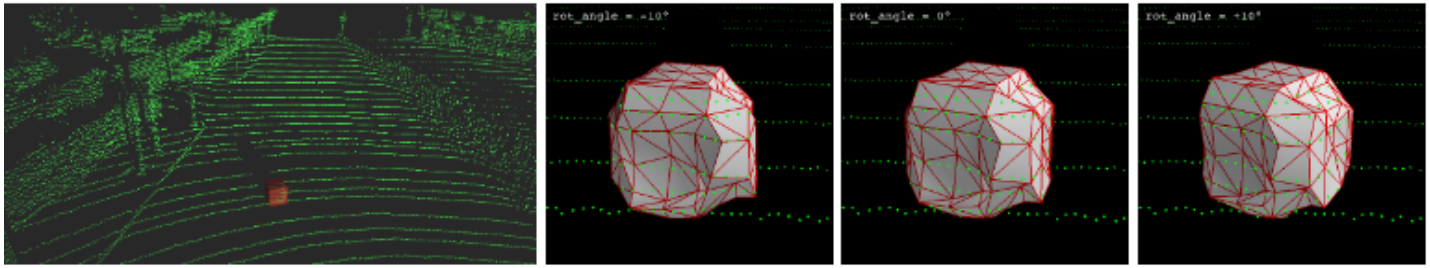


Pipeline of *LiDAR-adv*

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



Robust Adversarial Objects Under Different Viewpoints



(a) Benign frame (b) Adv (-10°) (c) Adv (0°) (d) Adv (10°)

The visualization of adversarial object with different angles.

Angle		-10°	-5°	0°	5°	10°
Objectness (Confid.)	Model	✓	✓	✓	✓	✓
	Apollo	✓	✓	✓	✓	✓

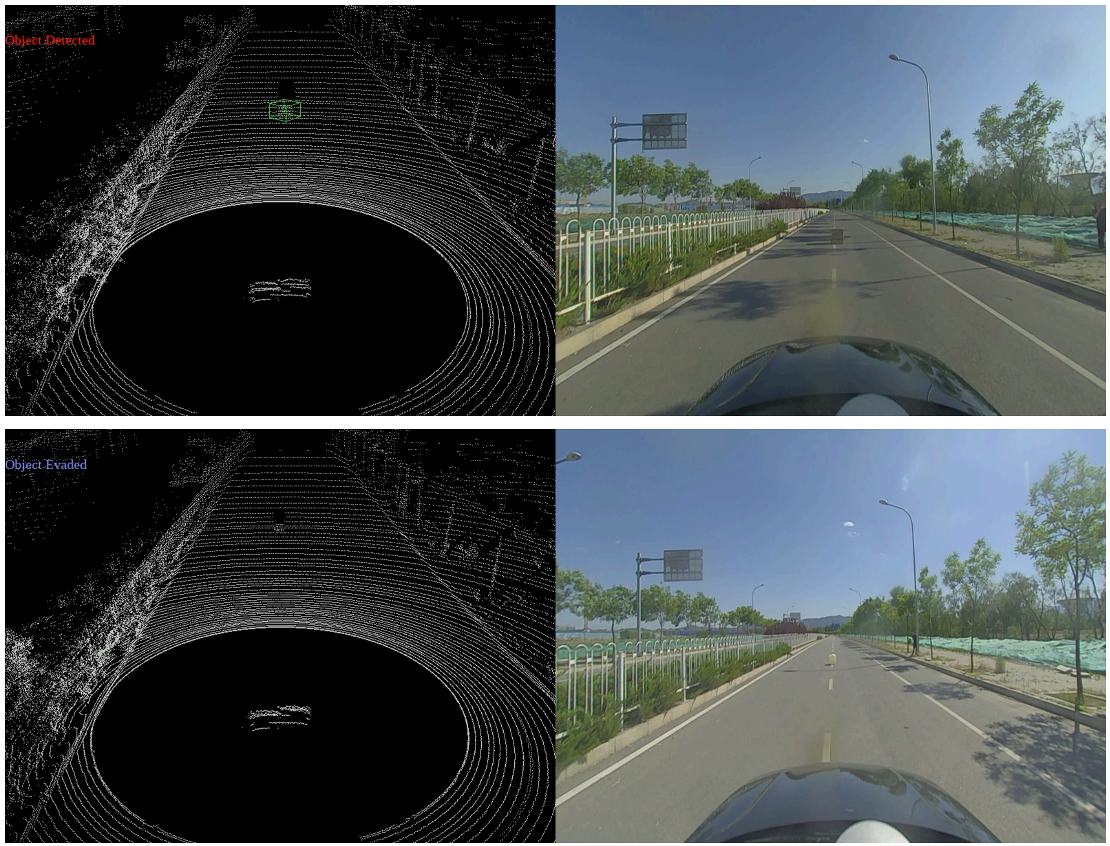
Robust Adversarial Object against different angles. The original confidence is x. Our success rate is 100%. (□represents no object detected)

