



Trustworthy Machine Learning: Robustness, Privacy, Generalization, and Their Interconnections

Bo Li

University of Illinois at Urbana-Champaign

Machine Learning in Physical World



Autonomous Driving



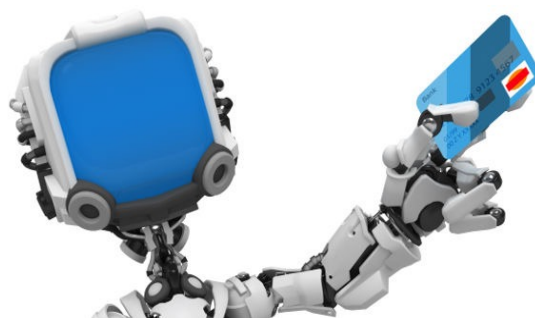
Healthcare



Smart City



Malware Classification



Fraud Detection



Biometrics Recognition

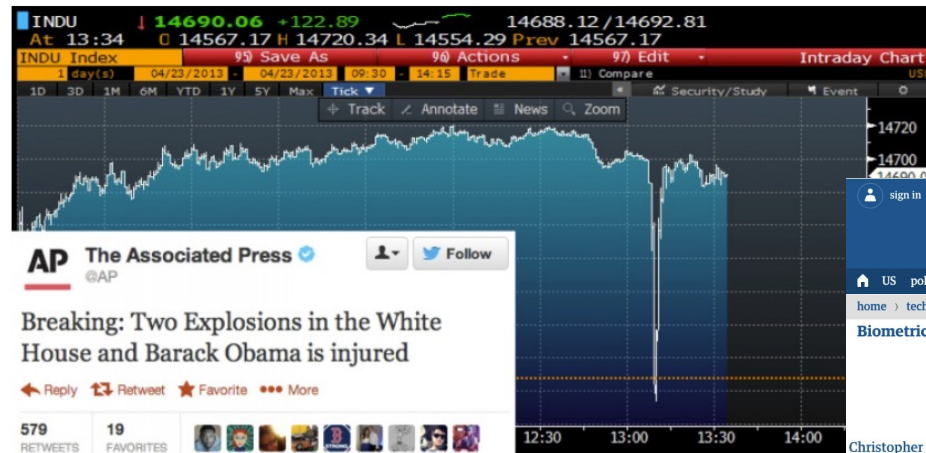
Security & Privacy Problems

Sections  

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

sign in become a supporter subscribe search jobs US edition

the guardian

US politics world opinion sports soccer tech arts lifestyle fashion business travel environment browse all sections

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

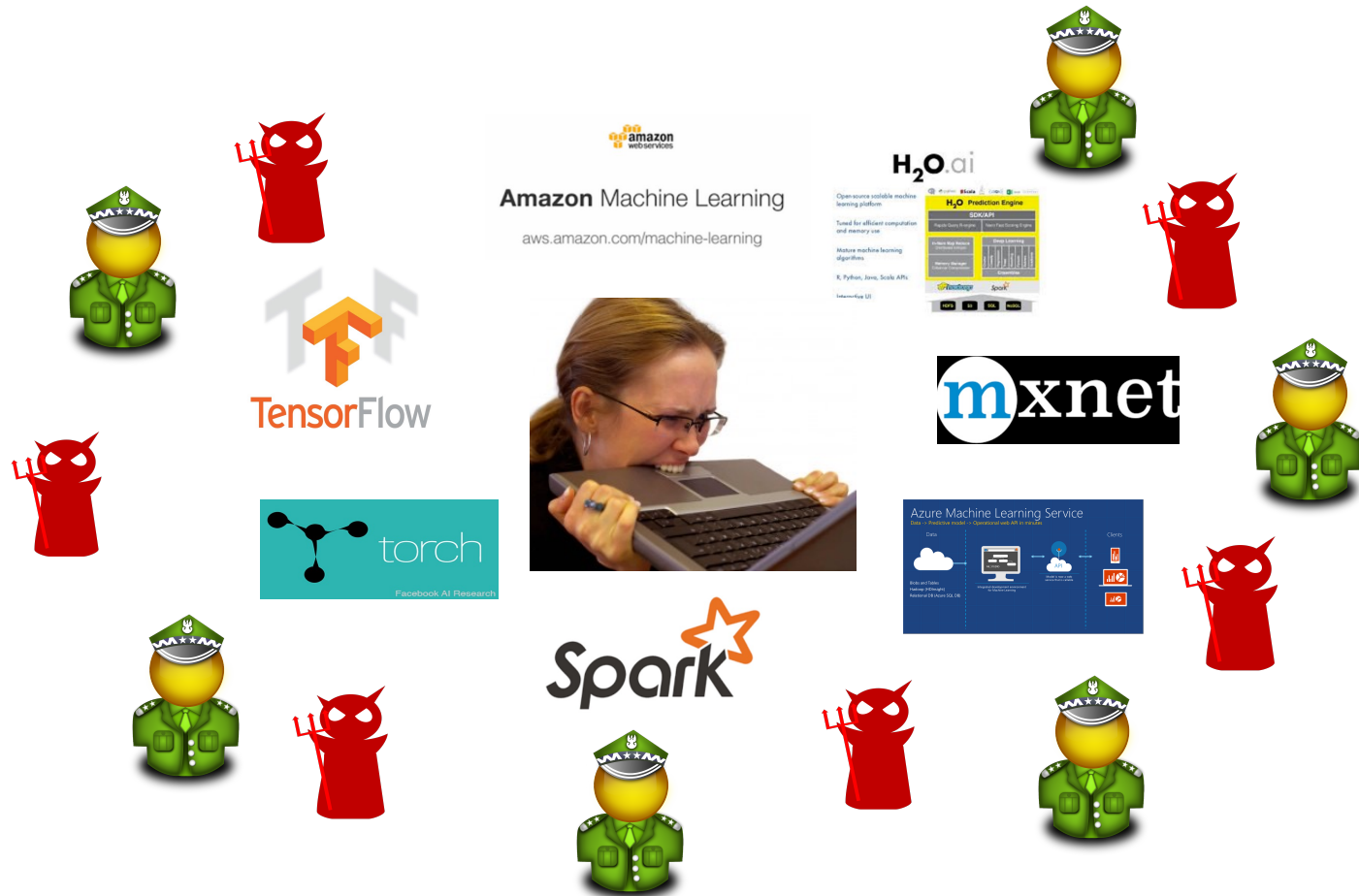
Christopher Knaus
Monday 23 January 2017 21:02 EST

237 146



Privacy Concerns

We Are in Adversarial Environments

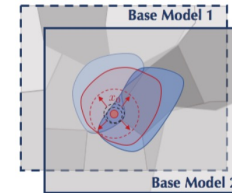


Goal of Secure Learning Lab: Design and certify **robust**, **private**, and **explainable** machine learning paradigms for real-world applications



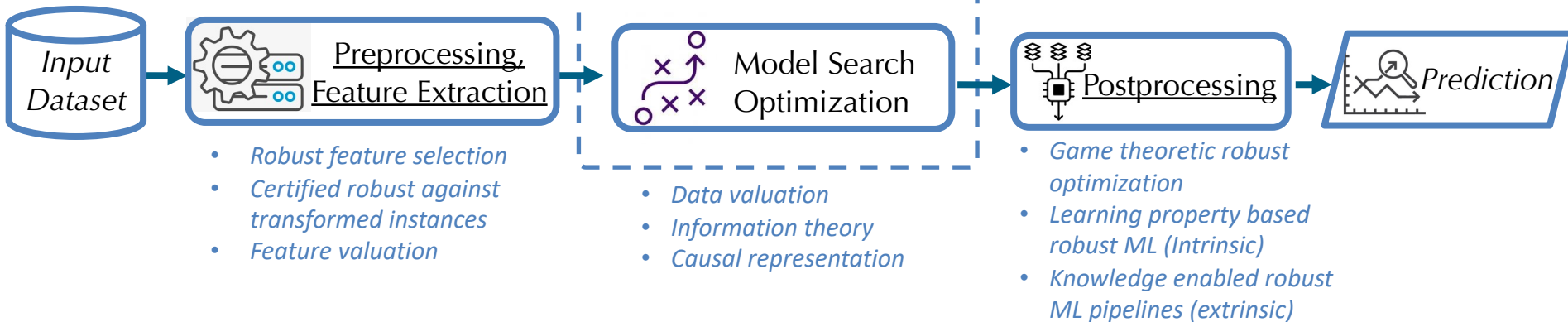
- First robust physical attack
- First spatial attack
- First distributed attack on FL
- ...

Adversarial Attacks

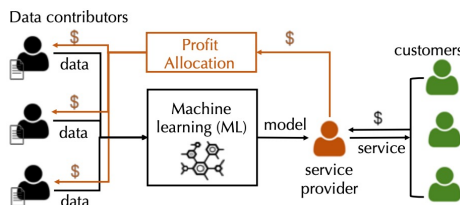


- First certified defense against semantic transformations
- Tight bounds for adv transferability
- Adv audio detection (Watson)
- ...

ML Robustness Enhancement and Certification



ML Explanation, Fairness

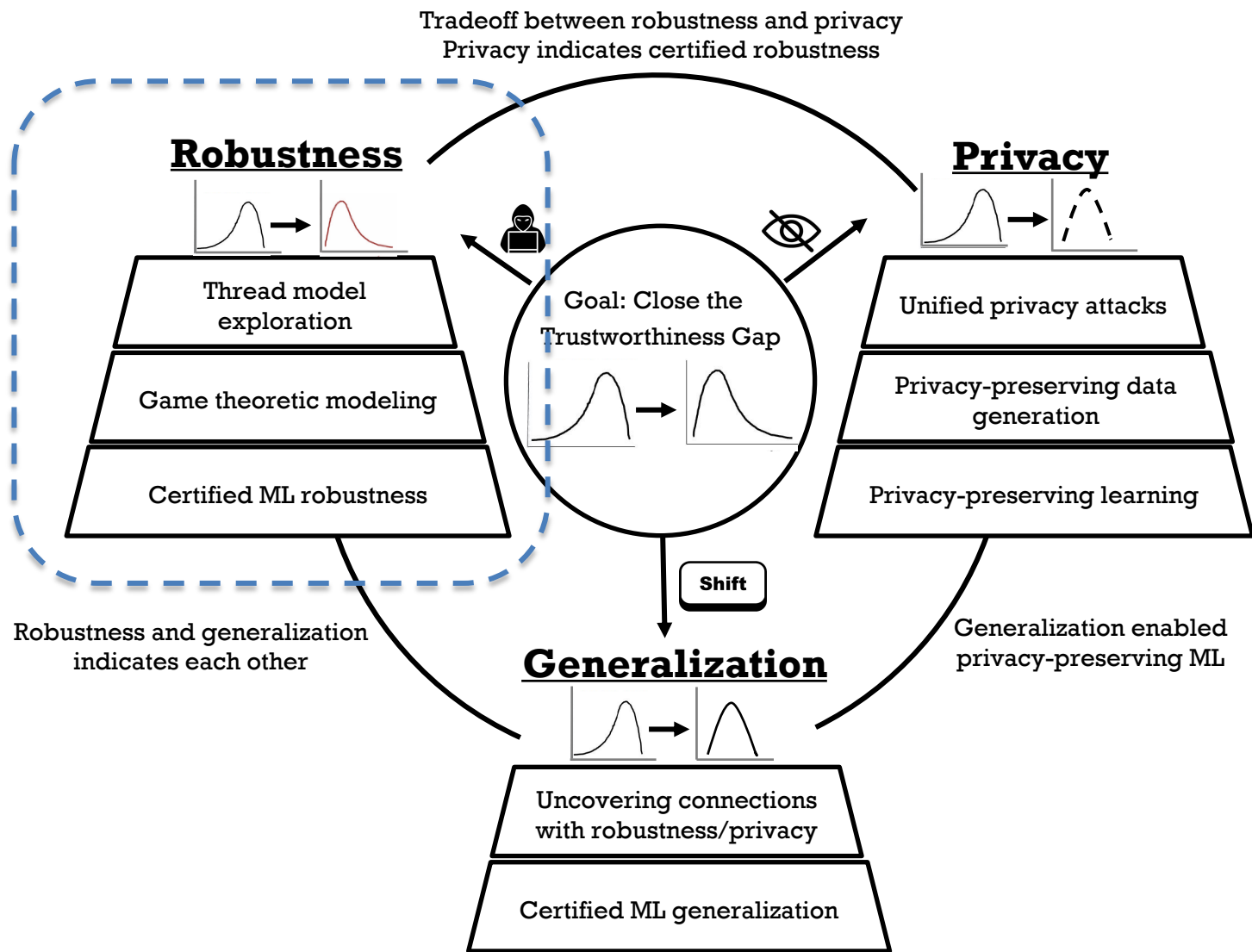


- First $O(n \log n)$ Shapley value
- First fairness on VFL
- First de-toxicity pip. on NLP
- ...

Privacy preserving ML Learning and Data Generation

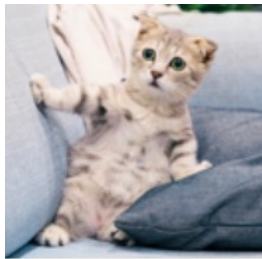


- First model inversion attack with partial info.
- First scalable DP data generative model
- ...



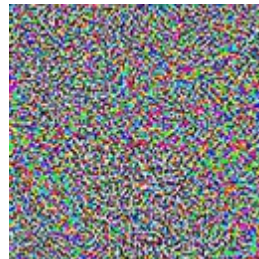
Physical Attacks In Practice

Digital World:



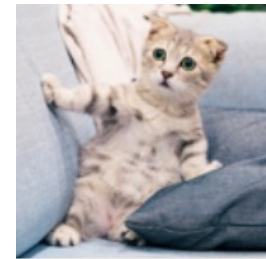
Predicted as
“cat”

+ 0.001 *



Small
Perturbation

=



Predicted as
“dog”

Physical World:



However, What We Can See Everyday...

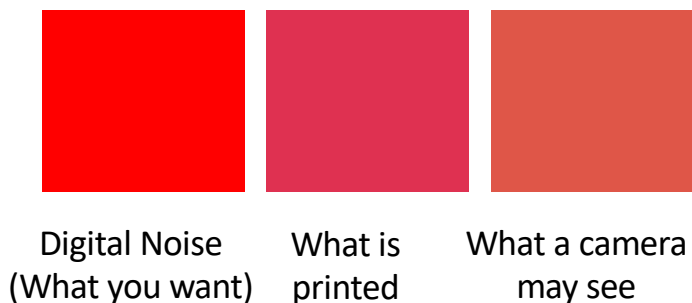


The Physical World Is... Messy

Varying Physical Conditions (Angle, Distance, Lighting, ...) Physical Limits on Imperceptibility



Fabrication/Perception Error (Color Reproduction, etc.)



