The Intriguing Aspects and Trends of Research on Security for Autonomous Vehicles

France-Japan Cybersecurity Workshop 2023

Tatsuya Mori

WASEDA University

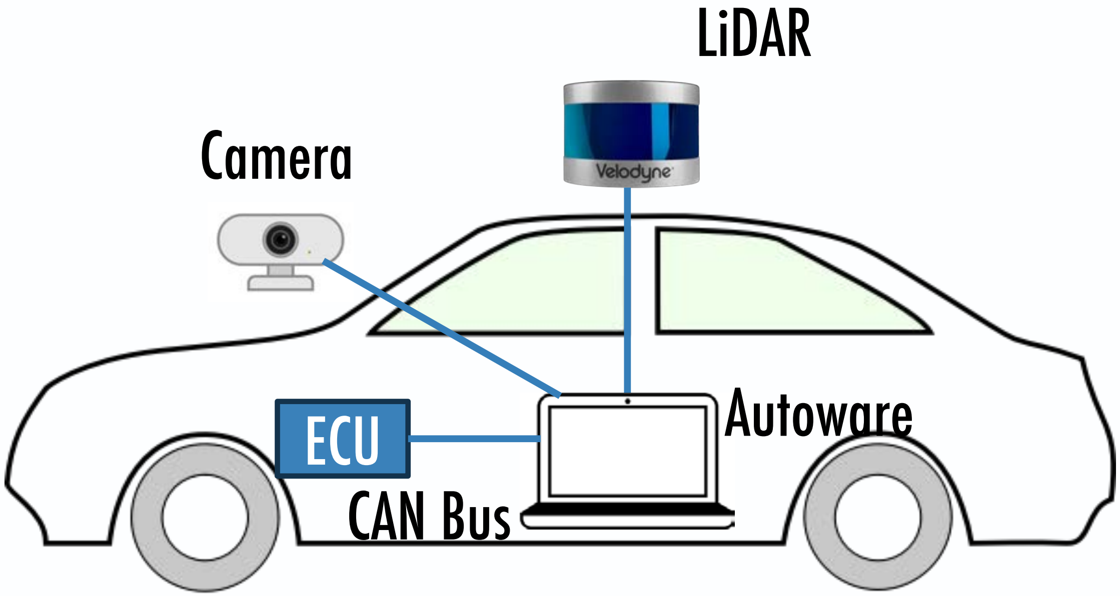
Why is autonomous driving security is an interesting research target?

Agenda

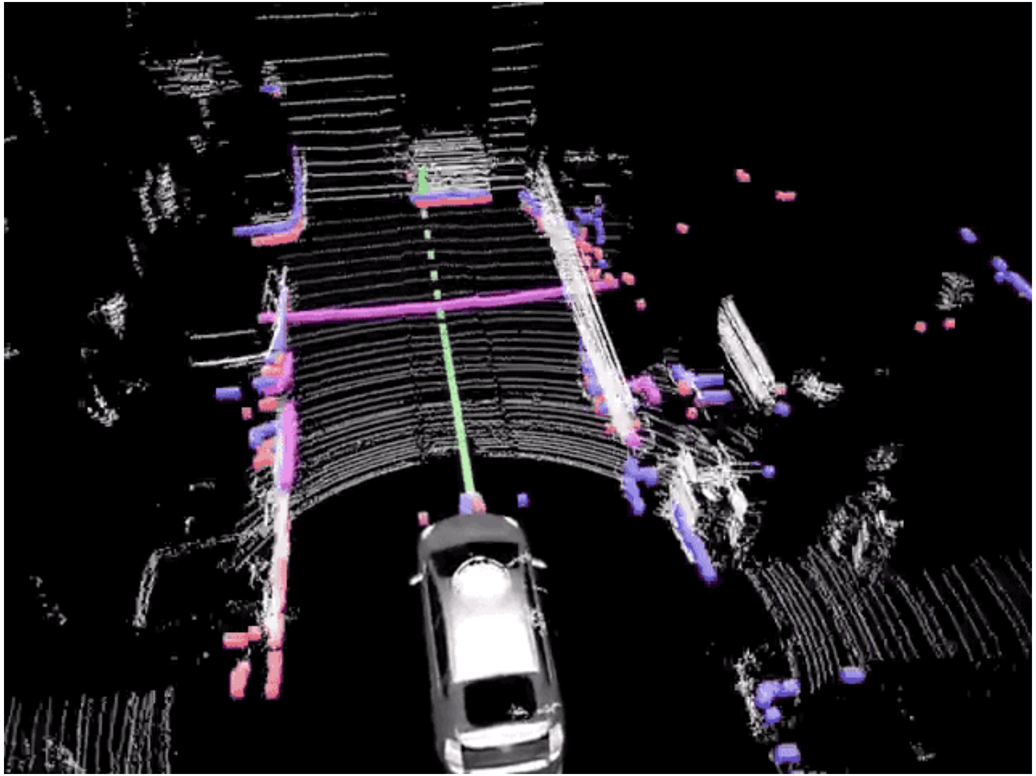
- Background: How Autonomous Vehicle Works
- Recent Trends in Autonomous Vehicle Security Research
- Future Research Directions
- Introduction to Our Research Project (JST CREST)

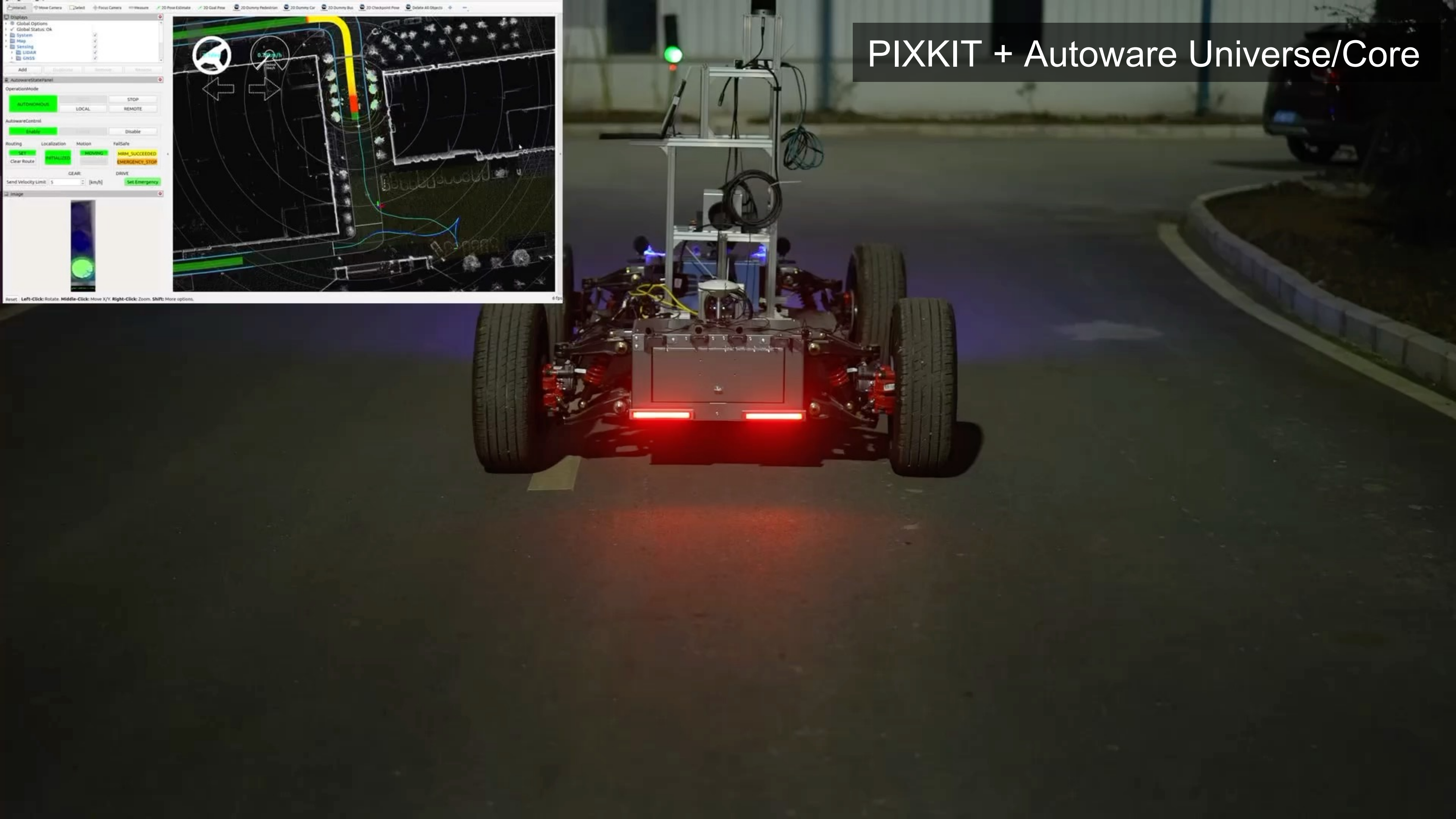
Background:
How an autonomous vehicle (AV) works