Physical Experiments

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

Benign Object

Adversarial Object

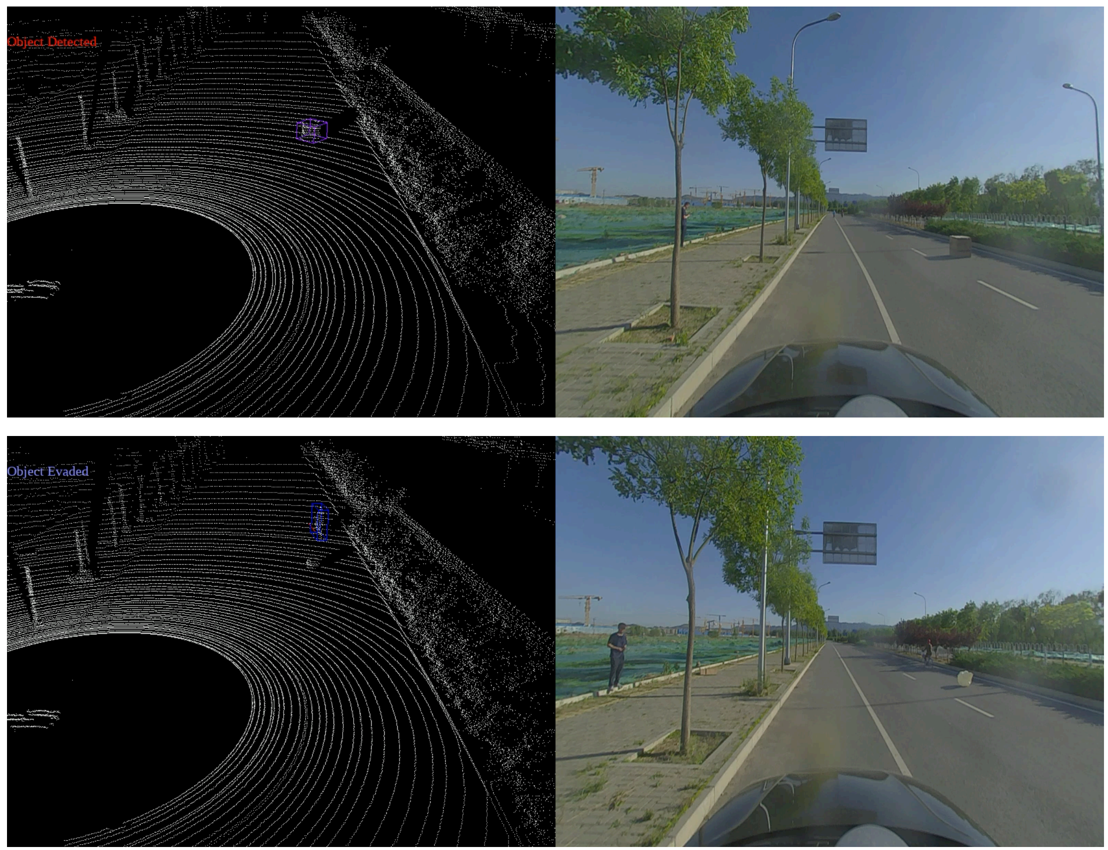


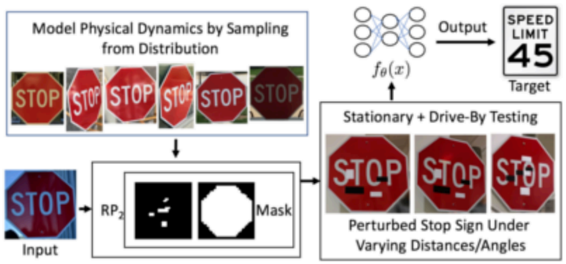
Physical Experiments

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

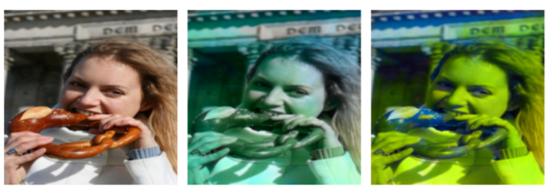
Benign Object

Adversarial Object

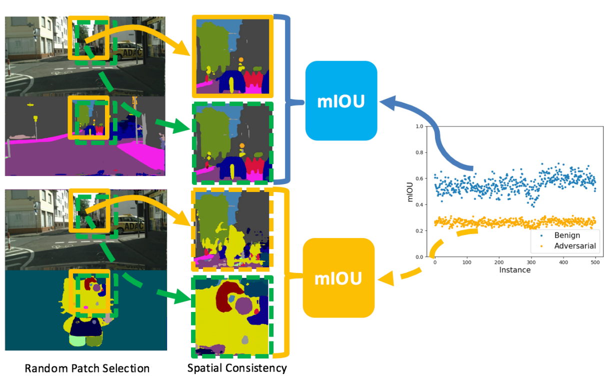




Real world attacks against **different sensors**

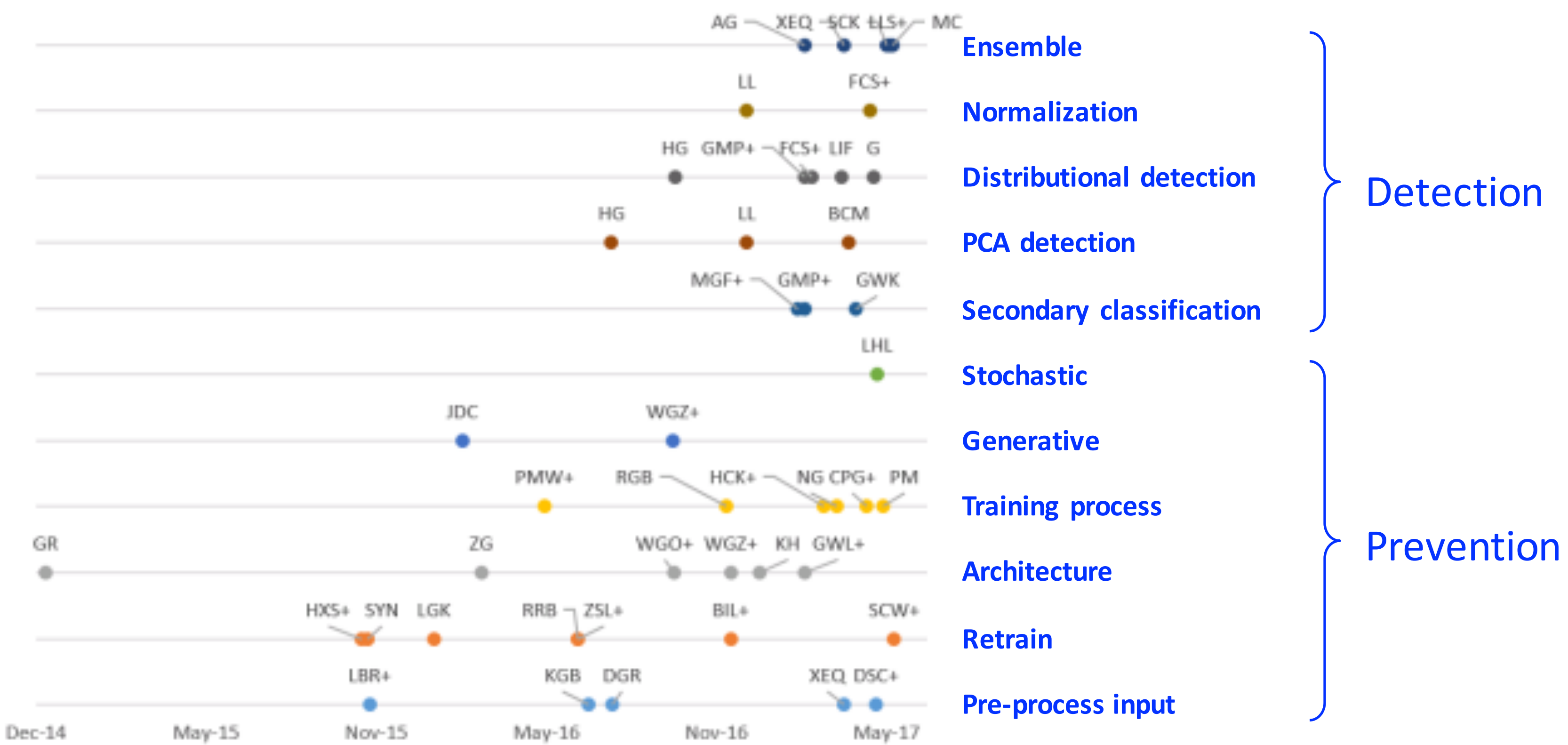


Potential **defenses** against adversarial behaviors via game theory



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

Numerous Defenses Proposed



Example of Evasion: Spam Filter V1.0

1. From: spammer@example.com
   Cheap mortgage now!!!

Feature Weights

2. cheap = 1.0
   mortgage = 1.5

3. Total score = 2.5 > 1.0 (threshold)

Spam

Example of Evasion: Spammer V1.0

1. From: spammer@example.com
Cheap mortgage now!!!
Joy Oregon

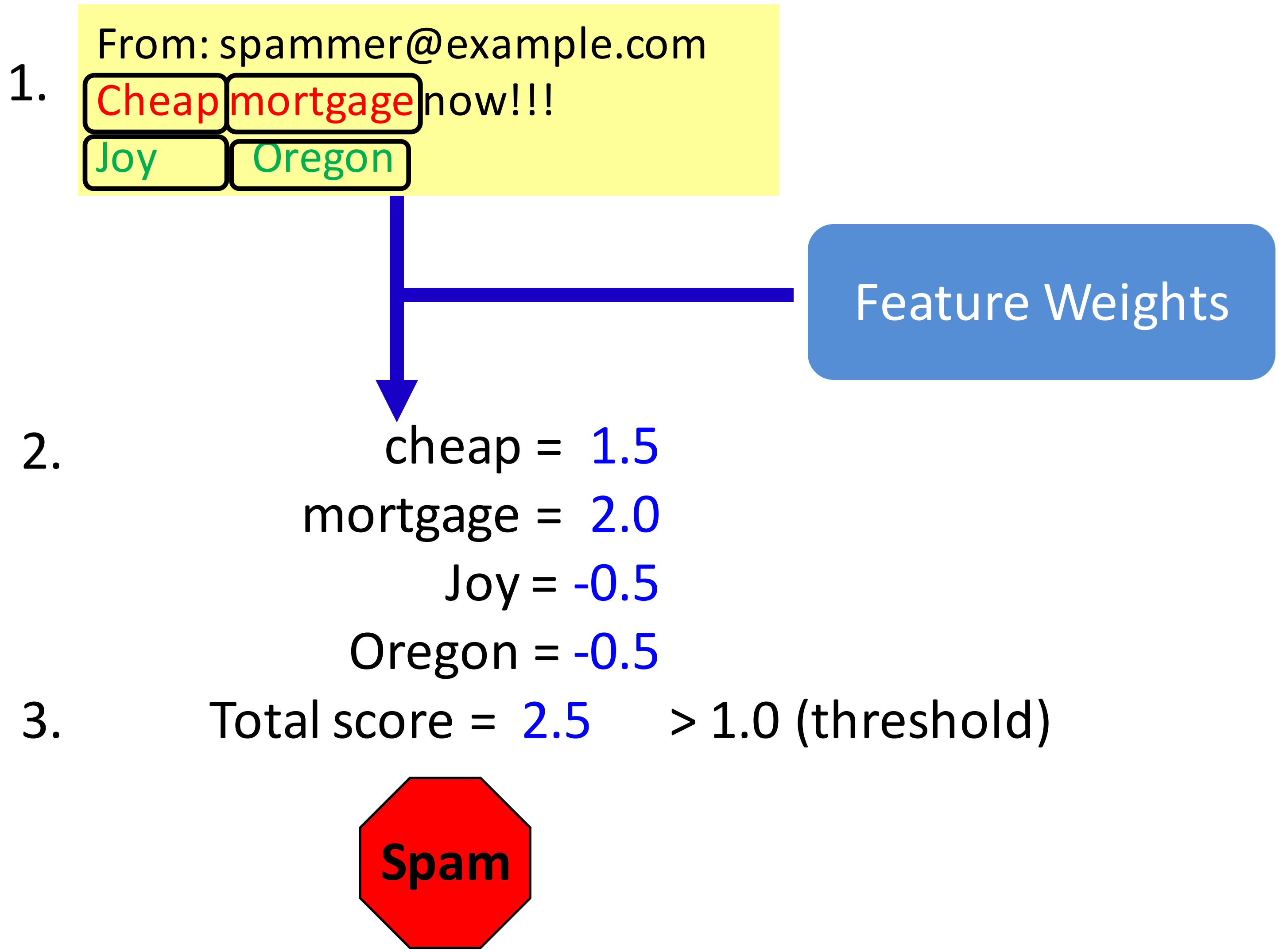
Feature Weights

2. cheap = 1.0
mortgage = 1.5
Joy= -1.0
Oregon = -1.0

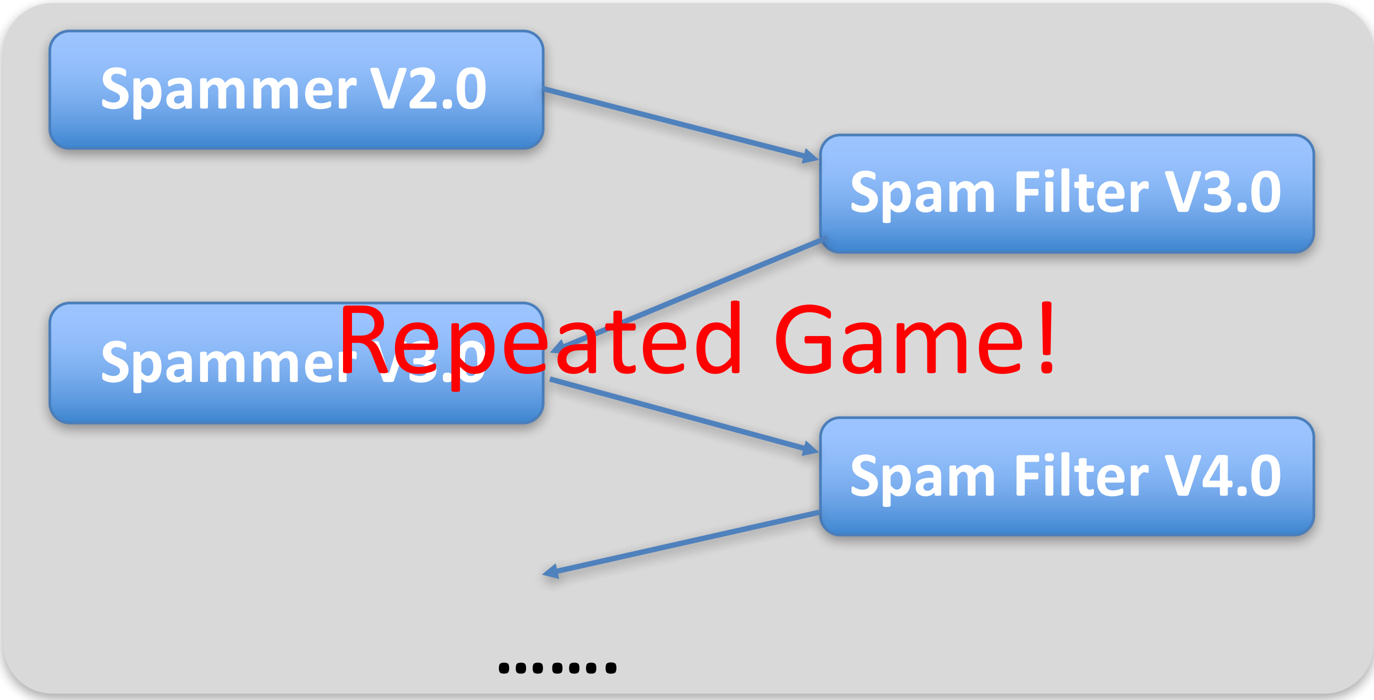
3. Total score = 0.5 < 1.0 (threshold)

OK

Example of Evasion: Spam Filter V2.0 (Retraining)



Challenge



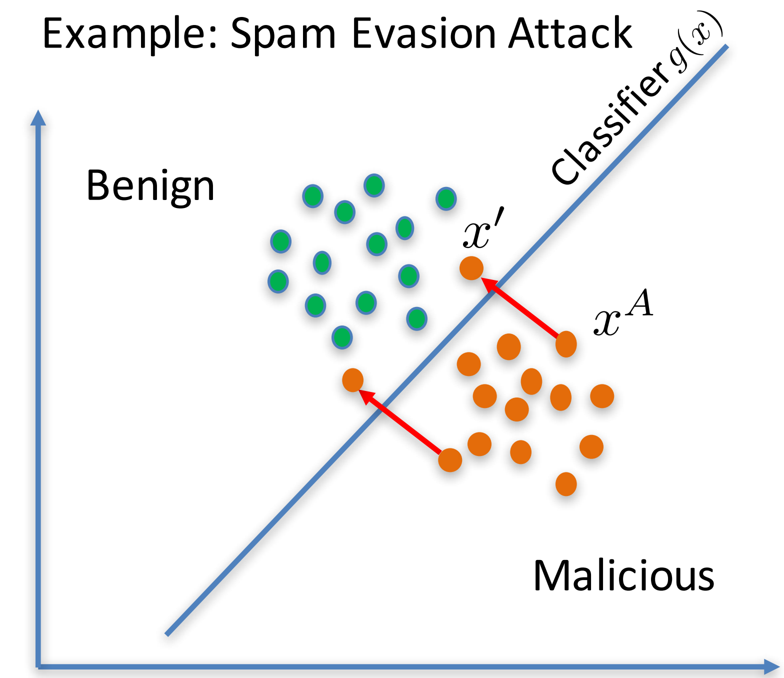
How to efficiently solve the game?

Stackelberg Game

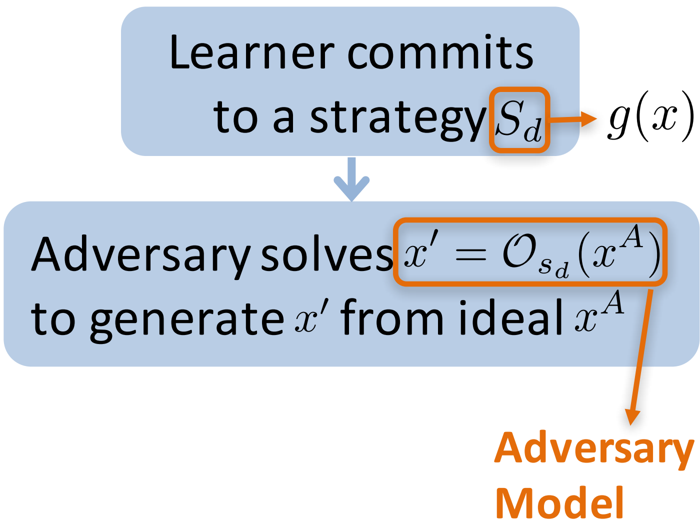
Stackelberg Game

- Learner: commits strategy S_d
- Adversary: best response based on S_d

Example: Spam Evasion Attack



x^A : adversarial instance



[NIPS 2014]

Defending Evasions via Stackelberg Game

Learner commits to a strategy S_d

Adversary solves $\boxed{x' = \mathcal{O}_{s_d}(x^A)}$ to generate x' from ideal x^A

Modeling Adversaries

Human Subject Experiments

$$S'_d = \min L(S_d, x' = \mathcal{O}_{s_d}(x^A))$$