Background Modifications* Image Courtesy, OpenAI



An Optimization Approach To Creating Robust Physical Adversarial Examples

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

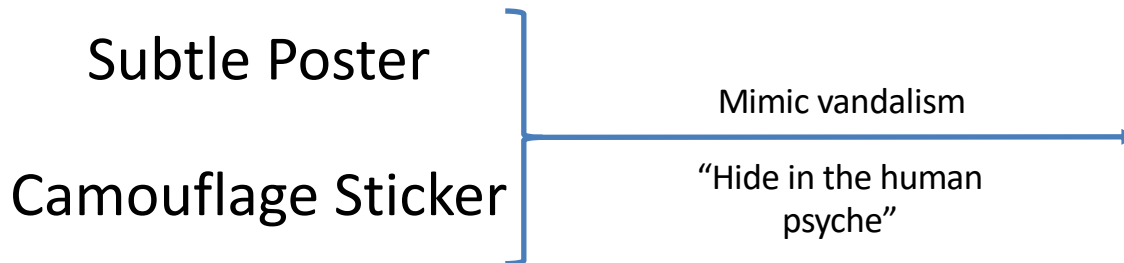
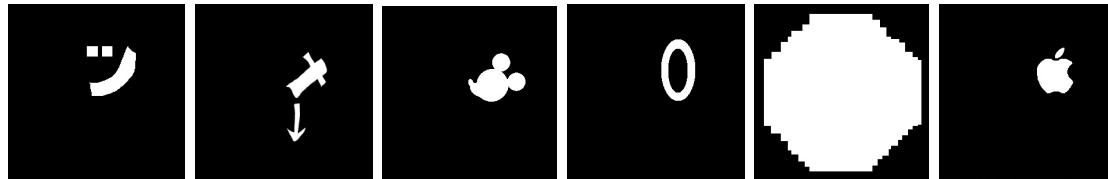
Perturbation/Noise Matrix \rightarrow δ \rightarrow $\|\delta\|_p$ \rightarrow Lp norm (L-0, L-1, L-2, ...) \rightarrow δ \rightarrow $f_{\theta}(x + \delta)$ \rightarrow Loss Function \rightarrow $J(f_{\theta}(x + \delta), y^*)$ \rightarrow Adversarial Target Label y^*

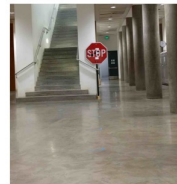
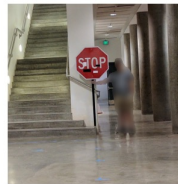
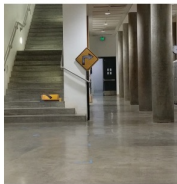
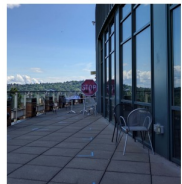
$$\underset{\delta}{\operatorname{argmin}} \lambda \|\delta\|_p + \frac{1}{k} \sum_{i=1}^k J(f_{\theta}(x_i + \delta), y^*)$$



Optimizing Spatial Constraints (Handling Limits on Imperceptibility)

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





Lab Test Summary (Stationary)

Adversarial Target:

Stop Sign -> Speed Limit 45

Right Turn -> Stop Sign

Subtle Poster

Subtle Poster

Camo Graffiti

Camo Art

Camo Art

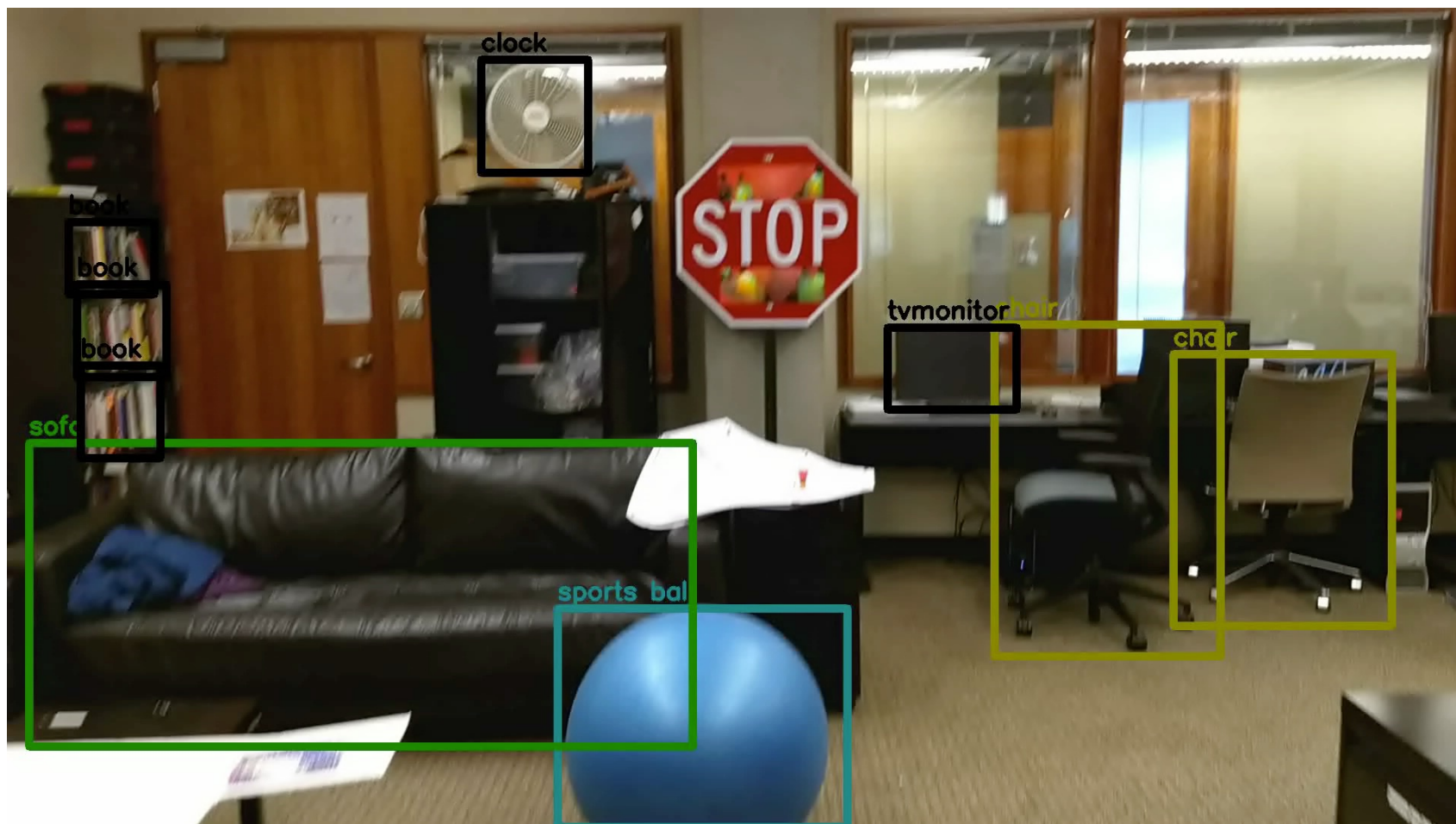
Art Perturbation



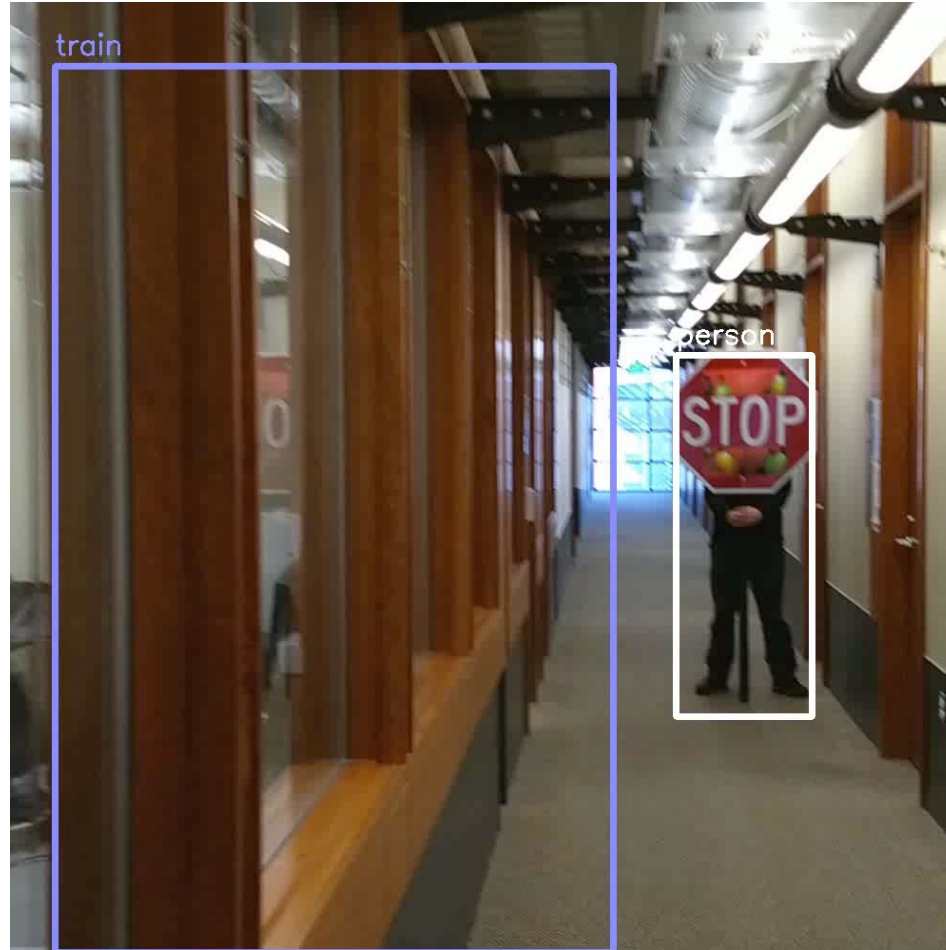
Subtle Perturbation



Physical Attacks Against Detectors



Physical Attacks Against Detectors



Physical Adversarial Stop Sign in the Science Museum of London



Physical Adversarial Attacks Against LiDAR Sensor

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



LiDAR-based perception

Adversarial Point Clouds

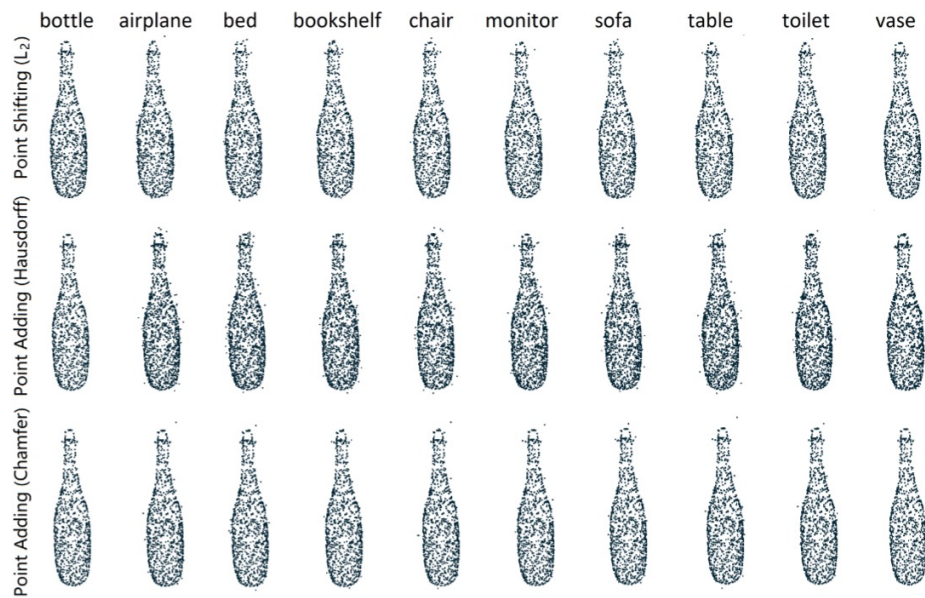
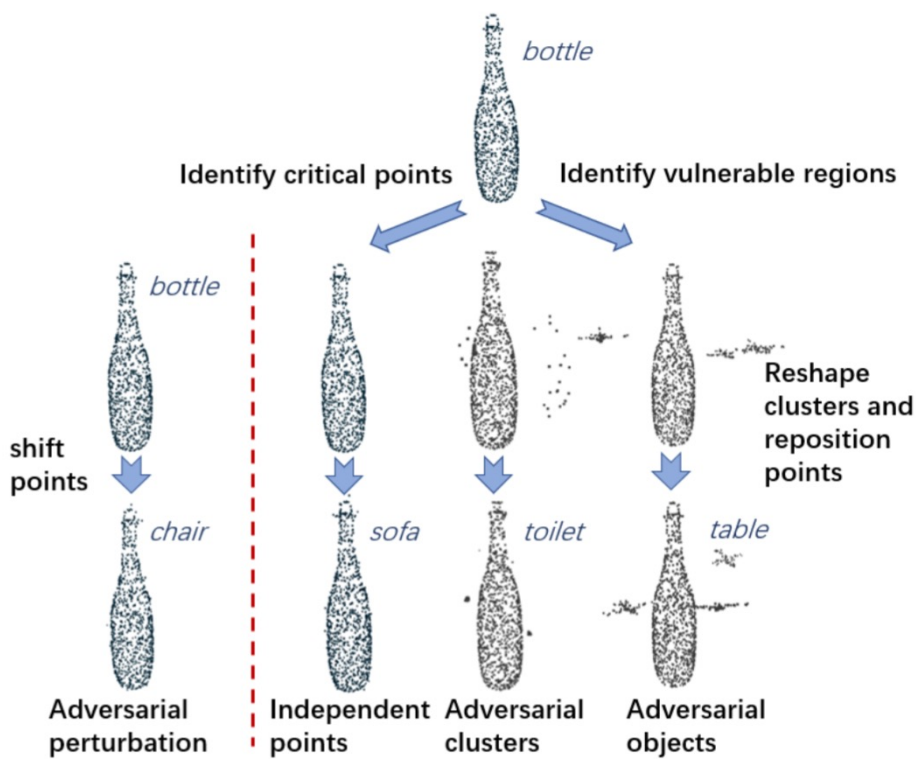
- PointNet is widely used including in autonomous driving systems to process Lidar point cloud data
- Perturbation on point cloud
 - Points shifting
 - Independent points adding
 - Adversarial clusters
 - Adversarial objects
- Adversarial objectives

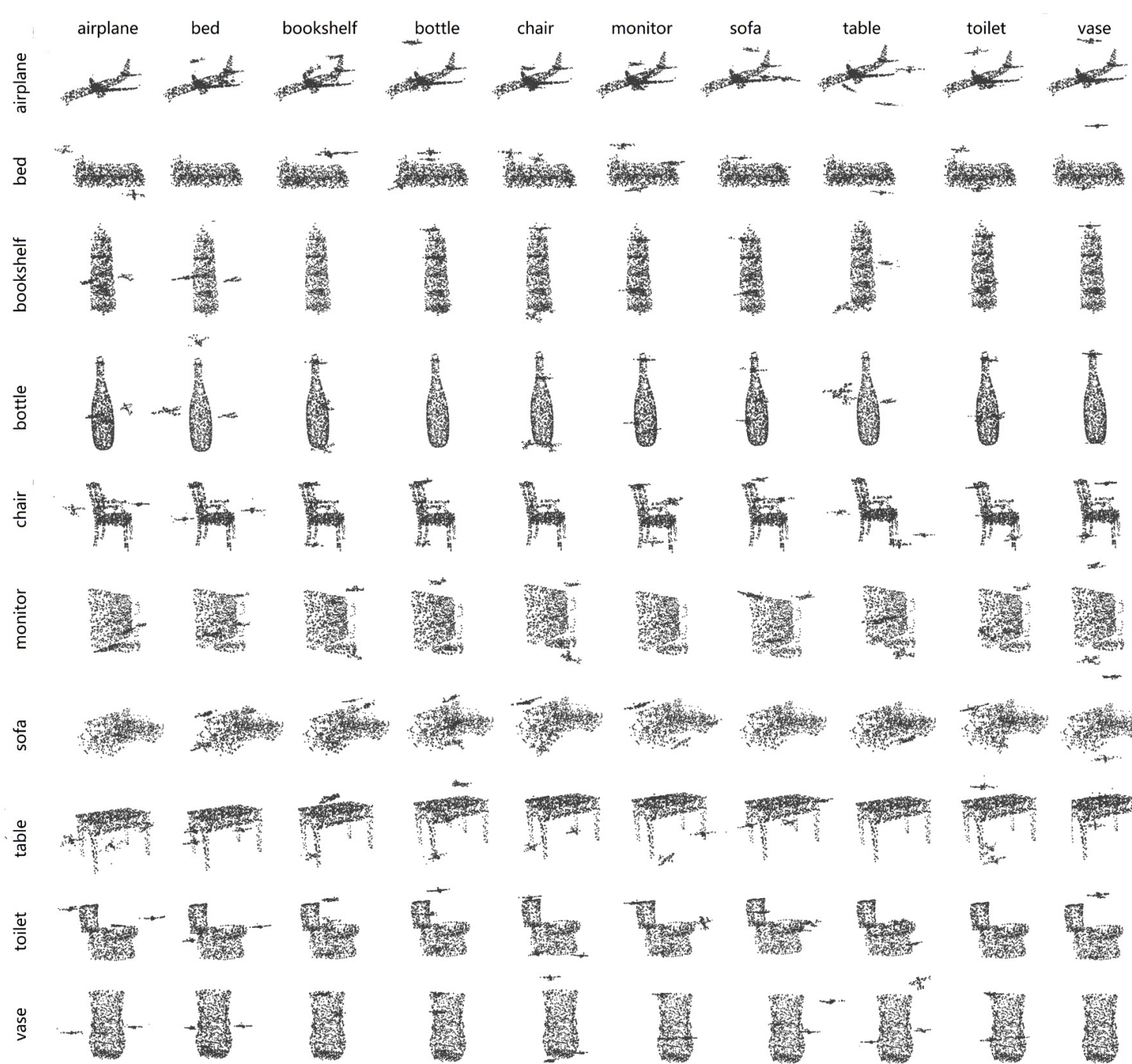
$$\min \mathcal{D}(x, x'), \quad s.t. \mathcal{F}(x') = t'$$

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

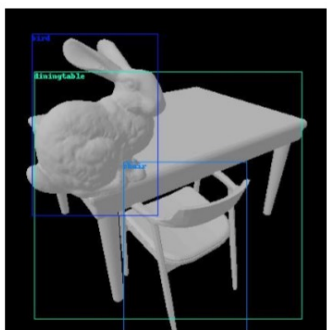
$$\mathcal{D}_H(\mathcal{S}, \mathcal{S}') = \max_{y \in \mathcal{S}'} \min_{x \in \mathcal{S}} \|x - y\|_2^2$$

