Securing Autonomy for Contested World

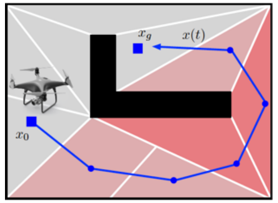
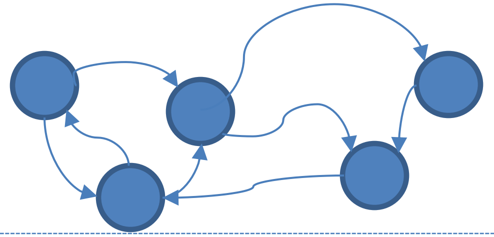
Miroslav Pajic

CPSL@Duke
Department of Electrical and Computer Engineering
Department of Computer Science
Department of Mechanical Engineering & Material Science
Duke University

Duke
PRATT SCHOOL of ENGINEERING

DUKE ROBOTICS

Security-Aware Autonomy
Vulnerability Analysis and Providing Resiliency

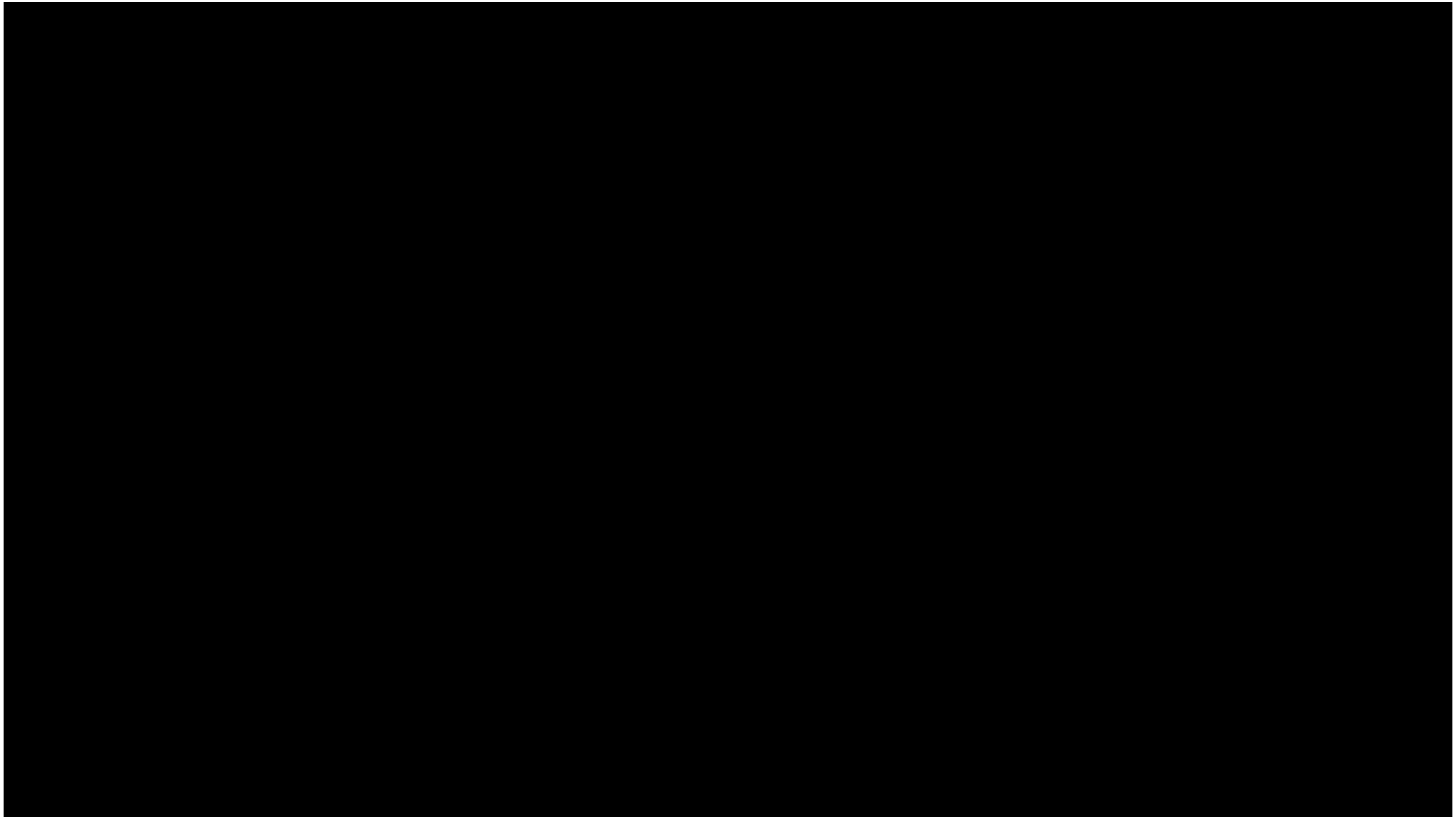
Control Stack	Control view	Modeling view
Mission Planner	Long-horizon views	
Tactical Planner	Short-horizon views	
Low-level Control	Continuous/discrete control with constraints	$f_r(x(t)) = \int_0^T \varrho(x(t), h(t))dt + \int_0^T \lVert x(t) \rVert^2 dt,$ $\min \; f_r(x_r(t)) + f_h(x_h(t))$ $\text{s.t.} \; x_r(t) = x_h(t), \;\; u_r(t) = u_h(t),$
Vehicle		

Our Goal: Add resiliency to controls across different/all levels of the autonomy stack

Non-Secure Vehicle Platooning

Secure Vehicle Platooning

With Intermittent Integrity Guarantees



Fundamental Autonomy Limitations Under Attack (low-level control)

Duke PRATT SCHOOL *of* ENGINEERING

What can we say in general about resiliency of *(perception-based) control & decision making*?

- Khazraei, H. Meng, and M. Pajic, "Stealthy Perception-based Attacks on Unmanned Aerial Vehicles", IEEE International Conference on Robotics and Automation (ICRA), pp. 3346-3352, June 2023.
- A. Khazraei, H. Meng, and M. Pajic, "Black-box Stealthy GPS Attacks on Unmanned Aerial Vehicles", 63rd IEEE Conf. on Decision and Control (CDC), Dec. 2024.
- Khazraei, H. Pfister, and M. Pajic, "Attacks on Perception-Based Control Systems: Modeling and Fundamental Limits", TAC, 2024.
- A. Khazraei, H. Pfister, and M. Pajic, "Attacks on Perception-Based Control Systems: Modeling and Fundamental Limits", IEEE TAC, revised.