Idea: model the adversary's behavior

- Adversary cannot find additional manipulations
- Adversary incur too high manipulation cost

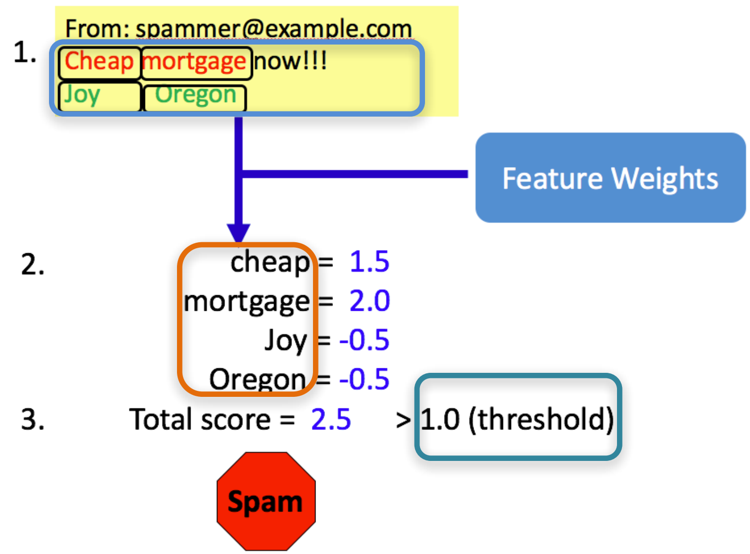
Modeling Evasion Attacks

- Adversary modifies x^A into instance x'
  - Cost $c(x', x^A)$
- Evasion attack:
  - $\boxed{\min_{x'} c(x', x^A)}$ s.t.: $\boxed{f(x')} \le \boxed{\delta}$

Cost Function

Feature Selection

Dynamic Operational Decision



Better Cost Functions → Better Performance

- Model the adversary's cost function
  - Traditional: Distance based cost function

$$c(x', x^A) = \sum_i a_i |x'_i - x^A_i|$$

Distance Based Cost Function Underestimates Adversary

Synonym

Spam

Ham

C = 1

Hello!
Are you ready to become more active and attractive than ever before?
Our final product for losing weight is on clearance now.
Follow the link and you will find he cheapest way to gain your body back.
http://www.ebay.com/application_form

Hello!
Are you happy to become more active and attractive than ever before?
Our final merchandises for losing weigbt is on claerance now.
Follow the link and you will find he deapest way to gain your body back.
http://www.ebay.com/application_form

Letter substitution

A Better Cost Function

- Model the adversarial cost function
  - Traditional: Distance based cost function

$$c(x', x^A) = \sum_i a_i |x'_i - x^A_i|$$

  - Equivalence based cost function

$$c(x', x^A) = \sum_i \min_{j \in \boxed{F_i} | x^A_j \oplus x'_j = 1} a_i |x'_j - x^A_i|$$

Feature Class

Semantic Based Distances

- Colorization and texture for images

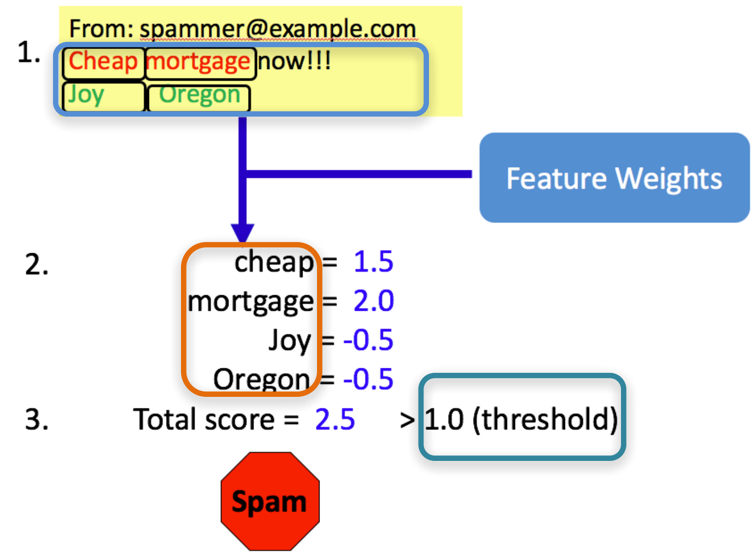
Modeling Evasion Attacks

- Adversary modifies x^A into instance x'
  - Modification cost $c(x', x^A)$
- Evasion attack:
  - $\boxed{\min_{x'} c(x', x^A)}$ s.t.: $\boxed{f(x')} \leq \boxed{\delta}$