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

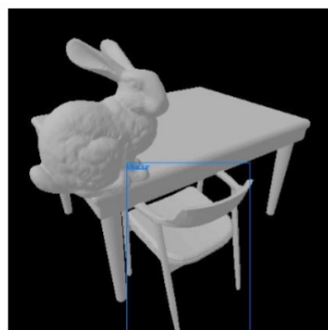




Adversarial Perturbation on Shape/Texture



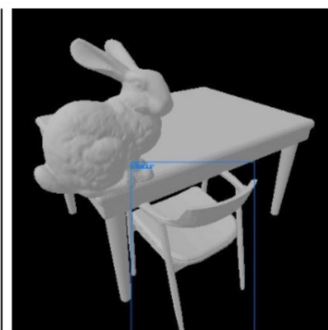
(a) Benign



(b) Table | Shape



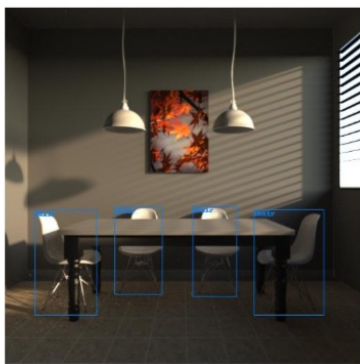
(c) All | Shape



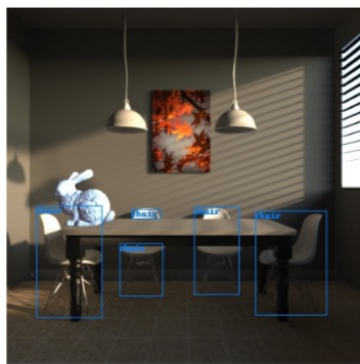
(d) Table | Texture



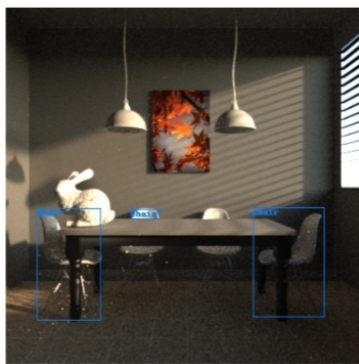
(e) All | Texture



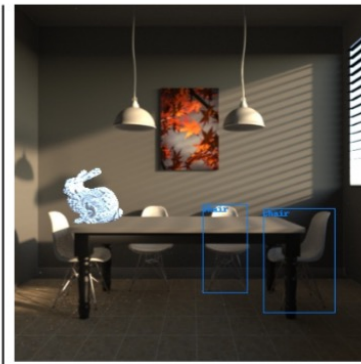
(a) Benign



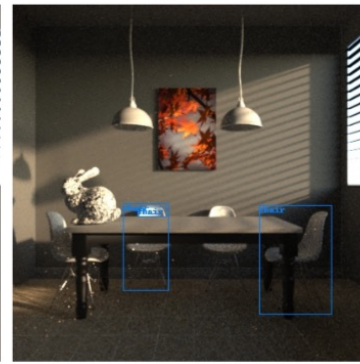
(b) S | NMR



(c) S | Mitsuba

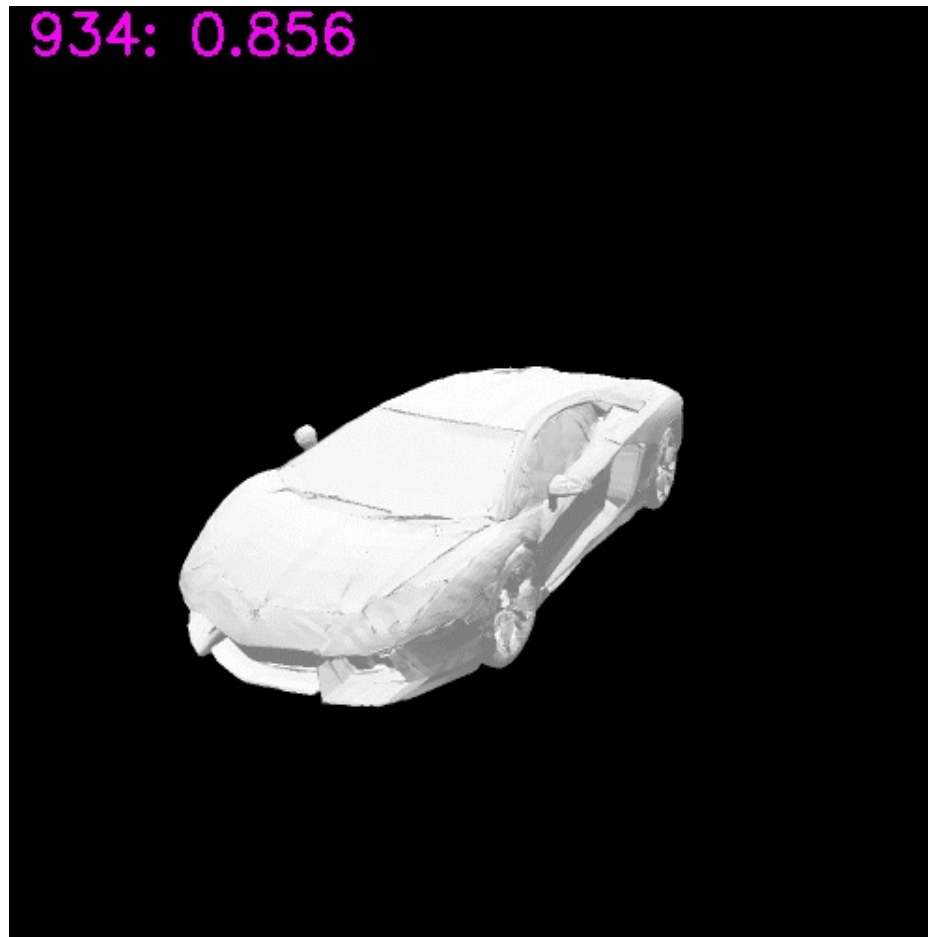


(d) S^{adv} | NMR



(e) S^{adv} | Mitsuba

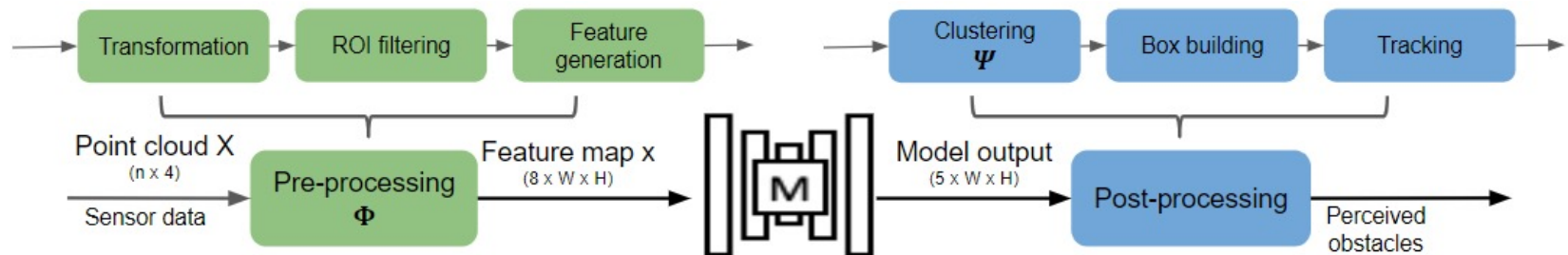
Adversarial 3D Meshes



- 934 : hot dog

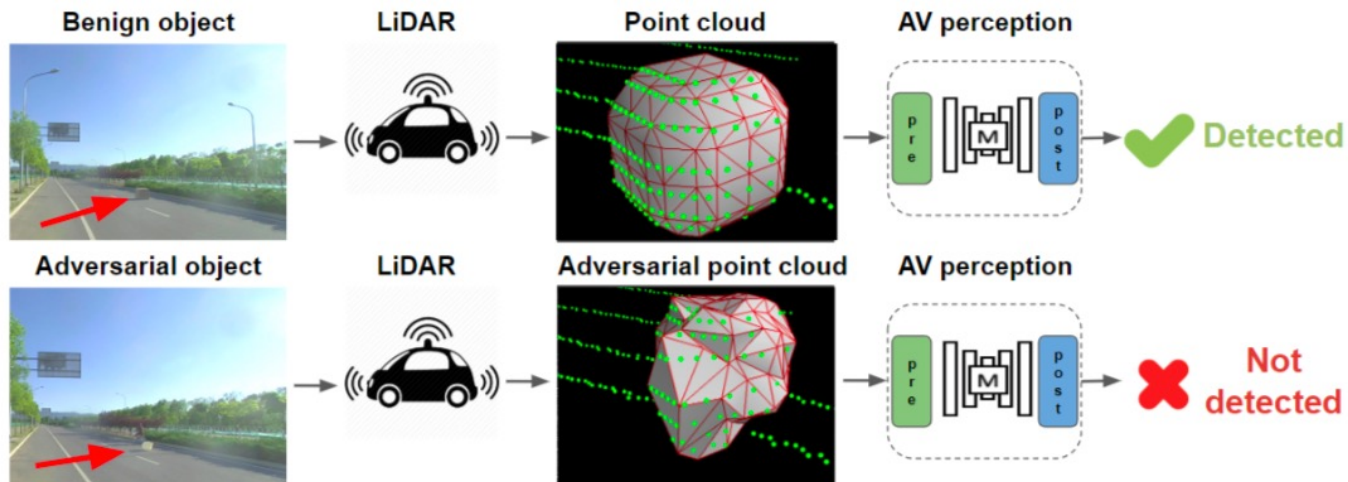
Real-world Challenges

- 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-World Adversarial Attack

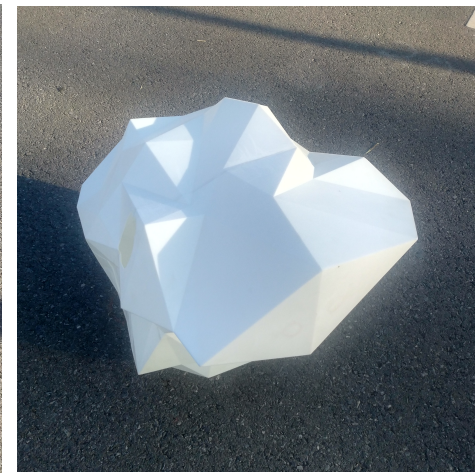
- Physical world experiment setup
 - A real vehicle equipped with a Velodyne HDL-64E LiDAR and camera



Road & car with LiDAR and camera



Benign

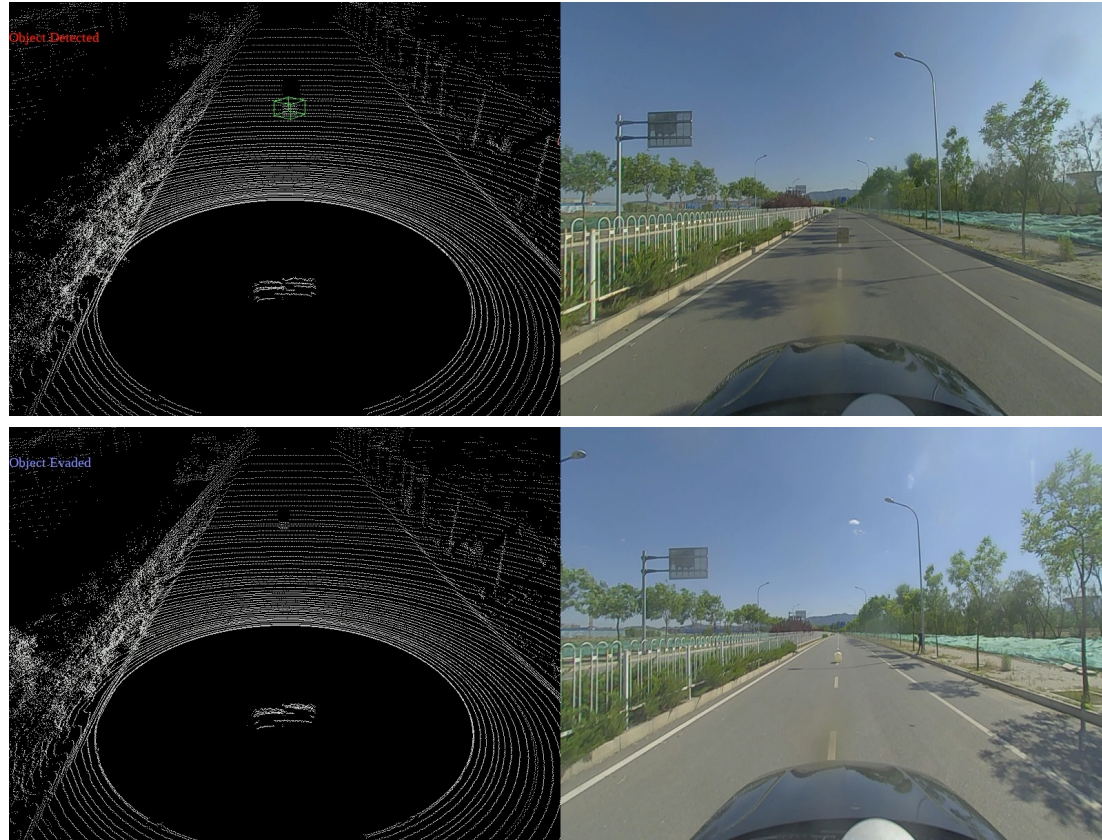


Adversarial

Adversarial object/benign box in the middle

Adversarial
Object

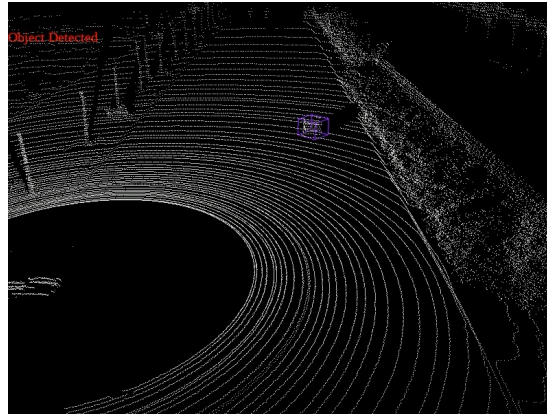
Benign
Object



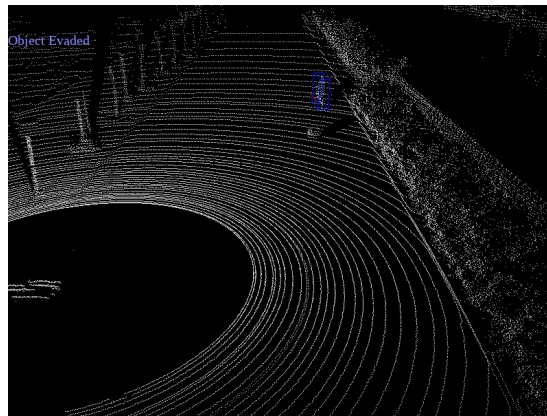
Physical Experiments

Adversarial object/benign box
on the right

Benign
Object



Adversarial
Object



MSF: widely recognized as a general defense strategy against existing attacks on AD perception

10.3.2 *Sensor-Level Defenses.* Several defenses could be adopted against spoofing attacks on LiDAR sensors:

Detection techniques. Sensor fusion, which intelligently combines data from several sensors to detect anomalies and improve performance, could be adopted against LiDAR spoofing attacks. AV systems are often equipped with sensors beyond LiDAR. Cameras, radars, and ultrasonic sensors provide additional information and redundancy to detect and handle an attack on LiDAR.

[Cao et al. CCS'19]

5.2 Potential Countermeasures

Redundancy and Fusion: If a vehicle is equipped with an overlapping field of view, the effect of saturating is mitigated to a certain extent. However, this is not a definitive solution because attackers can blind multiple sensors. Besides, it is also not easy to detect spoofing, when non-overlapped zones. Likewise, the fusion of multiple sensors is not an ultimate solution either. Radars [44], cameras [44] have all been revealed to be vulnerable to spoofing.

[Shin et al. CHES'17]

same physical variables in the presence of transient faults. The existing methods do not work well when an attacker wants to keep undetected by maximizing the interval of the sensor, for example, stealth attacks. We proposed a novel approach for attack detection which was presented based upon fusion intervals and past measurements. In this approach, we added a virtual sensor, and used pairwise inconsistencies between sensors to detect and identify attacks. The algorithm was evaluated on a real-world case study. The results demonstrated that the proposed algorithm outperforms the existing algorithms in various attack scenarios. Our future work is to explore how to further improve the recognition rate, especially for stealth attacks, and can identify as soon as

[Yang et al. FGCS'18]

As the system's autonomy increases, so does the concern about its security. In modern vehicles, a malicious attacker may deceive the controller into performing a dangerous action by altering the measurements of some sensors [1], [2]. Depending on the attacker's goal and capabilities, the consequences may range from minor disturbances in performance to crashes and loss of human lives. Consequently, performing attack-resilient sensor fusion is essential for the safety of such systems.

