Play the Imitation Game: Model Extraction Attack against Autonomous Driving Localization

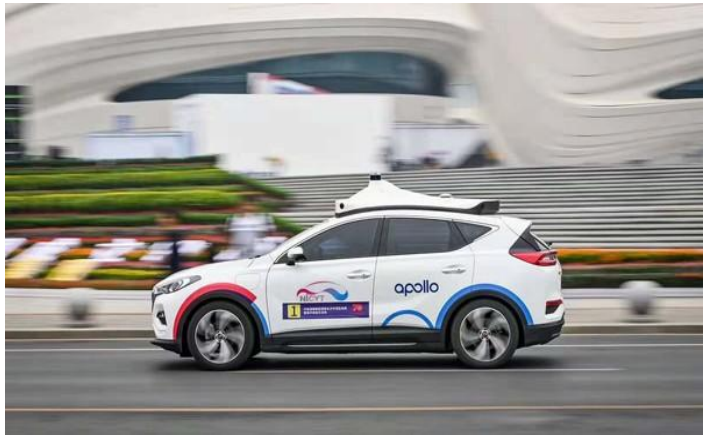
Qifan Zhang, Junjie Shen, Mingtian Tan,
Zhe Zhou, Zhou Li, Qi Alfred Chen and Haipeng Zhang

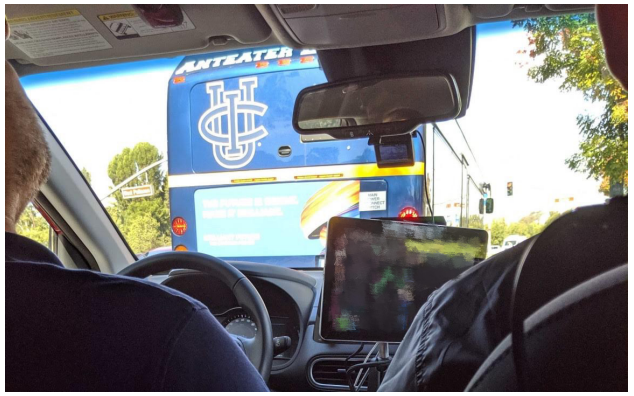
Presenter: **Qifan Zhang**, University of California, Irvine

