# Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations


Pengfei Jing, *The Hong Kong Polytechnic University and Keen Security Lab, Tencent;*
Qiyi Tang and Yuefeng Du, *Keen Security Lab, Tencent;*
Lei Xue and Xiapu Luo, *The Hong Kong Polytechnic University;*
Ting Wang, *Pennsylvania State University;* Sen Nie and Shi Wu, *Keen Security Lab, Tencent*



30TH USENIX SECURITY SYMPOSIUM


Presenter: Hyunmin Ju
14 Oct 24

SYSSEC

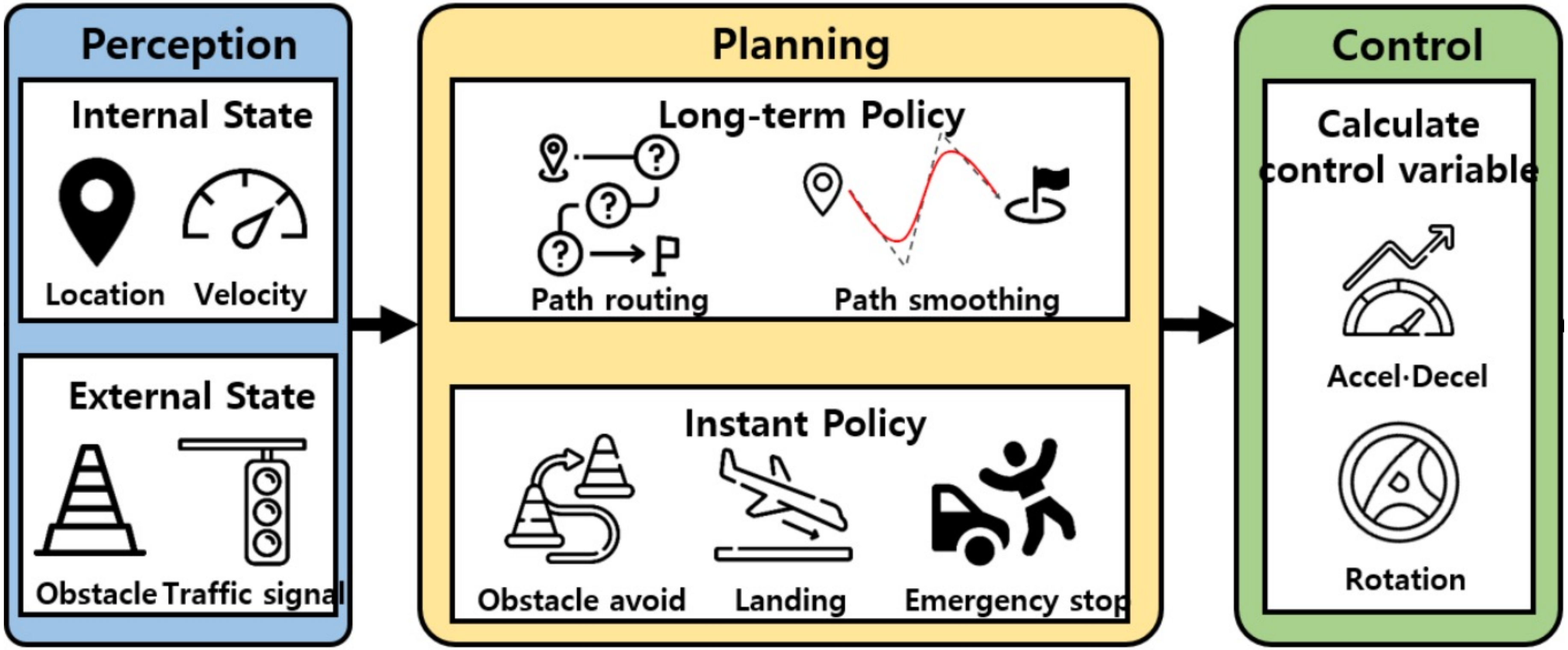
# Motivation

Autonomous driving system is still SAFETY-CRITICAL!


May 2020


Apr 2024

# Motivation


Perception
Internal State
Location
Velocity
External State
Obstacle
Traffic signal
Planning
Long-term Policy
Path routing
Path smoothing
Instant Policy
Obstacle avoid
Landing
Emergency stop
Control
Calculate control variable
Accel·Decel
Rotation

# Introduction

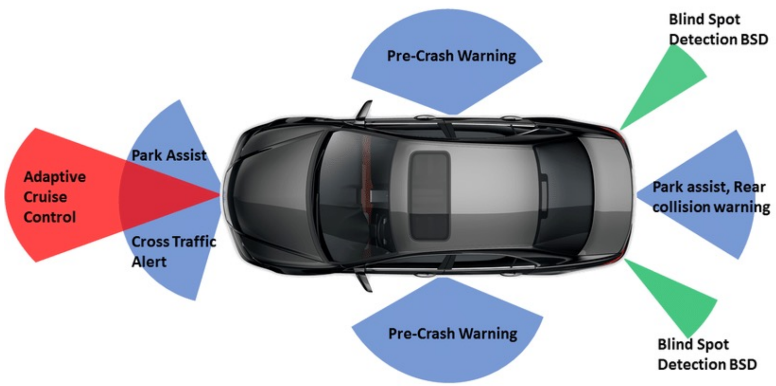
- **Tesla Autopilot (Target of the Attack)**