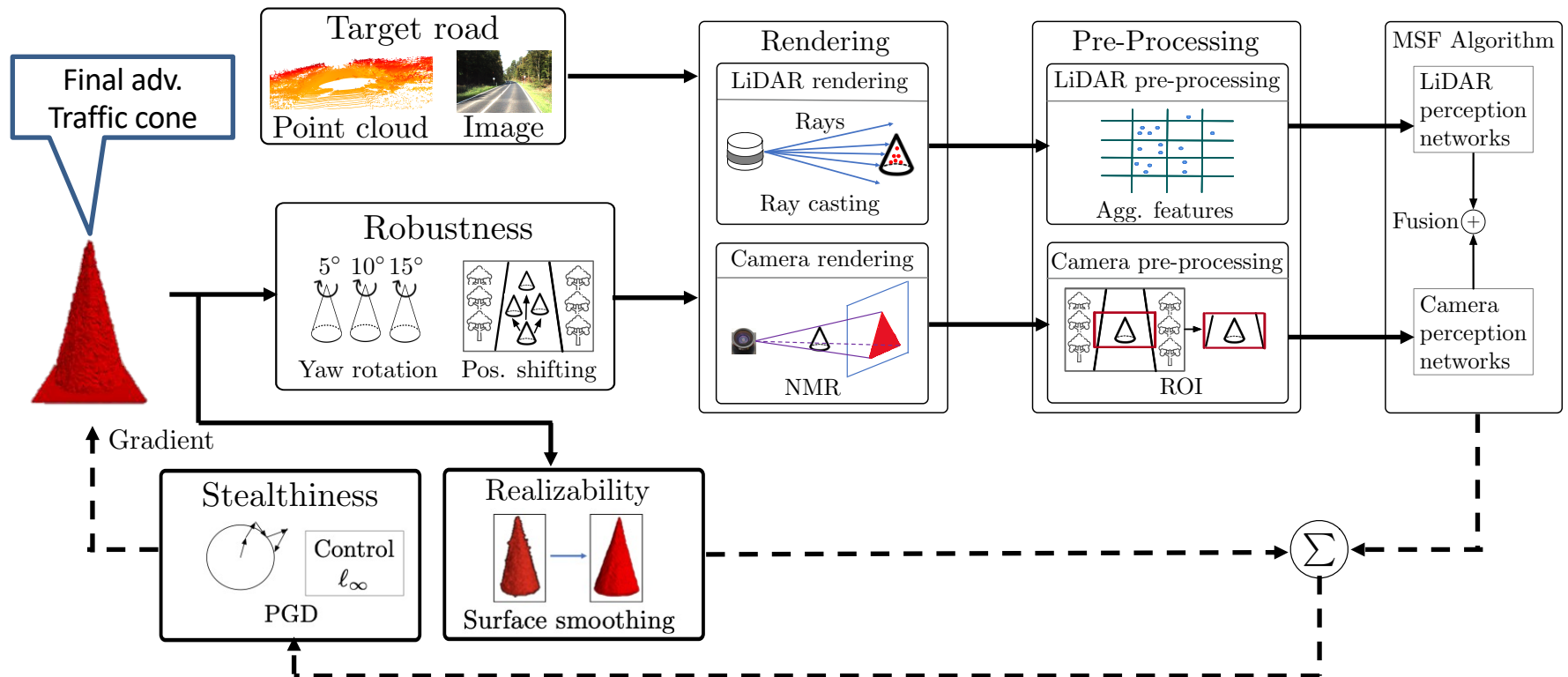
[Ivanov et al. DATE'14]

2.1 System Model and Current Approach

We consider a system with n sensors measuring the same physical variable. As mentioned above, we assume *abstract* sensors; therefore, each sensor provides the controller with an interval of all possible values. We assume the system queries all the sensors periodically such that a centralized estimator receives measurements from all sensors, and then performs attack detection/identification and sensor fusion (SF). We now explain the current approach to attack detection, referred to herein as a SF-based detector, before providing the improved version addressed in this paper.

[Park et al. ICCPS'15]

Attack Generation against MSF



End-to-End Attack Simulation

- Perform end-to-end attack evaluation on Baidu Apollo-5.0 and LGSVL simulator



Single lane road map



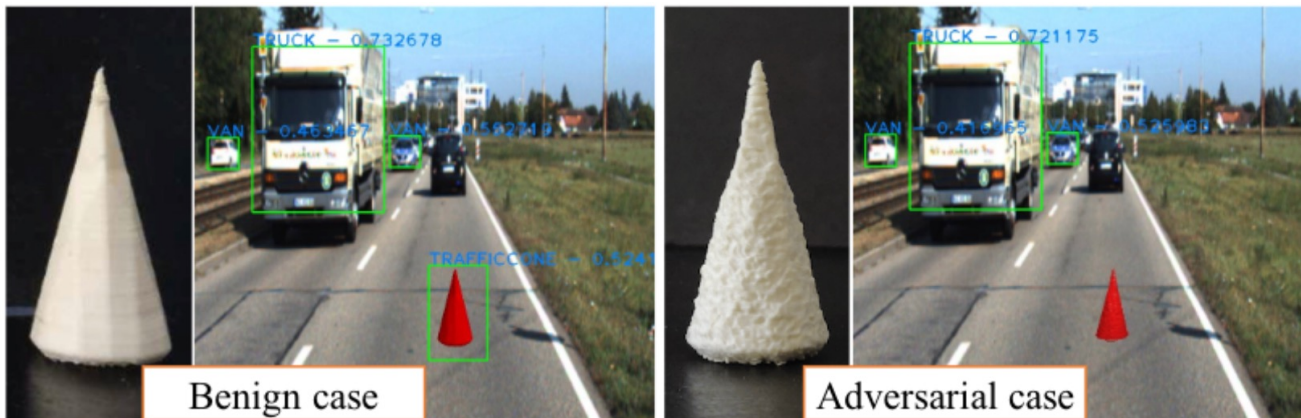
Vehicle

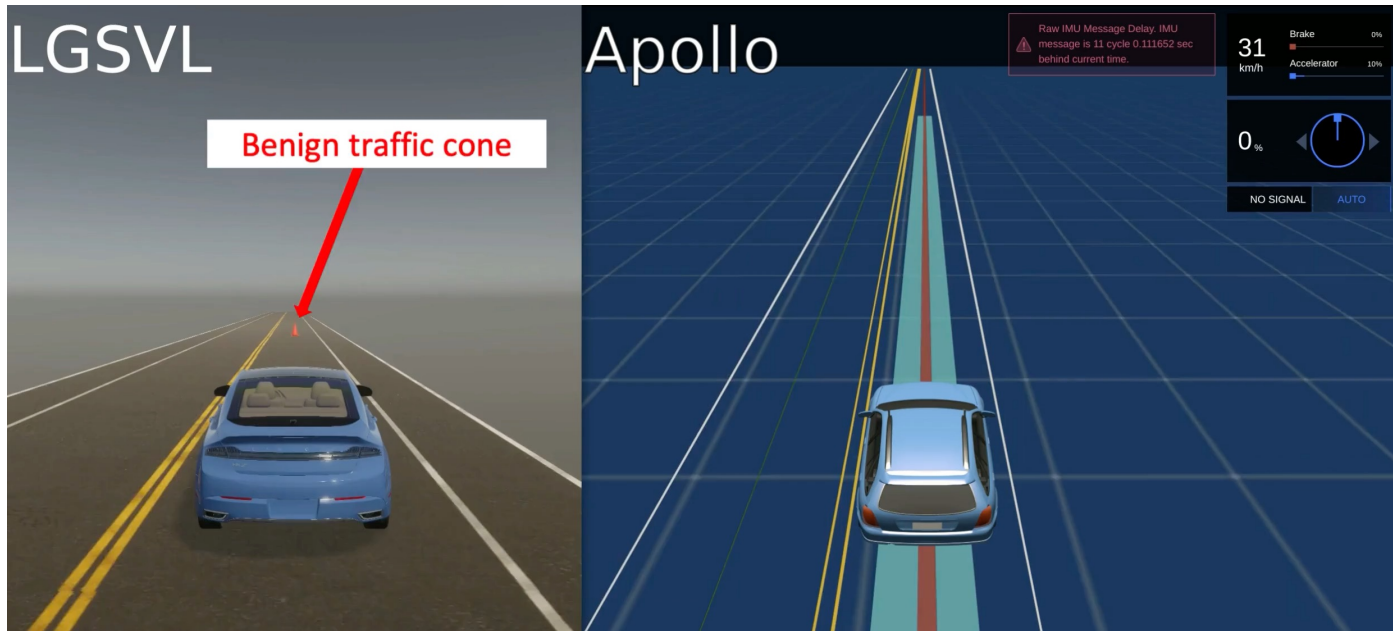


Benign

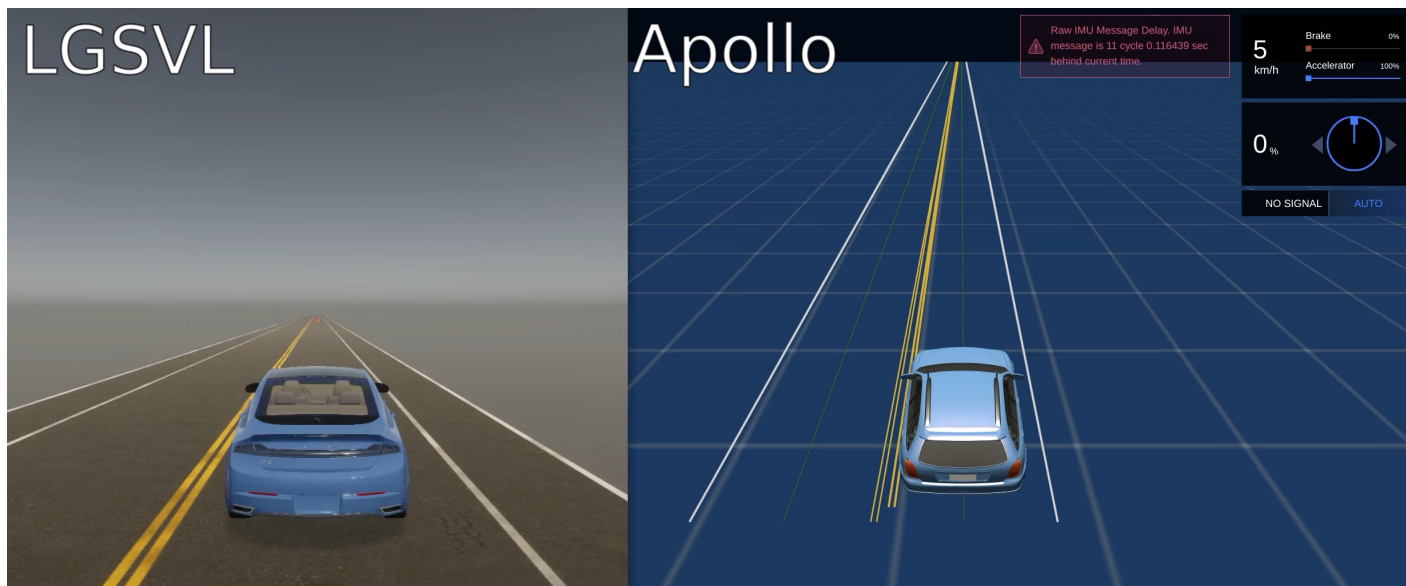


Adversarial



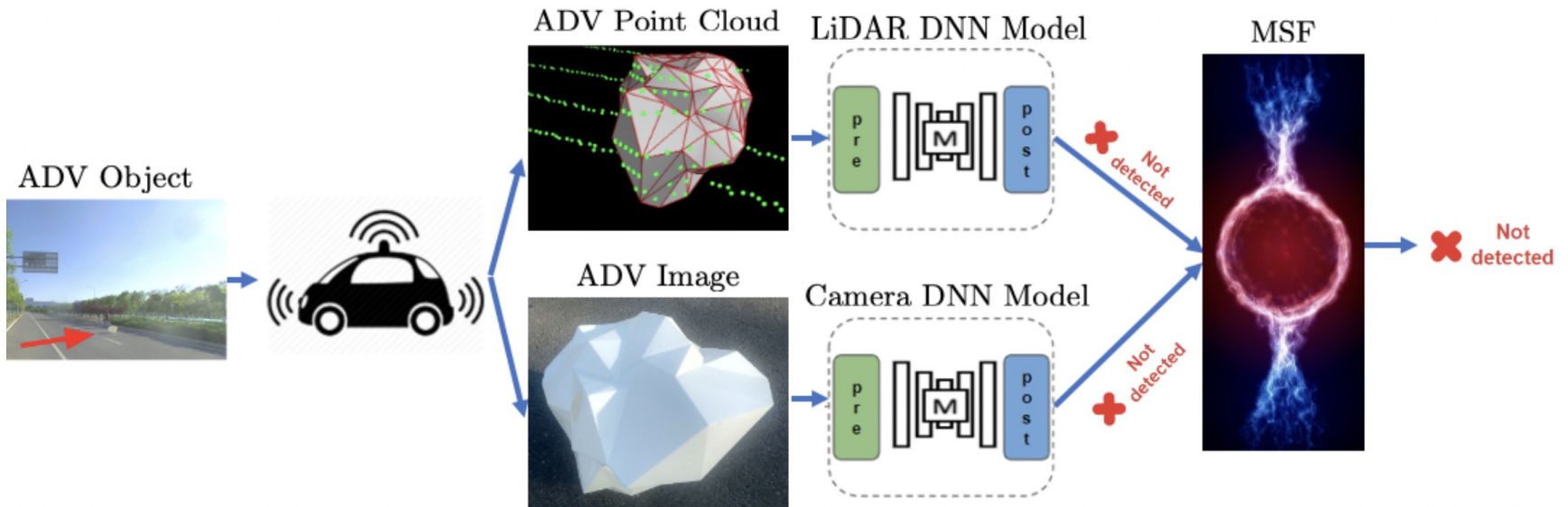


Control Experiment



Driving facing the adversarial object

Physical World MSF-based Attacks



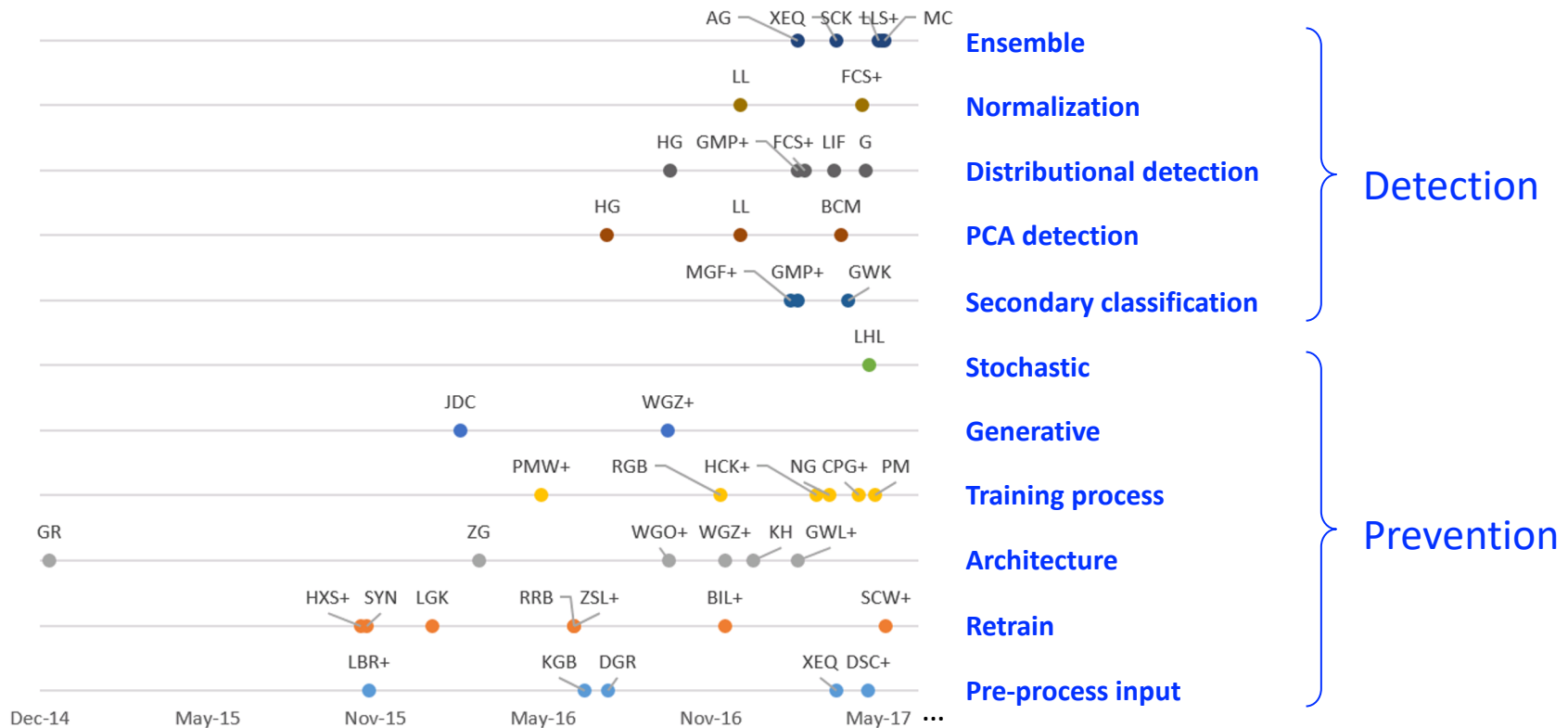
<https://aisecure.github.io/BLOG/MRF/Home.html>