December 7, 2022

Autonomous Driving

- **Autonomous Vehicles (AVs) are now on the road!**

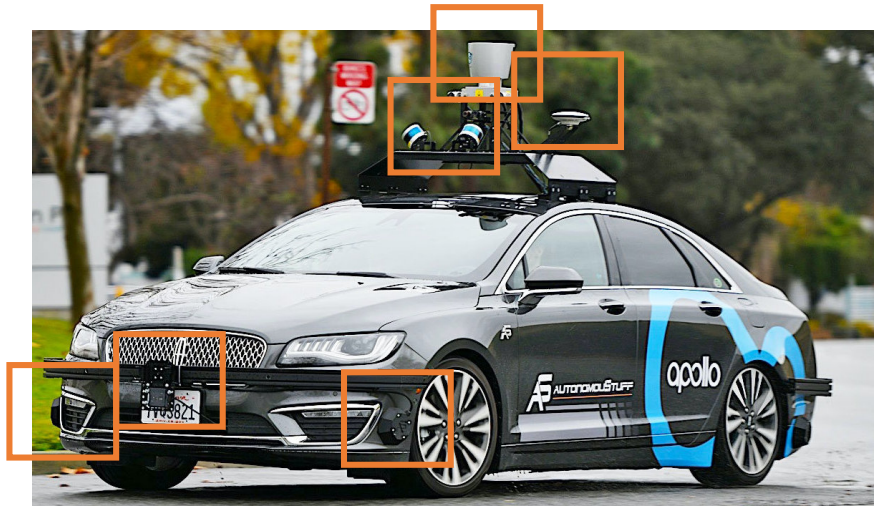




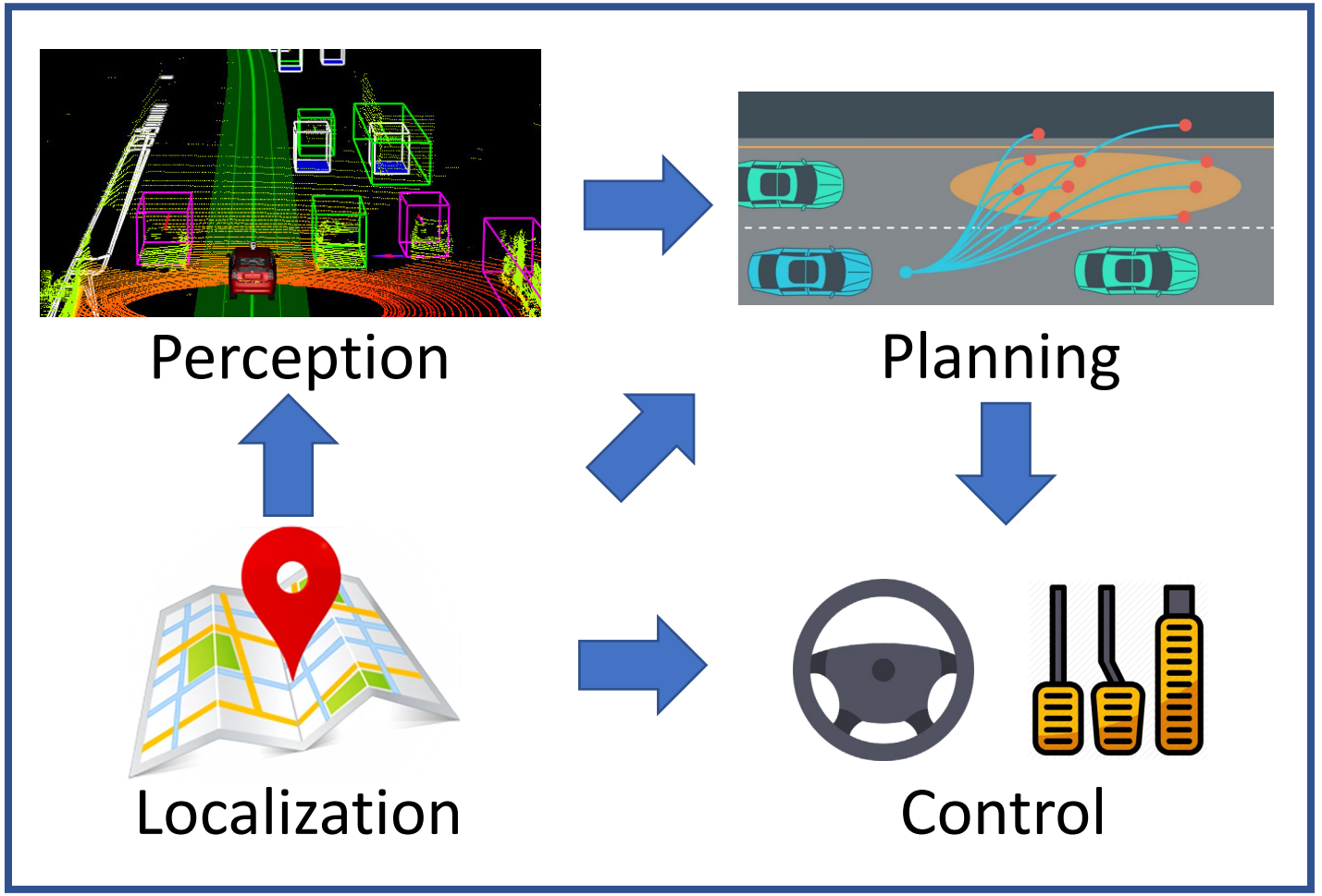
Autonomous Driving

➢High-level Autonomous Driving (AD) System

A typical Level-4 AV:

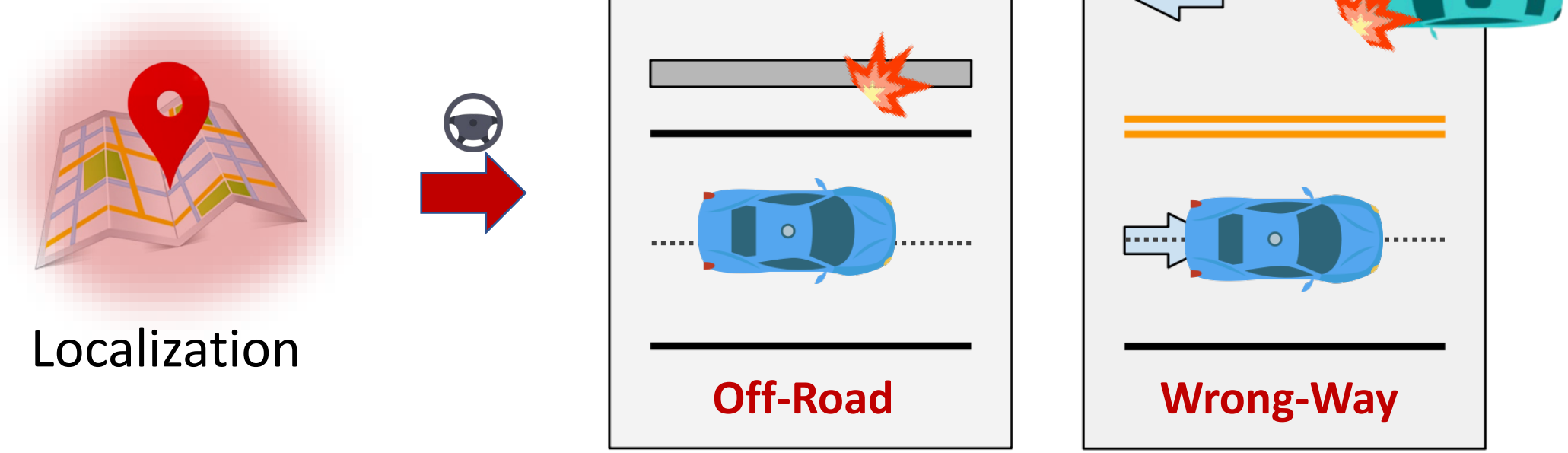


Abundant sensors:
LiDAR, GPS, IMU, Camera, Radar, etc.