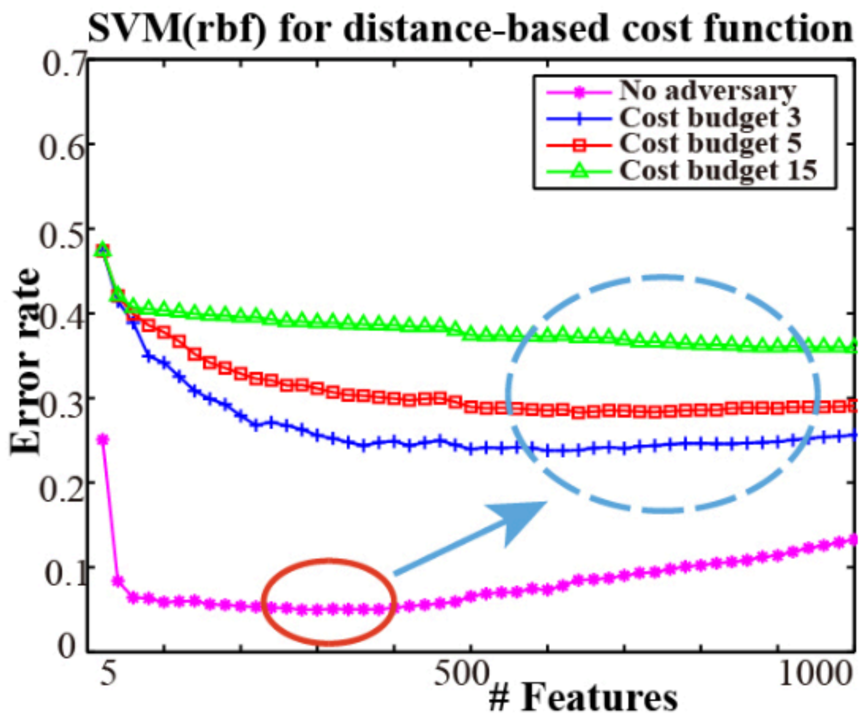
Cost Function

Feature Selection

Dynamic Operational Decision



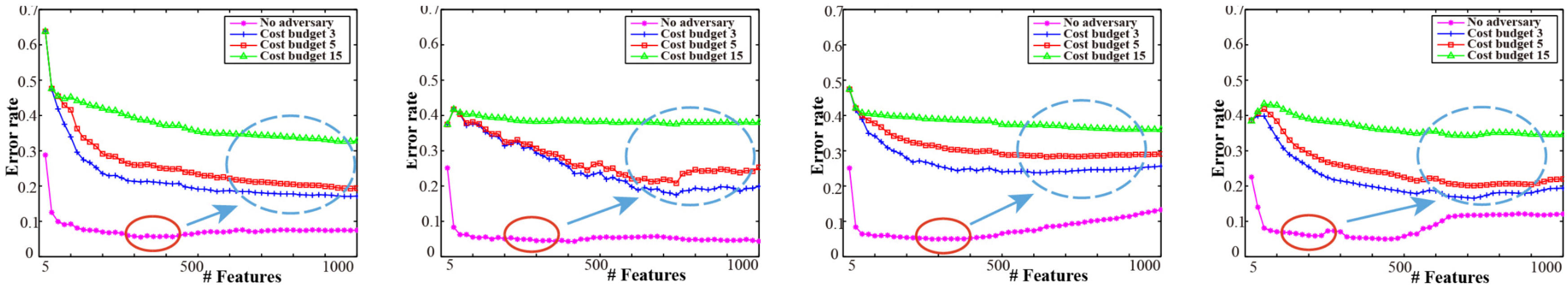
Dangers of Dimension Reduction



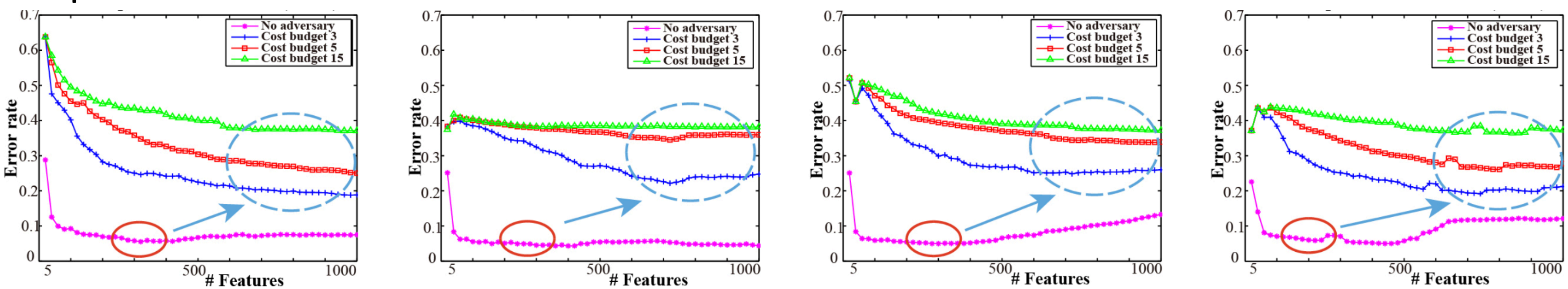
No Adversary: Dimension Reduction = Good
With Adversary: Dimension Reduction = Vulnerable

Vulnerability Across Learning Algorithms

Distance Based Cost Function



Equivalence Based Cost Function



Naïve Bayesian SVM(linear) SVM (rbf) Neural Networks

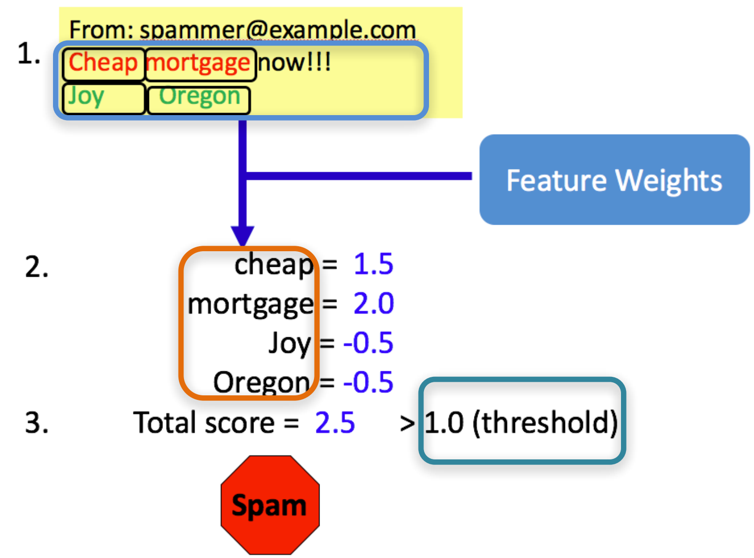
Modeling Evasion Attacks

- Adversary modifies x^A into instance x'
  - Modification cost $c(x', x^A)$
- Evasion attack:
  - $\min_{x'} c(x', x^A)$ s.t.: $f(x') \leq \delta$

Cost Function

Feature Selection

Dynamic Operational Decision



Scaling Optimization

- Adversary has a preferred malicious instance x^A
  - Modifying x^A into instance x' incurs a cost $c(x', x^A)$
- Evasion attack:
  - $\min_{x'} c(x', x^A)$ s.t.: $c(x', x^A) \leq B$, $f(x') \leq \boxed{\delta}$
  - Use $q(x') = Q(x', f(x'))$
  - Scale up: $q(x') = \sum_j \alpha_j \boxed{\phi_j(x')}$

Boolean Basis Functions

[AISTATS 2015]

Adversary’s Best Response is Hard!