Possible Vulnerability Disclosure

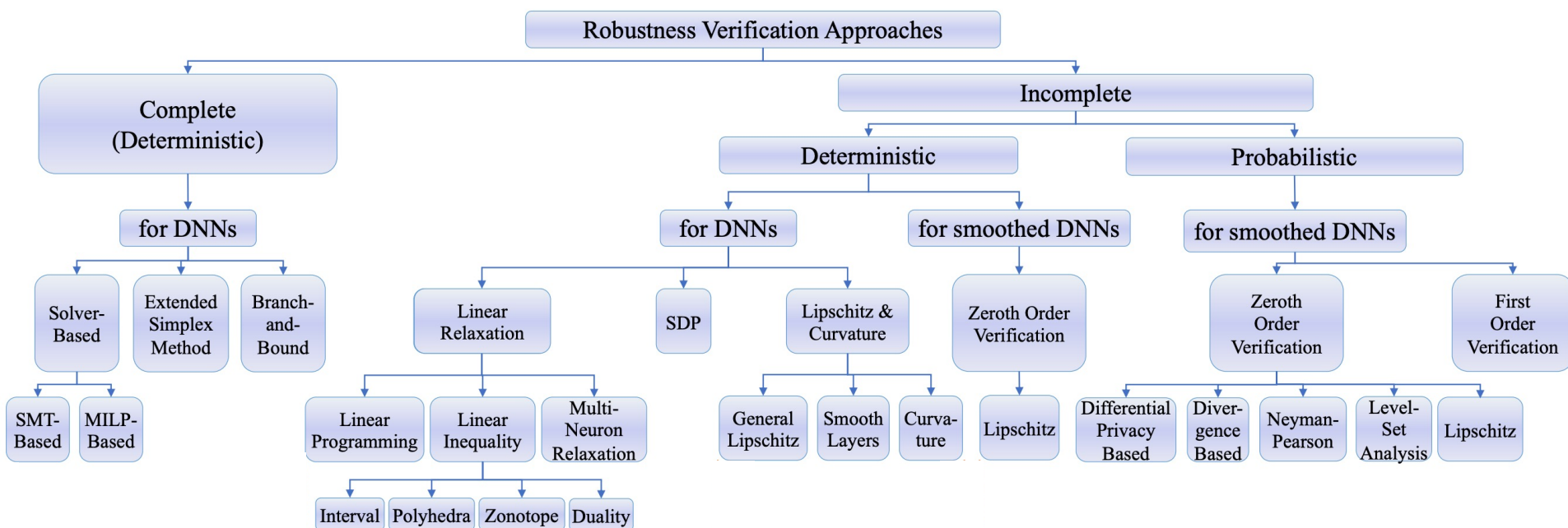
- As of 4/8/21, informed **32 companies** developing/testing AVs
 - **12** has replied so far and **have started investigation**



Numerous Defenses Proposed



Certified Robustness For ML



<https://sokcertifiedrobustness.github.io/>

STOA Certified Robustness on MNIST

- On MNIST
 - ℓ_∞ norm, $r = 0.3$
 - SOTA Certified Robust Accuracy: **93.09%**
 - *Towards Certifying ℓ_∞ Robustness using Neural Networks with ℓ_∞ -dist Neurons*
 - *ArXiv: 2102.05363*
 - SOTA Empirical Robust Accuracy (against existing attacks): **96.34%**
 - https://github.com/MadryLab/mnist_challenge
 - *Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples*
 - *ArXiv: 2010.03593*

➤ Not much difference

STOA Certified Robustness on CIFAR

- On CIFAR-10
 - ℓ_∞ norm, $r = 8/255$
 - SOTA Certified Robust Accuracy: **39.88%**
 - *Fast and Stable Interval Bounds Propagation for Training Verifiably Robust Models.*
ArXiv: 1906.00628
 - SOTA Empirical Robust Accuracy (against existing attacks): **65.87%**
 - Leaderboard: <https://robustbench.github.io/>
 - ℓ_∞ norm, $r = 2/255$
 - SOTA Certified Robust Accuracy: **68.2%**
 - *Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers.*
NeurIPS 2019
- Still a gap

STOA Certified Robustness on ImageNet

On ImageNet

- ℓ_2 norm, $r = 2.0$
- SOTA Certified Robust Accuracy: 27%
 - *Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. NeurIPS 2019*
- Still hard (also for empirical robustness)
- We maintain the SOTA certified robustness @
<https://github.com/AIsecure/Provable-Training-and-Verification-Approaches-Towards-Robust-Neural-Networks>

Robust ML Pipeline with Exogenous Information

- Vulnerabilities of statistical ML models: pure *data-driven* without considering exogenous information that cannot be modeled by data
 - Intrinsic information (e.g., spatial consistency)
 - Extrinsic information (e.g. domain knowledge)

COMMUNICATIONS
OF THE
ACM

Search















HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH | PRACTICE | CAREERS | ARCHIVE | VIDEOS

Home / Magazine Archive / February 2021 (Vol. 64, No. 2) / Polanyi's Revenge and AI's New Romance with Tacit... / Full Text

VIEWPOINT

Polanyi's Revenge and AI's New Romance with Tacit Knowledge

By Subbarao Kambhampati
Communications of the ACM, February 2021, Vol. 64 No. 2, Pages 31-32
10.1145/3446369
Comments (4)

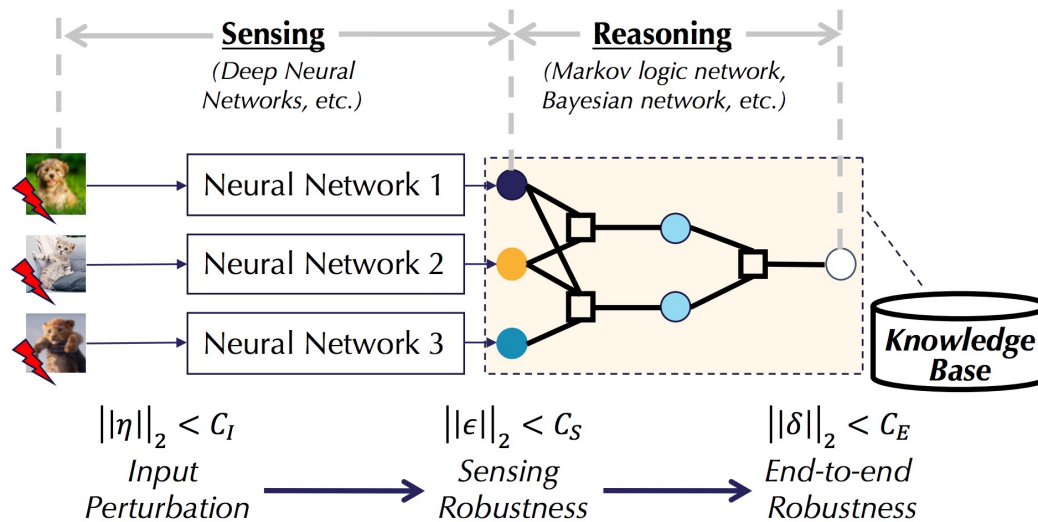
VIEW AS:       SHARE:        



The NetHack Challenge:
Dungeons, Dragons, and Tourists
NEURIPS 2021 REPORT

Certified Robustness for *Sensing-Reasoning* ML Pipelines

- Can we reason about the robustness of an end-to-end ML pipeline beyond a single ML model or ensemble?

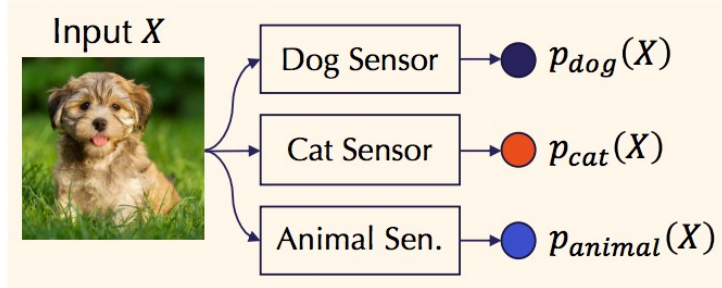


- Intuition: It is hard to attack every sensor in and still preserve their logical relationship
- Goal: Upper bound the end-to-end maximal change of the marginal probability of prediction
- Challenges: Solve the minmax for the pipeline

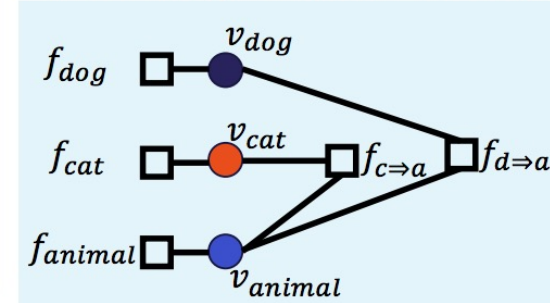
Challenges and Opportunities

- Challenges: Compared with neural networks whose inference can be executed in polynomial time, many reasoning models (e.g., MLN) can be #P-complete for inference.
- Opportunities: Many reasoning models define a probability distribution in the exponential family, which provides functional structures for solving the min-max problem.

(a) Sensing Component



(c) Reasoning Comp. (Factor Graph)



(b) MLN Program

<u>predicates</u>	
Dog(X); Cat(X); Animal(X)	
<u>weight</u>	<u>rule</u>
10.5	Dog(X) => Animal(X)
5.3	Cat(X) => Animal(X)

<u>factor</u>	<u>factor function</u>	<u>weight</u>	
f_{dog}	$f_{dog}(v) = v$	$\log \frac{p_{dog}(X)}{1 - p_{dog}(X)}$	w_{G_i}
$f_{d \Rightarrow a}$	$f_{d \Rightarrow a}(d, a) = 1 - d(1 - a)$	10.5	w_H
$f_{c \Rightarrow a}$	$f_{c \Rightarrow a}(c, a) = 1 - c(1 - a)$	5.3	

Two types of factors: Interface factors \mathcal{G} and Interior factors \mathcal{H}

$$\mathbb{E}[R_{MLN}(\{p_i(X)\}_{i \in [n]})] = \mathbf{Pr}[v = 1] = \frac{Z_1(\{p_i(X)\}_{i \in [n]})}{Z_2(\{p_i(X)\}_{i \in [n]})}$$

Marginal prediction probability

$$= \sum_{\sigma \in \Sigma \wedge \sigma(v)=1} \exp \left\{ \sum_{G_i \in \mathcal{G}} w_{G_i} \sigma(x_i) + \sum_{H \in \mathcal{H}} w_H f_H(\sigma(\bar{\mathbf{v}}_H)) \right\} \quad \frac{Z_1(\{p_i(X)\}_{i \in [n]})}{Z_2(\{p_i(X)\}_{i \in [n]})}$$

$$= \sum_{\sigma \in \Sigma} \exp \left\{ \sum_{G_i \in \mathcal{G}} w_{G_i} \sigma(x_i) + \sum_{H \in \mathcal{H}} w_H f_H(\sigma(\bar{\mathbf{v}}_H)) \right\}$$

Hardness

Definition 2 (COUNTING). Given input polynomial-time computable weight function $w(\cdot)$ and query function $Q(\cdot)$, parameters α , a real number $\varepsilon_c > 0$, a COUNTING oracle outputs a real number Z such that

$$1 - \varepsilon_c \leq \frac{Z}{\mathbf{E}_{\sigma \sim \pi_\alpha} [Q(\sigma)]} \leq 1 + \varepsilon_c.$$

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:

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

Theorem 4 (COUNTING \leq_t ROBUSTNESS). Given polynomial-time computable weight function $w(\cdot)$ and query function $Q(\cdot)$, parameters α and real number $\varepsilon_c > 0$, the instance of COUNTING, $(w, Q, \alpha, \varepsilon_c)$ can be determined by up to $O(1/\varepsilon_c^2)$ queries of the ROBUSTNESS oracle with input perturbation $\epsilon = O(\varepsilon_c)$.

Theorem 5 (MLN Hardness). Given an MLN whose grounded factor graph is $\mathcal{G} = (\mathcal{V}, \mathcal{F})$ in which the weights for interface factors are $w_{G_i} = \log p_i(X)/(1 - p_i(X))$ and constant thresholds δ, C , deciding whether

$$\begin{aligned} \forall \{\epsilon_i\}_{i \in [n]} \quad (\forall i. |\epsilon_i| < C) &\implies \\ |\mathbb{E} R_{MLN}(\{p_i(X)\}_{i \in [n]}) - \mathbb{E} R_{MLN}(\{p_i(X) + \epsilon_i\}_{i \in [n]})| &< \delta \end{aligned}$$

is as hard as estimating $\mathbb{E} R_{MLN}(\{p_i(X)\}_{i \in [n]})$ up to ε_c multiplicative error, with $\epsilon_i = O(\varepsilon_c)$.

Robustness of the Reasoning Component

Can we efficiently reason about the provable robustness for the reasoning component when given an [oracle](#) for the statistical inference?

Lemma 6 (MLN Robustness). Given access to partition functions $Z_1(\{p_i(X)\}_{i \in [n]})$ and $Z_2(\{p_i(X)\}_{i \in [n]})$, and a maximum perturbation C , $\forall \epsilon_1, \dots, \epsilon_n$, if $\forall i. |\epsilon_i| < C$, we have that $\forall \lambda_1, \dots, \lambda_n \in \mathbb{R}$,

$$\begin{aligned}
 & \max_{\{|\epsilon_i| < C\}} \ln \mathbb{E}[R_{MLN}(\{p_i(X) + \epsilon_i\}_{i \in [n]})] \\
 & \leq \max_{\{|\epsilon_i| < C\}} \widetilde{Z}_1(\{\epsilon_i\}_{i \in [n]}) - \min_{\{|\epsilon'_i| < C\}} \widetilde{Z}_2(\{\epsilon'_i\}_{i \in [n]}) \\
 & \min_{\{|\epsilon_i| < C\}} \ln \mathbb{E}[R_{MLN}(\{p_i(X) + \epsilon_i\}_{i \in [n]})] \\
 & \geq \min_{\{|\epsilon_i| < C\}} \widetilde{Z}_1(\{\epsilon_i\}_{i \in [n]}) - \max_{\{|\epsilon'_i| < C\}} \widetilde{Z}_2(\{\epsilon'_i\}_{i \in [n]})
 \end{aligned}$$

Oracle inference

where

$$\widetilde{Z}_r(\{\epsilon_i\}_{i \in [n]}) = \ln Z_r(\{p_i(X) + \epsilon_i\}_{i \in [n]}) + \sum_i \lambda_i \epsilon_i.$$

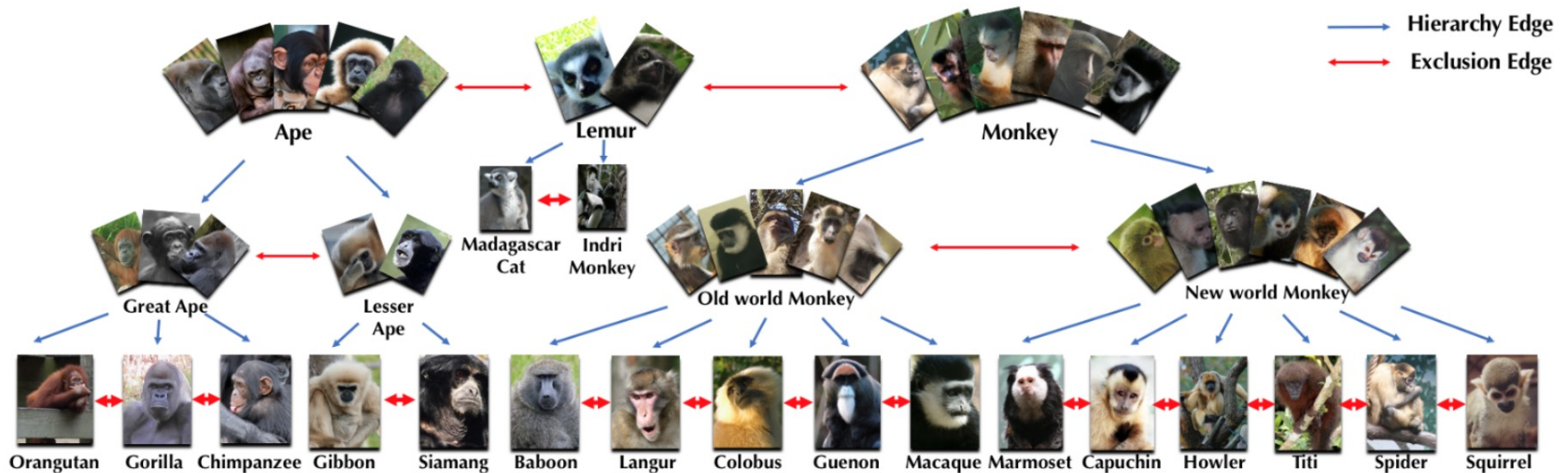
1. When $\lambda_i \geq 0$, $\widetilde{Z}_r(\{\epsilon_i\}_{i \in [n]})$ monotonically increases w.r.t. ϵ_i ; Thus, the maximal is achieved at $\epsilon_i = C$ and the minimal is achieved at $\epsilon_i = -C$. When $\lambda_i \leq -1$, $\widetilde{Z}_r(\{\epsilon_i\}_{i \in [n]})$ monotonically decreases w.r.t. ϵ_i ; Thus, the maximal is achieved at $\epsilon_i = -C$ and the minimal is achieved at $\epsilon_i = C$.
2. When $\lambda_i \in (-1, 0)$, the maximal is achieved at $\epsilon_i \in \{-C, C\}$, and the minimal is achieved at $\epsilon_i \in \{-C, C\}$ or at the zero gradient of $\widetilde{Z}_r(\{\tilde{\epsilon}_i\}_{i \in [n]})$ with respect to $\tilde{\epsilon}_i = \log \left[\frac{(1-p_i(X))(p_i(X)+\epsilon_i)}{p_i(X)(1-p_i(X)-\epsilon_i)} \right]$, due to the convexity of $\widetilde{Z}_r(\{\tilde{\epsilon}_i\}_{i \in [n]})$ in $\tilde{\epsilon}_i$, $\forall i$.