Autonomous Driving

- **Localization is critical to the safety of AV**

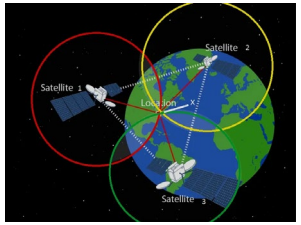


Reliability of Sensors

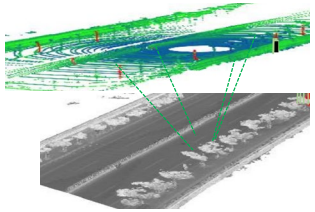
- **Sensors are not always reliable**
  - Atmosphere delays, multi-path effect, spoofing attacks, etc.
- **Sensors have different properties**
  - Different measurements, frequencies, precisions, etc.
- **Need to combine the measurements and make it robust!**
  - Normalize all the sensor data, and reduce errors
  - Multi-Sensor Fusion (MSF)

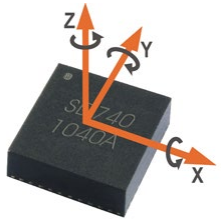
Multi-Sensor Fusion (MSF)

- **AD systems highly depend on MSF-based localization**
  - **Baidu Apollo**, [ICRA'18] [ITS'16] [IV'16] [Sensors'15], etc.
  - **Leverage strengths & compensate weaknesses** of different sensors to generally improve **accuracy & robustness**
    - Most popularly fuse from GPS, LiDAR, and IMU
    - Can achieve **5.4 cm** accuracy



GNSS


LiDAR locator



IMU

Multi-Sensor Fusion (MSF)

- **MSF are hard to tune**
  - Baidu Apollo ESKF confirmed that it takes **more than 6 months** for a specialized team to tune ESKF!
- **MSF models are important intellectual property!**
  - Competitors may want to steal for shorter tuning.
- **MSF models are heavily protected**
  - Could not be easily reverse engineered.

Baidu Apollo MSF

- **Baidu Apollo is one of the most popular AD systems**
- **Baidu Apollo MSF is heavily protected!**
  - Important intellectual property, close-sourced and obfuscated
- **Traditional reverse engineering is not workable**
  - Limited #queries to an AD system
  - Takes **more than 6 months** for a specialized team to tune
  - We tried to decompile the binary for **5 weeks** but failed

Is it possible to reverse engineer MSF? If so, what is needed?

Yes. We first need to know **MSF model structure**.

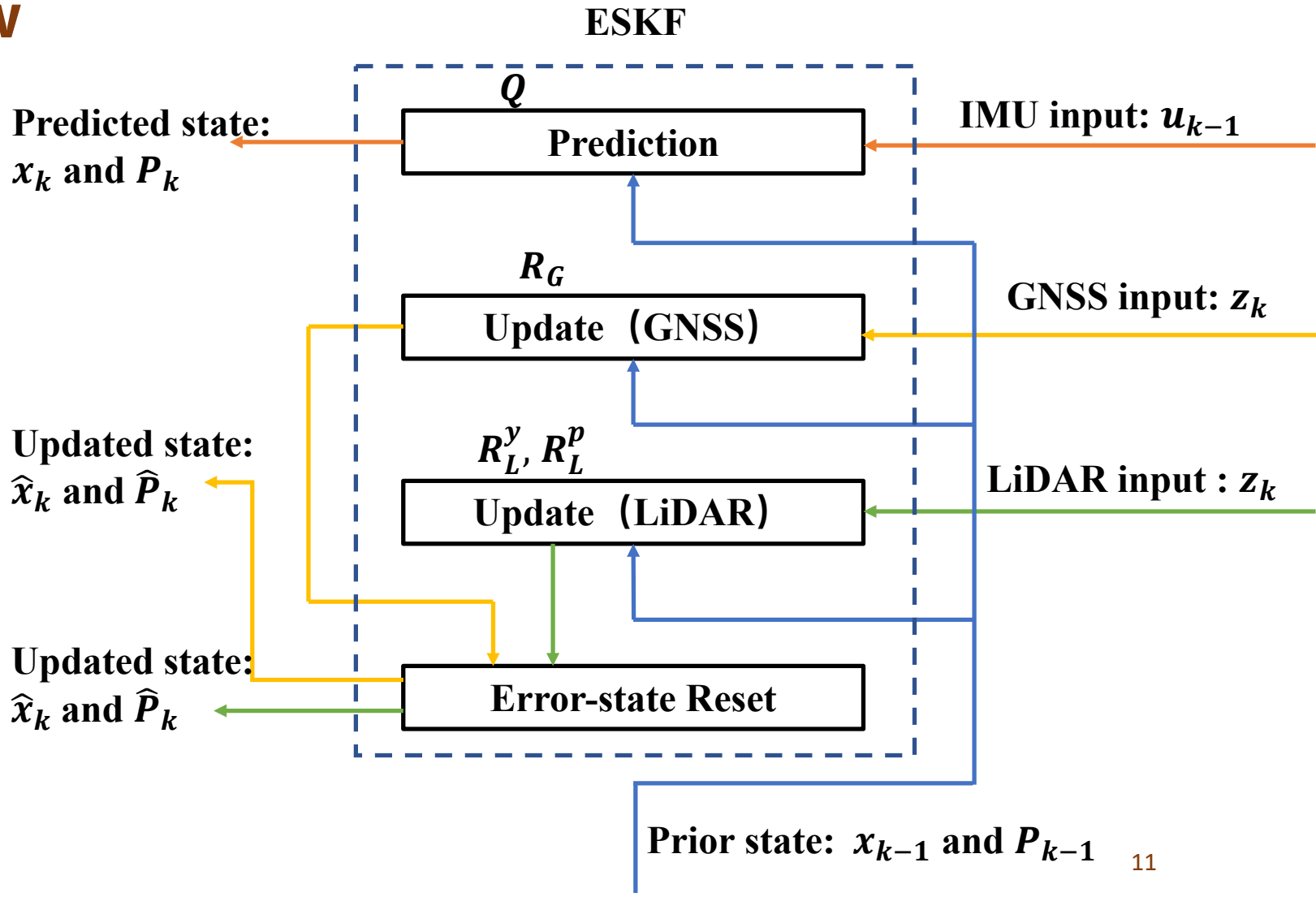
Error-State Kalman Filter (ESKF)