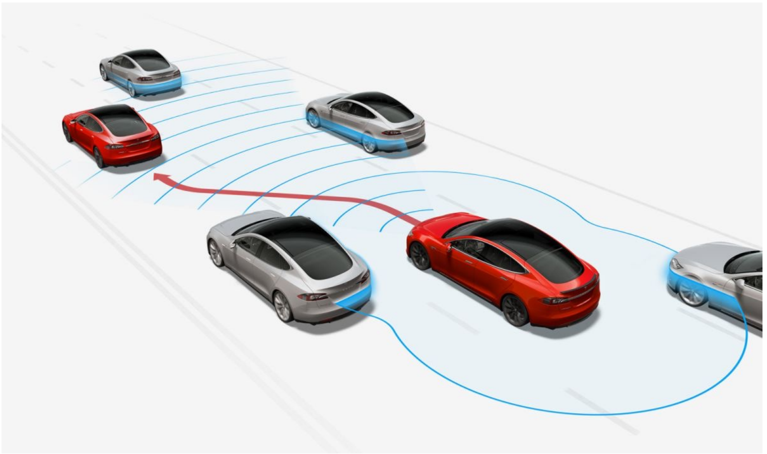
- **Advanced Driver-Assistance System (ADAS)**

  - Assists with steering, braking, and lane changing

- **Target service:  Lane Detection**

  - Directly influences steering by identifying and tracking lanes



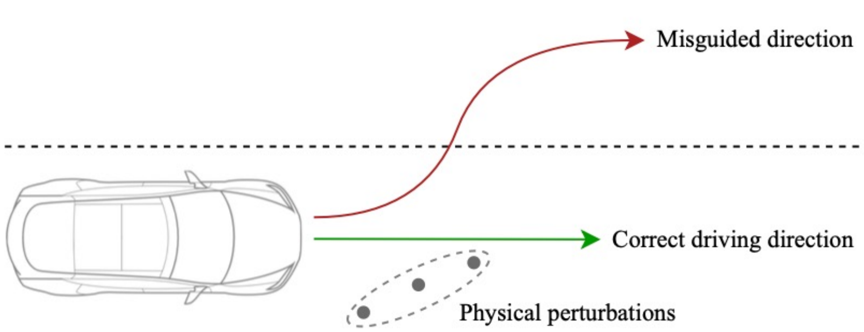



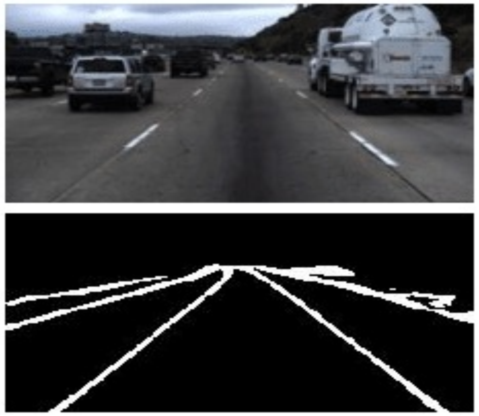
# Introduction

- **Lane detection module Vulnerability**: Over-sensitivity to subtle road stickers , misinterpreting them as valid lanes
- **Attack Goal**: Changing the lane detection result to misdirect the autonomous vehicle
- **Target**: Tesla Autopilot's lane detection module in auto-steering mode
- **How:**

    1) Reverse engineering on the firmware

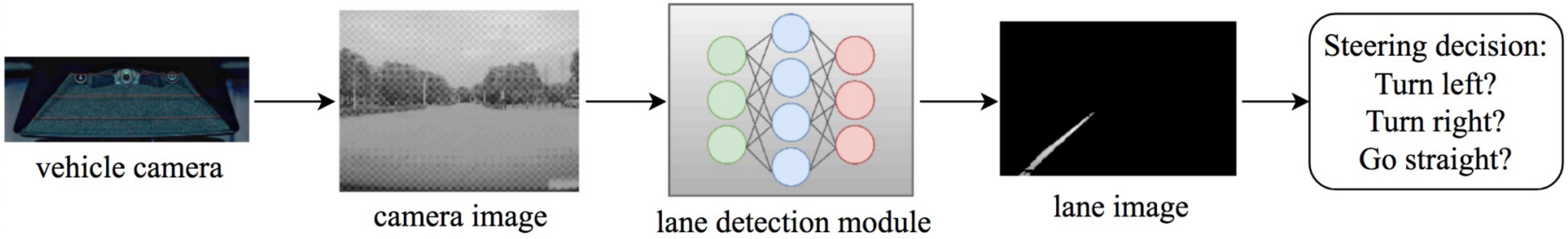
    2) Use optimal perturbation for creating fake lanes



# Background

❖ **Lane Detection Module**




1) Images are collected by **camera**
2) Lane detection module generates **corresponding lanes**
3) Autonomous vehicle behaves based on the **lane detection results**

**Changing the lane detection result can affect the steering decision**

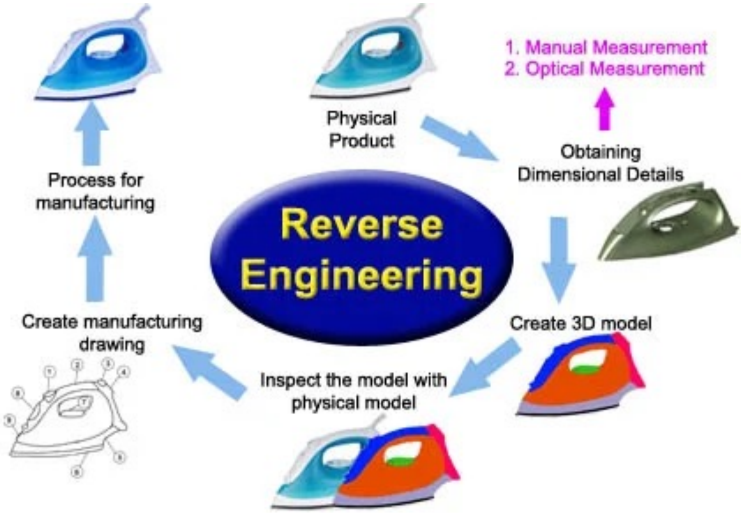
**(i.e., Exploiting its over-sensitivity to create a fake lane!)**