Beyond Markov Logic Networks

- Bayesian networks with tree structures
- Bayesian networks with binary tree structure or a 1-NN classifier – tight upper and lower bound of reasoning robustness

Example: PrimateNet (ImageNet)

PrimateNet. The knowledge structure of blue arrows represent the Hierarchical rules between different classes, and red arrows the Exclusive rules. (Some exclusive rules are omitted)



- **Hierarchy edge** $u \implies v$: If one object belongs to class u , it should belong to class v as well:

$$x_u \wedge \neg x_v = \text{False}$$

- **Exclusion edge** $u \oplus v$: One object couldn't belong to class u and class v at the same time:

$$x_u \wedge x_v = \text{False}$$

Table 1: *Benign* accuracy (i.e. $C_I = 0, \alpha = 0$) of models with and without knowledge under different smoothing parameter σ evaluated on PrimateNet.

σ	With knowledge	Without knowledge
0.12	0.9670	0.9638
0.25	0.9612	0.9554
0.50	0.9435	0.9371

Table 2: Certified Robustness and Certified Ratio with different perturbation magnitude C_I and sensing model attack ratio α on PrimateNet. The sensing models are smoothed with Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ with different smoothing parameter σ .

(a) $\sigma = 0.12$

		With knowledge		Without knowledge	
C_I	α	Cert. Robustness	Cert. Ratio	Cert. Robustness	Cert. Ratio
0.12	10%	0.8849	0.9419	0.5724	0.5724
	20%	0.8078	0.8609	0.5717	0.5717
	30%	0.7508	0.7988	0.5706	0.5706
	50%	0.6236	0.6647	0.5706	0.5706
0.25	10%	0.7888	0.8428	0.2342	0.2342
	20%	0.6226	0.6657	0.2320	0.2320
	30%	0.5225	0.5596	0.2309	0.2309
	50%	0.3594	0.3824	0.2268	0.2268

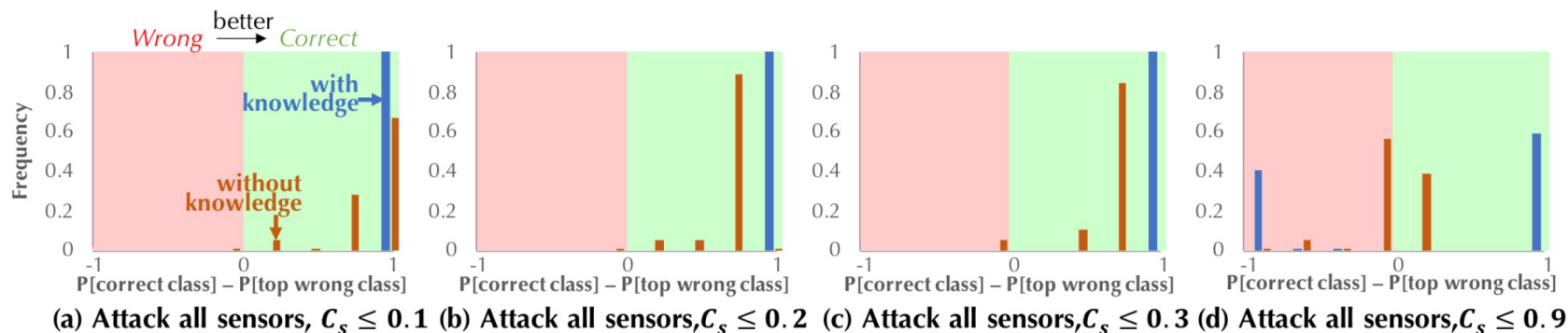
(c) $\sigma = 0.50$

		With knowledge		Without knowledge	
C_I	α	Cert. Robustness	Cert. Ratio	Cert. Robustness	Cert. Ratio
0.50	10%	0.8288	0.9449	0.4762	0.4762
	20%	0.7407	0.8488	0.4749	0.4749
	30%	0.6907	0.7968	0.4736	0.4736
	50%	0.5581	0.6395	0.4635	0.4635
1.00	10%	0.7307	0.8448	0.1679	0.1679
	20%	0.5285	0.6336	0.1615	0.1615
	30%	0.4347	0.5375	0.1612	0.1612
	50%	0.2624	0.3318	0.1584	0.1584

Example: (NLP) Relation Extraction Task

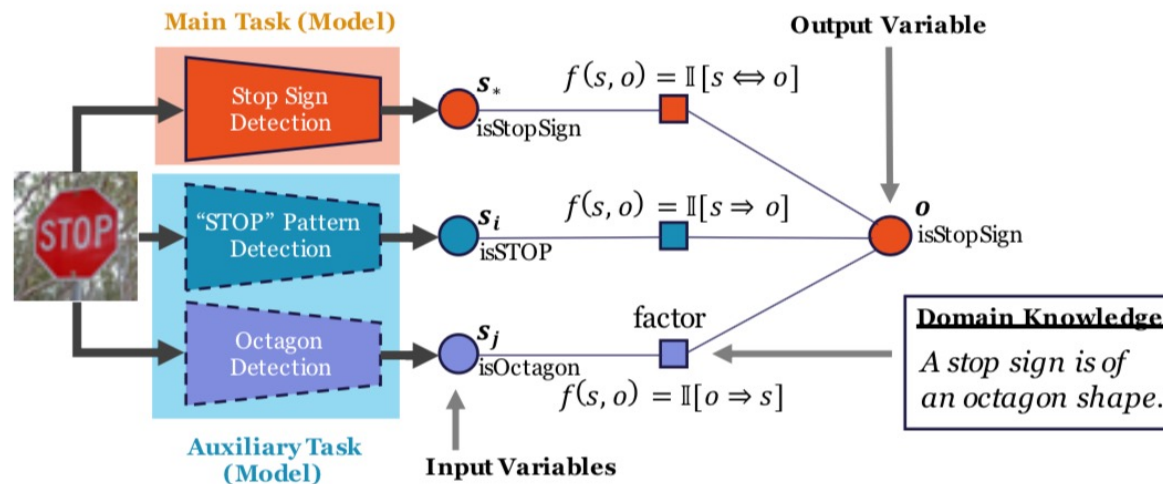
Table 3: (NLP) Certified Robustness and Certified Ratio for approaches when all sensing models are attacked.

	With knowledge		Without knowledge	
C_S	Cert. Robustness	Cert. Ratio	Cert. Robustness	Cert. Ratio
0.1	1.0000	1.0000	0.9969	0.9969
0.5	1.0000	1.0000	0.9474	0.9474
0.9	0.5882	0.5882	0.3839	0.3839



Example: Knowledge Enhanced ML Pipeline against *Diverse* Adversarial Attacks

- Example: Robust road sign recognition
- The output of ML models are modeled as input random variables for reasoning
- Permissive knowledge: $s \text{ infers } y$
- Preventive knowledge: $y \text{ infer } s$



Knowledge Enhanced ML Pipeline against Diverse Adversarial Attacks

- Lower bound of the pipeline accuracy

Theorem 1 (Convergence of $\mathcal{A}^{\text{KEMLP}}$). For $y \in \mathcal{Y}$ and $\mathcal{D} \in \{\mathcal{D}_b, \mathcal{D}_a\}$, let $\mu_{y,\mathcal{D}}$ be defined as in Lemma 1. Suppose that the modeling assumption holds, and suppose that $\mu_{d_{\mathcal{K}},\mathcal{D}} > 0$, for all $\mathcal{K} \in \{\mathcal{I}, \mathcal{J}\}$ and $\mathcal{D} \in \{\mathcal{D}_b, \mathcal{D}_a\}$. Then

$$\mathcal{A}^{\text{KEMLP}} \geq 1 - \mathbb{E}_{\mu_{y,\mathcal{D}}}[\exp(-2\mu_{y,\mathcal{D}}^2/v^2)],$$

where v^2 is the variance upper bound to $\mathbb{P}[o = y|y, \mathbf{w}]$ with

$$v^2 = 4 \left(\log \frac{\vee \alpha_*}{1 - \wedge \alpha_*} \right)^2 + \sum_{k \in \mathcal{I} \cup \mathcal{J}} \left(\log \frac{\vee \alpha_k (1 - \wedge \epsilon_k)}{\wedge \epsilon_k (1 - \vee \alpha_k)} \right)^2.$$

$\mu_{y,\mathcal{D}}$ consists of three terms: $\mu_{d_*,\mathcal{D}}$, $\mu_{\mathcal{I},\mathcal{D}}$, and $\mu_{\mathcal{J},\mathcal{D}}$ measuring the contributions from the main, permissive, and preventative sensors.

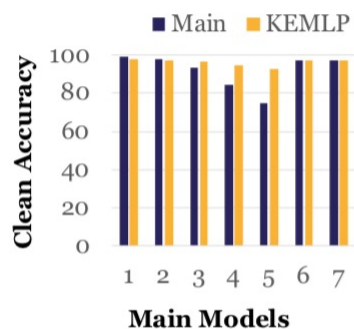
- The accuracy of pipeline is higher than that of the main sensor

Theorem 2 (Sufficient condition for $\mathcal{A}^{\text{KEMLP}} > \mathcal{A}^{\text{main}}$). Let the number of permissive and preventative models be the same and denoted by n such that $n := |\mathcal{I}| = |\mathcal{J}|$. Note that the weighted accuracy of the main model in terms of its truth rate is simply $\alpha_* := \sum_{\mathcal{D} \in \{\mathcal{D}_b, \mathcal{D}_a\}} \pi_{\mathcal{D}} \alpha_{*,\mathcal{D}}$. Moreover, let $\mathcal{K}, \mathcal{K}' \in \{\mathcal{I}, \mathcal{J}\}$ with $\mathcal{K} \neq \mathcal{K}'$ and for any $\mathcal{D} \in \{\mathcal{D}_b, \mathcal{D}_a\}$, let

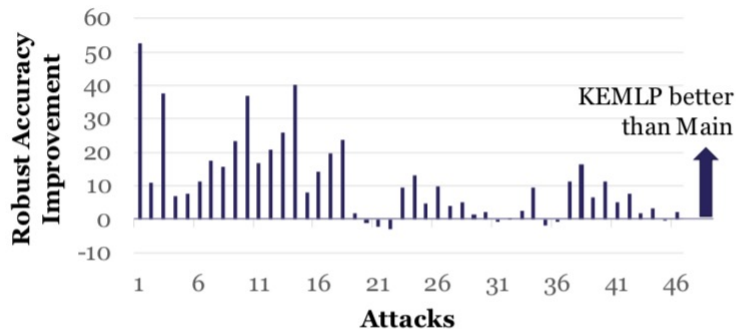
$$\gamma_{\mathcal{D}} := \frac{1}{n+1} \min_{\mathcal{K}} \left\{ \alpha_{*,\mathcal{D}} - 1/2 + \sum_{k \in \mathcal{K}} \alpha_{k,\mathcal{D}} - \sum_{k' \in \mathcal{K}'} \epsilon_{k',\mathcal{D}} \right\}.$$

If $\gamma_{\mathcal{D}} > \sqrt{\frac{4}{n+1} \log \frac{1}{1-\alpha_*}}$ for all $\mathcal{D} \in \{\mathcal{D}_b, \mathcal{D}_a\}$, then $\mathcal{A}^{\text{KEMLP}} > \mathcal{A}^{\text{main}}$.

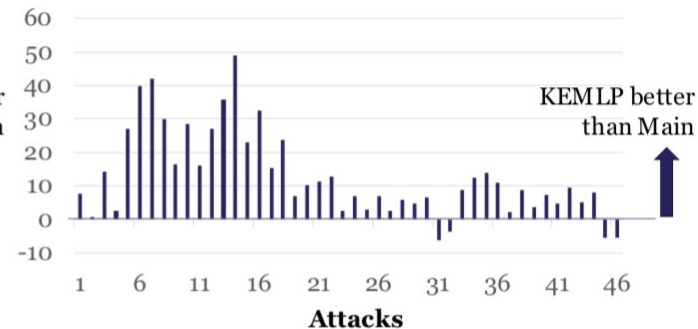
Experimental Results



(a) Clean Accuracy (KEMLP vs. Different Main Models)



(b) Robust Accuracy Improvement (KEMLP over AdvTrain ($\epsilon=8$))



(c) Robust Accuracy Improvement (KEMLP over DOA (5×5))

(a) Clean accuracy and (b) (c) robust accuracy improvement of KEMLP ($\alpha = 0.5$) over baselines against different attacks under both whitebox and blackbox settings.

Robustness of KEMLP against *Physical Attacks*

Model performance (%) under physical attacks ($\alpha = 0.4$). Performance **gain** and **loss** of KEMLP over baselines are highlighted.

	Main				KEMLP		
	Clean Acc	Robust Acc	W-Robust Acc		Clean Acc	Robust Acc	W-Robust Acc
GTSRB-CNN	100	5	52.5		100(± 0)	87.5(+82.5)	93.75(+41.25)
AdvTrain ($\epsilon = 4$)	100	12.5	56.25		100(± 0)	90(+77.5)	95(+38.75)
AdvTrain ($\epsilon = 8$)	97.5	37.5	67.5		100(+2.5)	90(+52.5)	95(+27.5)
AdvTrain ($\epsilon = 16$)	87.5	50	68.75		100(+12.5)	90(+40)	95(+26.25)
AdvTrain ($\epsilon = 32$)	62.5	32.5	47.5		100(+37.5)	90(+57.5)	95(+47.5)
DOA (5x5)	95	90	92.5		100(+5)	100(+10)	100(+7.5)
DOA (7x7)	57.5	32.5	45		100(+42.5)	100(+67.5)	100(+55)

Robustness of KEMLP against L_p Bounded Attacks

Table 2. Accuracy (%) under whitebox \mathcal{L}_∞ attacks ($\alpha = 0.8$)

Models		$\epsilon = 0$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 32$
GTSRB-CNN	Main	99.38	81.22	61.16	37.73	6.30
	KEMLP	97.38(-2.00)	90.33(+9.11)	77.88(+16.72)	60.44(+22.71)	35.52(+29.22)
AdvTrain ($\epsilon = 4$)	Main	97.94	87.99	69.34	42.44	20.29
	KEMLP	97.25(-0.69)	92.05(+4.06)	80.76(+11.42)	63.32(+20.88)	40.43(+20.14)
AdvTrain ($\epsilon = 8$)	Main	93.72	84.18	72.04	44.06	23.84
	KEMLP	96.48(+2.76)	92.10(+7.92)	84.08(+12.04)	63.58(+19.52)	40.66(+16.82)
AdvTrain ($\epsilon = 16$)	Main	84.54	78.55	71.99	57.87	26.13
	KEMLP	95.42(+10.88)	92.70(+14.15)	86.78(+14.79)	72.48(+14.61)	45.09(+18.96)
AdvTrain ($\epsilon = 32$)	Main	74.74	70.24	65.66	56.58	36.06
	KEMLP	94.86(+20.12)	91.69(+21.45)	86.39(+20.73)	76.05(+19.47)	54.78(+18.72)
DOA (5x5)	Main	97.43	57.97	29.84	9.44	3.01
	KEMLP	97.09(-0.34)	86.16(+28.19)	71.53(+41.69)	53.37(+43.94)	34.75(+31.74)
DOA (7x7)	Main	97.27	40.20	11.96	3.94	2.67
	KEMLP	96.99(-0.28)	84.52(+44.32)	70.47(+58.51)	56.58(+52.64)	45.73(+43.06)

Table 3. Accuracy (%) under whitebox unforeseen attacks ($\alpha = 0.8$)

		Clean	Fog-256	Fog-512	Snow-0.25	Snow-0.75	Jpeg-0.125	Jpeg-0.25	Gabor-20	Gabor-40	Elastic-1.5	Elastic-2.0
GTSRB-CNN	Main	99.38	59.65	34.18	56.58	24.54	55.74	27.01	57.25	32.41	44.78	24.31
	KEMLP	97.38(-2.00)	76.95(+17.30)	62.83(+28.65)	78.94(+22.36)	53.22(+28.68)	79.63(+23.89)	63.40(+36.39)	80.17(+22.92)	65.20(+32.79)	69.34(+24.56)	52.37(+28.06)
AdvTrain ($\epsilon = 4$)	Main	97.94	55.53	29.50	66.31	32.61	56.58	28.11	73.30	46.76	57.25	30.09
	KEMLP	97.25(-0.69)	76.08(+20.55)	61.96(+32.46)	80.45(+14.14)	57.84(+25.23)	84.23(+27.65)	68.57(+40.46)	81.48(+8.18)	65.77(+19.01)	71.19(+13.94)	50.33(+20.24)
AdvTrain ($\epsilon = 8$)	Main	93.72	50.03	23.56	63.71	34.93	57.56	26.16	76.72	53.76	48.25	24.46
	KEMLP	96.48(+2.76)	76.59(+26.56)	63.97(+40.41)	81.40(+17.69)	57.07(+22.14)	85.11(+27.55)	68.70(+42.54)	85.29(+8.57)	68.90(+15.14)	68.78(+20.53)	49.31(+24.85)
AdvTrain ($\epsilon = 16$)	Main	84.54	47.92	19.75	66.46	37.60	66.56	34.23	78.01	64.33	55.48	32.28
	KEMLP	95.42(+10.88)	77.13(+29.21)	64.38(+44.63)	81.64(+15.18)	58.20(+20.60)	86.99(+20.43)	70.40(+36.17)	87.42(+9.41)	72.61(+8.28)	67.31(+11.83)	50.28(+18.00)
AdvTrain ($\epsilon = 32$)	Main	74.74	48.71	22.84	61.78	38.91	63.58	43.49	70.37	65.20	54.58	39.45
	KEMLP	94.86(+20.12)	79.22(+30.51)	66.33(+43.49)	81.20(+19.42)	64.53(+25.62)	86.70(+23.12)	73.38(+29.89)	87.04(+16.67)	74.92(+9.72)	66.38(+11.80)	54.76(+15.31)
DOA (5x5)	Main	97.43	58.00	32.69	61.19	28.34	41.13	11.29	55.43	29.55	58.02	32.74
	KEMLP	97.09(-0.34)	76.85(+18.85)	63.07(+30.38)	78.78(+17.59)	56.76(+28.42)	78.60(+37.47)	61.78(+50.49)	80.25(+24.82)	63.89(+34.34)	72.69(+14.67)	57.51(+24.77)
DOA (7x7)	Main	97.27	59.88	38.01	62.47	30.17	23.46	3.65	54.58	27.29	56.33	30.97
	KEMLP	96.99(-0.28)	78.09(+18.21)	62.76(+24.75)	79.68(+17.21)	58.26(+28.09)	74.25(+50.79)	61.39(+57.74)	79.06(+24.48)	62.29(+35.00)	71.27(+14.94)	55.09(+24.12)