- Computing adversary’s best response
  - Theorem 1. Evasion is NP-complete

$$\sum_j \alpha_j \phi_j(x') \leq \lambda$$

s.t.: $\|x - x'\| \leq k$

- Approximation algorithm

The number of inputs in basis is bounded by c.
ApproxEvasion computes a solution x’ which achieves $\hat{\Delta} \geq \frac{\Delta}{1+\varepsilon}$ in $poly(n, \frac{1}{\varepsilon}, 2^c)$

  - Branch and bound
  - Greedy Heuristic
  - Approximation

Defending Evasions via Stackelberg Game

Loss for benign instances

Loss for malicious instances

$$\min_w \alpha \boxed{\sum_{i|y_i=0} l(w, x_i)} + (1-\alpha) \boxed{\sum_{i|y_i=1} l(w, x_i')} + \lambda \boxed{||w||_1}$$

Tradeoff between dimension reduction and robustness

$$s.t.: \quad \forall i : \boxed{y_i = 1},$$

x^A

$$z_i = \arg\min_{x|w^T x \leq 0} c(x, x_i),$$

$$x_i' = \begin{cases} z_i & c(z_i, x_i) \leq B \\ x_i & otherwise \end{cases}$$

Adversarial Strategies

Mixed Integer Linear Programming (MILP)

$$\min_{\omega, z, r} \alpha \sum_{i|y_i=0} D_i + (1-\alpha) \sum_{i|y_i=1} S_i + \lambda \sum_j K_j$$

$$\begin{aligned}
s.t.:\quad & \forall i, j : z_i(a), r(a) \in \{0,1\} \\
& \sum_a z_i(a) = 1 \\
& e_i = \sum_a m_i(a)(L_{ai}T_a + (1-L_{ai})x_i) \\
& \forall a, i, j : -Mz_i(a) \le m_{ij}(a) \le Mz_i(a) \\
& \omega_j - M(1-z_i(a)) \le m_{ij}(a) \le \omega_j + M(1-z_i(a)) \\
& \sum_j \omega_j T_{aj} \le 2\sum_j T_{aj} y_{aj} \\
& \forall a, j : -Mr_a \le y_{aj} \le Mr_a \\
& \omega_j - M(1-r_a) \le y_{aj} \le \omega_j + M(1-r_a) \\
& D_i = \max(0, 1-\omega^T x_i) \\
& S_i = \max(0, 1+e_i) \\
& K_j = \max(\omega_j, -\omega_j)
\end{aligned}$$

Solve the game: MILP!
Solved?

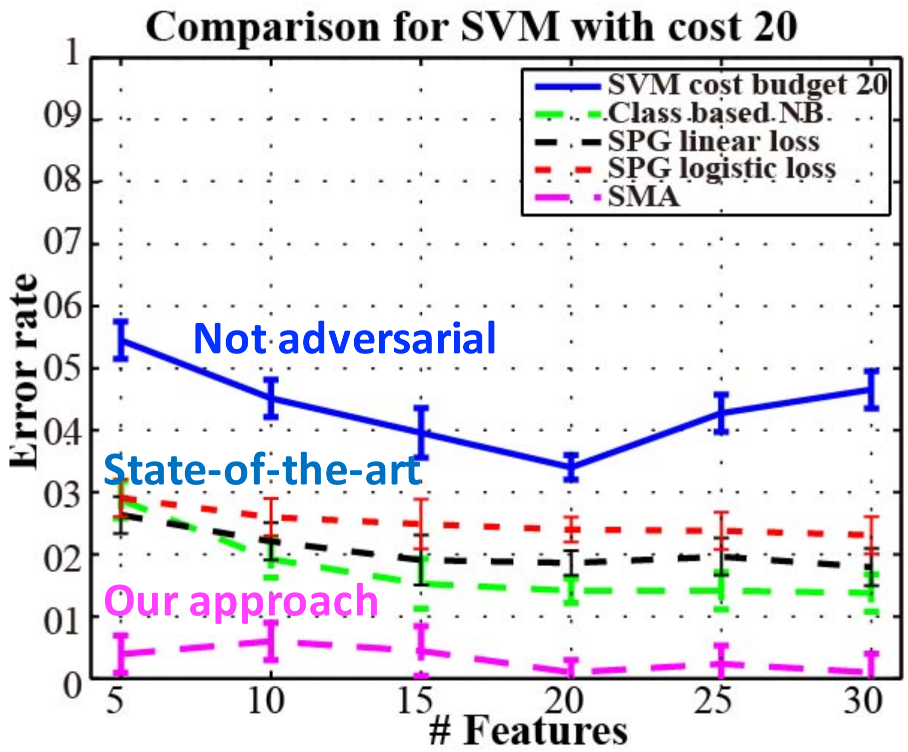
Two reasons for intractability:

- The large number of adversarial objective instances x^A
- Intractable amount of constraints for each attack action x'

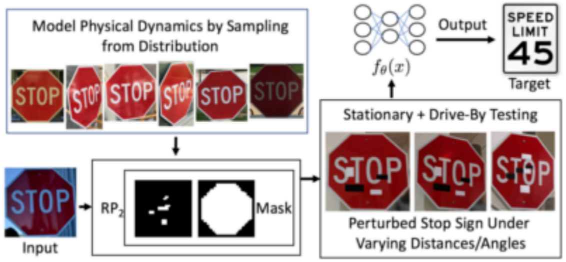
Solutions:

- Clustering attacks: cluster malicious feature vectors in training data
- Constraint generation: iteratively add "best response" attacks into MILP

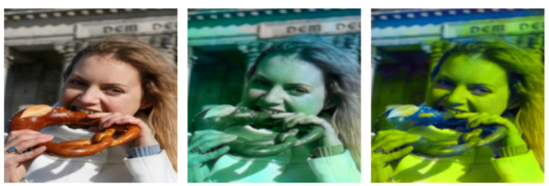
Our Solution (SMA) Outperforms



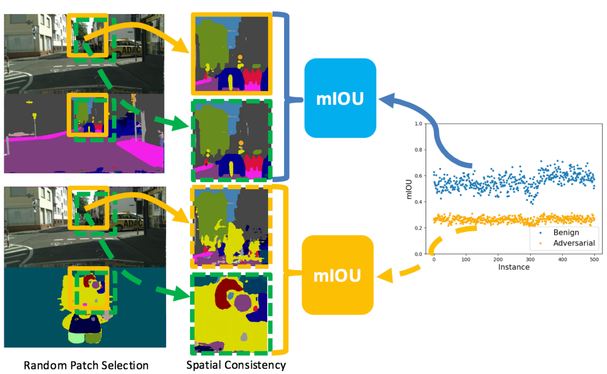
The Stackelberg multi-adversary model (SMA) significantly outperforms in adversarial environments with a range of selected dimensions



Real world attacks against **different sensors**



Potential **defenses** against adversarial behaviors via game theory



Potential **defenses** against adversarial behaviors based on 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

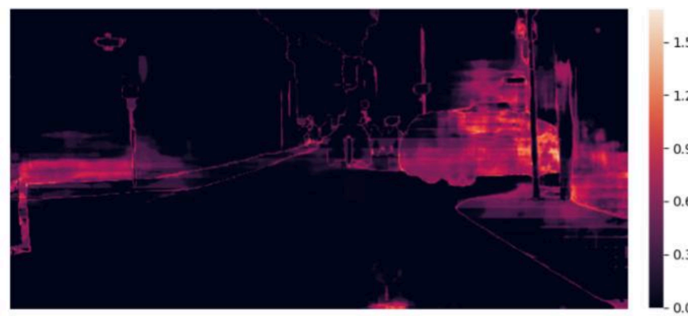
Spatial Context Information

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

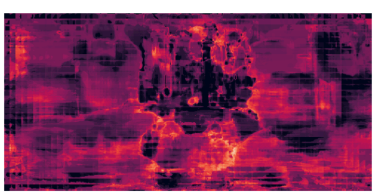
$$\mathcal{H}(m) = -\sum_{j} \mathcal{V}_m[j] \log \mathcal{V}_m[j]$$