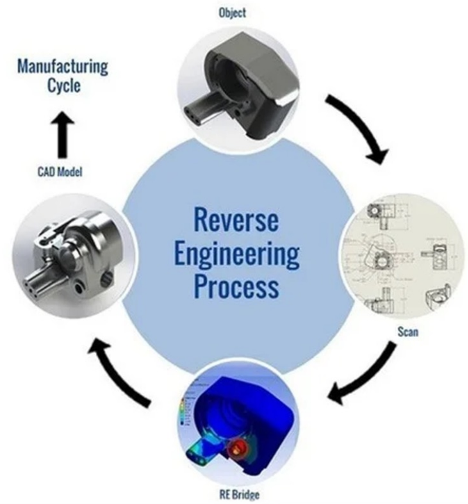
# Background

- **Reverse Engineering**

  - The act of breaking down  an object to see how it works

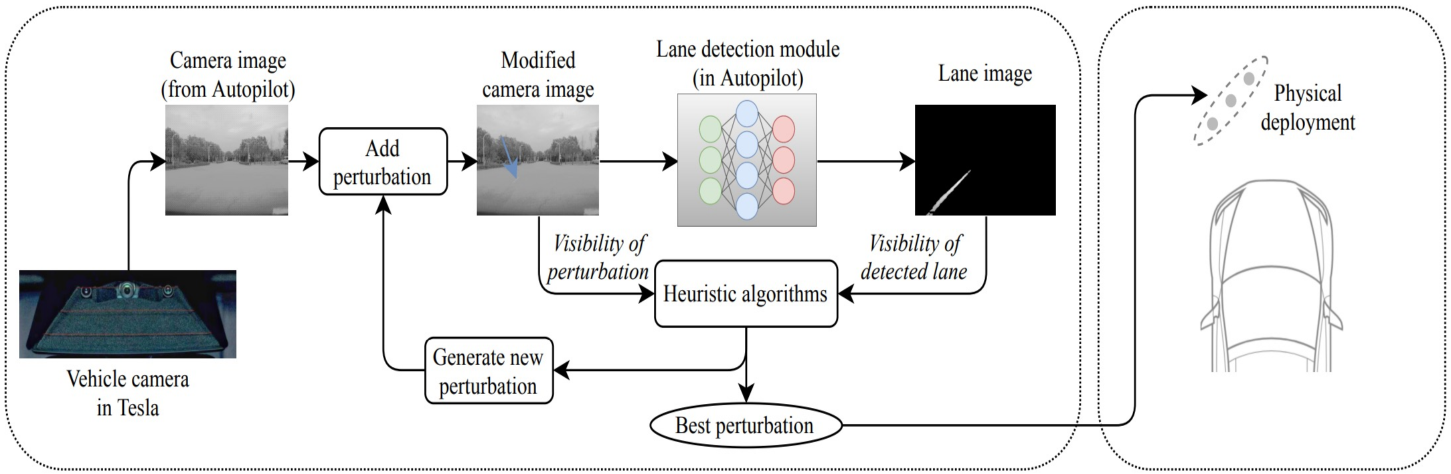
- **In this paper**: Analyzing Tesla's firmware to understand Autopilot's camera input and lane detection output



# Contribution

- Reveal that the **lane detection module's sensitivity** can be exploited to generate fake lanes
- **Reverse engineering** of Tesla Autopilot firmware and propose **a two-stage attack approach**
- Conduct **extensive experiments** on a Tesla vehicle (Tesla Model S)




# Threat Model

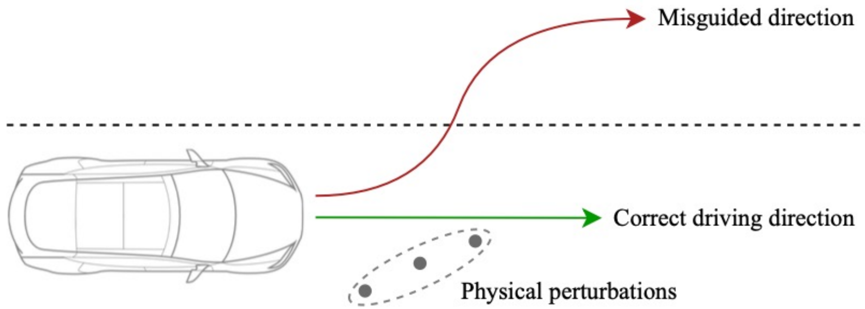
- Attacker has an autonomous vehicle with identical lane detection module. (black-box setting)
- Attacker aims to add **subtle marking** on the ground.

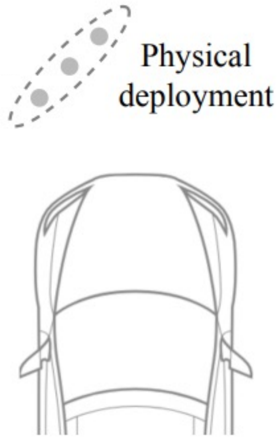
  - Change the position and the shape, and Repeat the try-and-error method