General Vulnerability Analysis for False Data-Injection Attacks on Control Systems (TAC24, TAC24b*)

An attack sequence

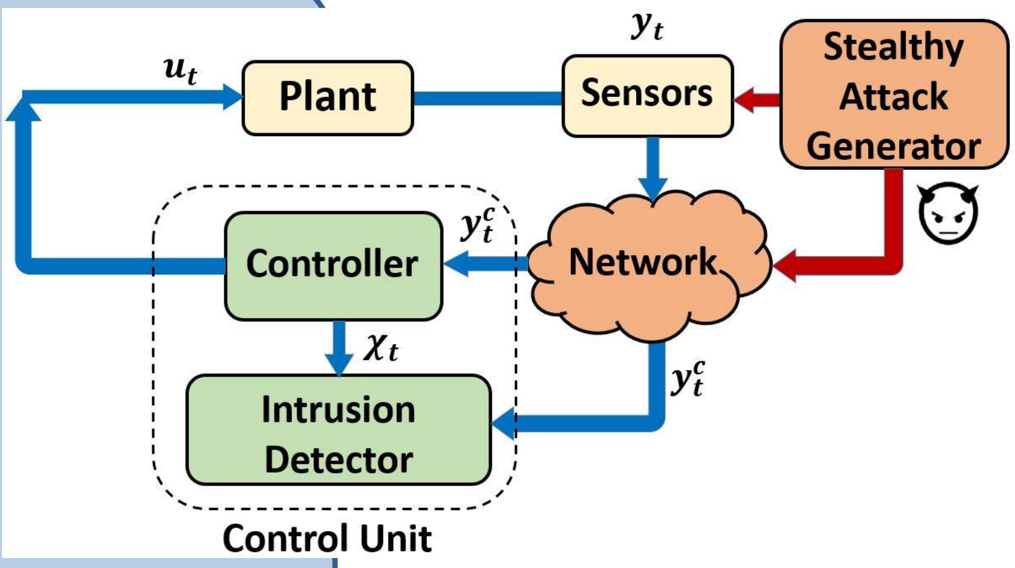
- is **strictly stealthy** iff

$$\boldsymbol{KL}\big(\boldsymbol{Q}(Y_{-\infty}^{-1}, Y_0^a : Y_t^a) || \boldsymbol{P}(Y_{-\infty} : Y_t)\big) = 0$$

for any $t \geq 0$,

- is **ϵ-stealthy** if

$$\boldsymbol{KL}\big(\boldsymbol{Q}(Y_{-\infty}^{-1}, Y_0^a : Y_t^a) || \boldsymbol{P}(Y_{-\infty} : Y_t)\big) \leq \log(\frac{1}{1-\epsilon^2}) \text{ for any } t \geq 0.$$



The system is (ϵ, α)-attackable for arbitrarily large α and arbitrarily small ϵ, if the closed-loop dynamics is incrementally exponentially stable (IES) in the set S and the open loop dynamics is incrementally unstable in the set S.

Attack Strategy I: Using Estimate of the Plant State

Attack injection

$$z_t^a = G(x_t^a - s_t)$$

$$y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s$$

Attack dynamics: $s_{t+1} = f(\hat{x}_t^a) - f(\hat{x}_t^a - s_t)$

Assumption: $\zeta = x_t^a - \hat{x}_t^a, \quad \|\zeta\| \le b_\zeta$

Idea:
Fake state $\mathrm{e} = x_t^a - s_t$,

Theorem: Assume that the functions f, f' and Π'(i.e., derivatives of f and Π) are Lipschitz with constants L_f, L_f' and L_Π', respectively, and let us define

$$L_1 = L_f'(b_x + 2b_\zeta + d), L_2 = min\{2L_f, L_f'(\alpha + b_x + b_\zeta\} \text{ and } L_3 = L_\Pi'(b_x + d + b_v).$$

Moreover, assume that b_x has the maximum value such that the inequalities

$$L_1 + L_3\|B\| < \frac{c_3}{c_4} \text{ and } L_2 b_\zeta < \frac{c_3 - (L_1 + L_3\|B\|)c_4}{c_4}\sqrt{\frac{c_1}{c_2}}\theta r \text{ for some } 0 < \theta < 1\text{, are satisfied.}$$

Then, the system is (ϵ, α)**-attackable** with probability $\delta(T(\alpha + b + b_x, s_0), b_x, b_v)$ for some $\epsilon > 0$, if $f \in \mathcal{U}_\rho$ with $\rho = 2L_f(b + b_x + b_\zeta)$ and $b = \frac{c_4}{c_3 - (L_1 + L_3\|B\|)c_4}\sqrt{\frac{c_2}{c_1}}\frac{L_2 b_\zeta}{\theta}$.

General Perception-Based Attacks

Stealthy & Effective Attacks (ICRA'23, TAC'24, TAC24*)