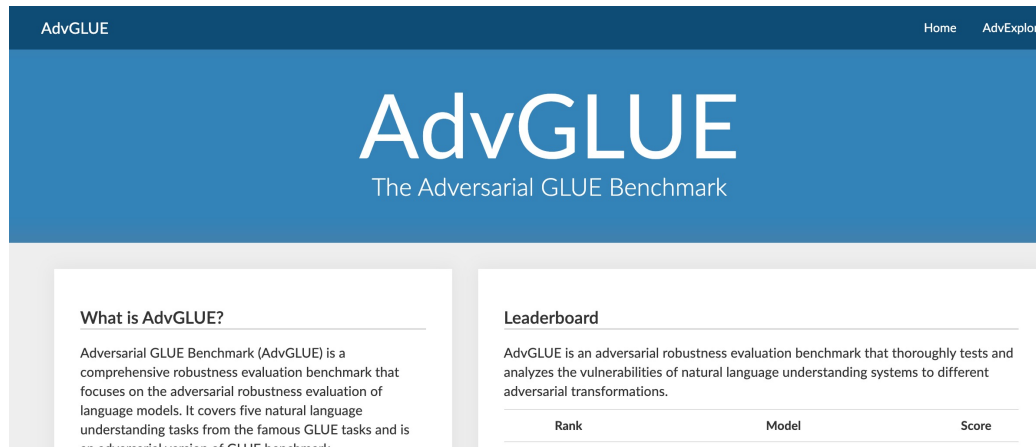
Robustness of KEMLP against *Common Corruptions*

Table 4. Accuracy (%) under common corruptions ($\alpha = 0.2$)

		Clean	Fog	Contrast	Brightness
GTSRB-CNN	Main	99.38	76.23	57.61	85.52
	KEMLP	98.28(-1.10)	78.14(+1.91)	72.43(+14.82)	89.58(+4.06)
AdvTrain ($\epsilon = 4$)	Main	97.94	63.81	42.31	78.47
	KEMLP	97.89(-0.05)	70.29(+6.48)	67.46(+25.16)	86.70(+8.23)
AdvTrain ($\epsilon = 8$)	Main	93.72	59.05	31.97	78.47
	KEMLP	96.79(+3.07)	67.41(+8.36)	66.69(+34.72)	85.91(+7.44)
AdvTrain ($\epsilon = 16$)	Main	84.54	56.58	34.31	78.01
	KEMLP	94.68(+10.14)	66.80(+10.22)	68.39(+34.08)	86.14(+8.13)
AdvTrain ($\epsilon = 32$)	Main	74.74	50.87	30.45	71.30
	KEMLP	91.46(+16.72)	64.94(+14.07)	68.31(+37.86)	83.20(+11.90)
DOA (5x5)	Main	97.43	73.95	62.24	83.92
	KEMLP	97.45(+0.02)	76.08(+2.13)	74.38(+12.14)	87.60(+3.68)
DOA (7x7)	Main	97.27	73.41	57.54	83.56
	KEMLP	97.22(-0.05)	76.00(+2.59)	72.40(+14.86)	87.78(+4.22)

Thorough Robustness Evaluation and Certification

- <https://adversarialglue.github.io/>



- <https://crop-leaderboard.me/>



A standardized benchmark for certified robustness of RL algorithms

The goal of **CROP-leaderboard** is to systematically certify the robustness of different RL algorithms based on certification criteria such as per-state action and the lower bound of cumulative reward. The related paper can be found [here](#).

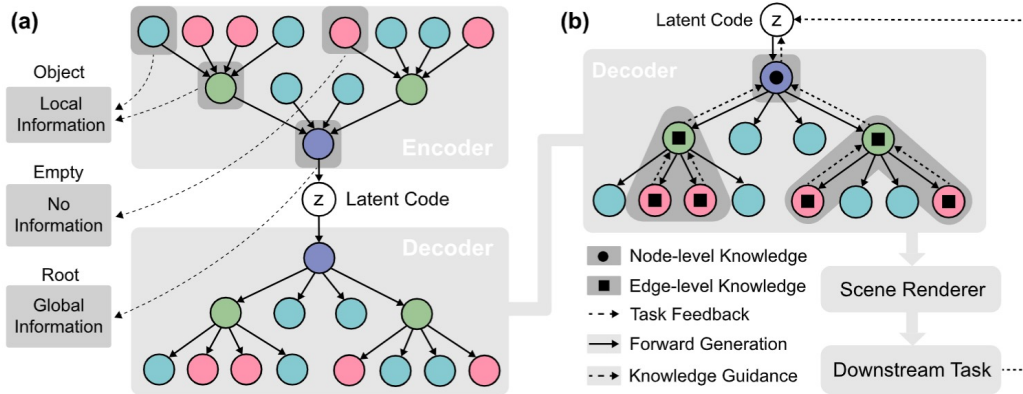
Available Leaderboards

Game: Certification strategy:

[CartPole-v0](#)

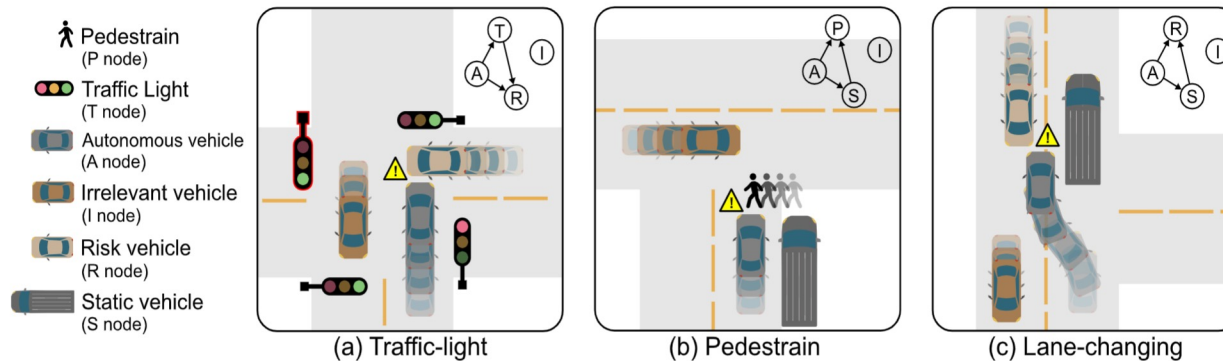


Real-world Case: Autonomous Driving Testing via Logic Reasoning



Knowledge enabled safety-critical traffic scenario generation

- (a) Train T-VAE model to learn the representation of structured data.
(b) Integrate node-level and edge-level knowledge for generation.

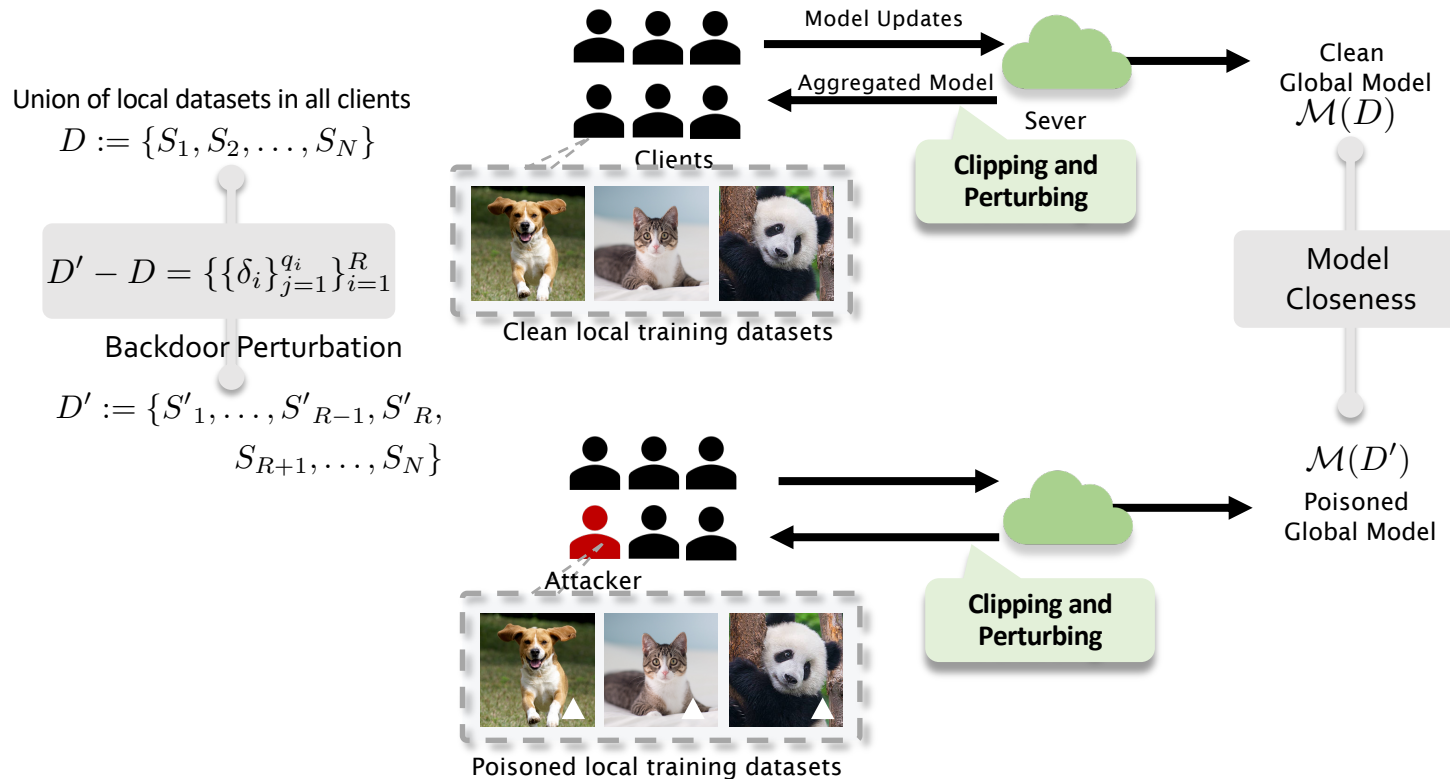


Causal relationship enabled safety-critical traffic scenario generation

The causal graphs are defined in the upper right for the three scenarios.

The generated safety-critical traffic scenarios can significantly improve the test efficiency of autonomous vehicles

Testing-time Adversary: Certifiably Robust FL (CRFL)



Robustness Certification

$$D' - D = \{\{\delta_i\}_{j=1}^{q_i}\}_{i=1}^R \quad \Leftarrow \quad D_f(\mu(\mathcal{M}(D)) || \mu(\mathcal{M}(D'))) \quad \Leftarrow \quad h_s(\mathcal{M}(D); x_{test}) = h_s(\mathcal{M}(D'); x_{test})$$

Backdoor Perturbation Model Closeness Prediction Consistency

General Robustness Condition

$$R \sum_{i=1}^R (p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}} \|\delta_i\|)^2 \leq \frac{-\log(1 - (\sqrt{p_A} - \sqrt{p_B})^2) \sigma_{t_{adv}}^2}{2L_Z^2 \prod_{t=t_{adv}+1}^T (2\Phi(\frac{\rho_t}{\sigma_t}) - 1)}$$

Our certification is in three levels:
feature, sample, and client.

Robustness Condition in Feature Level

$$\|\delta\| < \text{RAD}$$

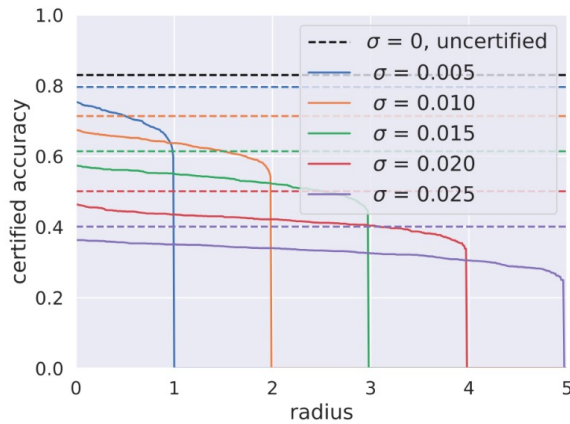
$$\text{RAD} = \sqrt{\frac{-\log(1 - (\sqrt{p_A} - \sqrt{p_B})^2) \sigma_{t_{adv}}^2}{2RL_Z^2 \sum_{i=1}^R (p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}})^2 \prod_{t=t_{adv}+1}^T (2\Phi(\frac{\rho_t}{\sigma_t}) - 1)}}$$

Certified radius

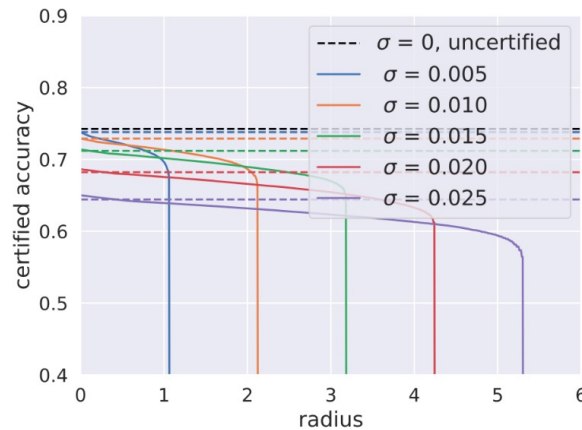
Empirical Results

$$\text{RAD} = \sqrt{\frac{-\log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right) \sigma_{t_{\text{adv}}}^2}{2RL_Z^2 \sum_{i=1}^R (p_i \gamma_i \tau_i \eta_i \frac{q_{B,i}}{n_{B,i}})^2 \prod_{t=t_{\text{adv}}+1}^T \left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1\right)}}$$

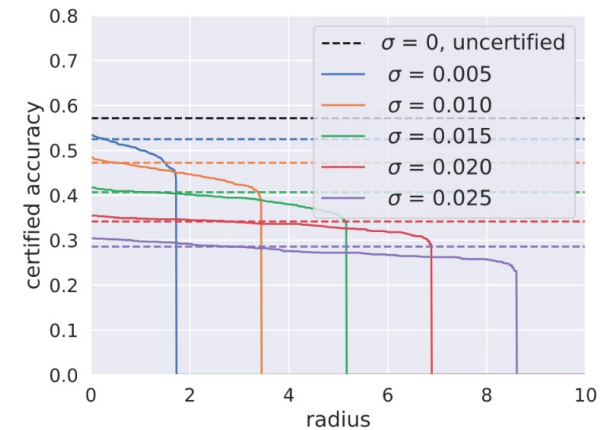
Varying noise levels



MNIST

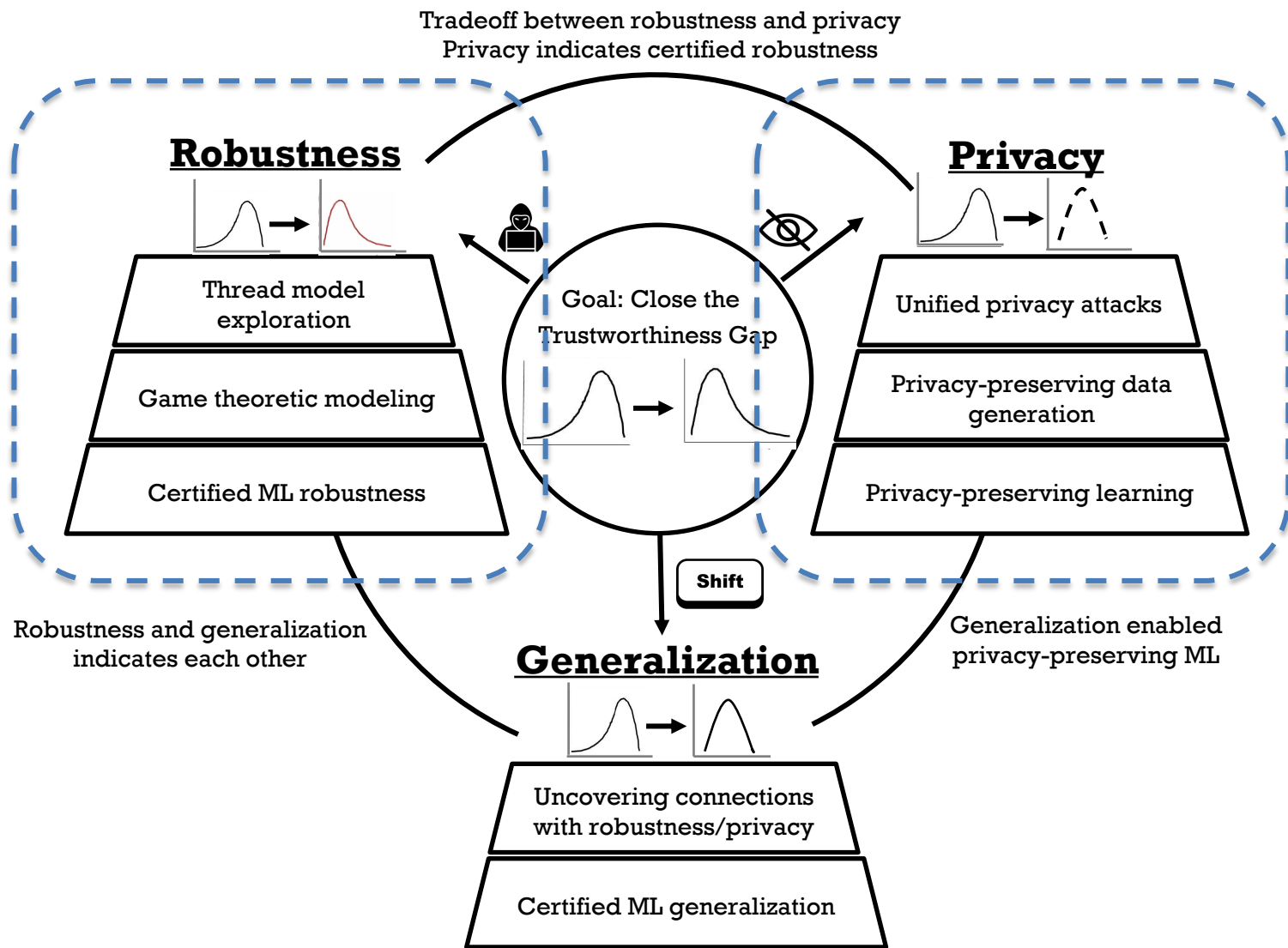


LOAN



EMNIST

- When noise level is large, large radius is certified but at a low accuracy.
- The smoothing noise level control the robustness–accuracy tradeoff.
- Comparing the solid line with the dashed line for each color, we can see that the parameter smoothing does not hurt the accuracy much.