  -> labor-intensive and error-prone ☹

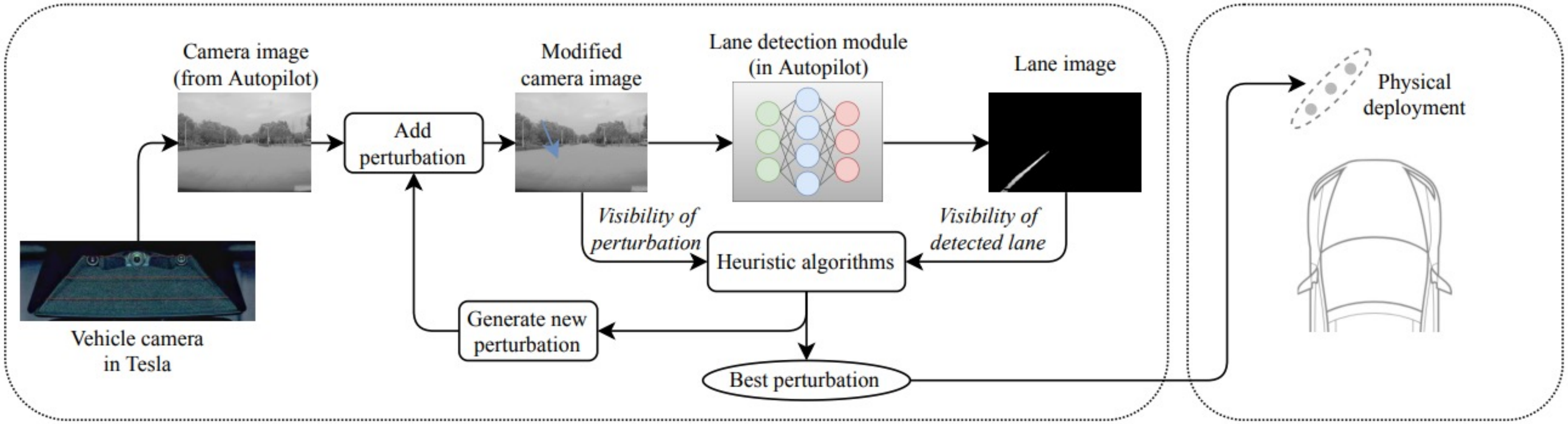
→ **Two-Stage Attack Approach**:

Efficient method to automate attack, overcoming limitations ☺



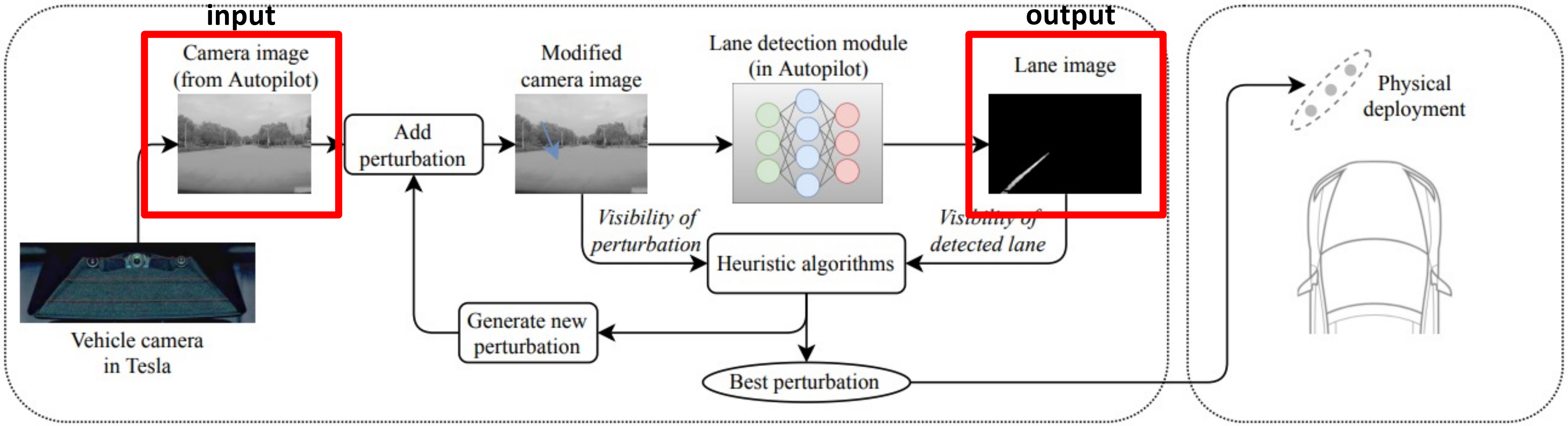



# Two-Stage Attack Approach




- **PRE-PROCESSING:** locate the input camera image to the lane detection module and the corresponding output lane image
- **Stage 1)** Finding the best digital perturbation
- **Stage 2)** Deployment in physical world