- Attack Strategy Ⅰ: Using an estimate of the plant state

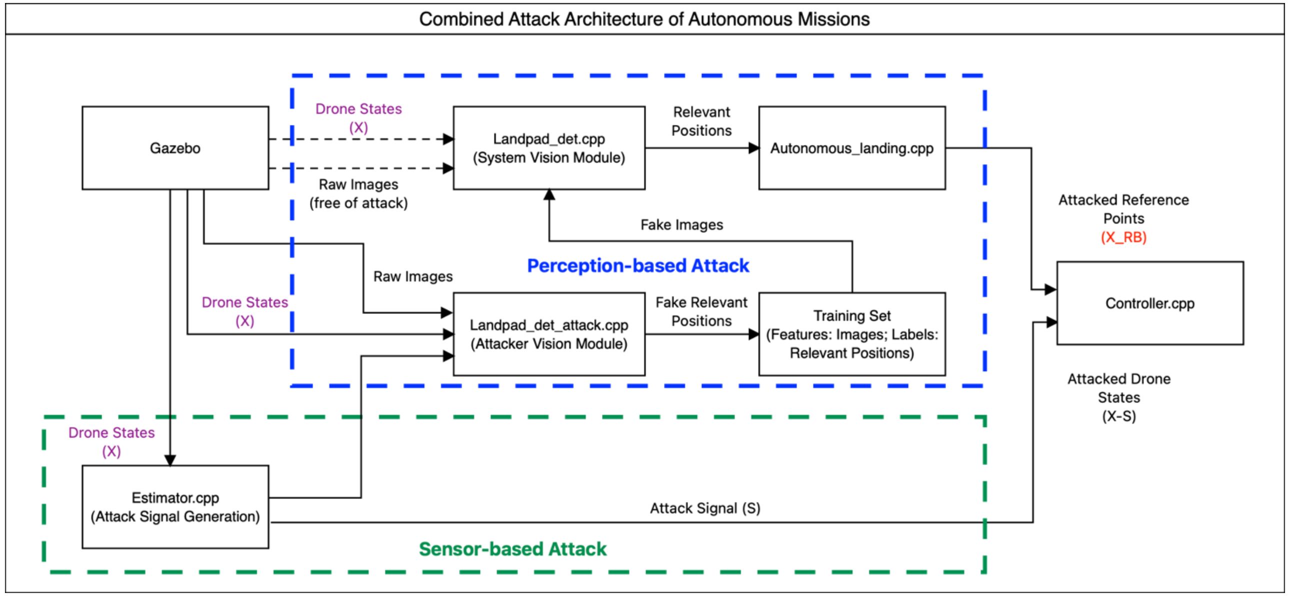
$$z_t^a = G(x_t^a - s_t) \qquad s_{t+1} = f(\hat{x}_t^a) - f(\hat{x}_t^a - s_t)$$

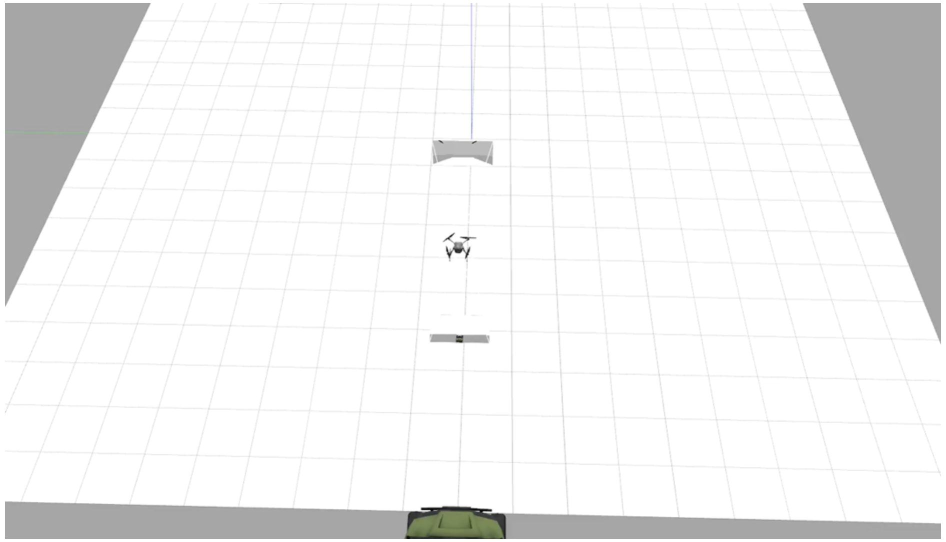
$$y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s \qquad \zeta = x_t^a - \hat{x}_t^a, \quad \|\zeta\| \le b_\zeta$$

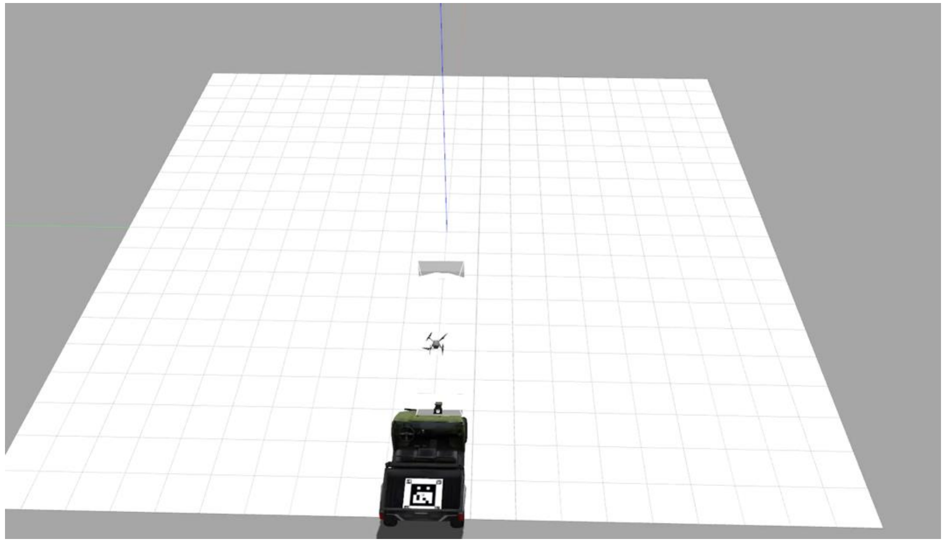
- Attack Strategy II: Without an estimate of the state