Primary components of an autoware-installed EV



How LiDAR sensor works



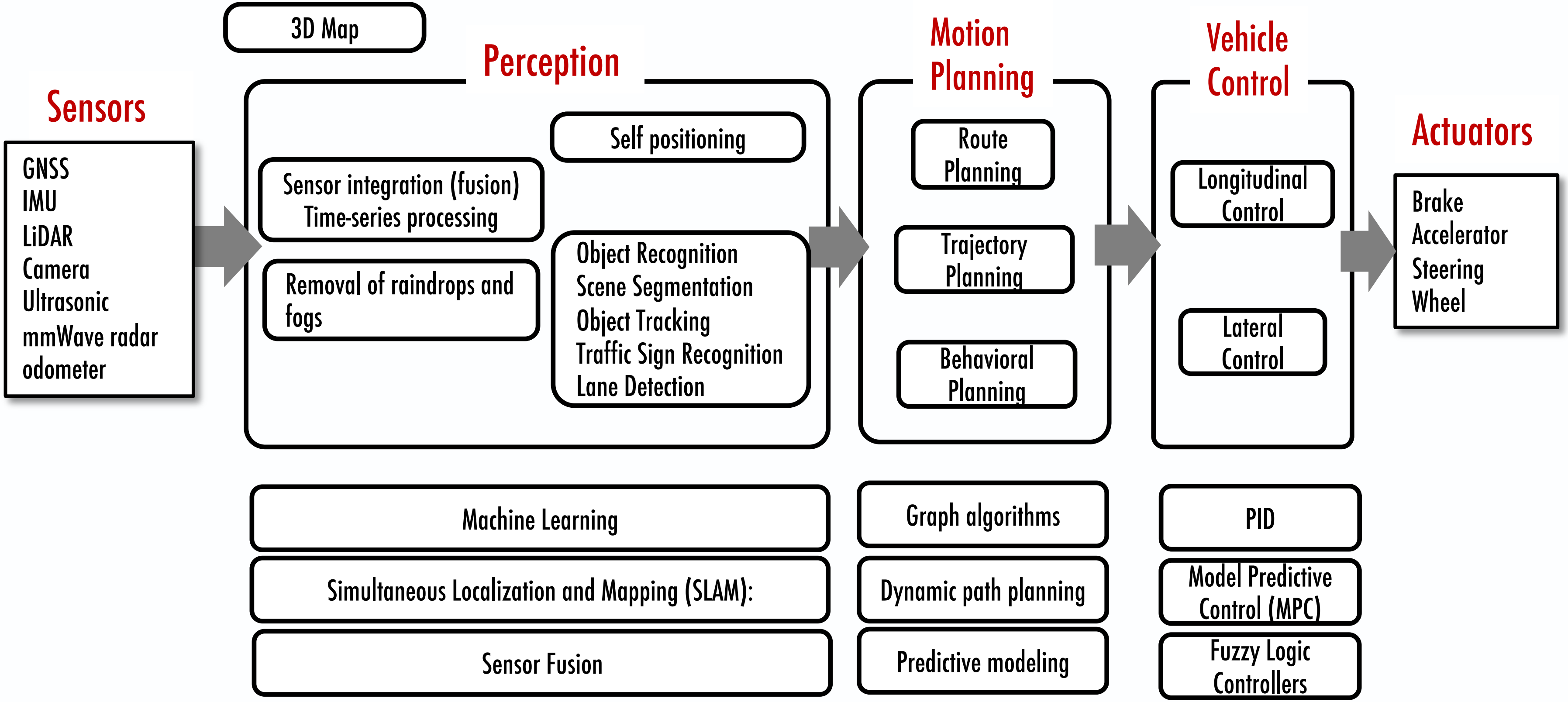

PIXKIT + Autoware Universe/Core

A brief overview of the AV system

Sensors → Perception → Motion Planning → Vehicle Control → Actuators

GM Cruise’s autonomous driving car
https://www.youtube.com/watch?v=IA5NVJf3K4Q

Integration of various technologies



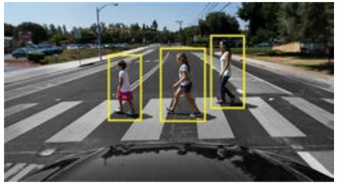
AI components used in AV systems

1. Perception and Object Recognition
2. Environmental Understanding and Decision Making
3. Predictive Analysis and Behavior Prediction
4. End-to-end autonomous driving

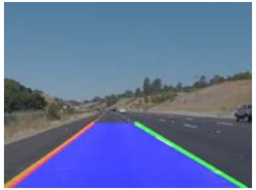
1. Perception and Object Recognition

- Traffic Sign Recognition:

- Pedestrian and Vehicle Detection:



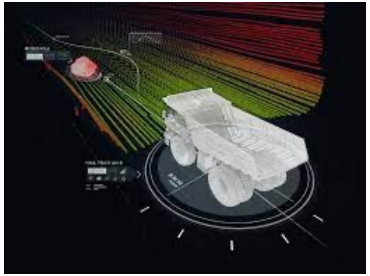
- Lane Detection:



- Traffic Light Recognition:

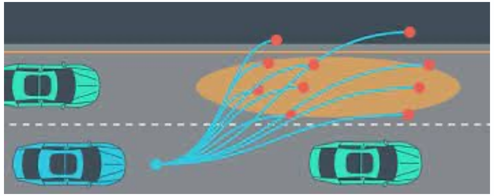
2. Environmental Understanding and Decision Making

- Obstacle and Hazard Detection



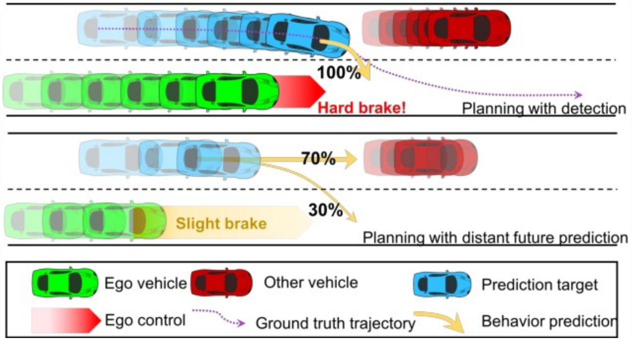
- Scene Segmentation

- Path Planning



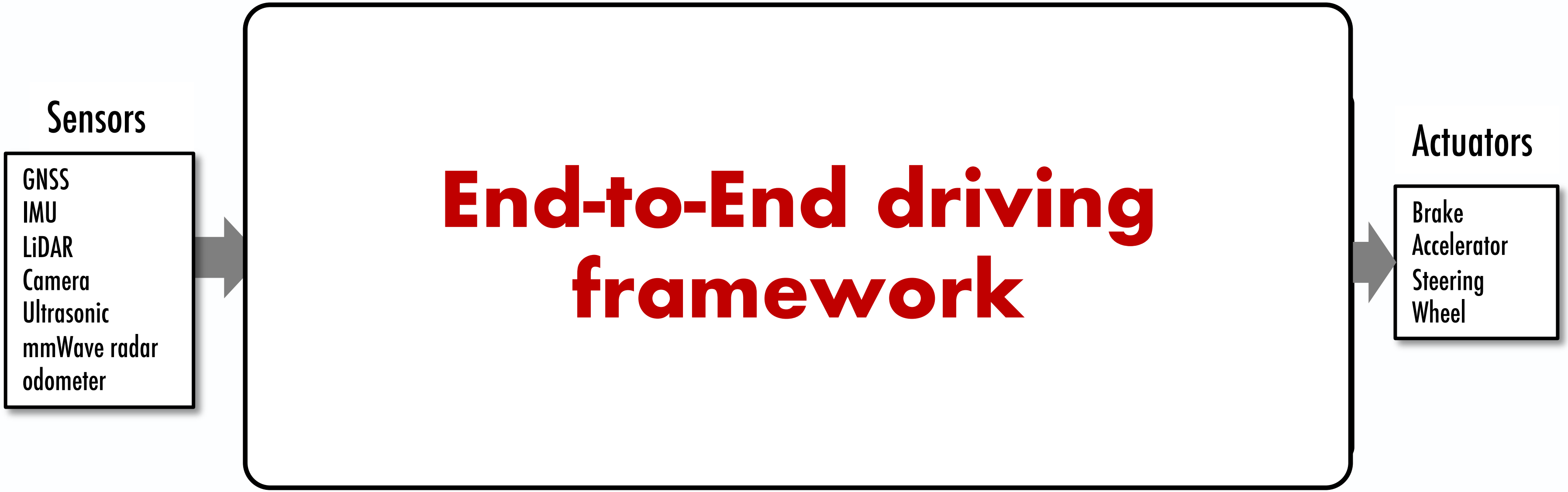
3. Predictive Analysis and Behavior Prediction

