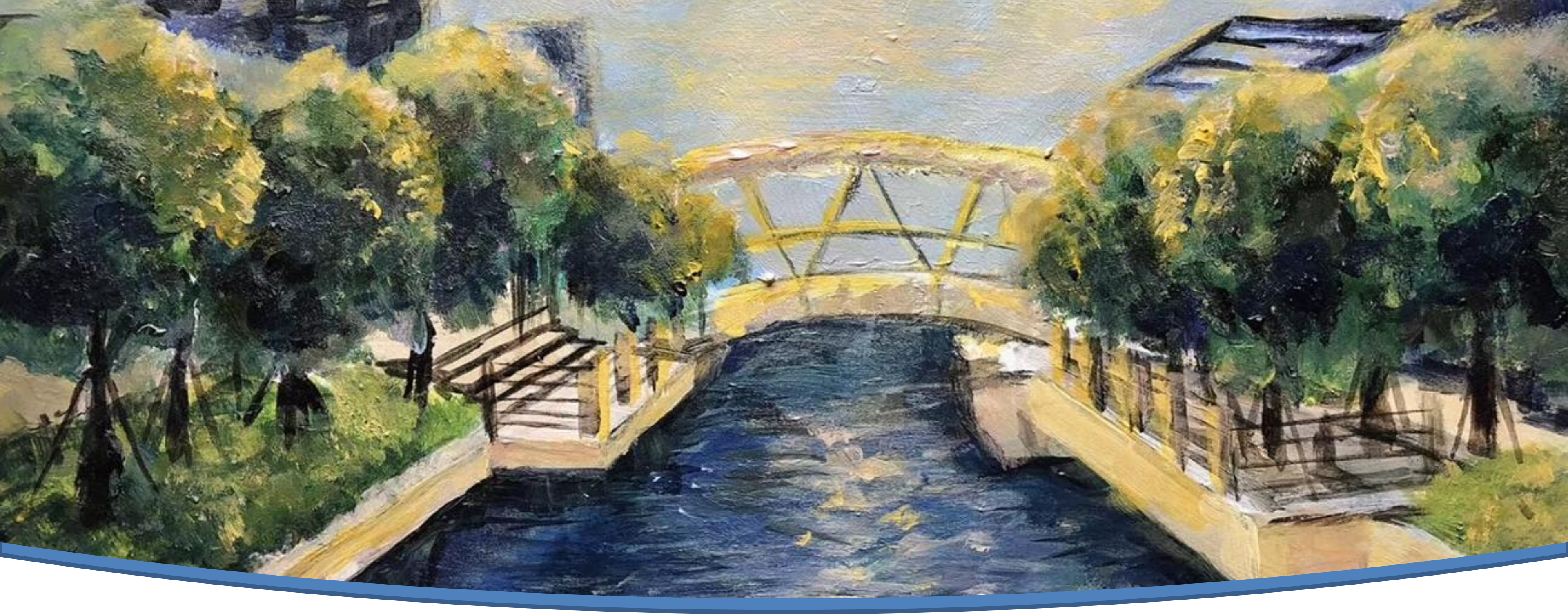


CS 253 Cyber Security
Automotive & Adversarial Learning

ShanghaiTech University

Admin

- Project 3 is released. Deadline 12/17 (Wed) 23:59.
- Guide and requirements can be found on course website (PDF) or Piazza Resrouces (MD). MD works better because it is copy-able.
- Don’t worry if you cannot finish everything. Points are given by how much you have completed.
- Do NOT cheat by copying code of others! If found, you will receive 0 points for Project 3 and one grade down for the overall course grade.

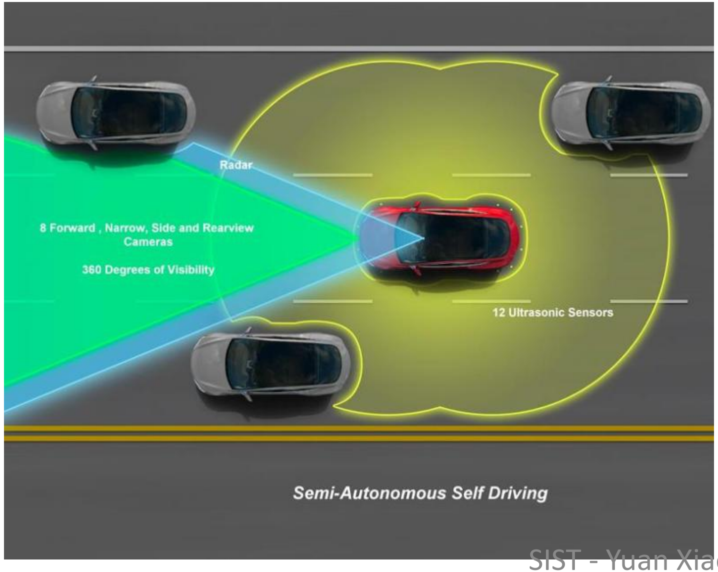
In The News: Crashing of Self-Driving Car (Uber 2018)