$$z_t^a = G(x_t^a - s_t) \qquad s_{t+1} = f(s_t)$$

$$y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s$$







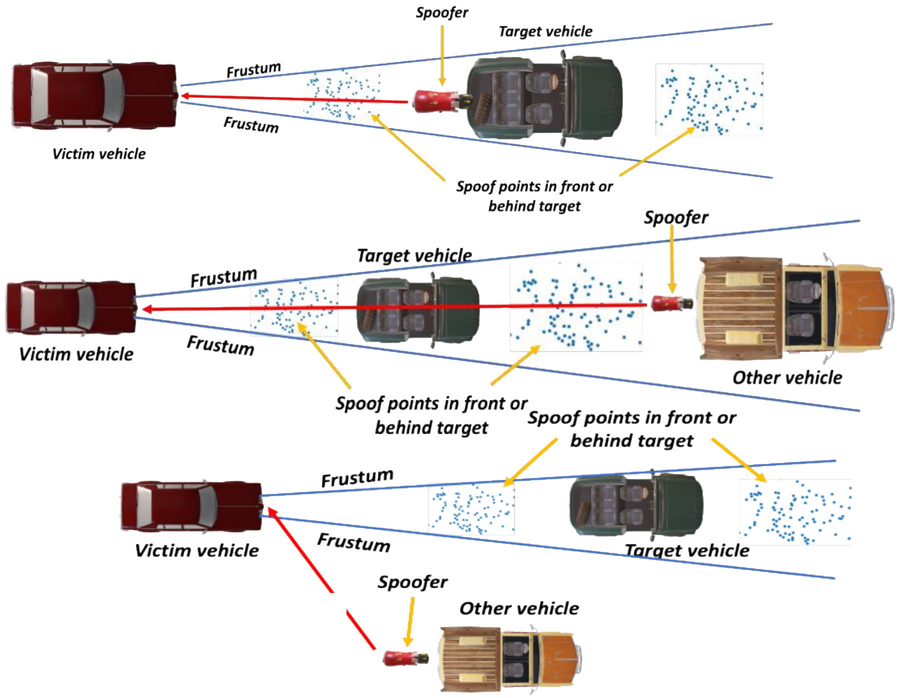
Attacking Camera-LiDAR Perception

- How feasible are such attacks?
  - Physical dynamics – time-series analysis!
- Beyond Naïve Attack: Novel Frustum Attack

Attack injection

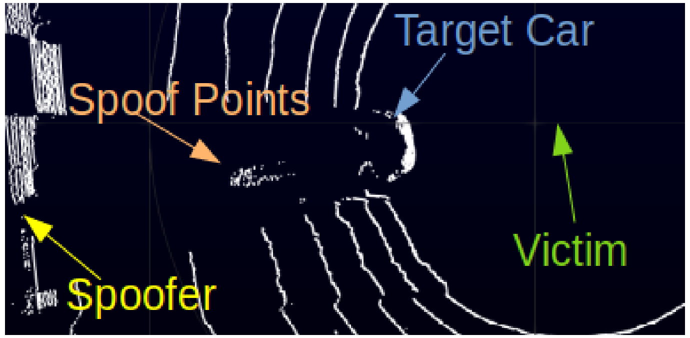
$$z_t^a = G(x_t^a - s_t)$$
$$y_t^{s,a} = C_s(x_t^a - s_t) + v_t^s$$

Three candidate realizations of the frustum attack. Additional configurations shown later



Target car in front of victim

Spoofer set behind target car



Stable spoof points placed in frustum

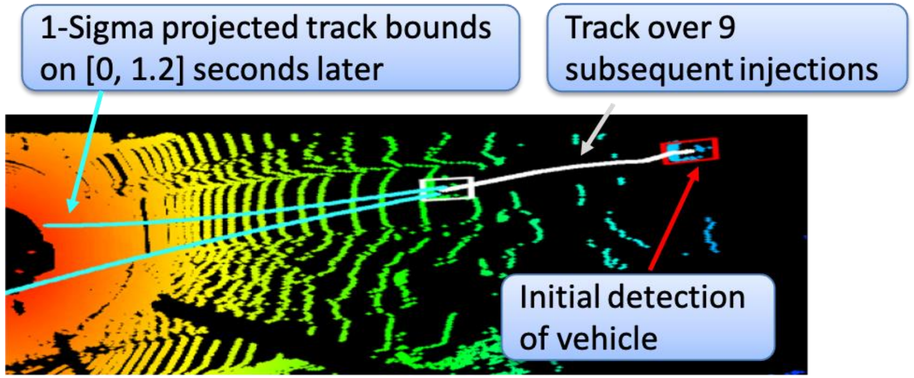
S. Hallyburton, Y. Liu, Y. Cao, Z. M. Mao, and M. Pajic, "Security Analysis of Camera-LiDAR Fusion Against Black-Box Attacks on Autonomous Vehicles", *31st* **USENIX SECURITY**, 2022.

Frustum Attacks on Camera-LiDAR Fusion (Usenix Security'22)