- Other Vehicle Behavior Prediction:



- Pedestrian Behavior Prediction:

4. End-to-End autonomous driving



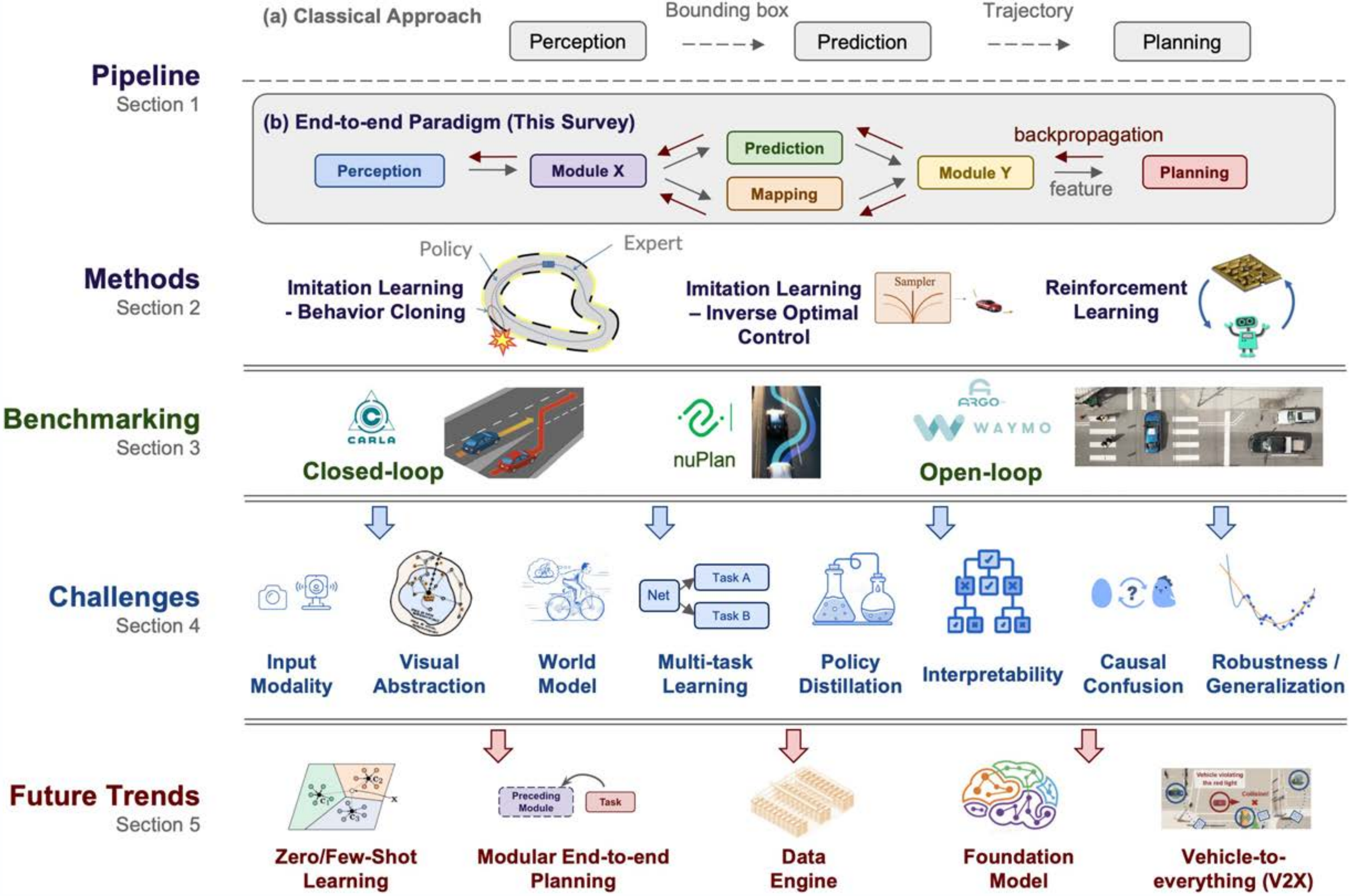
End-to-end Autonomous Driving: Challenges and Frontiers

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger and Hongyang Li

Abstract—The autonomous driving community has witnessed a rapid growth in approaches that embrace an end-to-end algorithm framework, utilizing raw sensor input to generate vehicle motion plans, instead of concentrating on individual tasks such as detection and motion prediction. End-to-end systems, in comparison to modular pipelines, benefit from joint feature optimization for perception and planning. This field has flourished due to the availability of large-scale datasets, closed-loop evaluation, and the increasing need for autonomous driving algorithms to perform effectively in challenging scenarios. In this survey, we provide a comprehensive analysis of more than 250 papers, covering the motivation, roadmap, methodology, challenges, and future trends in end-to-end autonomous driving. We delve into several critical challenges, including multi-modality, interpretability, causal confusion, robustness, and world models, amongst others. Additionally, we discuss current advancements in foundation models and visual pre-training, as well as how to incorporate these techniques within the end-to-end driving framework. To facilitate future research, we maintain an active repository that contains up-to-date links to relevant literature and open-source projects at https://github.com/OpenDriveLab/End-to-end-Autonomous-Driving.

Index Terms—Autonomous Driving, End-to-end System Design, Policy Learning, Simulation.

https://arxiv.org/abs/2306.16927



https://github.com/OpenDriveLab/End-to-end-Autonomous-Driving

Recent Trends in Autonomous Vehicle Security Research

Possible attack spots on AV systems

- Sensors
- $\boxed{\text{AI}}$
- Motion Planning
- Software / Firmware
- V2X communication
- ECU / CAN Bus

SoK: On the Semantic AI Security in Autonomous Driving

Junjie Shen, Ningfei Wang, Ziwen Wan, Yunpeng Luo, Takami Sato, Zhisheng Hu[†], Xinyang Zhang[†], Shengjian Guo[†], Zhenyu Zhong[†], Kang Li[†], Ziming Zhao[‡], Chunming Qiao[‡], Qi Alfred Chen

{junjies1, ningfei.wang, ziwenw8, yunpel3, takamis, alfchen}@uci.edu,
[†]{zhishenghu, xinyangzhang, sjguo, edwardzhong, kangli01}@baidu.com, [‡]{zimingzh, qiao}@buffalo.edu
UC Irvine, [†]Baidu Security, [‡]University at Buffalo

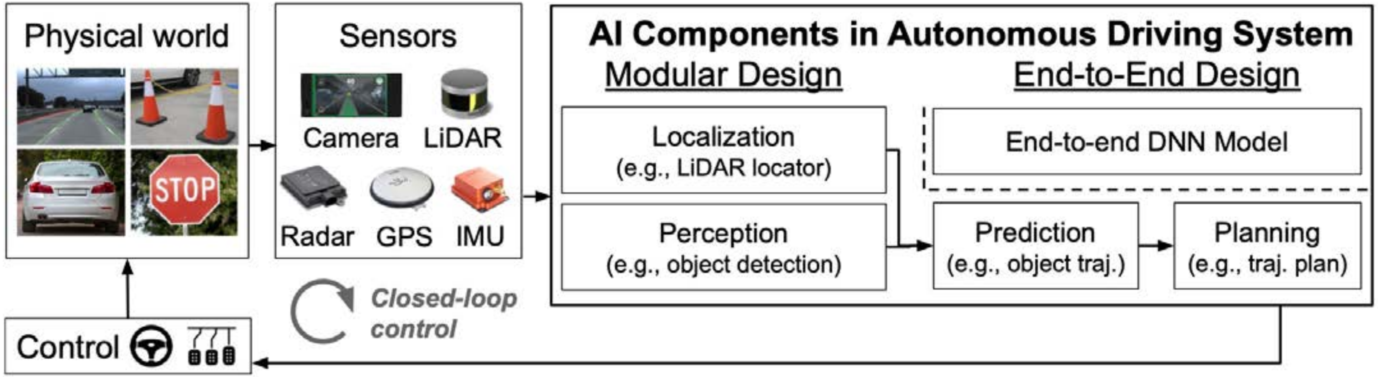


Figure 2. Overview of AD system designs and the roles of AD AI components.