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

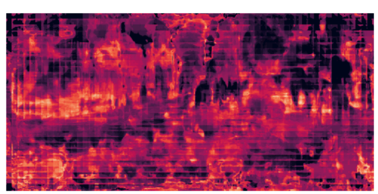
(a) Benign example



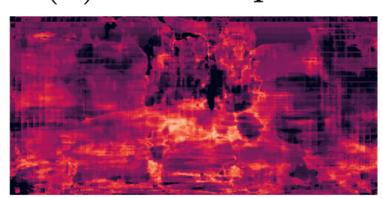
(b) Heatmap of benign image



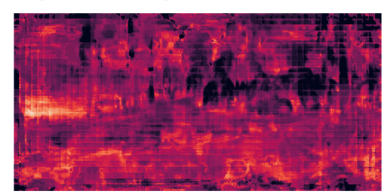
(c) DAG | Kitty



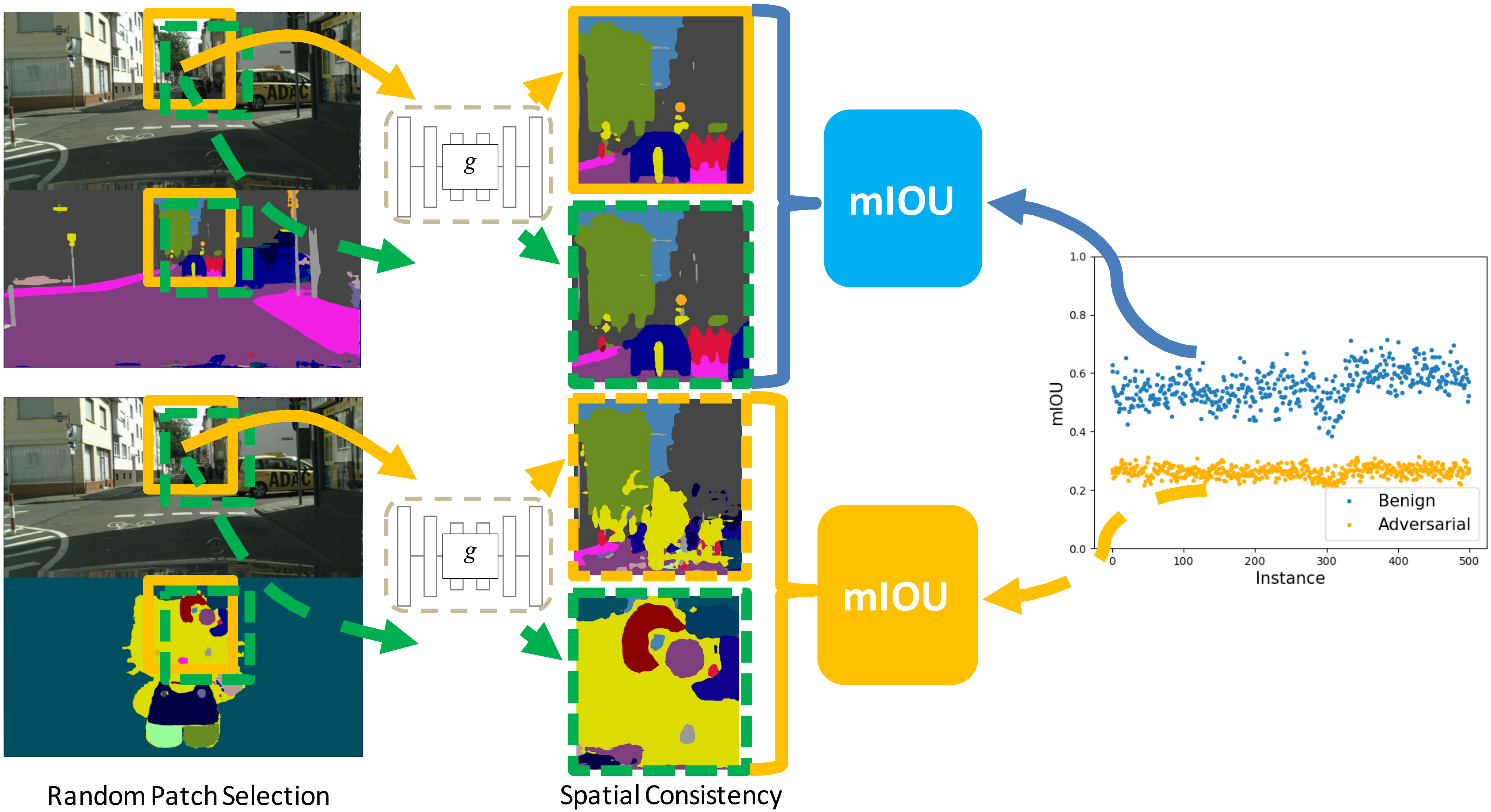
(d) DAG | Pure



(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 DAG Pure	Detection DAG Kitty	Detection Houdini Pure	Detection Houdini Kitty	Detection Adap DAG Pure	Detection Adap DAG Kitty	Detection Adap Houdini Pure	Detection Adap Houdini 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

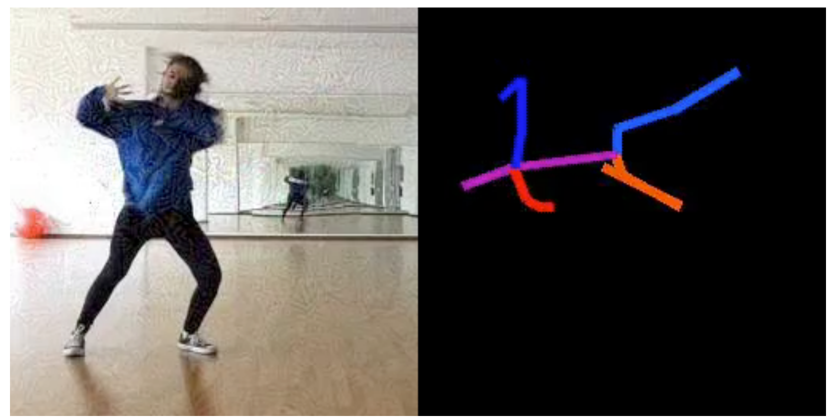
Spatial consistency information can be potentially applied to help distinguish benign and adversarial instances against segmentation models.

Temporal consistency?