- **ESKF is a variant of Kalman Filter.**
  - Two parts: **Prediction** and **Update**
- **Baidu Apollo MSF is ESKF-based.**
  - It fuses IMU, LiDAR and GNSS sensor data
- **State components:**
  - Position, velocity, heading direction, accelerometer bias and gyrometer bias

Error-State Kalman Filter (ESKF)

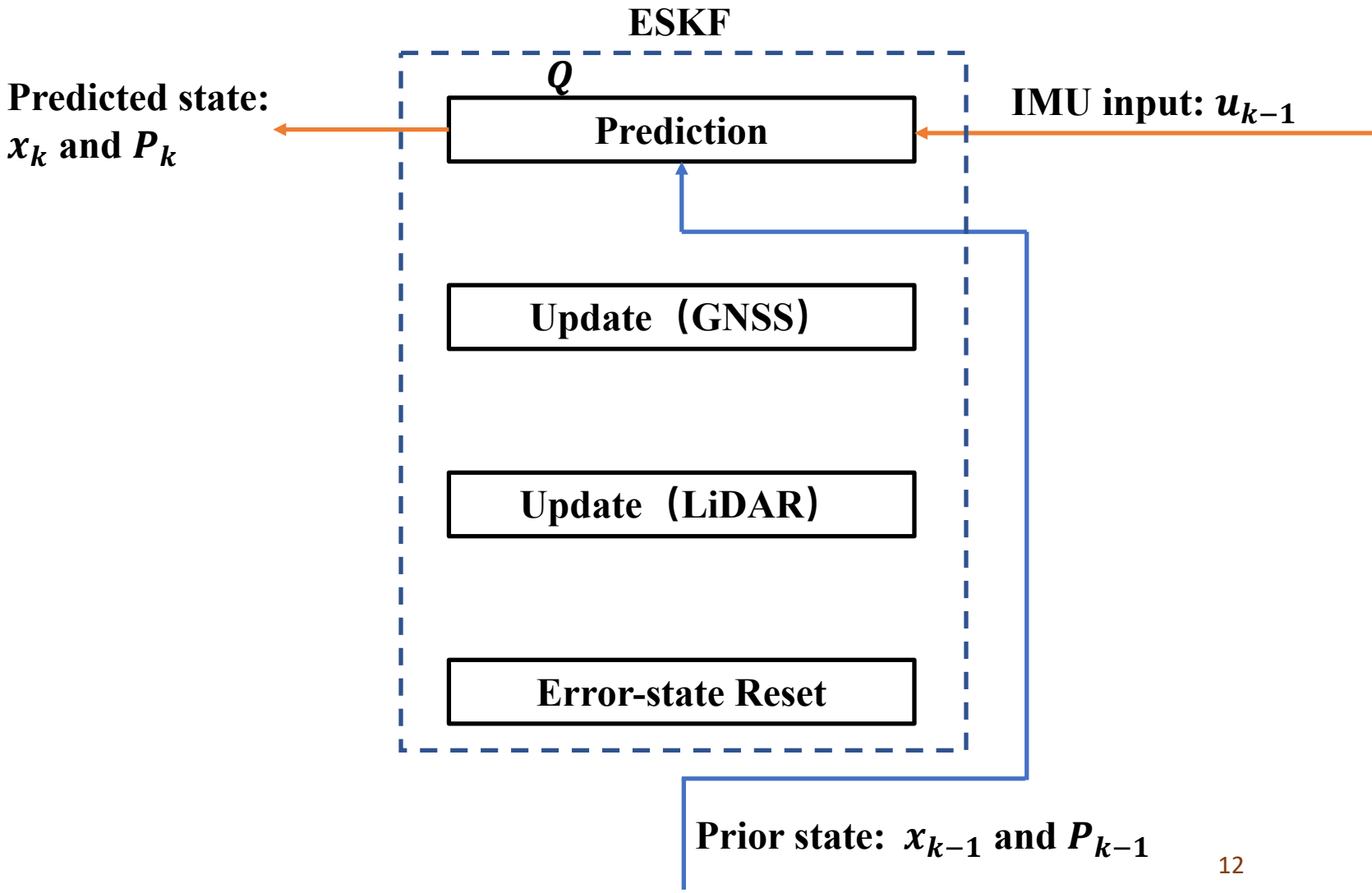
➢Structure overview

- ➢Q and R are covariance matrices for injected noises.
- ➢They are **secrets** of ESKF.