https://arxiv.org/abs/2203.05314

Targeted AI component		Paper																				
Camera perception	Object detection	Lu et al. [54]	'17	V	✓			✓											○	✓		
		Eykholt et al. [18]	'18	S	✓			✓											○	✓		
		Chen et al. [37]	'18	M	✓			✓											○	✓		✓
		Zhao et al. [26]	'19	S	✓			✓											○	✓		
		Xiao et al. [55]	'19	V	✓			✓	✓										○	✓		
		Zhang et al. [56]	'19	M	✓			✓											◐	✓		✓
		Nassi et al. [57]	'20	S	✓			✓											○	✓	✓	
		Man et al. [58]	'20	S	✓									✓					○	✓		✓
		Hong et al. [59]	'20	S	✓													✓	○		✓	
		Huang et al. [60]	'20	V	✓			✓											○	✓		✓
		Wu et al. [61]	'20	V	✓			✓											○	✓		
		Xu et al. [62]	'20	V	✓			✓											○	✓		
		Hu et al. [63]	'20	V	✓			✓											◐	✓		
		Hamdi et al. [64]	'20	M	✓				✓										◐	✓		
		Ji et al. [65]	'21	S	✓										✓				◐	✓		
		Lovisotto et al. [66]	'21	S	✓			✓											●	✓		✓
		Wang et al. [67]	'21	S	✓									✓					◐	✓		
		Köhler et al. [68]	'21	S	✓									✓					●	✓		
		Wang et al. [69]	'21	S	✓			✓											●	✓		✓
		Zolfi et al. [70]	'21	V	✓											✓			○	✓		
		Wang et al. [71]	'21	V	✓			✓											◐	✓		✓
		Zhu et al. [72]	'21	M	✓									✓					○	✓		
	Semantic segmentation	Nakka et al. [73]	'20	V	✓			✓											○	✓		
		Nesti et al. [74]	'22	V	✓			✓											○	✓		
	Object tracking	Jha et al. [75]	'20	S	✓													✓	○		✓	
		Jia et al. [17]	'20	M	✓			✓											○	✓		✓
		Ding et al. [76]	'21	M	✓			✓											○	✓		
		Chen et al. [77]	'21	M	✓			✓											○	✓		
	Lane detection	Sato et al. [78]	'21	S	✓			✓											○	✓	✓	✓
		Jing et al. [79]	'21	S	✓			✓											◐	✓	✓	
	Traffic light detection	Wang et al. [67]	'21	S	✓									✓					○	✓		
		Tang et al. [80]	'21	S	✓						✓								○		✓	
LiDAR perception	Object detection	Cao et al. [19]	'19	S	✓							✓							○	✓	✓	
		Sun et al. [81]	'20	S	✓							✓							●	✓		
		Hong et al. [59]	'20	S	✓													✓	○		✓	
		Tu et al. [82]	'20	V	✓				✓										○	✓		
		Zhu et al. [83]	'21	S	✓					✓									◐	✓		
		Yang et al. [84]	'21	S	✓				✓										◐	✓	✓	
		Hau et al. [85]	'21	S	✓							✓							○	✓		
		Li et al. [86]	'21	V	✓						✓								○	✓		✓
		Zhu et al. [87]	'21	O	✓					✓									◐	✓		
	Semantic segmentation	Tsai et al. [88]	'20	M	✓				✓										○	✓		✓
		Zhu et al. [87]	'21	O	✓					✓									◐	✓		
RADAR perception	Obj. detection	Sun et al. [89]	'21	S	✓								✓						◐	✓	✓	
MSF perception		Cao et al. [38]	'21	S	✓				✓										○	✓	✓	✓
		Tu et al. [90]	'21	O	✓				✓										○	✓		
LiDAR localization		Luo et al. [91]	'20	S		✓												✓	○		✓	
MSF localization		Shen et al. [92]	'20	S	✓						✓								○	✓	✓	
Camera localization		Wang et al. [67]	'21	S	✓									✓					○	✓		
Chassis		Hong et al. [59]	'20	S		✓												✓	○		✓	
End-to-end driving		Liu et al. [93]	'18	S	✓			✓									✓		○	✓	✓	✓
		Kong et al. [94]	'20	V	✓			✓											○	✓		✓
		Hamdi et al. [64]	'20	M	✓				✓										◐	✓		
		Boloor et al. [95]	'20	O	✓			✓											●	✓	✓	✓

Field: S = Security, V = Computer Vision, M = ML/AI, O = Others, e.g., Robotics, arXiv;
Attacker's knowledge: ○ = white-box, ◐ = gray-box, ● = black-box