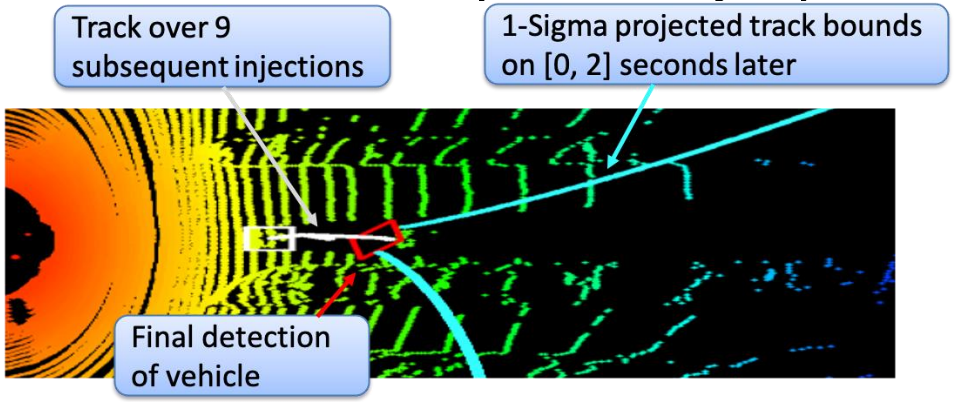
Duke
PRATT SCHOOL of ENGINEERING

Evaluation of Multi-Frame Tracking

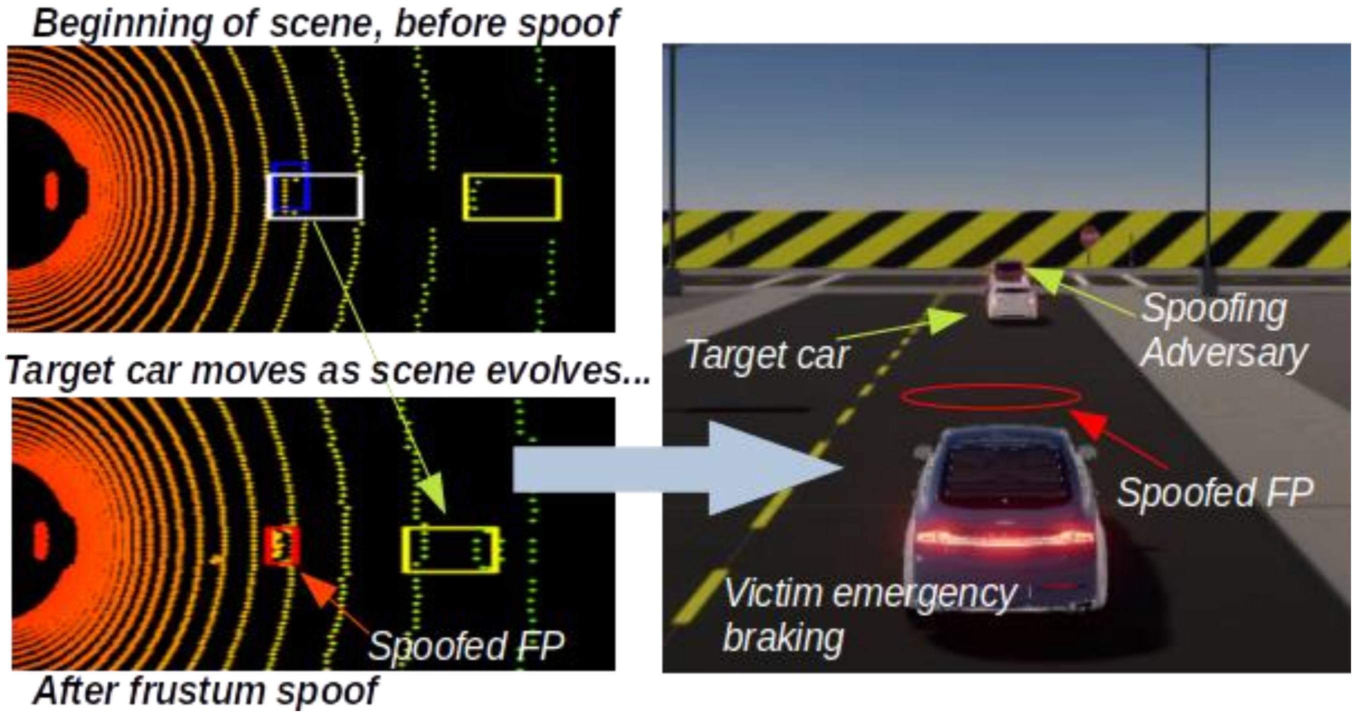
False positive car accelerates towards victim



Translation attack shows real object accelerating away from victim



Evaluation on industry-grade AVs: Baidu's Apollo + SVL



Tracking case studies show only few compromised frames cause safety-critical predicted outcomes

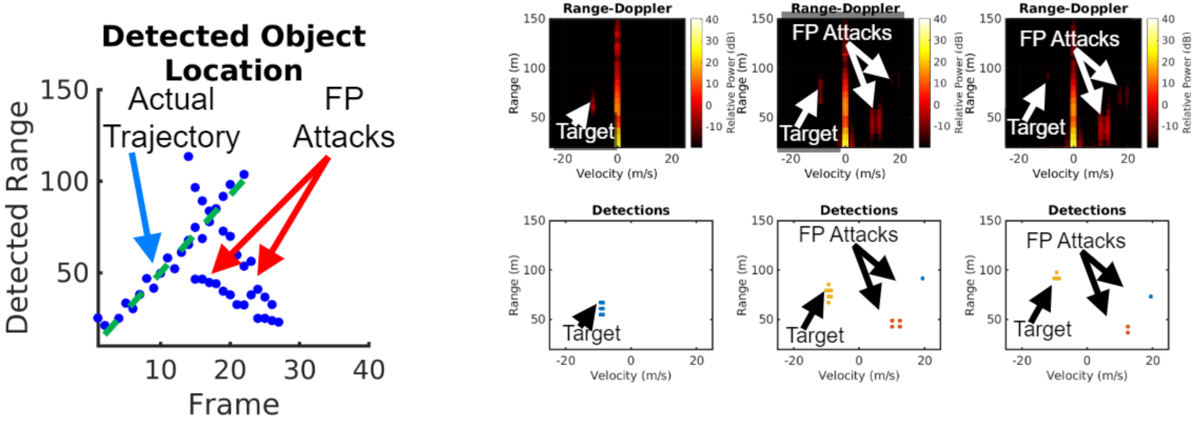
Baidu case study shows even industry-level AVs are vulnerable to frustum attack

Vulnerability Analysis of mmWave Radars

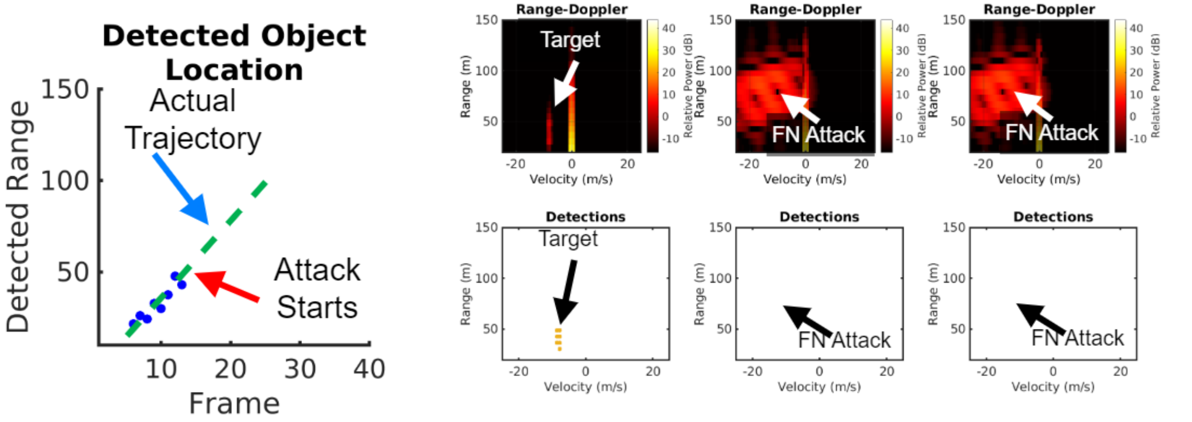
MadRadar: A Black-Box Physical Layer Attacks (NDSS'24)