Error-State Kalman Filter (ESKF)

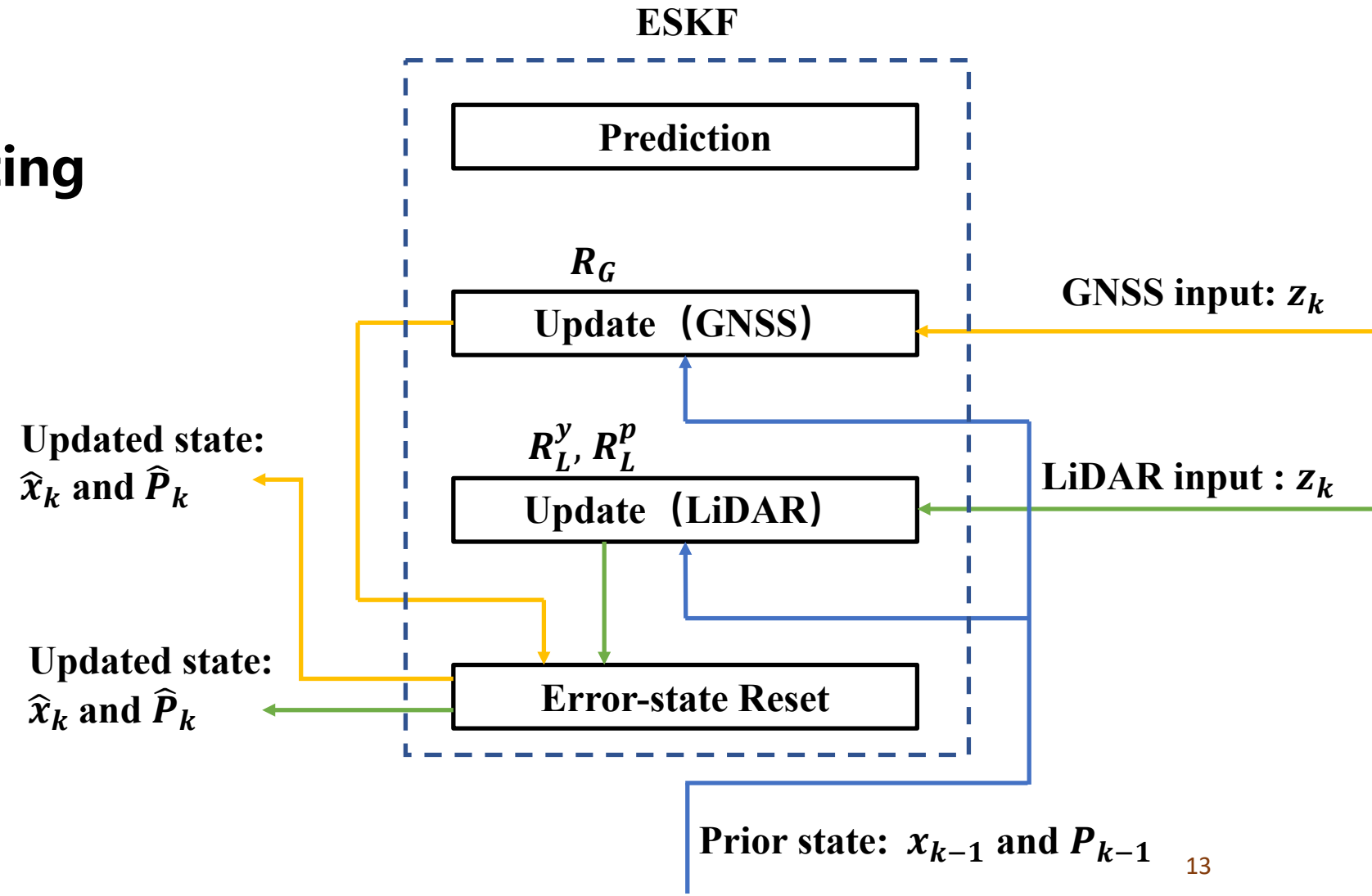
➢Prediction part



Error-State Kalman Filter (ESKF)

➢Update part

➢Reset part: avoid **observation drifting**



How to reverse engineer MSF?