- “Inadequate safety risk assessment procedures”
- The system is not trained to react to pedestrians crossing the street outside of designated crosswalks
- Vehicle operator distracted by personal cellphone

https://www.theverge.com/2019/11/19/20972584/uber-fault-self-driving-crash-ntsb-probable-cause

In The News: Crashing of Self-Driving Car (Tesla 2016)

- Car's cameras failed to pick out a **white trailer** against a **bright sky** in Florida



01

PART ONE

Adversarial Examples to Attack Vision Sensors

Perils of Stationary Assumption

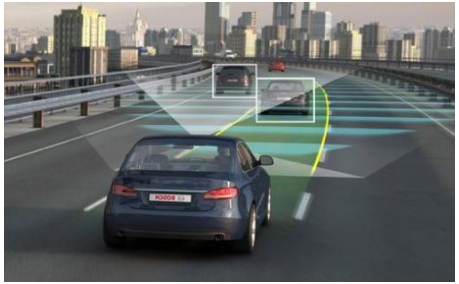
Traditional machine learning approaches assume

Training Data 

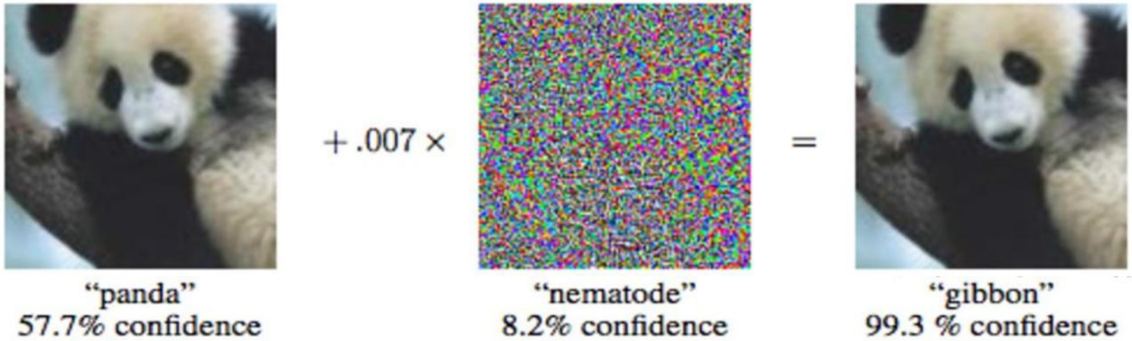
≈

Testing Data 

Autonomous Driving in Practice



Adversarial Examples



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *ICLR 2014.*

Adversarial Perturbation In ML

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

Model parameters

Input feature vector

label

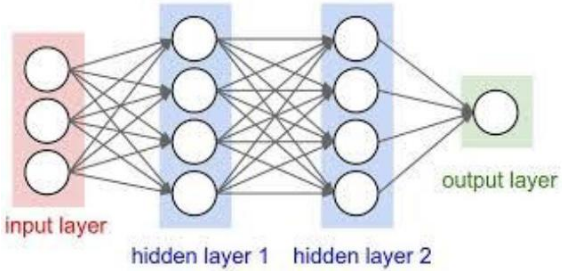
$$\max_{\epsilon} J(\theta, x + \boxed{\epsilon}, y)$$