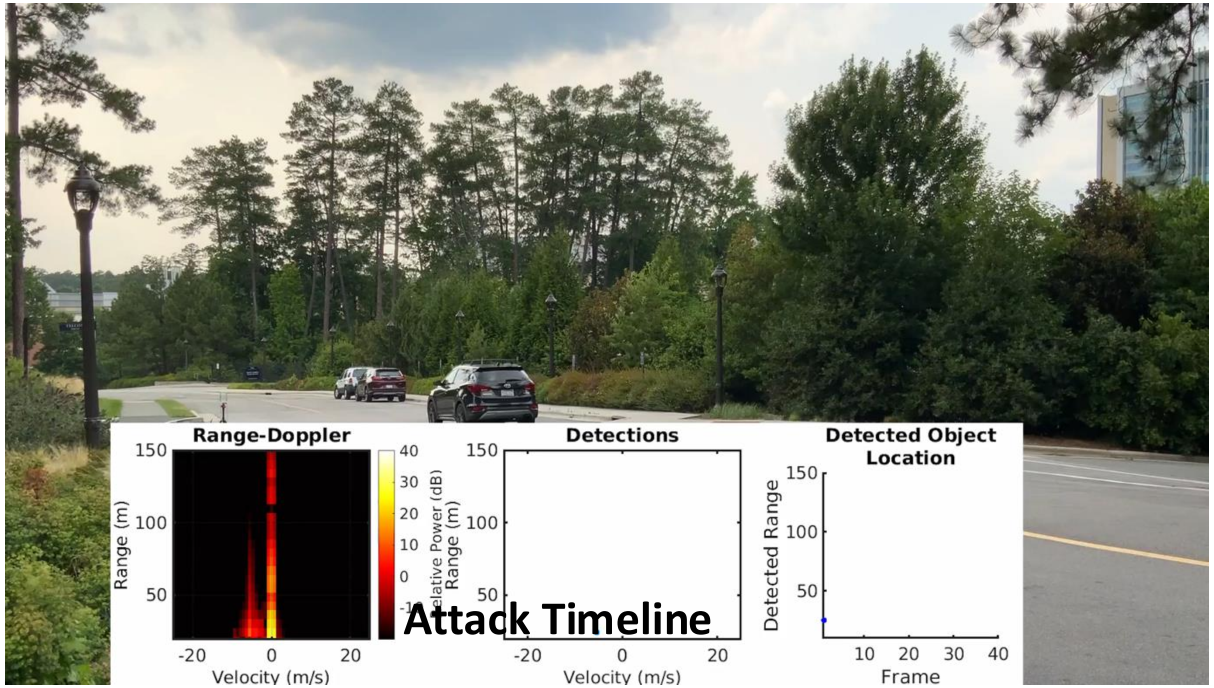
Duke
PRATT SCHOOL of ENGINEERING

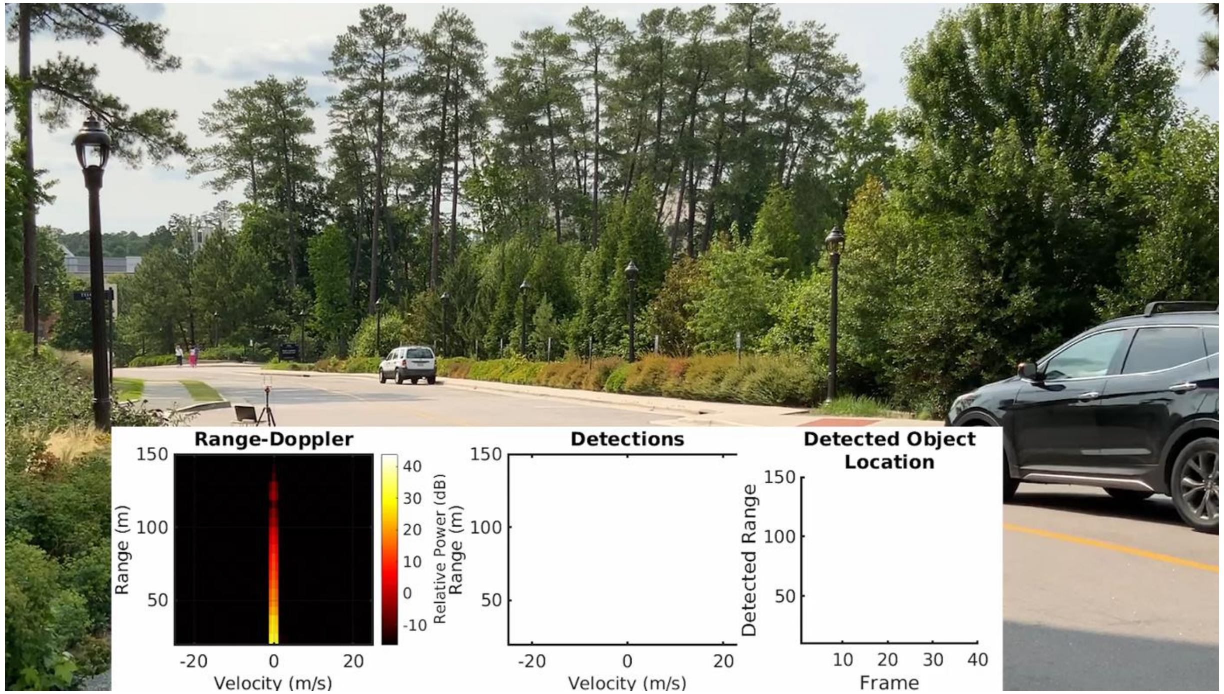
False Positive Attacks



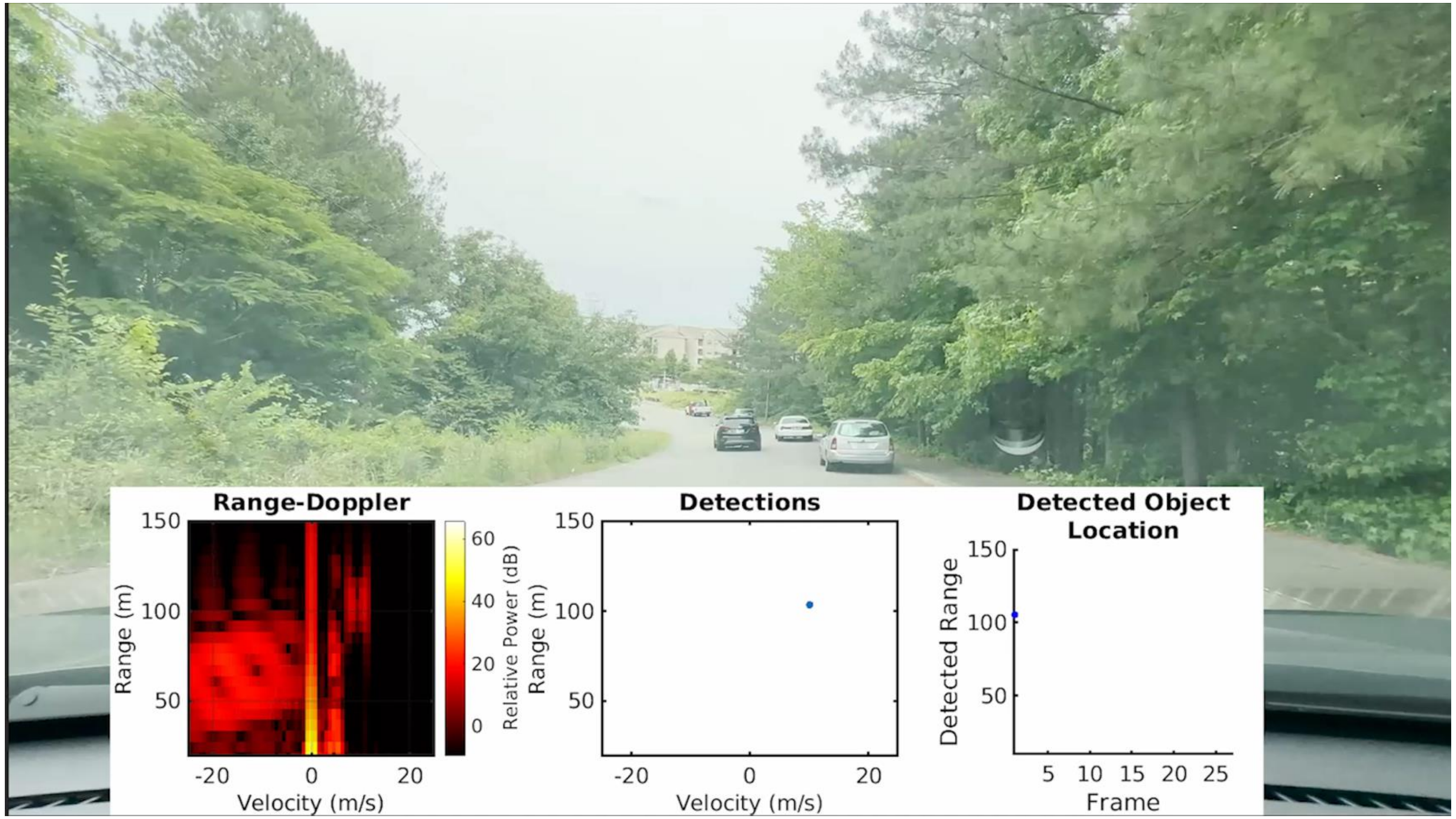
False Negative Attacks







Additional Attacks