Camera perception

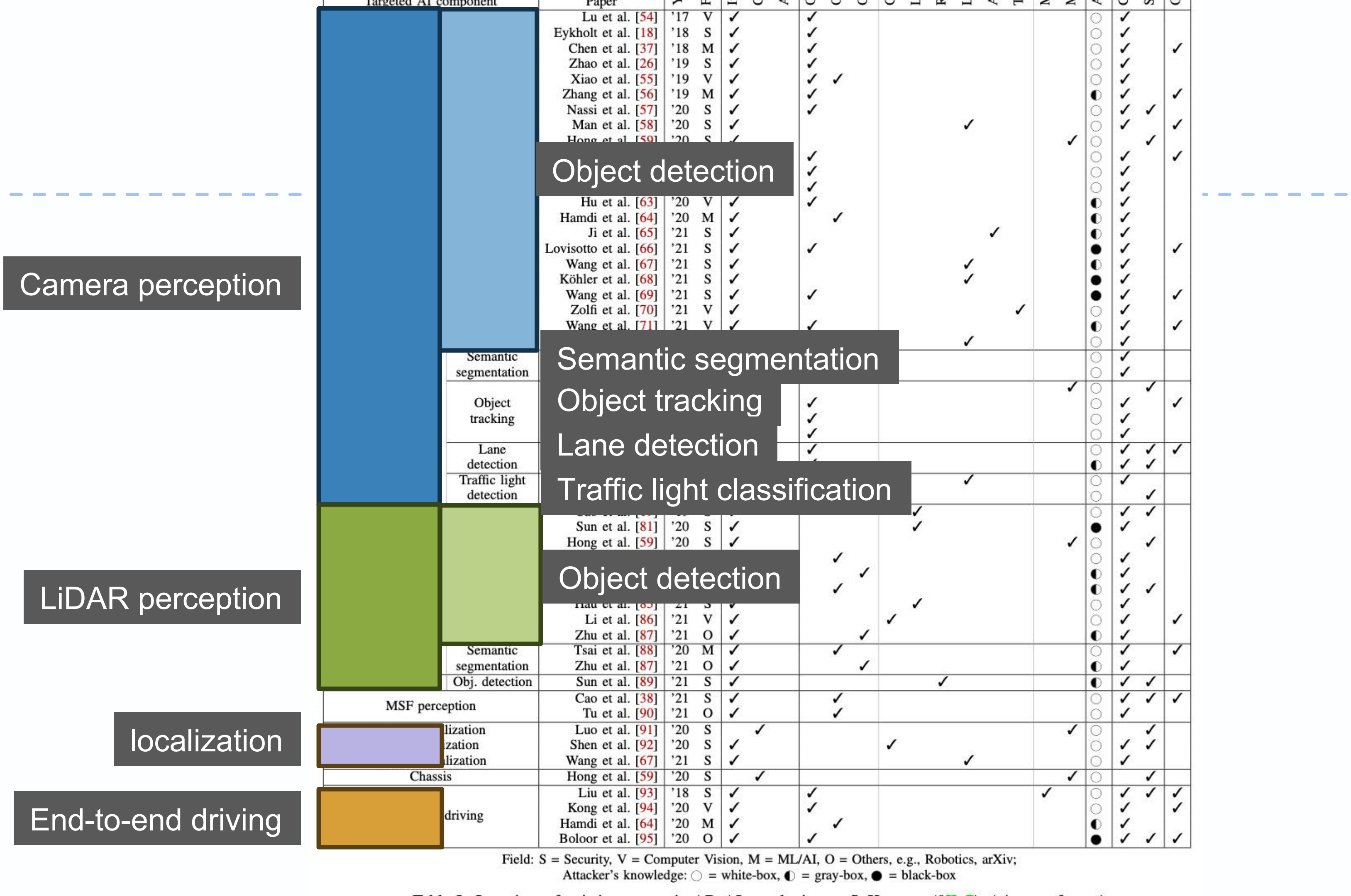
LiDAR perception

localization

End-to-end driving

Targeted AI component	Paper																				
Object detection	Lu et al. [54]	'17	V	✓			✓											○	✓		
	Eykholt et al. [18]	'18	S	✓			✓											○	✓		
	Chen et al. [37]	'18	M	✓			✓											○	✓		✓
	Zhao et al. [26]	'19	S	✓			✓											○	✓		
	Xiao et al. [55]	'19	V	✓			✓	✓										○	✓		
	Zhang et al. [56]	'19	M	✓			✓											◐	✓		✓
	Nassi et al. [57]	'20	S	✓			✓											○	✓	✓	
	Man et al. [58]	'20	S	✓									✓					○	✓		✓
	Hong et al. [59]	'20	S	✓													✓	○		✓	
	Huang et al. [60]	'20	V	✓			✓											○	✓		✓
	Wu et al. [61]	'20	V	✓			✓											○	✓		
	Xu et al. [62]	'20	V	✓			✓											○	✓		
	Hu et al. [63]	'20	V	✓			✓											◐	✓		
	Hamdi et al. [64]	'20	M	✓				✓										◐	✓		
	Ji et al. [65]	'21	S	✓										✓				◐	✓		
	Lovisotto et al. [66]	'21	S	✓			✓											●	✓		✓
	Wang et al. [67]	'21	S	✓									✓					◐	✓		
	Köhler et al. [68]	'21	S	✓									✓					●	✓		
	Wang et al. [69]	'21	S	✓			✓											●	✓		✓
	Zolfi et al. [70]	'21	V	✓											✓			○	✓		
	Wang et al. [71]	'21	V	✓			✓											◐	✓		✓
	Zhu et al. [72]	'21	M	✓									✓					○	✓		
Semantic segmentation	Nakka et al. [73]	'20	V	✓			✓											○	✓		
	Nesti et al. [74]	'22	V	✓			✓											○	✓		
Object tracking	Jha et al. [75]	'20	S	✓													✓	○		✓	
	Jia et al. [17]	'20	M	✓			✓											○	✓		✓
	Ding et al. [76]	'21	M	✓			✓											○	✓		
	Chen et al. [77]	'21	M	✓			✓											○	✓		
Lane detection	Sato et al. [78]	'21	S	✓			✓											○	✓	✓	✓
	Jing et al. [79]	'21	S	✓			✓											◐	✓	✓	
Traffic light detection	Wang et al. [67]	'21	S	✓									✓					○	✓		
	Tang et al. [80]	'21	S	✓						✓								○		✓	
Object detection	Cao et al. [19]	'19	S	✓							✓							○	✓	✓	
	Sun et al. [81]	'20	S	✓							✓							●	✓		
	Hong et al. [59]	'20	S	✓													✓	○		✓	
	Tu et al. [82]	'20	V	✓				✓										○	✓		
	Zhu et al. [83]	'21	S	✓					✓									◐	✓		
	Yang et al. [84]	'21	S	✓				✓										◐	✓	✓	
	Hau et al. [85]	'21	S	✓							✓							○	✓		
	Li et al. [86]	'21	V	✓						✓								○	✓		✓
	Zhu et al. [87]	'21	O	✓					✓									◐	✓		
Semantic segmentation	Tsai et al. [88]	'20	M	✓				✓										○	✓		✓
	Zhu et al. [87]	'21	O	✓					✓									◐	✓		
Obj. detection	Sun et al. [89]	'21	S	✓								✓						◐	✓	✓	
MSF perception	Cao et al. [38]	'21	S	✓				✓										○	✓	✓	✓
	Tu et al. [90]	'21	O	✓				✓										○	✓		
ization	Luo et al. [91]	'20	S		✓												✓	○		✓	
zation	Shen et al. [92]	'20	S	✓						✓								○	✓	✓	
lization	Wang et al. [67]	'21	S	✓									✓					○	✓		
Chassis	Hong et al. [59]	'20	S		✓												✓	○		✓	
driving	Liu et al. [93]	'18	S	✓			✓									✓		○	✓	✓	✓
	Kong et al. [94]	'20	V	✓			✓											○	✓		✓
	Hamdi et al. [64]	'20	M	✓				✓										◐	✓		
	Boloor et al. [95]	'20	O	✓			✓											●	✓	✓	✓

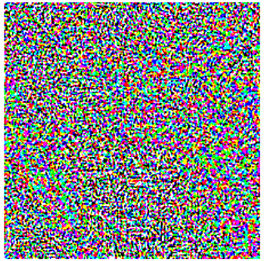
Field: S = Security, V = Computer Vision, M = ML/AI, O = Others, e.g., Robotics, arXiv;
Attacker's knowledge: ○ = white-box, ◐ = gray-box, ● = black-box


Camera perception
LiDAR perception
localization
End-to-end driving
Object detection
Semantic segmentation
Object tracking
Lane detection
Traffic light classification
Object detection
Field: S = Security, V = Computer Vision, M = ML/AI, O = Others, e.g., Robotics, arXiv;
Attacker's knowledge: ○ = white-box, ◐ = gray-box, ● = black-box

Three attack vectors against AI

- **Adversarial Example (AE)**
  - Generate input data (tiny noise injection) that induces misclassification of machine learning algorithms
- **Model Extraction**
  - Estimating (private) machine learning models from input and output results
- **Model Inversion**
  - Estimated original data used to train (private) machine learning algorithms

Adversarial Example (AE)

	Original Image		Adversarial Noise		Generated AE
Human perception		$+ .007 \times$		$=$	
	$\boldsymbol{x}$		$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$		$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

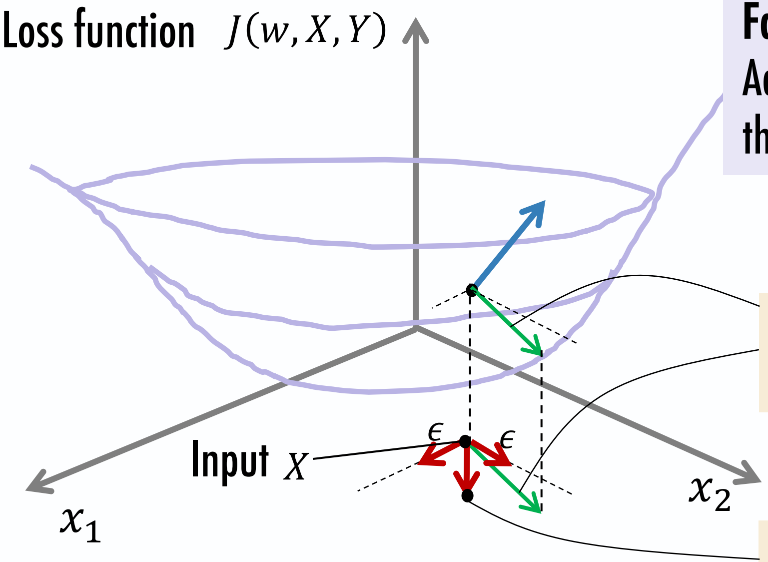
ML algorithm

F(x):
Detect as "panda"
With 57.7% of
confidence level

F(x+noise):
Detect as "gibbon"
With 99.3% of
confidence level

Goodfellow et al., Explaining and Harnessing Adversarial Examples https://arxiv.org/abs/1412.6572

Idea of generating AE (FGSM)



Fast Gradient Sigh Method:
Add a perturbation in the direction that maximizes the loss function under the max-norm constraints

Gradient

$$\nabla_X J(w, X, Y) = \left(\frac{\partial J}{\partial x_1}, \frac{\partial J}{\partial x_2}\right)$$

AE

$$X + \epsilon \; \mathrm{sign}(\nabla_X J)$$

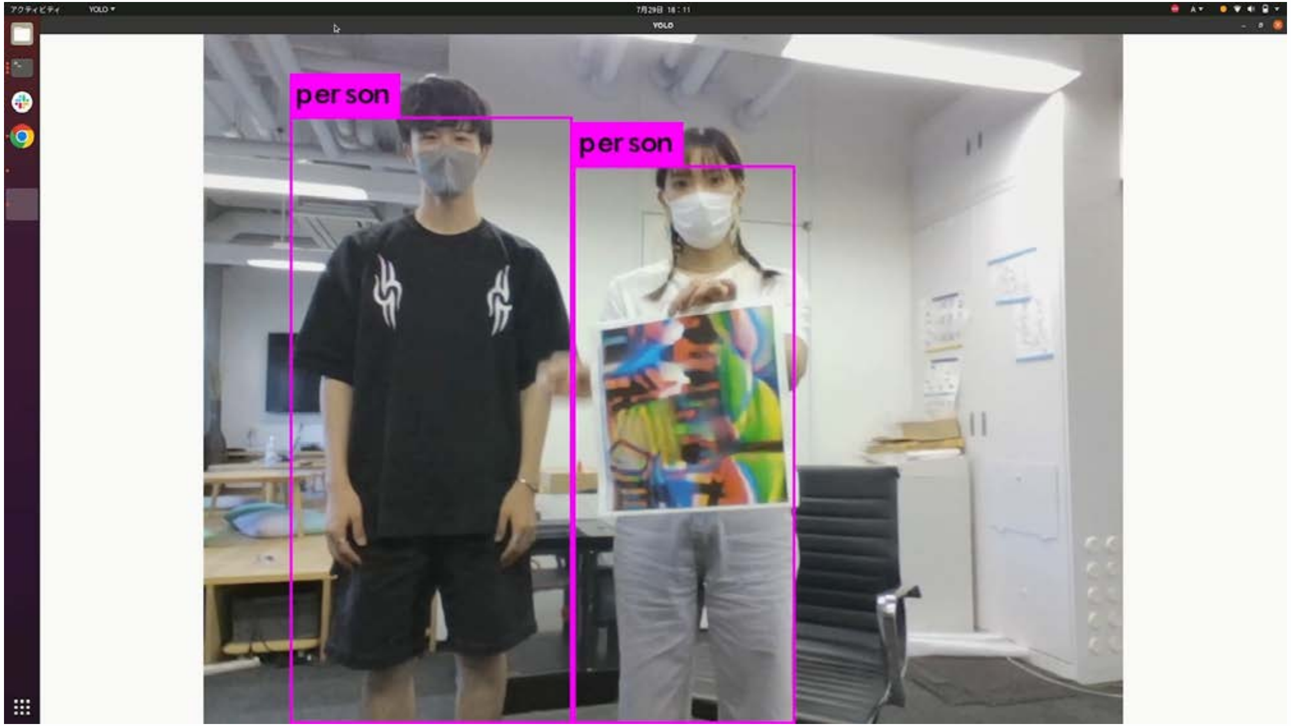
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R., "Intriguing properties of neural networks," arXiv:1312.6199v4 [cs.CV], Feb 2014.