Gradient-Descent based method
LSTM/RNN-like structure

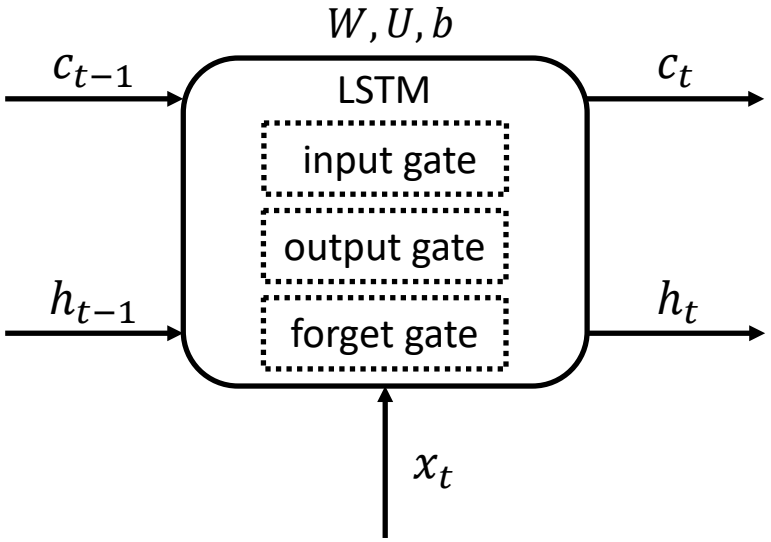
TaskMaster

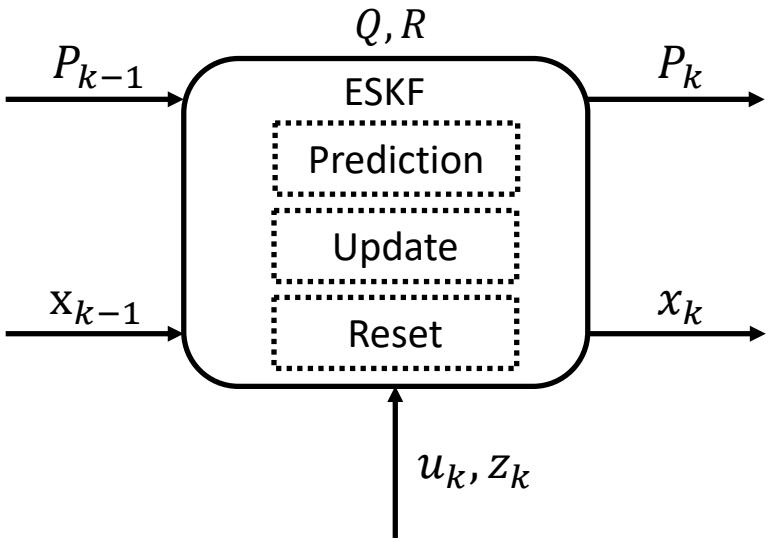
- **A Marvel fictional character**
  - Have the natural ability to ”copy” other heroes’ abilities
- **Our model could “absorb knowledge” from driving data**
  - Recover the parameters, by observing the input and output to the targeted AD system
  - High precision with 25 seconds AD sensor data for training

TaskMaster

➢Training strategy of ESKF

➢Derive the idea of training from LSTM/RNN





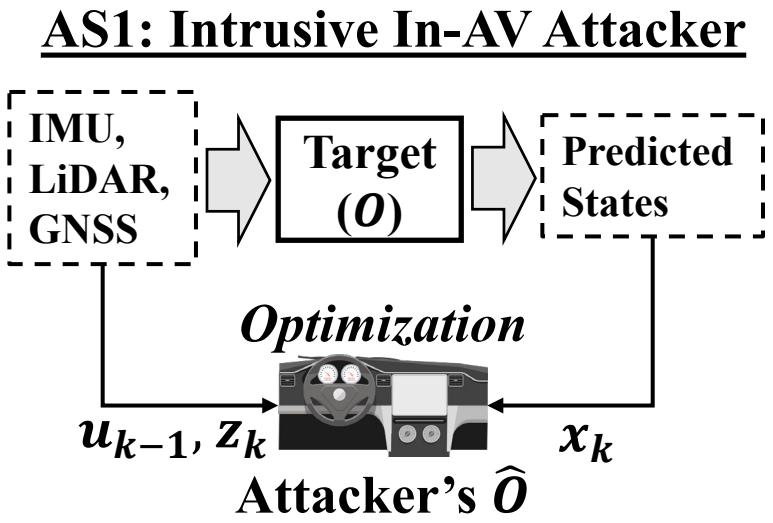
What attack scenarios should be considered?

3 scenarios, 2 models, 5 traces

Adversarial Model

➢AS1: Intrusive In-AV Attacker

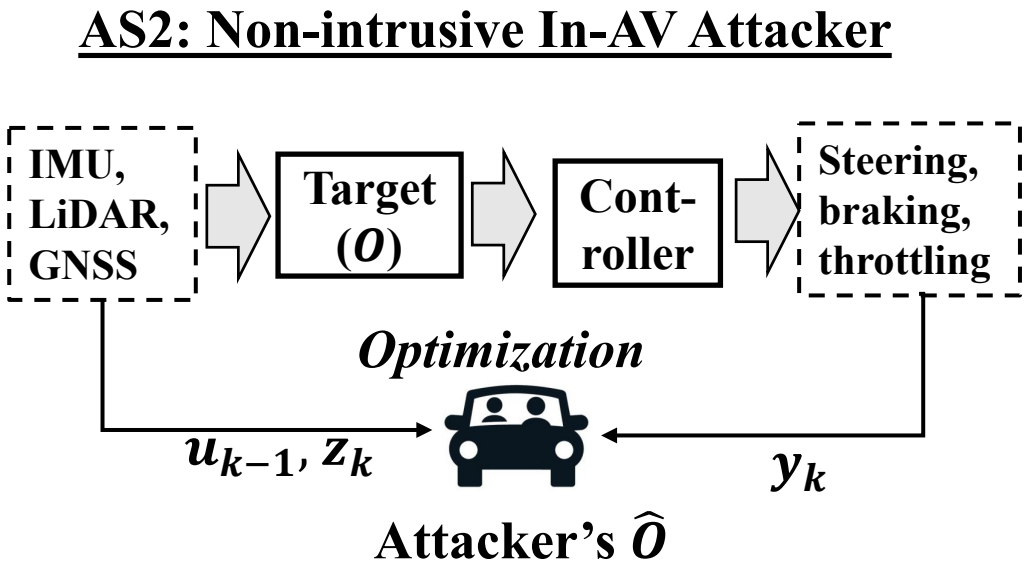
- ➢Attacker has **exclusive physical access** to the targeted AV
- ➢Observation of attackers:
  - ➢the input to ESKF, and the output from ESKF