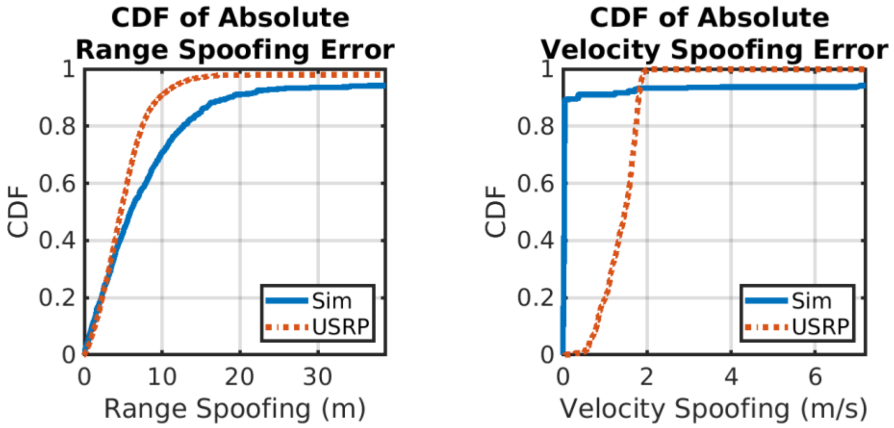


TABLE VI: Absolute Error of the Attack Spoofing Accuracy

Metric	Mean Absolute Error	90th Percentile
Range	7.53 m	9.67 m
Velocity	1.42 m/s	1.80 m/s

Duke
PRATT SCHOOL of ENGINEERING

How can we provide resiliency?