# Challenges and Solutions Ⅰ




**CHALLENGE 1:** How to locate the input camera image and output lane image in the vehicle?


SysSec
System Security Lab

# Challenges and Solutions Ⅰ

**SOLUTION**

**->** Reverse engineer the firmware of Tesla Autopilot

**1) Firmware under examination**

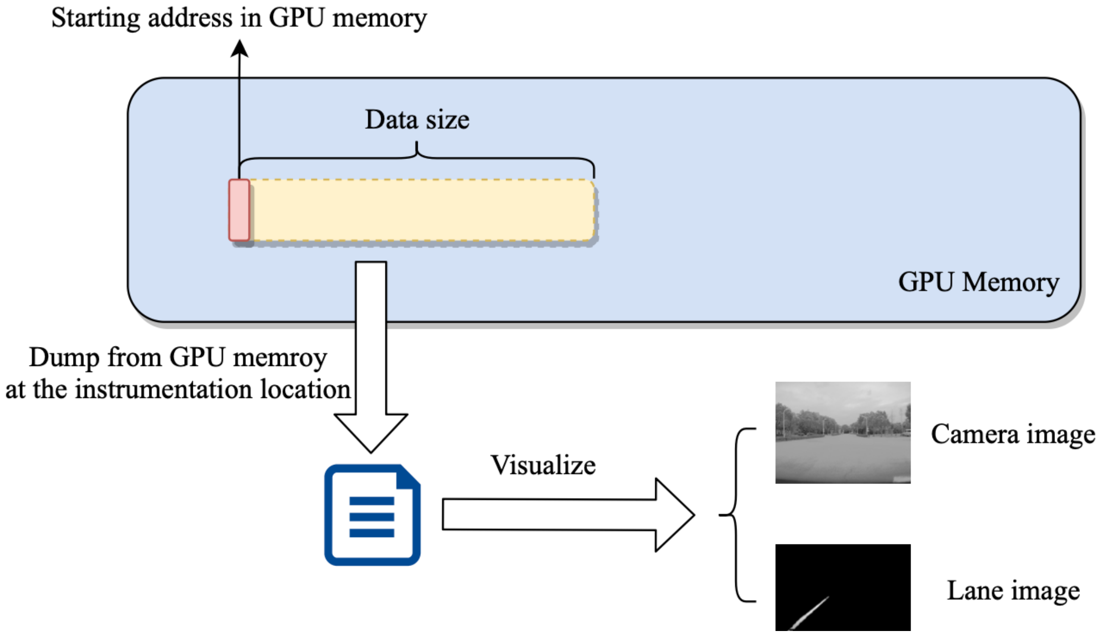
- Tesla Model S, with the Autopilot hardware version of 2.5 and software version of 2018.6.1.
- Running an AArch64 Linux OS and uses NVDIA GPU for deep learning computation.

**2) CUDA**

- Memory management functions: *cudaMalloc, cudaMemcpy*

NVIDIA CUDA

# Challenges and Solutions Ⅰ




**SOLUTION**

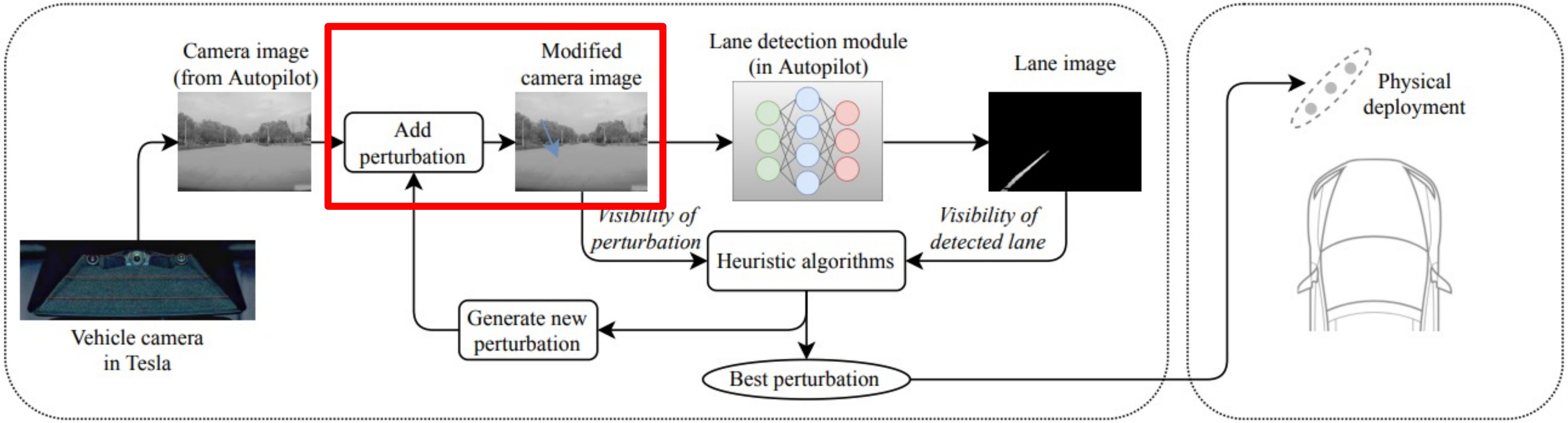
**3) Static and Dynamic analysis**

Find (1) source address, (2) destination address, (3) data size, and (4) mode of transfer



SysSec
System Security Lab

# Challenges and Solutions II




**CHALLENGE 2:** How to add perturbations to input camera image, considering the distortion of the lens?



SysSec
System Security Lab

# Challenges and Solutions II

| Parameters | Explanation |
|---|---|
| $len$ | Length of a single perturbation |
| $wid$ | Width of a single perturbation |
| $D_1$ | Longitudinal distance from the vehicle camera to the edge of the first perturbation |
| $D_2$ | Lateral distance from the vehicle camera to the edge of the first perturbation |
| $D_3$ | Distance between adjacent perturbations |
| $\Delta G$ | Increment of grayscale value of the perturbed pixels |
| $\theta$ | Rotation angle of the perturbation |
| $n$ | Number of the perturbations |