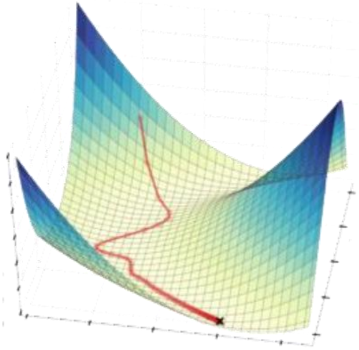
Adversarial perturbation

How to solve the adversary strategy

- Local search
- Combinatorial optimization
- Convex relaxation

Deep Neural Networks





Gradient Descent

Optimization Based Attack

Large probability of x+δ belonging to
a target class t

Small perturbation

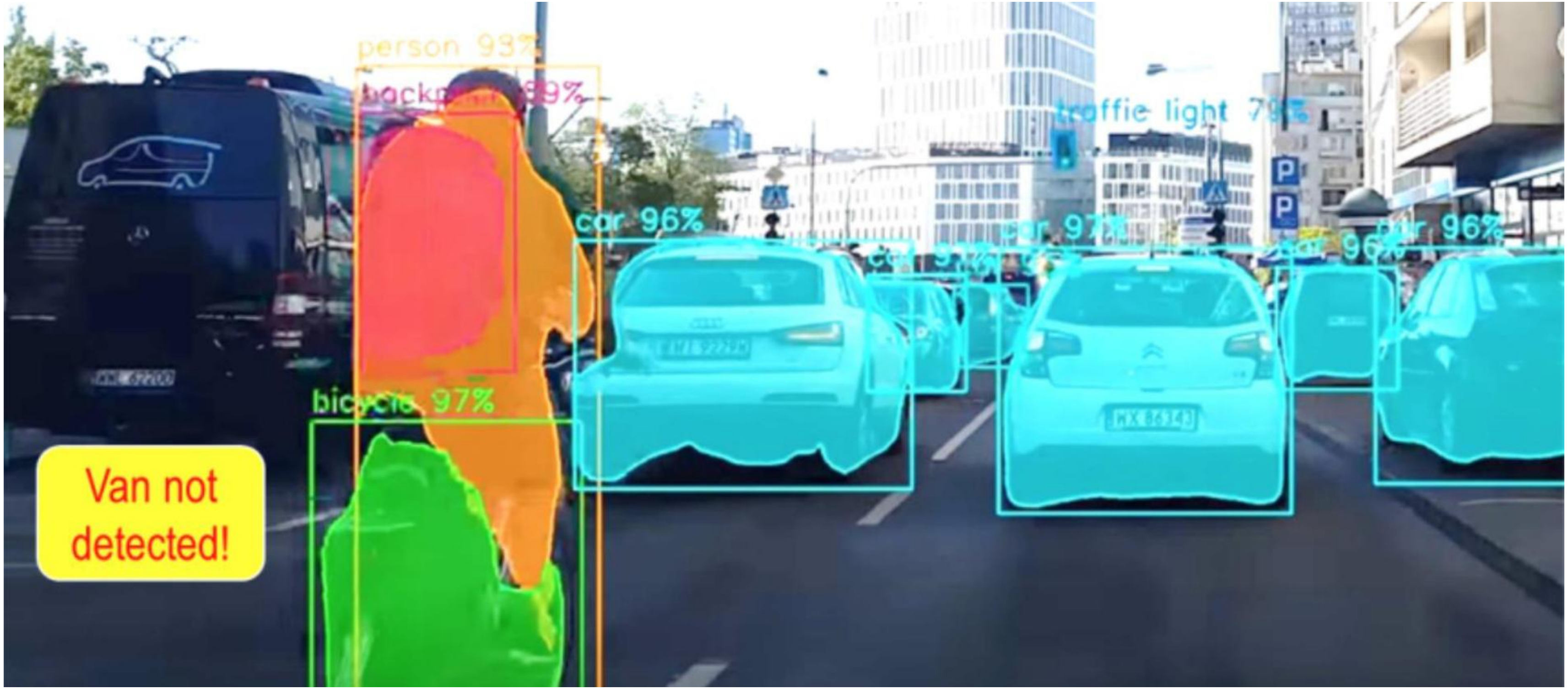
$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & C(x+\delta) = t \\ & x+\delta \in [0,1]^n \end{aligned}$$

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \\ \text{such that} \quad & x+\delta \in [0,1]^n \end{aligned}$$

	Best Case				Average Case				Worst Case			
	MNIST		CIFAR		MNIST		CIFAR		MNIST		CIFAR	
	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob
Our L_0	8.5	100%	5.9	100%	16	100%	13	100%	33	100%	24	100%
JSMA-Z	20	100%	20	100%	56	100%	58	100%	180	98%	150	100%
JSMA-F	17	100%	25	100%	45	100%	110	100%	100	100%	240	100%
Our L_2	1.36	100%	0.17	100%	1.76	100%	0.33	100%	2.60	100%	0.51	100%
Deepfool	2.11	100%	0.85	100%	—	-	—	-	—	-	—	-
Our L_∞	0.13	100%	0.0092	100%	0.16	100%	0.013	100%	0.23	100%	0.019	100%
Fast Gradient Sign	0.22	100%	0.015	99%	0.26	42%	0.029	51%	—	0%	0.34	1%
Iterative Gradient Sign	0.14	100%	0.0078	100%	0.19	100%	0.014	100%	0.26	100%	0.023	100%

[Carlini, Wagner, Towards robustness of neural networks. 2017]

Vulnerabilities of Perceptron Systems of Automobiles



Robust Physical-World Attacks on Deep Learning Visual Classification. CVPR, 2018

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

[Evtimov, Eykholt, Fernandes, Kohno, Li, Prakash, Rahmati, and Song, *2017*]

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-2, …)

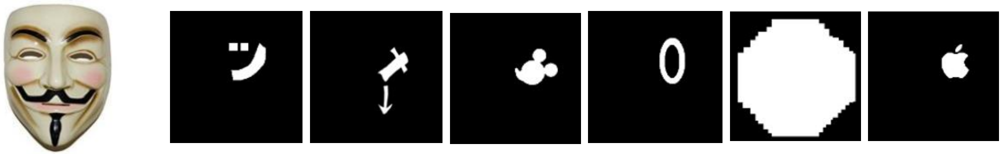
Loss Function

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"