Challenges for the "Physical" Adversarial Examples

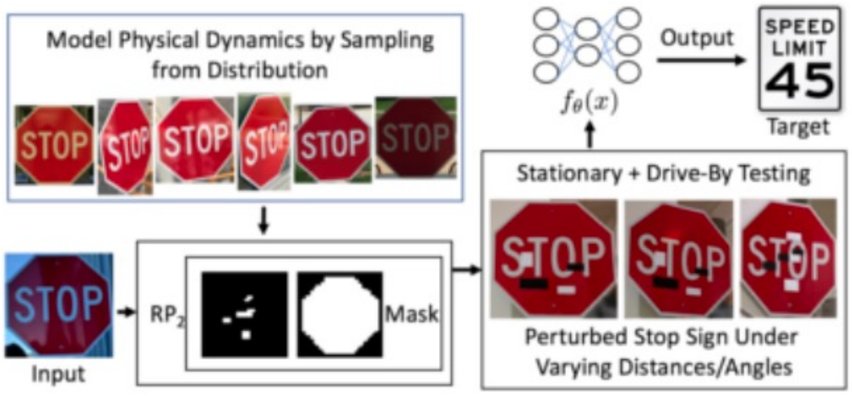
- Needs to add the adversarial perturbation as an <u>analog signal</u>
- It should be robust against various noises / environmental factors
- It should be "realizable" e.g., printable or projectable

- In many cases, "adversarial patch" works well
  - A universal pattern that satisfies the above conditions.

An example of adversarial patch



Adversarial road signs



Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Eykholt, Evtimov, Fernandes, Li, Rahmati, Xiao, Prakash, Kohno, and Song, "Robust Physical- World Attacks on Deep Learning Models," arXiv:1707.08945v5 [cs.CR], April 2018, pp. 1–11.

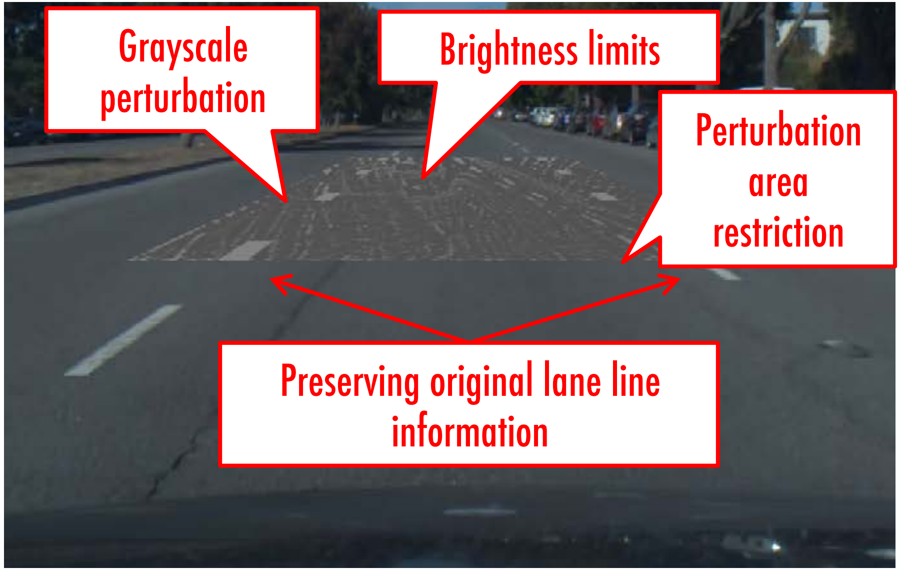
Recent Studies from Our Team

- Attacks against AI
  - Dirty Road Patch Attack: Sato (USENIX SEC 22)
  - Infrared Laser Reflection Attack: Sato, Sugawara (NDSS 24)
  - Retroreflector Attack: Tsuruoka, Sato, <u>Mori</u> (WIP)
- Attacks against sensors
  - Lidar physical removal attack: Sugawara (USENIX SEC 23)
  - Lidar practical removal attack: Sato, Yoshioka (NDSS 24)
  - Adversarial fog Attack: Tanaka , <u>Mori</u> (WIP)

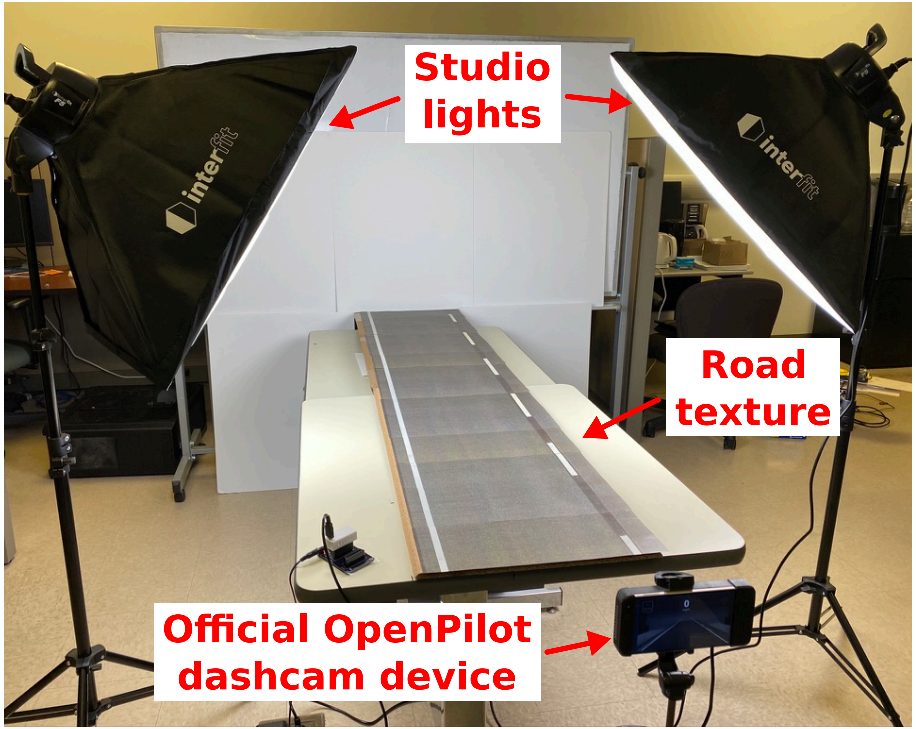
AI (1): Dirty Road Patch (DRP) [Sato et al., Usenix Security '21]

Key idea

- DRP attack pretends to be benign road patch but the surface patterns are designed for adversarial attack
  - Attacker can print malicious perturbation on patch and quickly deploy it



Attack demo 1: Miniature-scale physical-world setup




Attack
MAX
N/A
0
mph

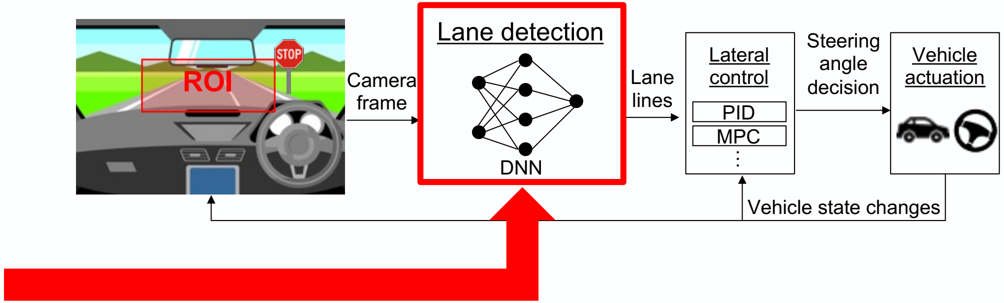
Attack Demo 2

Software-in-the-Loop Simulation with LGSVL

Target ALC: OpenPilot v0.6.6
Scenario: Local Road at 45 mph (72 km/h)

Attack demo 3: Safety impact on real vehicle

- We inject attack trace into real-world driving
  to see if other driving assistance features (e.g., AEB) can prevent crash



Replace model output with the one obtained in the simulator

* We obey California's road of conduct

Target of our study: OpenPilot

- Open-sourced production ACC with representative design: DNN-based camera lane detection
- Close performance to Tesla AutoPilot and GM Super Cruise*


Pre-collision alert starts 0.74 sec before the crash
Alert Only. Pre-collision braking is enabled but not applied.