Parameters determining the added perturbation

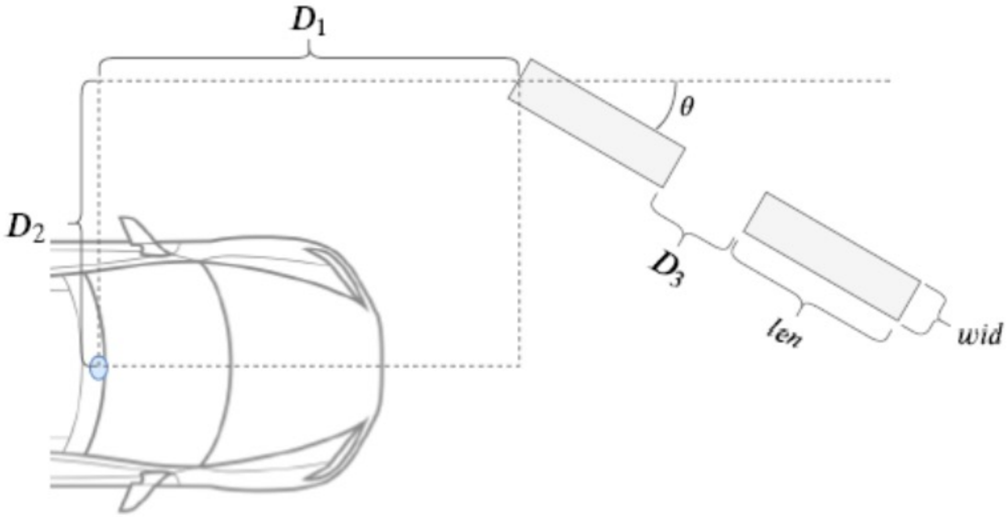


Illustration of the parameters

**SOLUTION:** Using a vector containing metrics from the physical world to represent the perturbations in digital world.

# Challenges and Solutions III

| Parameters | Explanation |
|---|---|
| $p$ | One single pixel in the image |
| $lane_o(x)$ | Lane pixels in the output image |
| $perturb_i(x)$ | Pixels on the added perturbations |
| $G_p$ | Grayscale value of pixel $p$ |
| $V_{lane}(x)$ | Visibility of the fake lane created by $x$ |
| $V_{perturb}(x)$ | Visibility of the perturbations added by $x$ |
| $S(x)$ | Overall score of the parameter $x$ |

1) $V_{lane}(x) = \sum_{p \epsilon lane(x)} G_p$

2) $V_{perturb}(x) = \sum_{p \epsilon perturb(x)} \Delta G$

3) $S(X) = \frac{V_{lane}(x)}{V_{perturb}(x)}$

4) $x^* = \max_{x \epsilon X} S(x)$

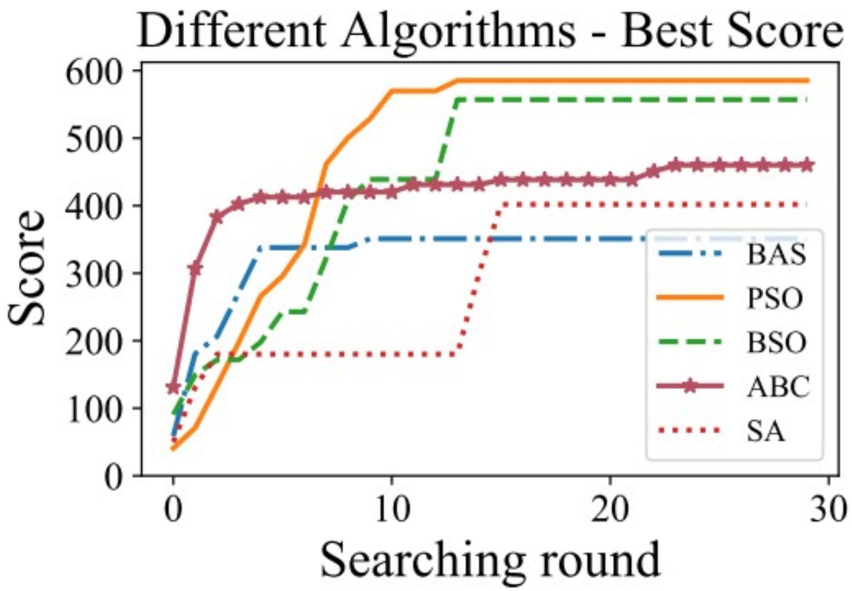
**CHALLENGE 3:** How to find the “best” perturbations?
**SOLUTION:** Designing two metrics to quantify (1) the visibility of the perturbation and (2)visibilit y of the corresponding detected lane. The visibility of lane should be high (effective), and the visi bility of perturbation should be low (unnoticeable).

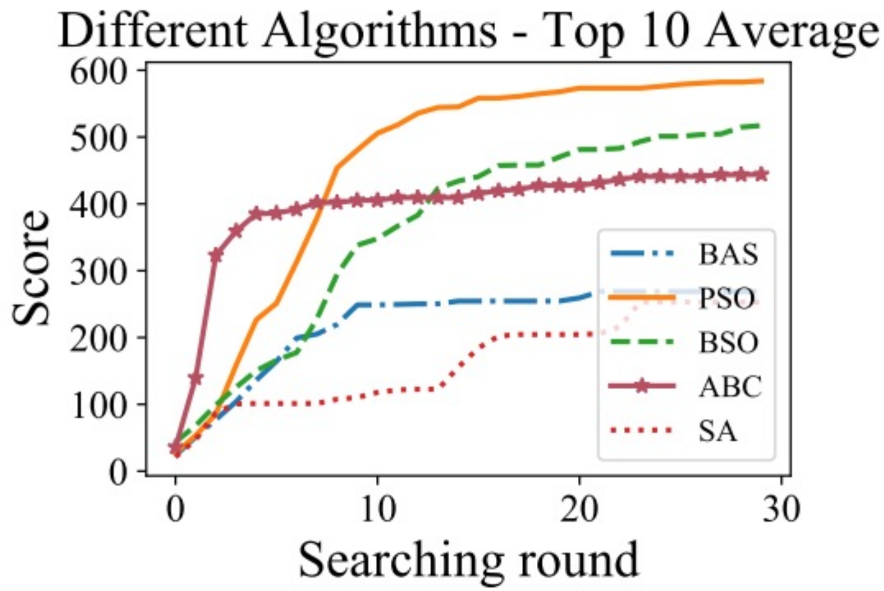
# Evaluation

- RQ1: How efficient are the heuristic algorithms to find the best perturbation?
- RQ2: How do the perturbation number $n$ and the rotation angle θ affect the best perturbation?
- RQ3: How is the performance of our approach given different input camera images?
- RQ4: What are the common characteristics of the best perturbations?
- RQ5: How effective is the attack in physical world?
- RQ6: Can we misguide the vehicle in physical world?