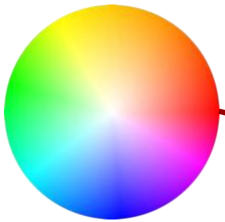
Handling Fabrication/Perception Errors

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

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

P is a set of printable RGB triplets

Color Space

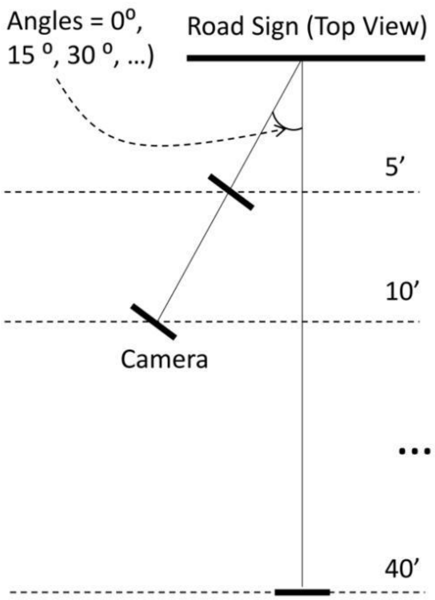


Sampled Set of RGB Triplets

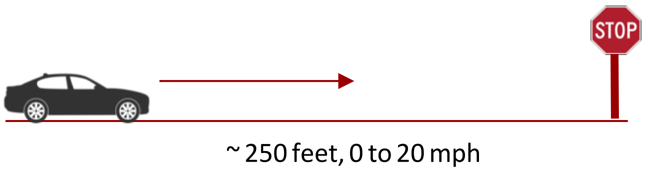
NPS based on Sharif et al., “Accessorize to a crime,” CCS 2016

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

Subtle Poster | Subtle Poster | Camo Graffiti | Camo Art | Camo Art

Lab Test (Stationary)

Target Class:

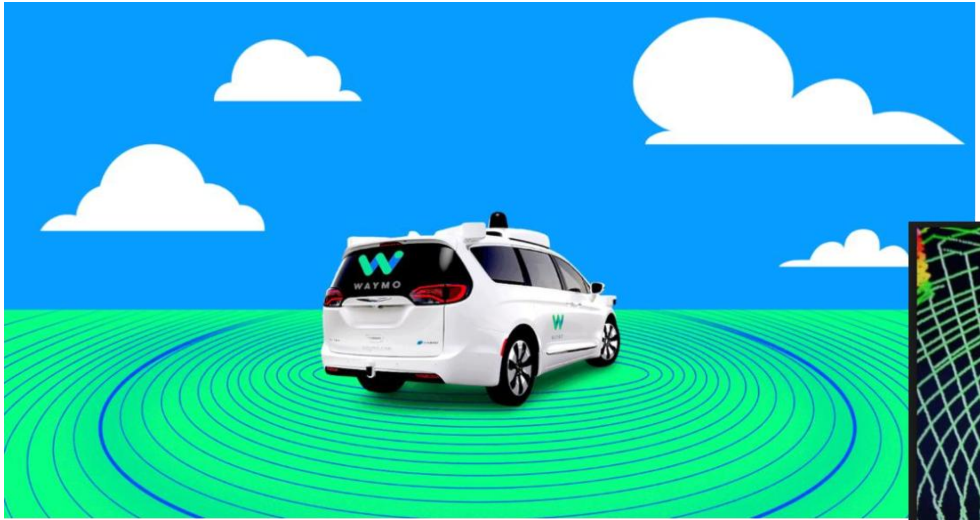
Speed Limit 45

Art Perturbation

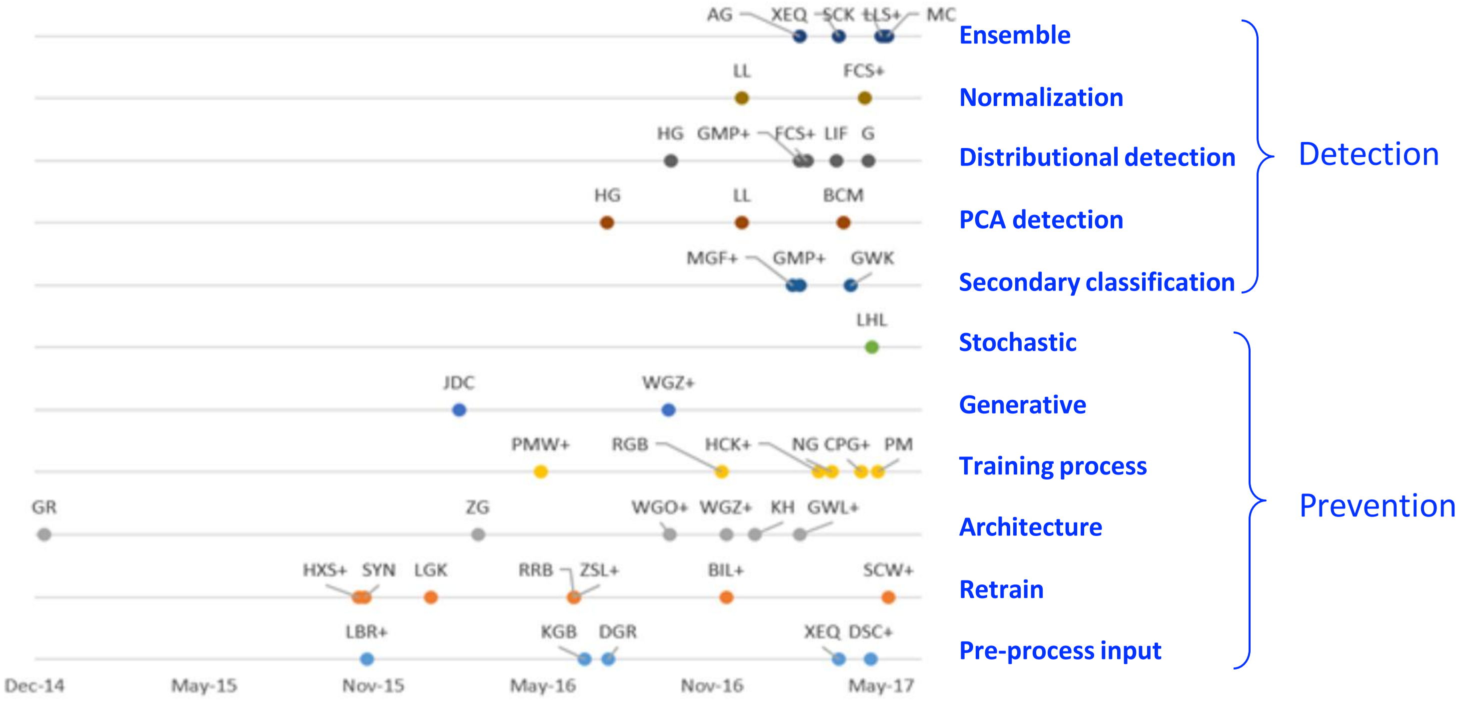
Subtle Perturbation

Thinking more about Physical objects

Similar attack against LiDAR sensors



Numerous Defenses Proposed



02

PART TWO

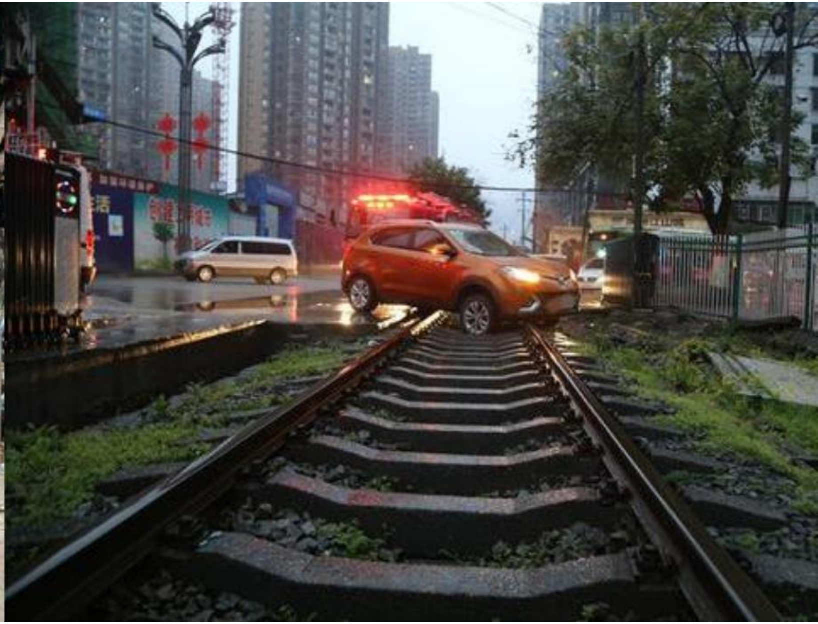
Attacking GPS Sensors

GPS Navigation Systems used by 1+Billion Users

- GPS navigation is widely used by drivers around the world

- Self-driving cars rely on GPS for navigation and on-road decisions

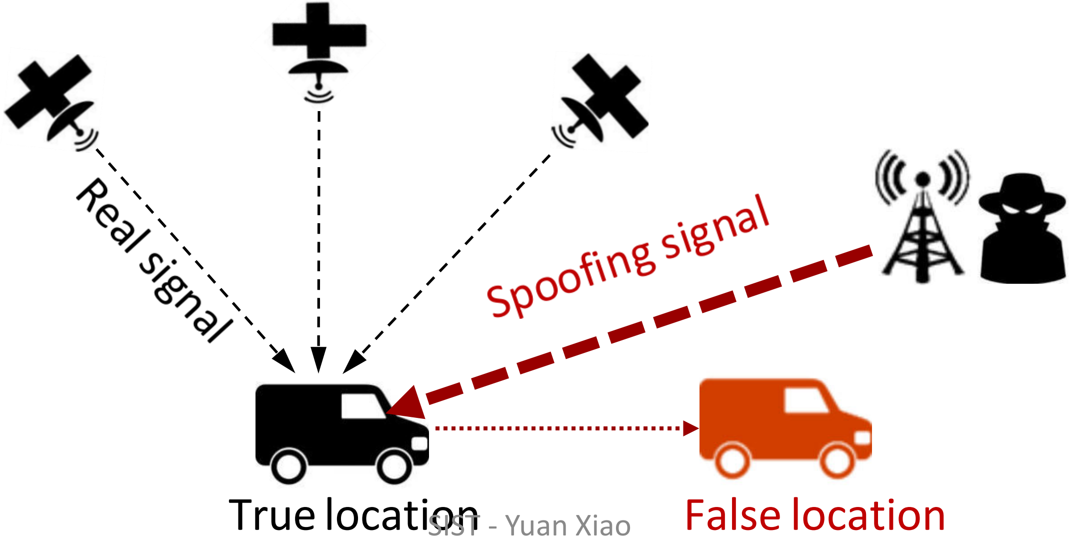
GPS Navigation Systems used by 1+Billion Users