Adversarial Model

➢AS2: Non-intrusive In-AV Attacker

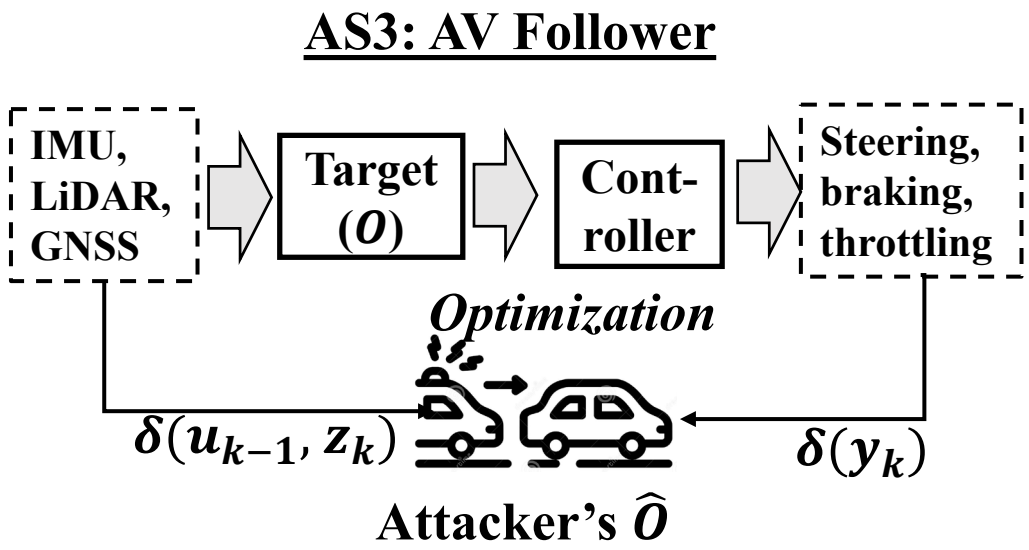
- ➢Attacker plugs onto the AV's Universal Serial Bus (USB)
- ➢Observation of attackers:
  - ➢output of the controllers
  - ➢We use PID and Stanley controller



Adversarial Model

➢AS3: AV Follower

- ➢Attacker drives another car to follow the AV in close vicinity
- ➢Observation of attackers:
  - ➢output of the controllers **with noise** ($\delta(\cdot)$ in the figure)



Evaluation

- **Dataset:**
  - **KAIST Complex Urban sensor traces[1]**
    - Dataset collects urban sensor data while driving in Seoul.
- **2 ESKF models**
  - **Sola-ESKF[2]** and Baidu **Apollo-ESKF**
- **Evaluation metrics:**
  - Parameter Error Rate (**PER**)
    - Distance between **parameters**
  - State Error Rate (**SER**)
    - Distance between predicted/updated **states**

[1]: Jeong, Jinyong, et al. "Complex urban dataset with multi-level sensors from highly diverse urban environments." *The International Journal of Robotics Research* 38.6 (2019): 642-657.
[2]: Sola, Joan. "Quaternion kinematics for the error-state Kalman filter." arXiv preprint arXiv:1711.02508 (2017).

How to optimize the reverse engineering process?

Search-space reduction
Multi-Stage optimization
Controller simulation

Attack Implementation

- **Search-space reduction**
  - Q and R are set as **diagonal matrices** based on KF properties
- **Multi-stage optimization**
  - Adopt different **#epochs** and **learning rates** for Q and R
  - Optimize Q and R **separately**
- **Controller simulation**
  - Controllers in Baidu Apollo are close sourced
  - We use **PID** and **Stanley** controllers to stimulate the original one.

How does it perform?

cm-level on AS1 and AS2, Sola-ESKF
dm-level on AS1, Apollo-ESKF and AS2
meter-level on AS3

Evaluation

➢Centimeter-level precision in AS1 and AS2, Sola-ESKF

AS	Training	PER	Testing (SER)				
			lo08	*lo07*	*lo31*	*hi06*	*hi17*
AS1	*lo08*	0.01347	-	0.067	0.042	0.076	0.089
	hi17	0.01297	0.043	0.029	0.073	0.038	-
AS2	*lo08*	0.00978	-	0.018	0.028	0.026	0.036
	hi17	0.00206	0.024	0.013	0.018	0.036	-
AS3E	*lo08*	4.5431	-	1.76	1.24	1.33	3.15
	hi17	4.6247	1.43	2.88	2.28	4.19	-
AS3G	*lo08*	3.9916	-	1.53	1.89	2.57	1.21
	hi17	4.3111	1.51	0.58	1.43	5.67	-
AS3R	*lo08*	5.8869	-	4.24	2.10	1.11	1.33
	hi17	4.2485	2.12	1.87	4.11	0.88	-

Table 1: Evaluation result on `Sola-ESKF`. The result of each testing trace is represented as SER. PER is the same for each training trace. PER could be larger than 1 if the error between the extracted value and the ground truth is larger than the ground truth itself. AS3E, AS3G and AS3R are AS3 under Exponential, Gamma and Rayleigh noises. "lo" and "hi" are short for "local" and "highway".