# Evaluation– Q1

❖ How efficient are the heuristic algorithms to find the best perturbation?



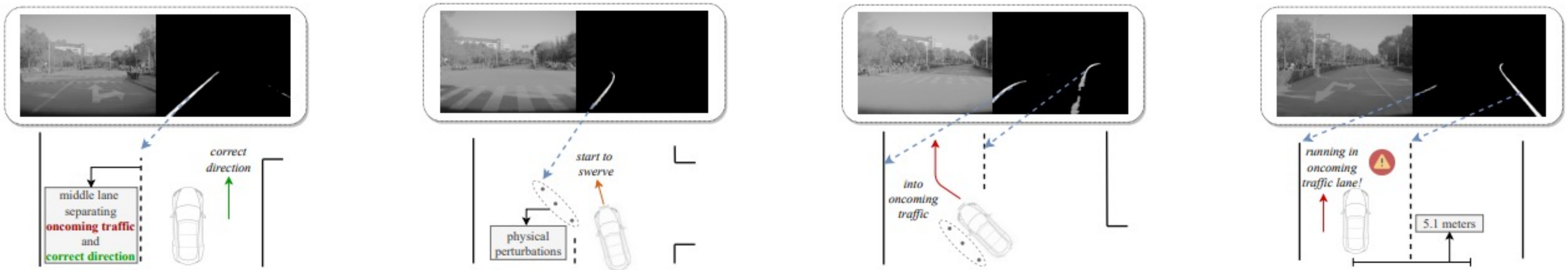





SysSec
System Security Lab

# Evaluation– Q2

- Can we misguide the vehicle in physical world?




(a) Vehicle is running on the correct direction.

(b) Fake lane is detected and vehicle starts to swerve.

(c) Vehicle follows the fake lane into oncoming traffic.

(d) Vehicle finally runs in the oncoming traffic lane!

- Demo Video

# Evaluation– Q2

**USENIX Security 2021 Paper**

Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations

Video demonstration - timeline:
0:05 ~ 1:05: Research Question 5 - Investigate effectiveness of perturbations in physical world.
1:06 ~ 1:15: Research Question 6 - Misguide the vehicle to the oncoming traffic.

*The Hong Kong Polytechnic University & Tencent Security Keen Lab*

# Defense

- **Better lane detection module** to distinguish craft perturbations
- **Better control policy**: more considerable elements, multi-sensor fusion
- **Personal idea**: more difficult model extraction, more robust algorithm

# Limitation and Discussion

❖ **Limitation**

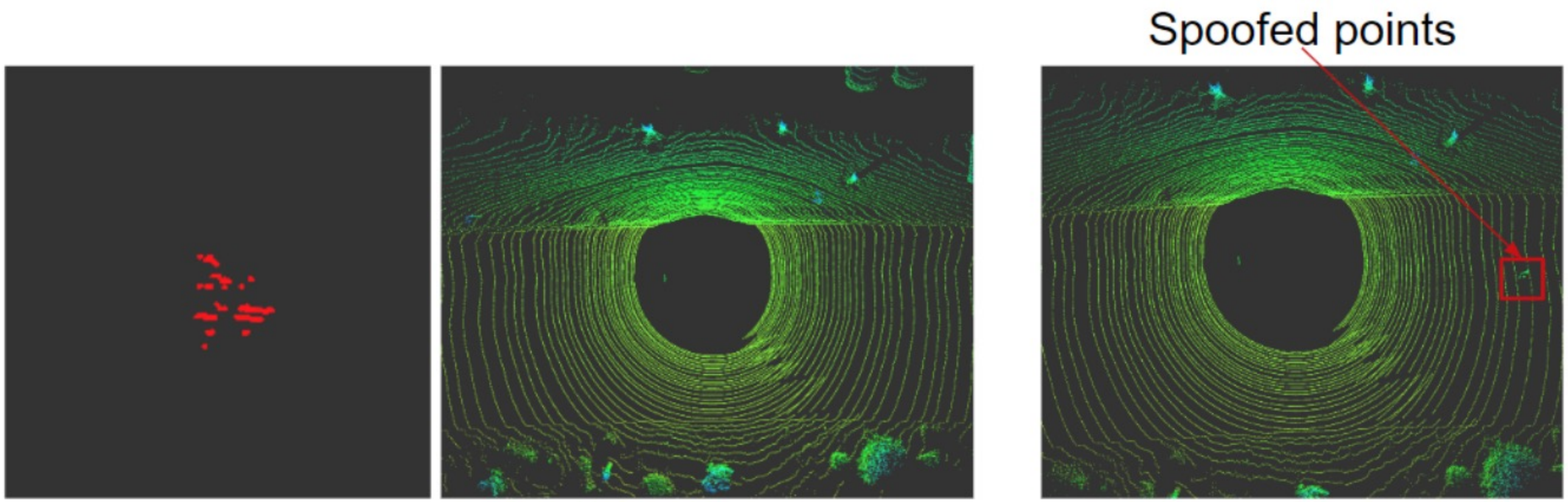
- A physical set up process, and it must be installed at a specific point.
- Cannot be completely invisible (a driver may notice)

❖ **Future Work**

- Same vulnerability in other autonomous driving systems (e.g., Apollo, Openpilot, etc.)
- Launching attacks on real lanes (e.g., dark markings to cover, etc.)