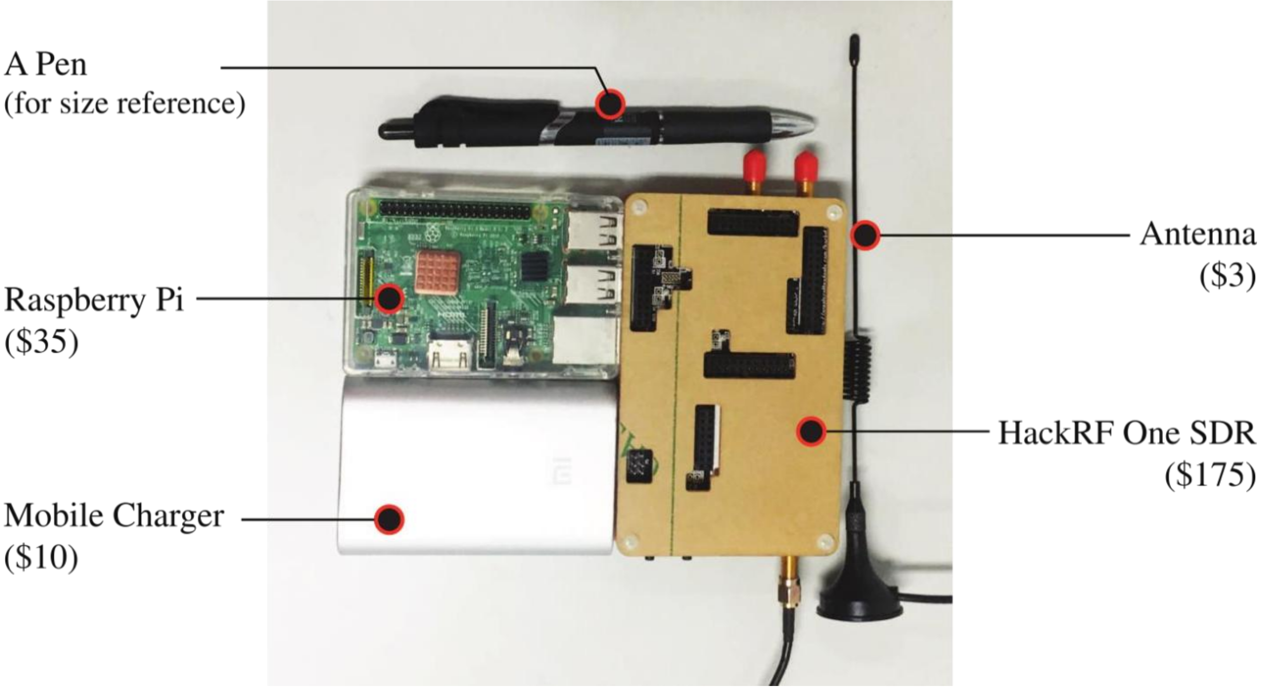
GPS malfunction can lead to real-world consequences

Known Threat: GPS Spoofing

- Civilian GPS is vulnerable to spoofing attacks
  - A lack of authentication of signal source
- Take over victim GPS via brute-force jamming or smooth methods



Portable GPS Spoofer is Affordable ($223)



GPS Spoofing in Free Space (Air, Water)

In 2012, a drone was diverted in White Sands, New Mexico

In 2013, a yacht was diverted on the way from Monaco to Greece

Spoofing Road Navigation: More Challenging

Physical World 

Digital Map 

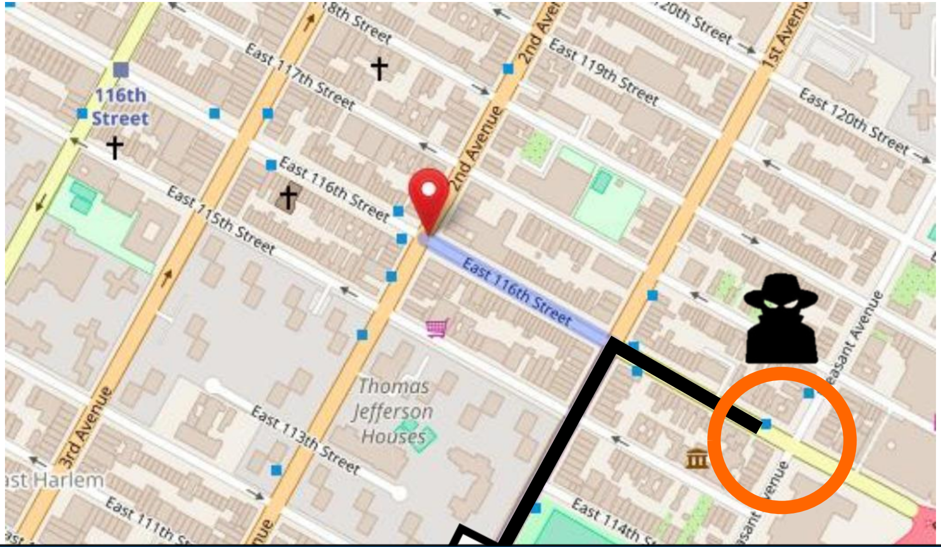
“Turn left” - physically impossible instruction!