Adversarial Frames In Videos

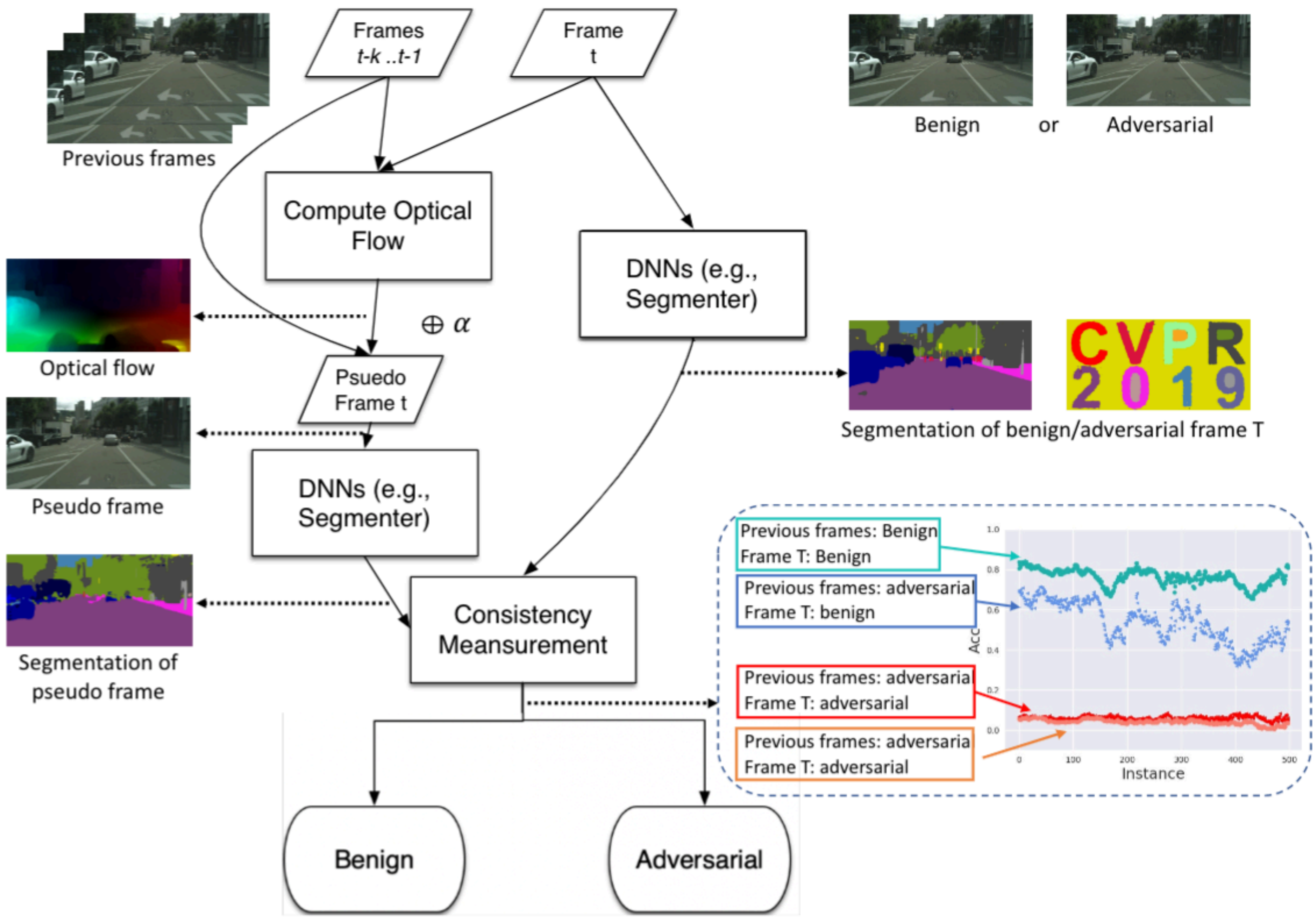
Attacks on segmentation

Attacks on pose estimation



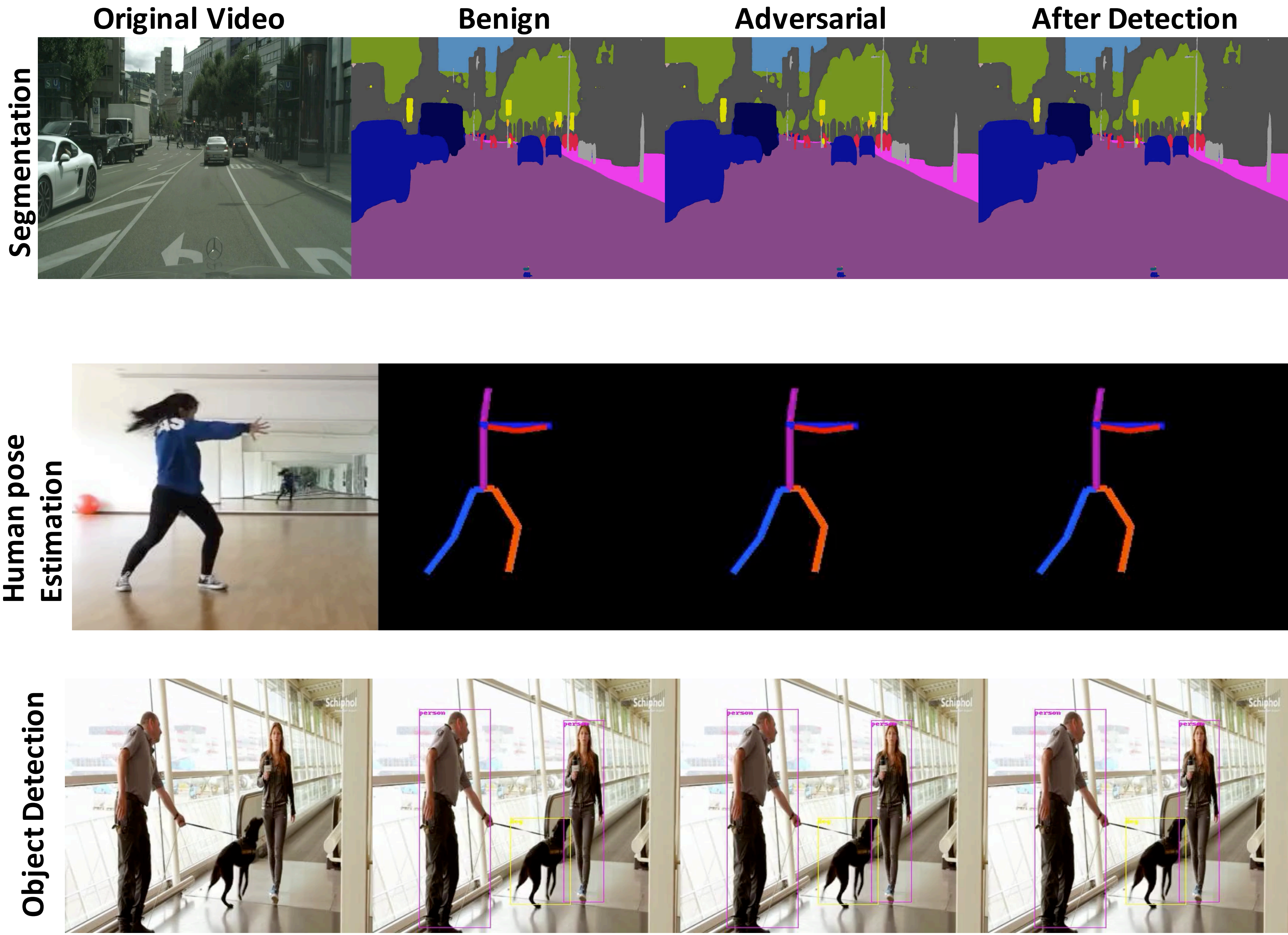
Attacks on object detection

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%

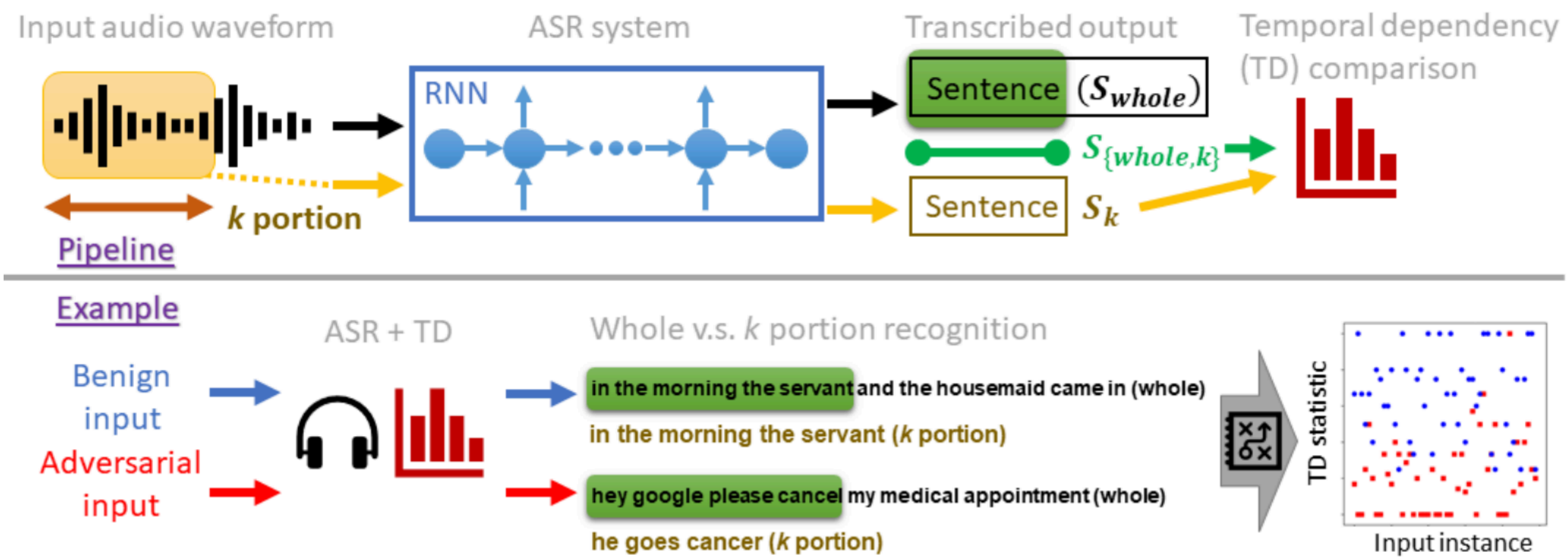
- The results show that choosing more random patches can improve detection rate while k=5 is enough to achieve AUC 100%
- The spatial consistency based detection is robust against strong adaptive attackers due to the randomness in patch selection



Original Video
Benign
Adversarial
After Detection
Segmentation
Human pose Estimation
Object Detection

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

Strong Adaptive Attacks

Segment Attack: Attack only the first k length S_K

Type	Transcribed results
Original	and he leaned against the wa lost in reveriey
the first half of Original	and he leaned against the wa
Adaptive attack target	this is an adversarial example
Adaptive attack result	this is an adversarial losin ver
the first half of Adv.	this is a agamsa
Adaptive attack target	okay google please cancel my medical appointment
Adaptive attack result	okay google please cancel my medcalosinver
the first half of Adv.	okay go please

Concatenate Attack: attack different segments individually and concatenate them

Type	Transcribed results
Original	why one morning there came a quantity of people and set to work in the loft
Attack target	this is an adversarial example
S_k	this is an
S_{k-}	adversarial example
$S_k + S_{k-}$	this is a quantity of people and set to work in a lift
S_k	this is an adversarial example
S_{k-}	*sil*
$S_k + S_{k-}$	this is an adernari eanquatete of pepl and sat to work in the loft

Strong Adaptive Attacks

Combination Attack: attack both individual sections and whole sentence

Combination Attack	Detection Parameter k_D	TD metrics WER	CER	LCP
$k_A = \{\frac{1}{2}\}$	1/2	0.607	0.518	0.643
	2/3	0.957	0.965	0.881
	3/4	0.943	0.951	0.875
	Rand(0.2, 0.8)	0.889	0.882	0.776
$k_A = \{\frac{2}{3}\}$	1/2	0.932	0.912	0.860
	2/3	0.611	0.543	0.604
	3/4	0.956	0.944	0.872
	Rand(0.2, 0.8)	0.879	0.890	0.762
$k_A = \{\frac{1}{2}, \frac{2}{3}\}$	1/2	0.633	0.690	0.552
	2/3	0.536	0.615	0.524
	3/4	0.942	0.974	0.934
	Rand(0.2, 0.8)	0.801	0.880	0.664
$k_A = \{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$	1/2	0.665	0.682	0.604
	2/3	0.653	0.664	0.564
	3/4	0.633	0.653	0.601
	Rand(0.2, 0.8)	0.785	0.832	0.642
$k_A = \{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}\}$	1/2	0.701	0.712	0.615
	2/3	0.684	0.701	0.583
	3/4	0.681	0.693	0.613
	Rand(0.2, 0.8)	0.742	0.811	0.623

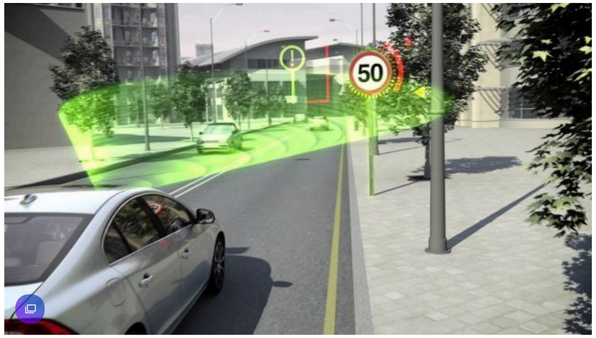
Conclusion: Strong adaptive attack seldom succeeds

YAHOO! NEWS

Search

Researchers demonstrate the limits of driverless car technology

AFP Relax 7 August 2017



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

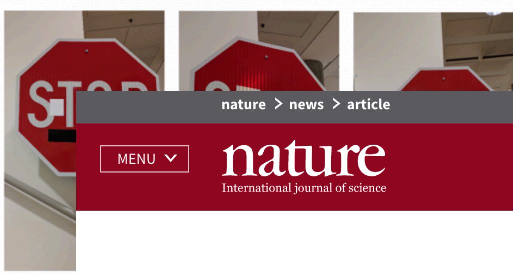
Researchers Show How Simple Stickers Could Trick Self-Driving Cars

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

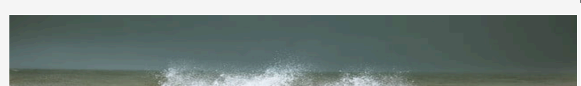
NEWS · 10 MAY 2019

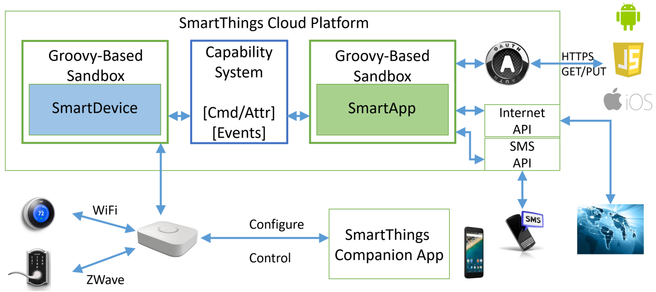
AI can now defend itself against malicious messages hidden in speech

Computer scientists have thwarted programs that can tri malicious audio as safe.