AI (2) Infrared Laser Reflection Attack [Sato, Sugawara et al., NDSS 24]

Limitations of Existing Attacks: Visibility for Human

[Eykholt et al., 2018]

[Chen et al., 2019]

[Zhao et al., 2019]

[Jia et al., 2022]

Existing attacks against vision-based traffic sign recognition are generally <u>visible to human eyes</u>

Our Attack: Infrared Laser Reflection (ILR) Attack

To human eye (normal camera)

A camera used in autonomous driving (AD)

<u>Idea: Project an IR laser onto traffic signs.</u>

- The IR laser's path is completely invisible to the human eye.
- It can disturb a large area on the traffic sign without compromising stealth.
- However, the trace may appear as a simple shape with a uniform purplish hue.

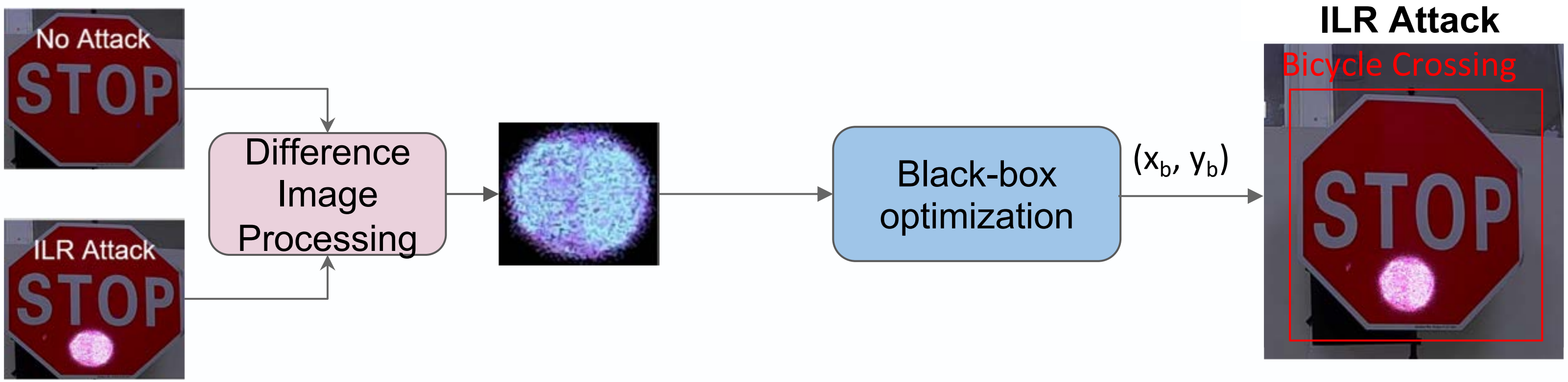
Trace Modeling and Optimization

Technical Challenges

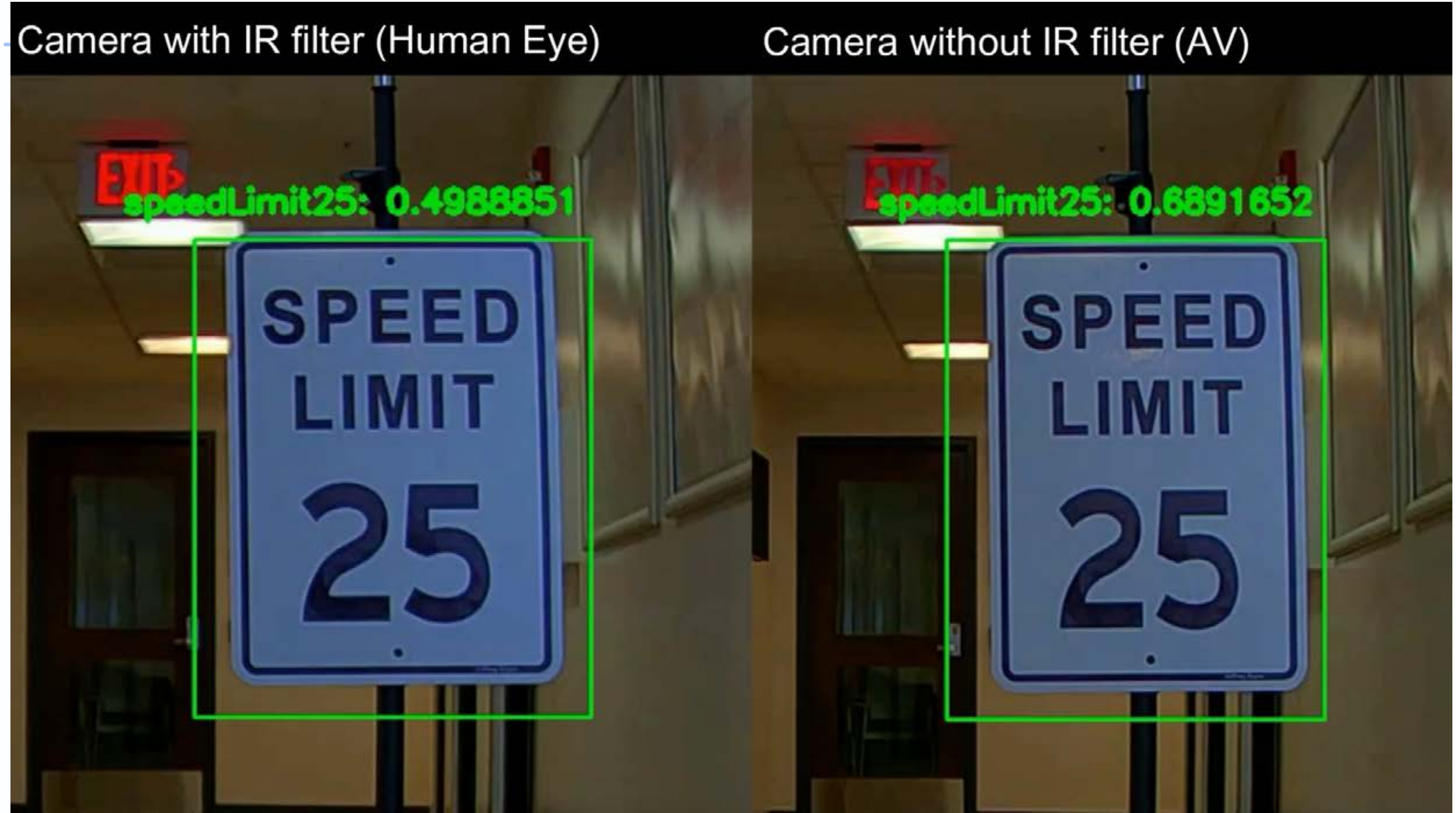
1. Accurate IR laser reflection modeling
2. Effective optimization of attack parameters

1. Image Difference-based IR Trace Modeling

2. Optimization Trace Position (x_b, y_b)



ILR Attack Demonstration



AI (3) Retroreflection Attack [Tsuruoka, Sato, Mori et al., WIP]

Retroreflector

Invisible in day time

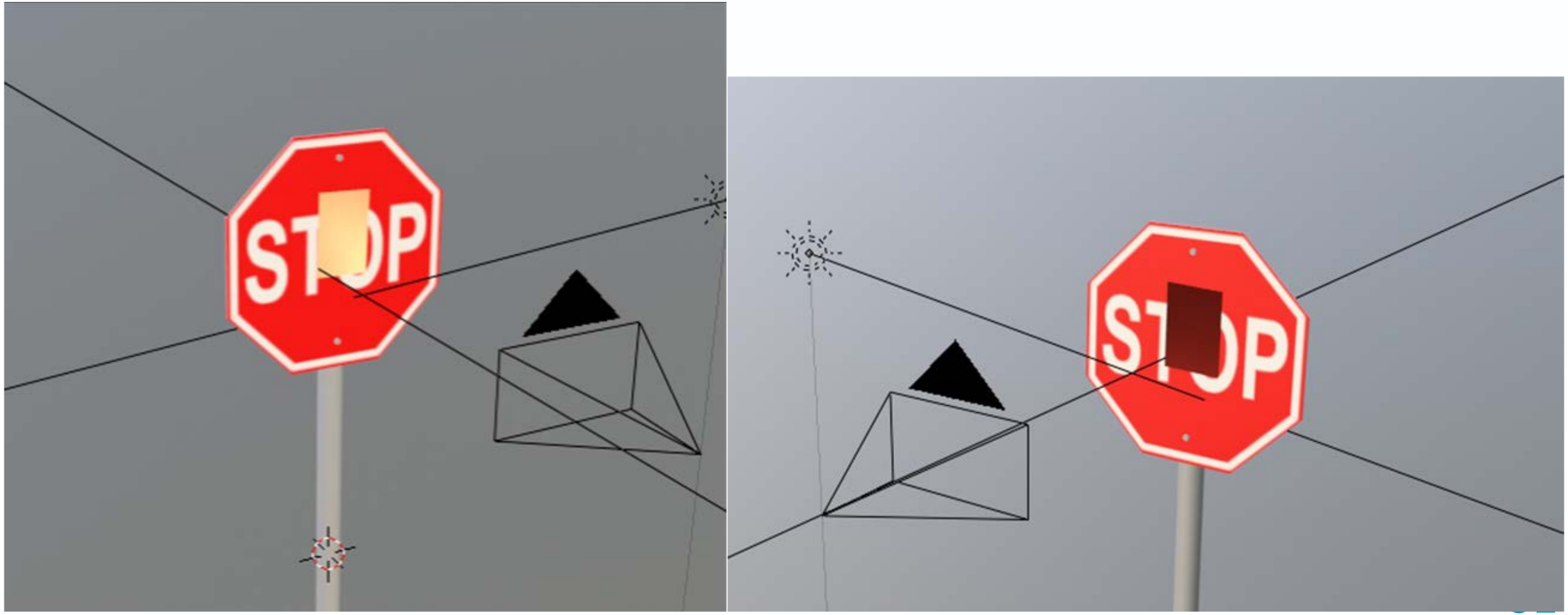
Visible in night (with light)