AS	Training	Testing (SER)				
		lo08	*lo07*	*lo31*	*hi06*	*hi17*
AS1	*lo08*	-	0.37	0.42	0.91	1.18
	hi17	0.95	0.72	0.68	0.89	-
AS2	*lo08*	-	1.26	1.01	1.03	0.62
	hi17	1.12	0.96	0.88	0.79	-
AS3E	*lo08*	-	1.72	1.82	2.03	1.96
	hi17	1.92	1.48	2.41	1.27	-
AS3G	*lo08*	-	1.45	1.63	2.11	3.07
	hi17	1.26	1.43	1.55	1.90	-
AS3R	*lo08*	-	1.78	2.08	2.13	1.49
	hi17	1.71	1.82	1.56	1.33	-

Table 2: Evaluation result on `Apollo-ESKF`. The result in each cell is represented as SER, as `Apollo-ESKF` is blackbox.

Evaluation

➢Decimeter-level precision in AS1 & AS2, Apollo-ESKF

AS	Training	PER	Testing (SER)				
			lo08	*lo07*	*lo31*	*hi06*	*hi17*
AS1	*lo08*	0.01347	-	0.067	0.042	0.076	0.089
	hi17	0.01297	0.043	0.029	0.073	0.038	-
AS2	*lo08*	0.00978	-	0.018	0.028	0.026	0.036
	hi17	0.00206	0.024	0.013	0.018	0.036	-
AS3E	*lo08*	4.5431	-	1.76	1.24	1.33	3.15
	hi17	4.6247	1.43	2.88	2.28	4.19	-
AS3G	*lo08*	3.9916	-	1.53	1.89	2.57	1.21
	hi17	4.3111	1.51	0.58	1.43	5.67	-
AS3R	*lo08*	5.8869	-	4.24	2.10	1.11	1.33
	hi17	4.2485	2.12	1.87	4.11	0.88	-

Table 1: Evaluation result on `Sola-ESKF`. The result of each testing trace is represented as SER. PER is the same for each training trace. PER could be larger than 1 if the error between the extracted value and the ground truth is larger than the ground truth itself. AS3E, AS3G and AS3R are AS3 under Exponential, Gamma and Rayleigh noises. "lo" and "hi" are short for "local" and "highway".

AS	Training	Testing (SER)				
		lo08	*lo07*	*lo31*	*hi06*	*hi17*
AS1	*lo08*	-	0.37	0.42	0.91	1.18
	hi17	0.95	0.72	0.68	0.89	-
AS2	*lo08*	-	1.26	1.01	1.03	0.62
	hi17	1.12	0.96	0.88	0.79	-
AS3E	*lo08*	-	1.72	1.82	2.03	1.96
	hi17	1.92	1.48	2.41	1.27	-
AS3G	*lo08*	-	1.45	1.63	2.11	3.07
	hi17	1.26	1.43	1.55	1.90	-
AS3R	*lo08*	-	1.78	2.08	2.13	1.49
	hi17	1.71	1.82	1.56	1.33	-

Table 2: Evaluation result on `Apollo-ESKF`. The result in each cell is represented as SER, as `Apollo-ESKF` is blackbox.

Evaluation

➢Meter-level precision in AS3

AS	Training	PER	Testing (SER)				
			lo08	*lo07*	*lo31*	*hi06*	*hi17*
AS1	*lo08*	0.01347	-	0.067	0.042	0.076	0.089
	hi17	0.01297	0.043	0.029	0.073	0.038	-
AS2	*lo08*	0.00978	-	0.018	0.028	0.026	0.036
	hi17	0.00206	0.024	0.013	0.018	0.036	-
AS3E	*lo08*	4.5431	-	1.76	1.24	1.33	3.15
	hi17	4.6247	1.43	2.88	2.28	4.19	-
AS3G	*lo08*	3.9916	-	1.53	1.89	2.57	1.21
	hi17	4.3111	1.51	0.58	1.43	5.67	-
AS3R	*lo08*	5.8869	-	4.24	2.10	1.11	1.33
	hi17	4.2485	2.12	1.87	4.11	0.88	-

Table 1: Evaluation result on `Sola-ESKF`. The result of each testing trace is represented as SER. PER is the same for each training trace. PER could be larger than 1 if the error between the extracted value and the ground truth is larger than the ground truth itself. AS3E, AS3G and AS3R are AS3 under Exponential, Gamma and Rayleigh noises. "lo" and "hi" are short for "local" and "highway".

AS	Training	Testing (SER)				
		lo08	*lo07*	*lo31*	*hi06*	*hi17*
AS1	*lo08*	-	0.37	0.42	0.91	1.18
	hi17	0.95	0.72	0.68	0.89	-
AS2	*lo08*	-	1.26	1.01	1.03	0.62
	hi17	1.12	0.96	0.88	0.79	-
AS3E	*lo08*	-	1.72	1.82	2.03	1.96
	hi17	1.92	1.48	2.41	1.27	-
AS3G	*lo08*	-	1.45	1.63	2.11	3.07
	hi17	1.26	1.43	1.55	1.90	-
AS3R	*lo08*	-	1.78	2.08	2.13	1.49
	hi17	1.71	1.82	1.56	1.33	-

Table 2: Evaluation result on `Apollo-ESKF`. The result in each cell is represented as SER, as `Apollo-ESKF` is blackbox.

Conclusion

- **First study on confidentiality issues in AD control models**
- **Design a novel optimization-based framework to infer the secret parameters**
  - By observing the input and output of an AD
- **Achieve high accuracy for Sola-ESKF and comparable accuracy for Apollo-ESKF**

Thanks for listening!

Any question?

Qifan Zhang, University of California, Irvine
qifan.zhang@uci.edu