DataLens: Scalable Privacy Preserving Training via Gradient Compression and Aggregation

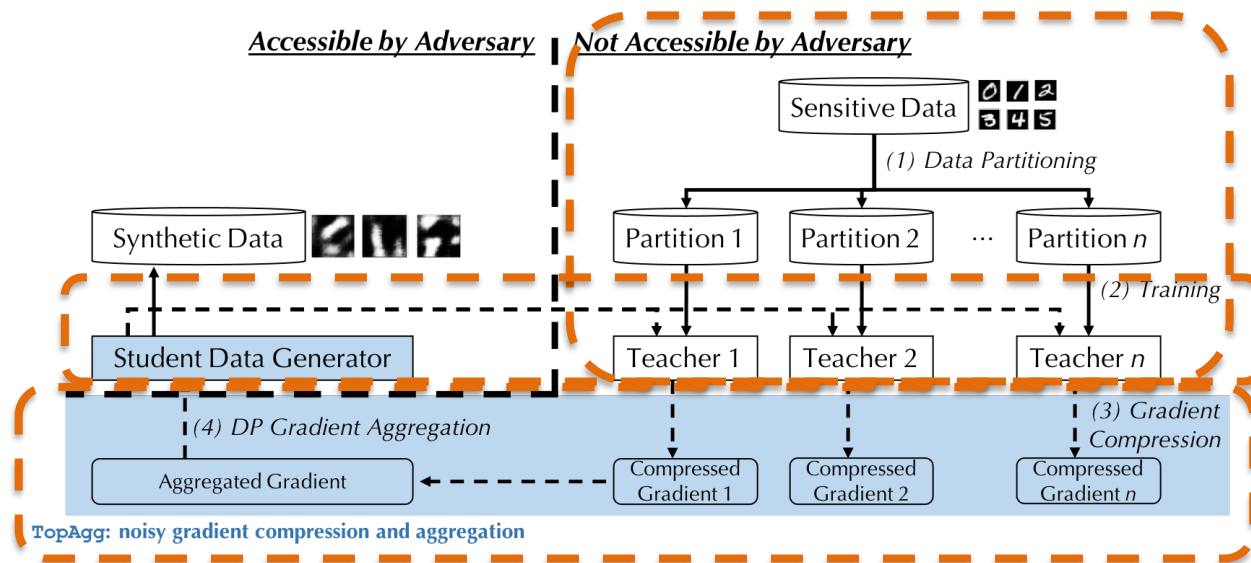
Goal: Differentially private data generative model for high-dimensional data

Overview:

1. Split the sensitive data into non-overlapped partitions to train teacher discriminators
2. Calculate the gradients of the teacher discriminators based on generated data
3. Differentially private gradient **compression** and **aggregation**
4. Train the student generator with the aggregated gradient

High dimensionality

Differential privacy



DataLens –TopAgg: Gradient Compression

- Gradients from different teacher discriminators

$$\mathbf{g}_j \leftarrow (\mathbf{g}_j^{(1)}, \mathbf{g}_j^{(2)}, \dots, \mathbf{g}_j^{(N)})$$

- For each teacher gradient $\mathbf{g}_j^{(i)}$, TopAgg performs Gradient Compression that compresses its dense, real-valued gradient vector into a sparse sign vector with k nonzero entries:
 - 1) Select top- k dimensions, and set the remaining dimensions to 0
 - 2) Clip the gradient at each dimension with threshold c
 - 3) Normalize the top- k gradient vector to get $\hat{\mathbf{g}}_j^{(i)}$
 - 4) Stochastic gradient sign quantization

$$\tilde{g}_j^{(i)} = \begin{cases} 1, & \text{with probability } \frac{1+\hat{g}_j^{(i)}}{2} \\ -1, & \text{with probability } \frac{1-\hat{g}_j^{(i)}}{2} \end{cases}$$

Privacy Bound for DataLens

- At each training step, calculate the data-independent RDP bound

Lemma 1. For any neighboring top- k gradient vector sets $\tilde{\mathcal{G}}, \tilde{\mathcal{G}}'$ differing by the gradient vector of one teacher, the ℓ_2 sensitivity for f_{sum} is $2\sqrt{k}$

Theorem 1. The TopAgg algorithm guarantees $(\lambda, 2k\lambda/\sigma^2) - \text{RDP}$, and thus guarantees $(\frac{2k\lambda}{\sigma^2} + \frac{\log 1/\delta}{\lambda-1}, \delta)$ -differential privacy for all $\lambda \geq 1$ and $\delta \in (0, 1)$

- Calculate the overall RDP by the Composition Theorem.
- Convert RDP to DP.

Convergence Analysis

- Each teacher model performs: $f(x) = \frac{1}{N} \sum_{n \in [N]} F_n(x)$
- Update rule: $x_{t+1} = x_t - \frac{\gamma}{N} \sum_{n \in [N]} (Q(\text{clip}(\text{top-k}(F'_n(x_t)), c), \xi_t) + \mathcal{N}(0, Ak))$

Theorem: (Convergence of top-K Mechanism w/ w/o Gradient Quantization)
after T updates using learning rate γ , one has:

$$\left(\frac{\min\{c, 1\}}{d+2} \right) \frac{1}{T} \sum_{t \in [T]} \min\left\{ \mathbb{E} \|\nabla f(x_t)\|^2, \mathbb{E} \|\nabla f(x_t)\|_1 \right\} \leq \underbrace{\min\{\tau_k M^2, c(d-k)M\}}_{\text{Bias of Top-K compression}} + \underbrace{L\gamma Ak}_{\text{Tradeoff}} + \underbrace{(f(x_0) - f(x^*)) / (T\gamma)}_{\text{DP noise}} + \max\{\|\sigma\|^2 + \|\sigma\|M, 2\|\sigma\|_1\} + 2L\gamma(\tilde{\sigma}^2 + \min\{c^2, M^2\})$$

DP Generated Data Utility

Table 1: Performance of different differentially private data generative models on Image Datasets: Classification accuracy of the model trained on the generated data and tested on real test data under different ϵ ($\delta = 10^{-5}$).

Dataset \ Methods	DC-GAN ($\epsilon = \infty$)	ϵ	DP-GAN	PATE-GAN	G-PATE	GS-WGAN	DataLens
MNIST	0.9653	$\epsilon = 1$	0.4036	0.4168	0.5810	0.1432	0.7123
		$\epsilon = 10$	0.8011	0.6667	0.8092	0.8075	0.8066
Fashion-MNIST	0.8032	$\epsilon = 1$	0.1053	0.4222	0.5567	0.1661	0.6478
		$\epsilon = 10$	0.6098	0.6218	0.6934	0.6579	0.7061
CelebA-Gender	0.8149	$\epsilon = 1$	0.5330	0.6068	0.6702	0.5901	0.7058
		$\epsilon = 10$	0.5211	0.6535	0.6897	0.6136	0.7287
CelebA-Hair	0.7678	$\epsilon = 1$	0.3447	0.3789	0.4985	0.4203	0.6061
		$\epsilon = 10$	0.3920	0.3900	0.6217	0.5225	0.6224
Places365	0.7404	$\epsilon = 1$	0.3200	0.3238	0.3483	0.3375	0.4313
		$\epsilon = 10$	0.3292	0.3796	0.3883	0.3725	0.4875

DataLens achieves the state-of-the-art data utility on high-dimensional image datasets

Data Utility (small privacy budget)

- $\epsilon \leq 1$

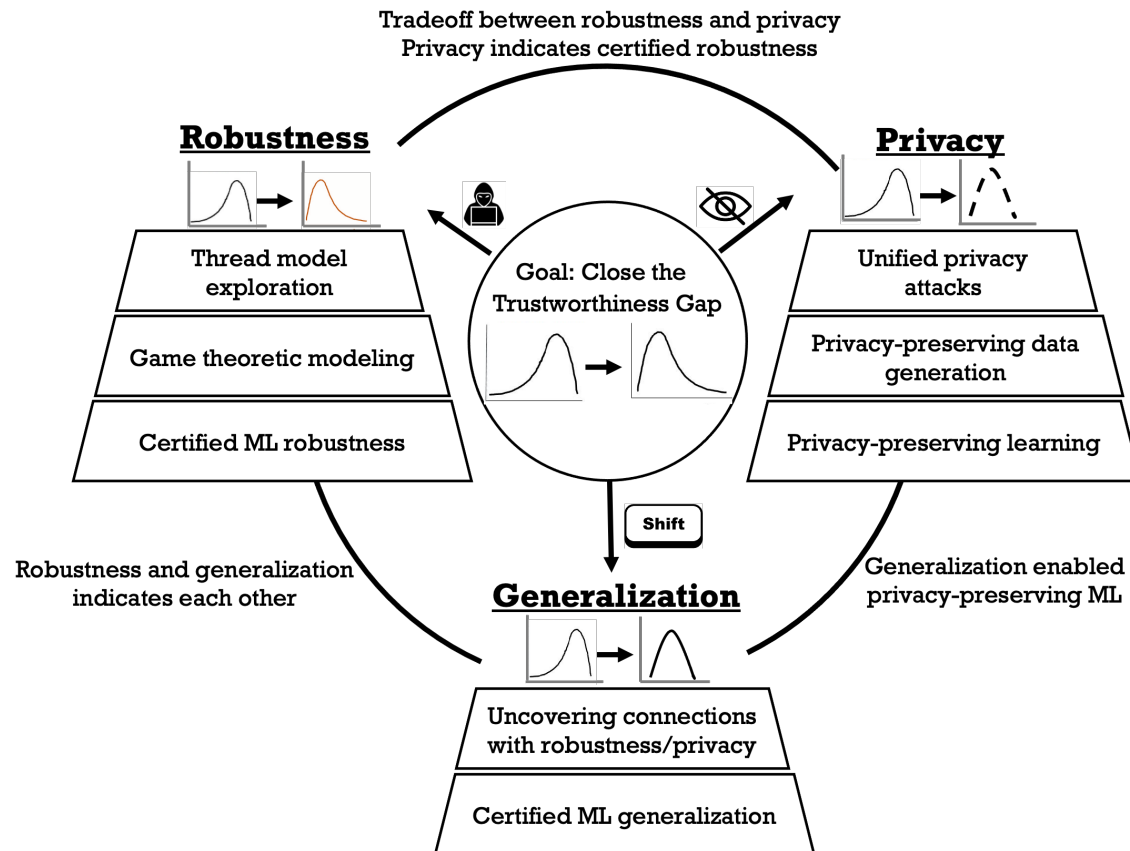
Table 2: Performance Comparison of different differentially private data generative models on Image Datasets under small privacy budget which provides strong privacy guarantees ($\epsilon \leq 1$, $\delta = 10^{-5}$).

ϵ	MNIST					Fashion-MNIST				
	DP-GAN	PATE-GAN	G-PATE	GS-WGAN	DataLens	DP-GAN	PATE-GAN	G-PATE	GS-WGAN	DataLens
0.2	0.1104	0.2176	0.2230	0.0972	0.2344	0.1021	0.1605	0.1874	0.1000	0.2226
0.4	0.1524	0.2399	0.2478	0.1029	0.2919	0.1302	0.2977	0.3020	0.1001	0.3863
0.6	0.1022	0.3484	0.4184	0.1044	0.4201	0.0998	0.3698	0.4283	0.1144	0.4314
0.8	0.3732	0.3571	0.5377	0.1170	0.6485	0.1210	0.3659	0.5258	0.1242	0.5534
1.0	0.4046	0.4168	0.5810	0.1432	0.7123	0.1053	0.4222	0.5567	0.1661	0.6478

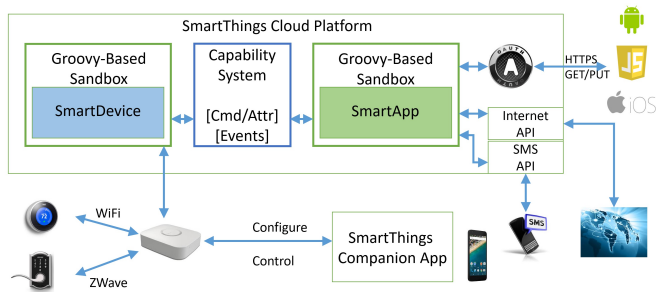
Faster convergence when the privacy budget is small



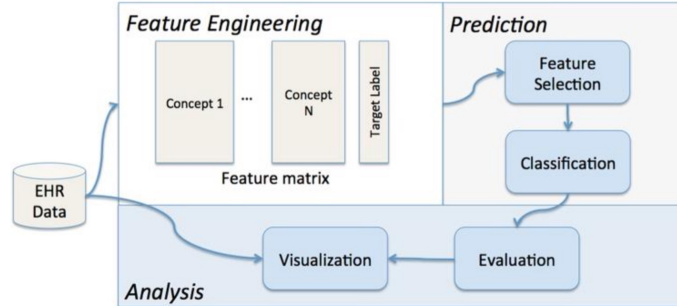
Summary



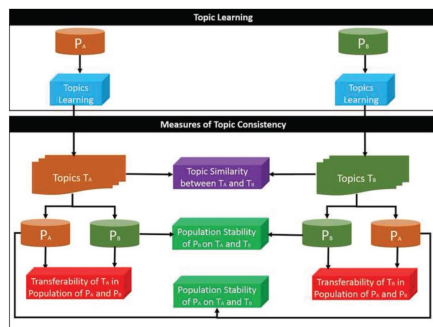
Closing today's trustworthiness gap requires us to tackle these three grappled problems in a holistic framework, driven by fundamental research focusing on not only each problem but more importantly their interactions.



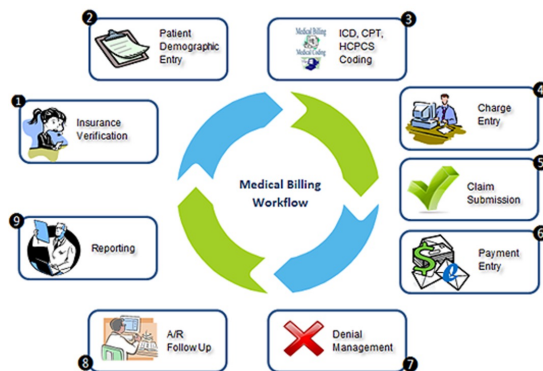
Robust Smart Home



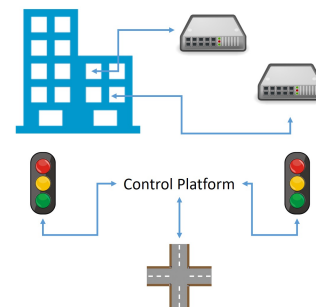
Privacy-Preserving Data Analysis



Topic of Workflow Analysis



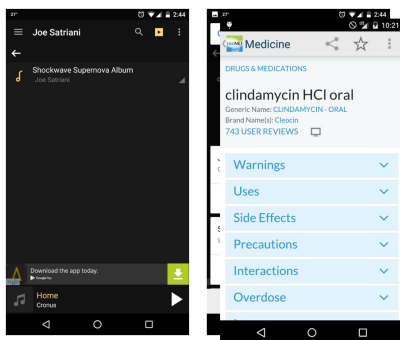
Game Theoretic Auditing System for EMR



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/>