# Related Work (Before)

**This paper conduct the *first investigation* on the lane detection module**




Attack trace

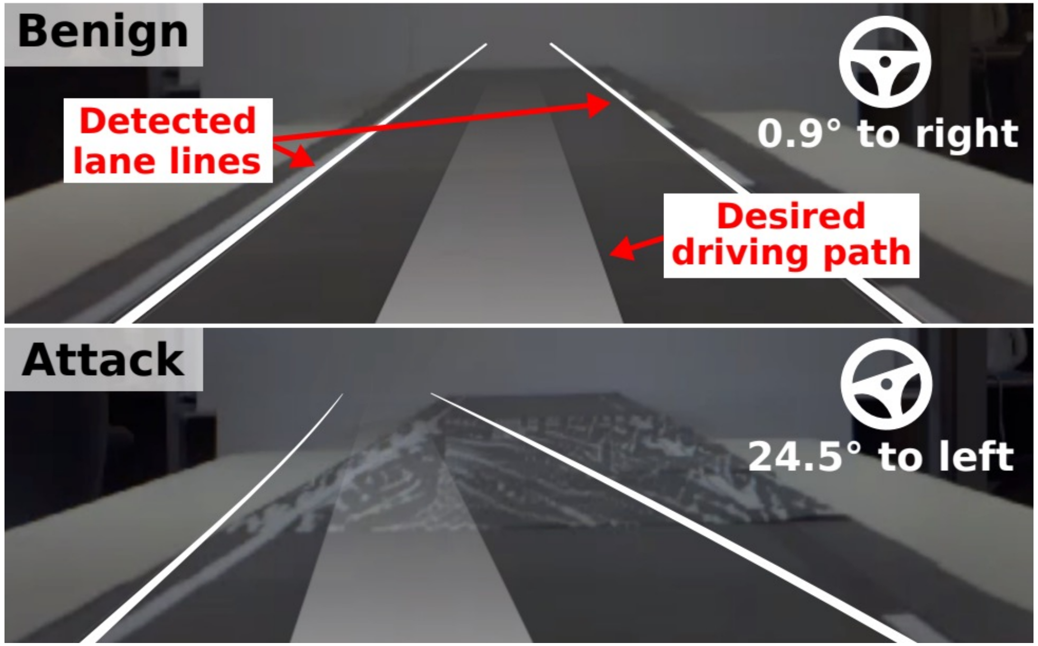
Pristine 3D point cloud

Attacker-perturbed 3D point cloud

CCS 2019 - Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving

# Related Work (After)

- **USENIX Security**





Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack (2021)

# Related Work (After)

❖ **CCS**

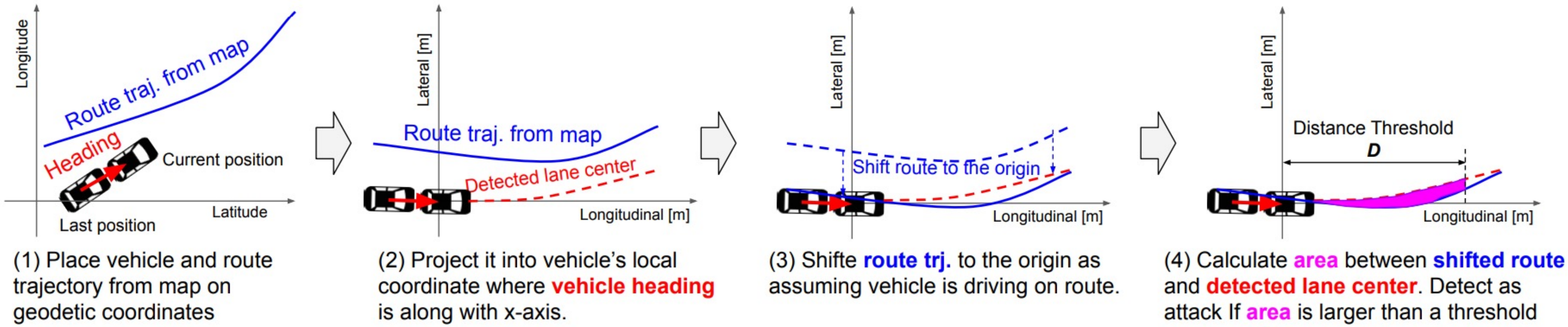
Benign

Backdoored

Physical Backdoor Attacks to Lane Detection Systems in Autonomous Driving (2022)

# Related Work (After)

## ❖ VehicleSec




A Cross-Verification Approach with Publicly Available Map for Detecting Off-Road Attacks against Lane Detection Systems (2024)

# Conclusion

- **Two-stage approach to generate the optimal perturbations**
  - Reverse engineering to access data
  - Misguide the vehicle into oncoming lane
  - Extensive evaluation

- **Need more reliable self-driving system**
  - Safety critical system
  - Standards and policies


Thank you
Thank you
Thank you
Thank you
Thank you

# Good Questions

- How vulnerable is the sensor fusion process itself to manipulation?
- Would this attack still be effective in environments where the lane markings do not disappear?
- Could a similar approach be used to attack the image detection module, causing it to perform sp ecific actions by manipulating road surface markings (e.g. STOP sign) through perturbations?
- If this sensitivity must be decreased to make it more robust to adversarial attacks, how can high accuracy be maintained?
- How transferable are these adversarial attacks to other aspects of autonomous driving systems, such as object recognition, or do they target lane detection specifically?

# Best Questions

- **Younghyo Kang:** Currently, techniques like code obfuscation, encryption, a nd packing are used to protect against reverse engineering. Assuming that t hese techniques are well-applied, is it still realistic to consider attacks in a bl ack-box environment as demonstrated in the paper?
- **Zunnoor Fayyaz Awan:** The authors of the paper suggest "multi-sensor fusi on" as a defense. However, Tesla in 2021 began removing the use of radar and switching fully to vision. Does this not make their autonomous driving sy stem less secure?
- **Munim Hasan Wasi:** Can these adversarial perturbations be applied in a wa y that is physically undetectable, such as by altering environmental condition s (e.g., manipulating shadows or reflections)?