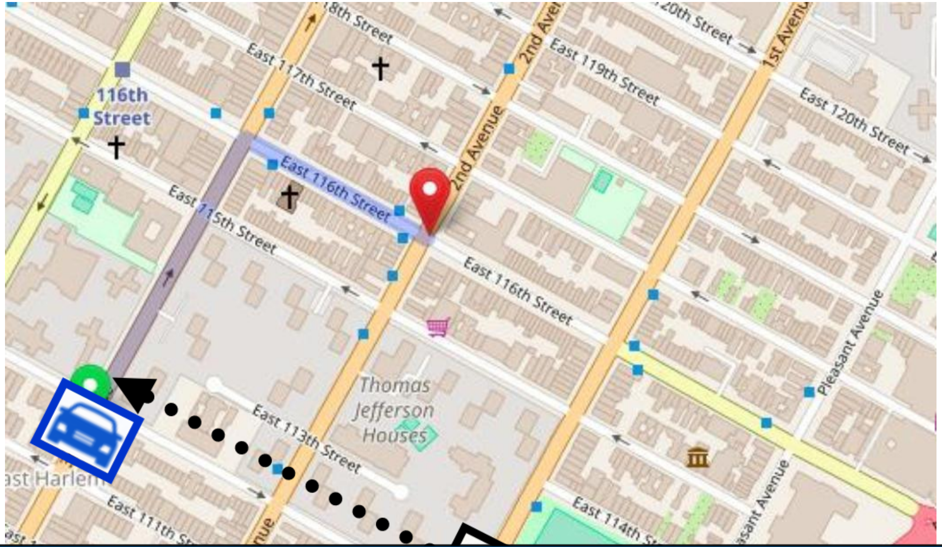
Easily alert human drivers

Making the Attack More Stealthy

Physical World 

Digital Map 

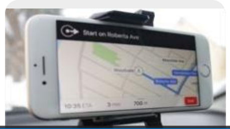




Navigation instructions lead to attacker's pre-defined location

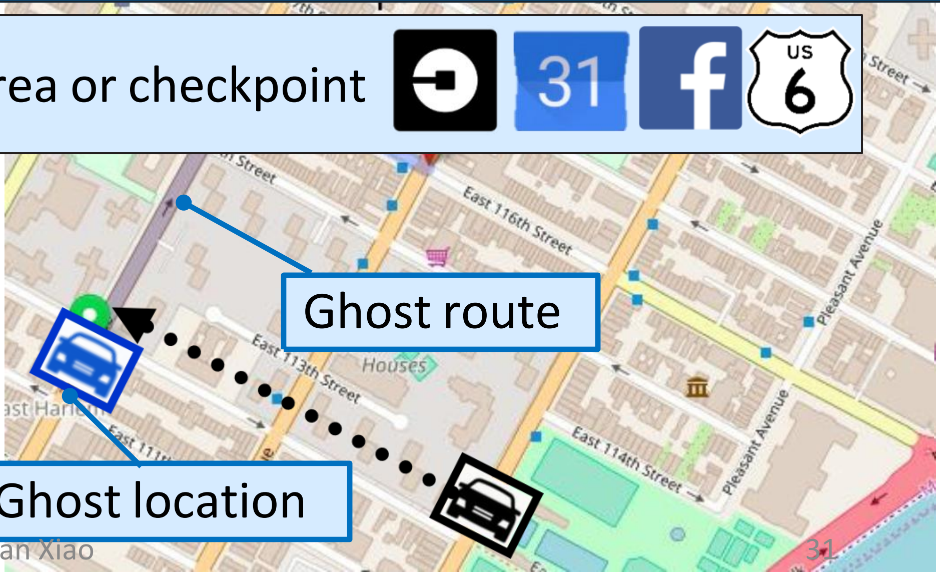
Core Idea: Calculate Spoofing Location and Timing