Adversarial Attack only effective in night

Without attack: detected as a stop sign

With the attack: nothing detected

Simulation evaluation



Future Research Directions

(1) End-to-End Perspective

- An End-to-End Perspective is essential!
  - End-to-End vs. Modular-based
  - Beyond the element-focused reductionism

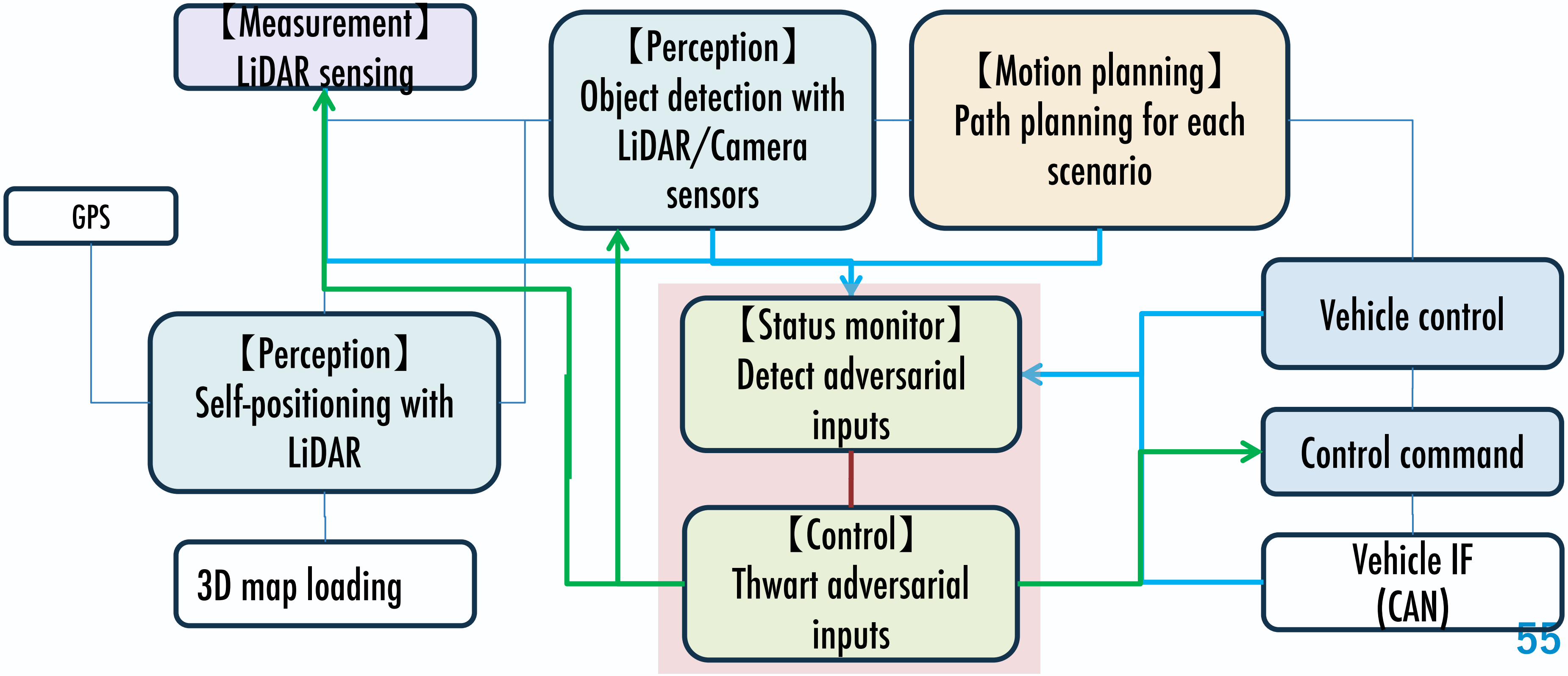
- To succeed the attack against a complex system like AV, it is necessary to optimize the attack for the whole system, not for a subsystem.

- Full self-driving simulation and experiments with real vehicles are essential.

(2) Realistic Test/Benchmark Environment

- Catalog of Attack Scenarios
  - A reference list of potential adversarial strategies targeting AV sensors and AI, crucial for structured security assessments.
- Benchmark Development
  - Quantitative standards to measure AV defenses against the cataloged attacks, identifying weaknesses and guiding enhancements.
- Testing Protocols for Realism
  - Procedures that apply these benchmarks in simulations and real-world tests to ensure AV systems can withstand practical security challenges.

(3) Integrated Software-Defined Defense



Introduction to our project

JST CREST

- **Funding agency:** JST (Japan Science and Technology Agency) 国立研究開発法人 科学技術振興機構 Japan Science and Technology Agency
- **Program:** CREST

CREST is a funding program for team-oriented research with the aim of achieving the strategic goals set forth by the government.
The objective is to create revolutionary technological seeds for science and technology innovation.

Our Project

- **Research area**: Creation of System Software for Society 5.0 by Integrating Fundamental Theories and System Platform Technologies
- **Project theme**: Security Evaluation and Countermeasure Infrastructure for AI-Driven Cyber-Physical System (AI-CPS)
- **Period**: Oct 2023 – Mar 2029 (5.5 years)
- **Budget**: 300,000,000 JPY (1,875,000 EUR)

The goal and work packages (WP)

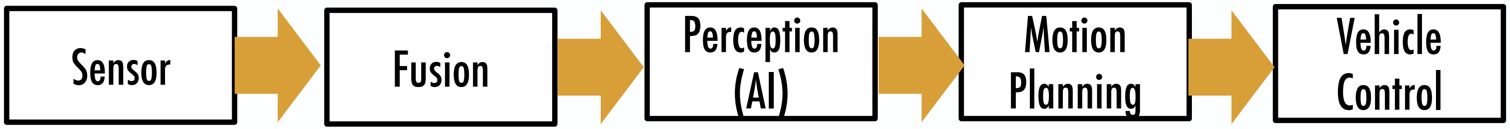
Goal: Realization of **Security by Design** to preemptively prevent the threat of adversarial inputs against AI-CPS (Achieving robustness against adversarial inputs)

WP1: Assessment and countermeasure technology for adversarial inputs **against elemental technologies**

WP2: Assessment and countermeasure technology for adversarial inputs **across the entire system**

WP3: Building software that implements security countermeasure technologies

End-to-End Perspective



- A system where multiple components work in coordination.
  - Adversarial inputs to cameras and sensors ripple through to subsequent processes: recognition, path planning, and control.
  - How they ripple through is not self-evident.
- As vehicles moves, the surrounding environment also changes.
  - The feedback loop of the entire system is essential.
  - It is necessary to deal with models that dynamically change input data and conditions to sensors and AI (such as angle, distance, illumination, and speed).

Our Team

Core PIs

Tatsuya Mori (Waseda U) System security

Kentaro Yoshioka (Keio U) Sensor integration

Takeshi Sugawara (UEC) Physical measurement

Jun Sakuma (Titech) Machine learning

Kenji Sawada (UEC) Vehicle control

Collaborators

Shunsuke Aoki (Turing / NII)

Takami Sato (UCI)

Qi Alfred Chen (UCI)

Yohei Akimoto (Tsukuba U)

Yu Zhe (RIKEN AIP)

Katsuhiko Yamafuji (NISSAN)

Osamu Kaneko (UEC)

Koichi Kobayashi (UEC)

Graduate students

Partners

TURING

Brain IV Intelligent Vehicle

WE ARE HIRING

Resources

Vehicles for experiments

Sensors

Measurement equipment

GPU servers

RIKEN AIP (RAIDEN)