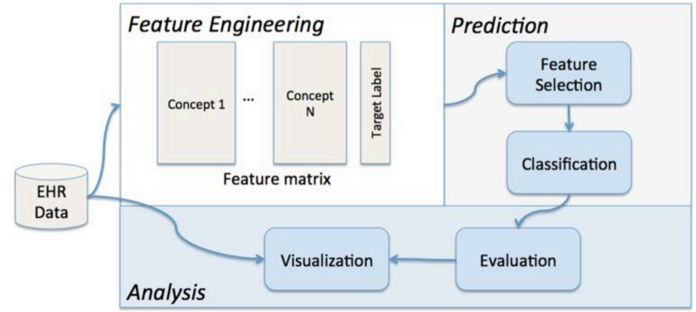
CARS

Stickers on street signs can confuse self-driving cars, researchers show

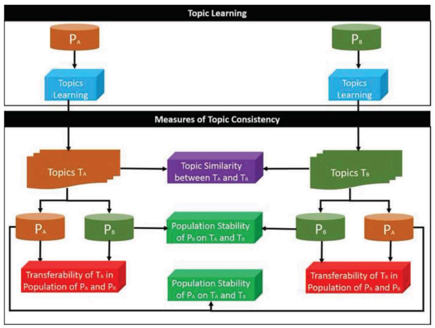




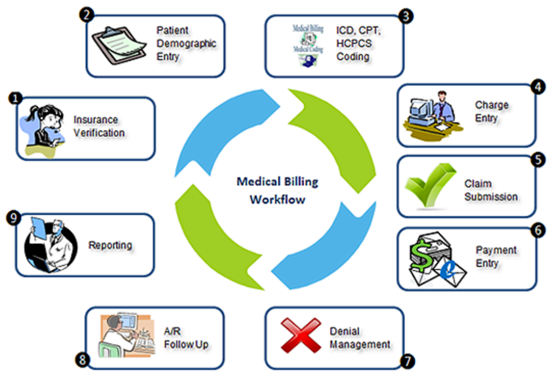
Robust Smart Home



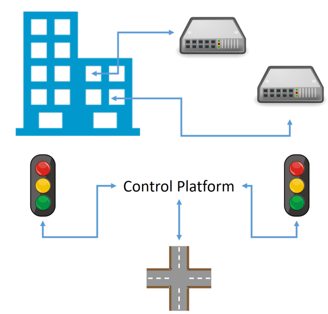
Privacy-Preserving Data Analysis



Topic of Workflow Analysis

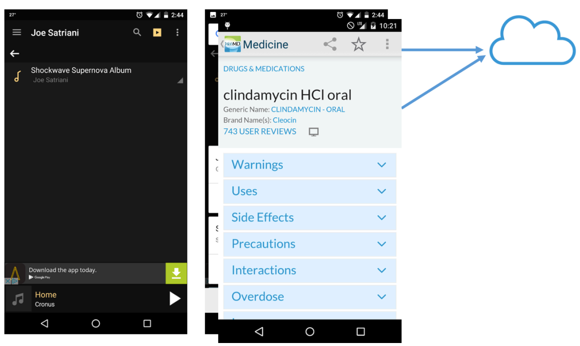


Game Theoretic Auditing System for EMR

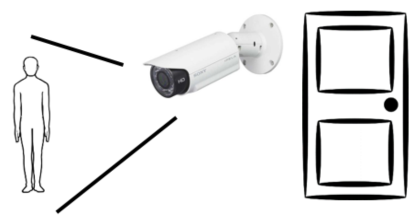


Large-Scale Auditing Game With Human In the Loop

Robust Learning



Privacy Protected Mobile Healthcare



Robust Face Recognition Against Poisoning Attack

Thank You!

Bo Li
lbo@illinois.edu

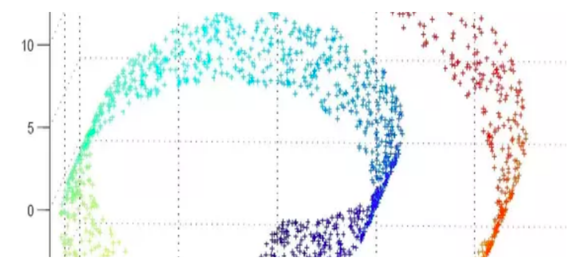
http://boli.illinois.edu/

Beyond the Min-max Game

- What if we have more knowledge about our learning tasks?
  - Properties of learning tasks and data
  - General understanding about ML models

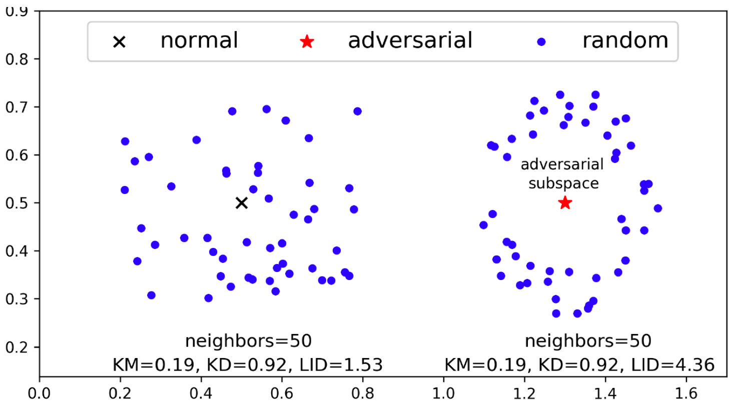
Important Concept: data manifold

- Data Manifold theory:
  - Manifold: the subspace that has local Euclidean space properties
  - The data we observed were actually mapped from a low-dimensional space
  - We use PCA/autoencoders etc. to “unwrap” the manifold
  - We assume the data points from testset and trainset are all from a same manifold
  - Not the case if we consider adversaries



[ICLR 2018]

Previous Measures



- K-means distance
  - Distance to k nearest neighbors
- Kernel density
  - non-parametric
  - estimate the pdf (probability density function) of a random variable

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - x_i}{h})$$

- Can fail to distinguish the sub-manifold that a test case lies in

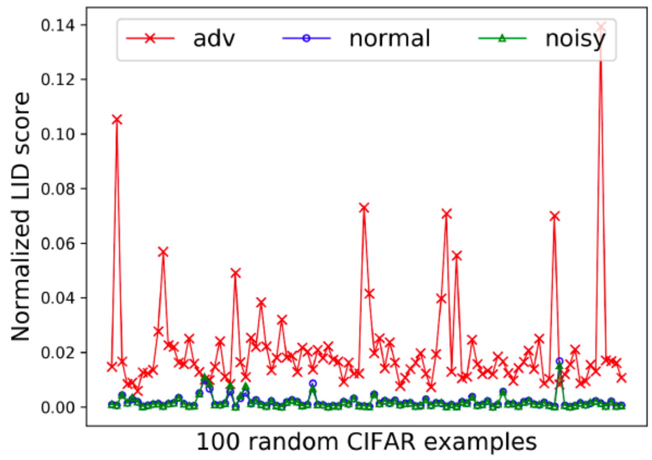
Estimation of Local Intrinsic Dimensionality (LID)

- The sub-manifolds are not parametric
  - given by data points instead
- We use estimation
  - Sample a small set of size larger than k
  - compute their distance to x, take closest k
  - $r_k(x)$ is the maximum of the neighbor distances

$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k}\log\frac{r_i(x)}{r_k(x)}\right)^{-1}$$

Use LID to characterize the sub-manifold

- LID of benign x
  - The dimension of S (the sub-manifold x lies in)
  - Should be small since S is under some intrinsic constraints
- LID of adversarial x':
  - Full degrees of freedom afforded by the representational dimension of the data domain
  - Attacks generally allow modification of all pixels



Characterizing Adversarial Examples

AUC of different detection methods against various attacks

Dataset	Feature	FGM	BIM-a	BIM-b	JSMA	Opt
MNIST	KD	78.12%	99.14%	98.61%	68.77%	95.15%
	BU	32.37%	91.55%	25.46%	88.74%	71.29%
	KD+BU	82.43%	99.20%	98.81%	90.12%	95.35%
	LID	**96.89%**	**99.60%**	**99.83%**	**92.24%**	**99.24%**
CIFAR-10	KD	64.92%	68.38%	98.70%	85.77%	91.35%
	BU	70.53%	81.60%	97.32%	87.36%	91.39%
	KD+BU	70.40%	81.33%	98.90%	88.91%	93.77%
	LID	**82.38%**	**82.51%**	**99.78%**	**95.87%**	**98.93%**
SVHN	KD	70.39%	77.18%	99.57%	86.46%	87.41%
	BU	86.78%	84.07%	86.93%	91.33%	87.13%
	KD+BU	86.86%	83.63%	99.52%	93.19%	90.66%
	LID	**97.61%**	**87.55%**	**99.72%**	**95.07%**	**97.60%**

Attack Failure Rate of Strong Adaptive Attacks Against LID Detector

	MNIST	CIFAR-10	SVHN
Attack Failure Rate (one-layer)	100%	95.7%	97.2%
Attack Failure Rate (all-layer)	100%	100%	100%