Physical World 

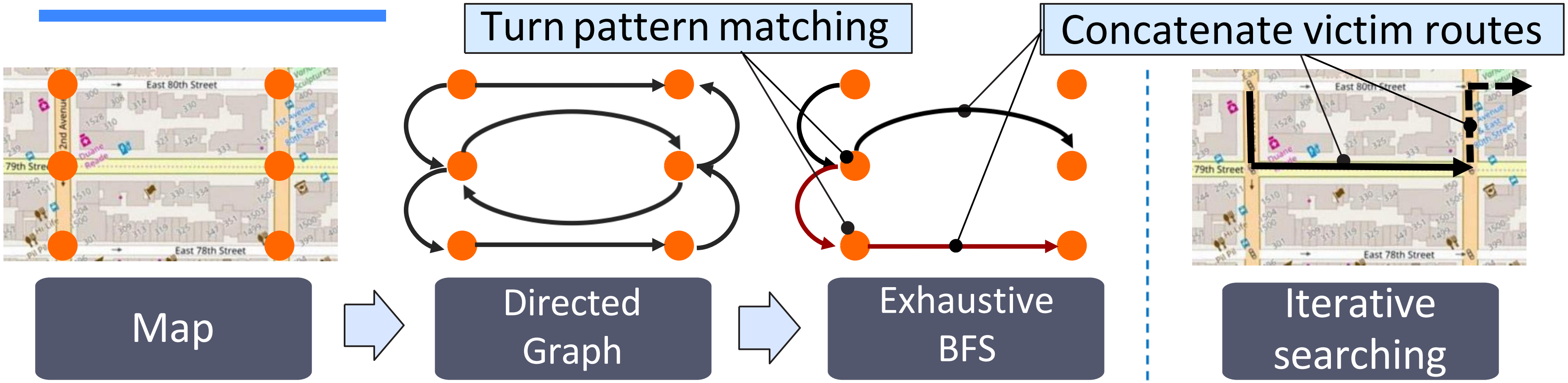
Digital Map 

Goal: find ghost route to mimic the shape of victim route

Assumption: know rough destination area or checkpoint



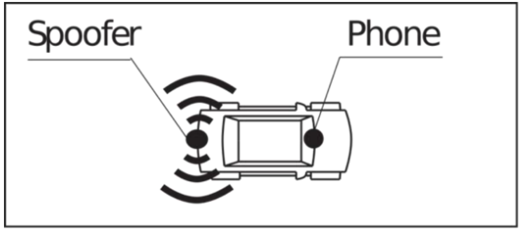
Route Searching Algorithm



Trace Driven Evaluation

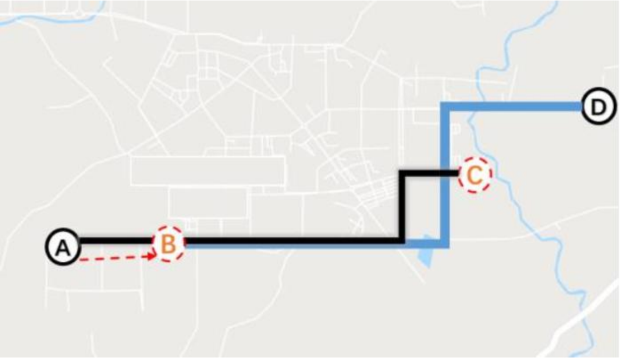
- 600 real-world trips from the taxi datasets of New York City and Boston
- Deviating attack: 3,507 qualified victim routes per trip
- Endangering attack: 599 out of 600 (99.8%) contains wrong-way

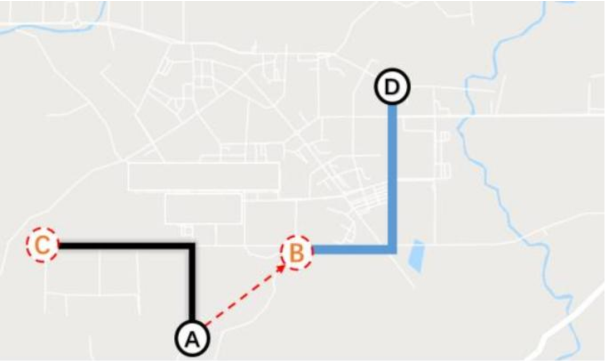
Real-world Experiments



- Experiments with legal permission from local authority and IRB approval
  - After midnight, spoofing signals do not affect outside (-127.41 dBm)

Trigger instructions in time and divert to 2.1 & 2.5 km away





Can Human Users Detect the Attack?

- Let users drive in a simulator
  - Play truck drivers to “deliver packages” from location A to B
  - Advertise the study as a usability study, spoof locations in real time
  - Post-study interview to know why users can/cannot detect the attack

Experiment setup

Simulator view

Google Street View

User Study Results

- Attack success rate: 95% (38 out of 40)
  - Two users detect it by cross-checking surrounding environment and the map to find inconsistency (Highway vs. local way)
- Users are more likely to use GPS in unfamiliar areas
  - Not enough pre-knowledge/time to check the inconsistency
  - Heavily rely on voice and turn-by-turn instructions
- Most users experienced GPS malfunction in real life
  - Unstable GPS signal does not alert users

Take-aways: Learning from Users

- It is feasible to manipulate road navigation systems
  - Advanced GPS spoofing strategies
  - Works even when humans are in the loop

- Defense ideas inspired by the user study
  - Cross-check data from digital and physical worlds
    - Computer vision-based localization
  - GPS-free localization & navigation