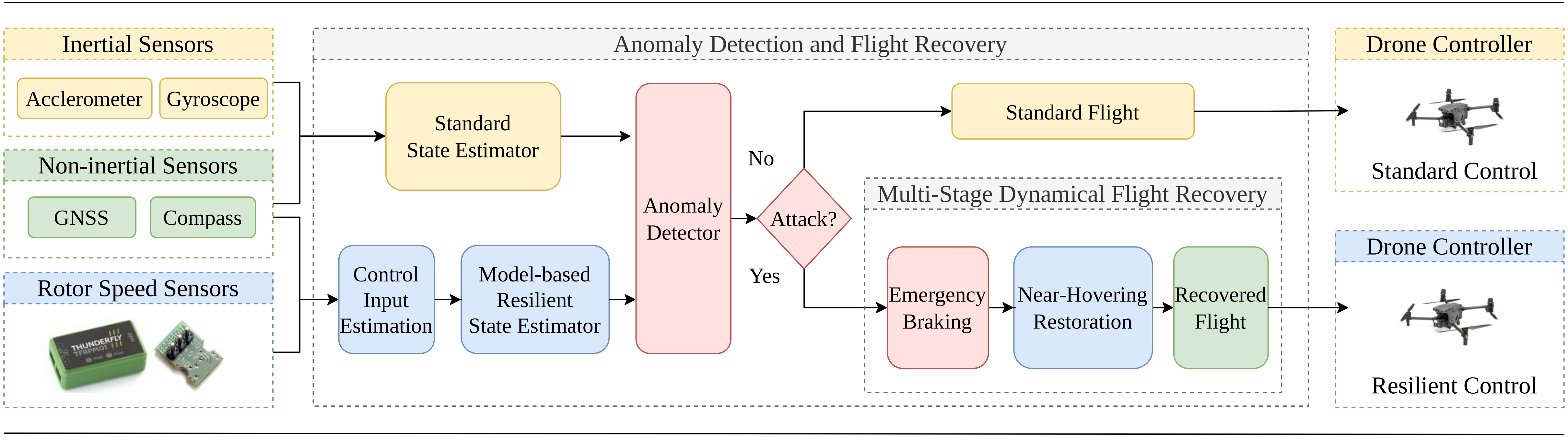
- Use new sensing modalities
- Platform aware use of security primitives

Defending UAVs Against Acoustic Attacks

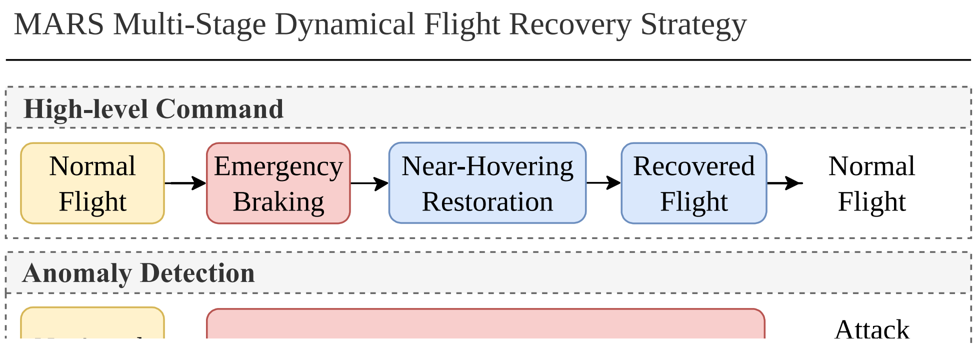
Duke PRATT SCHOOL of ENGINEERING

- Defending Unmanned Aerial Vehicles From Attacks on Inertial Sensors with Model-based Anomaly Detection and Recovery

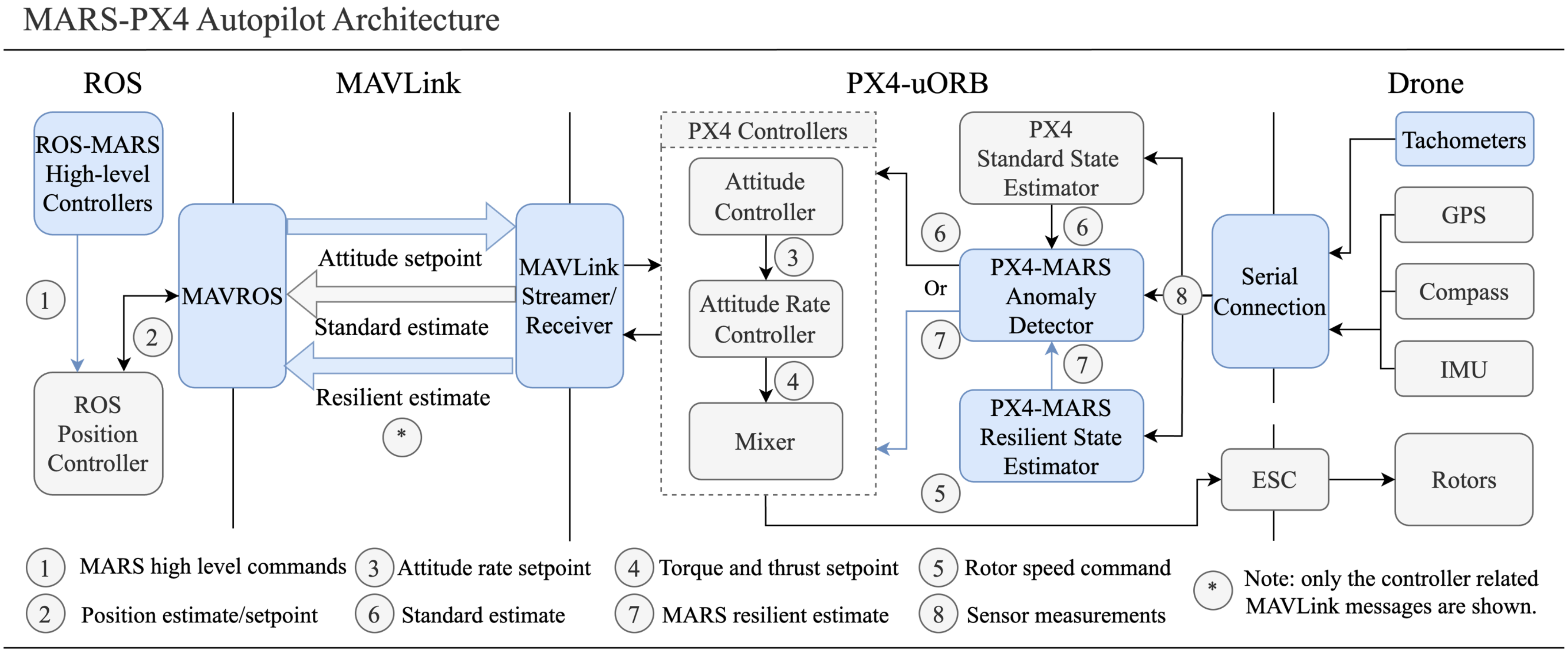
Model-based Anomaly Detection and Recovery System (MARS)



Approach



MARS Multi-Stage Dynamical Flight Recovery Strategy
High-level Command
Normal Flight
Emergency Braking
Near-Hovering Restoration
Recovered Flight
Normal Flight
Anomaly Detection
Attack



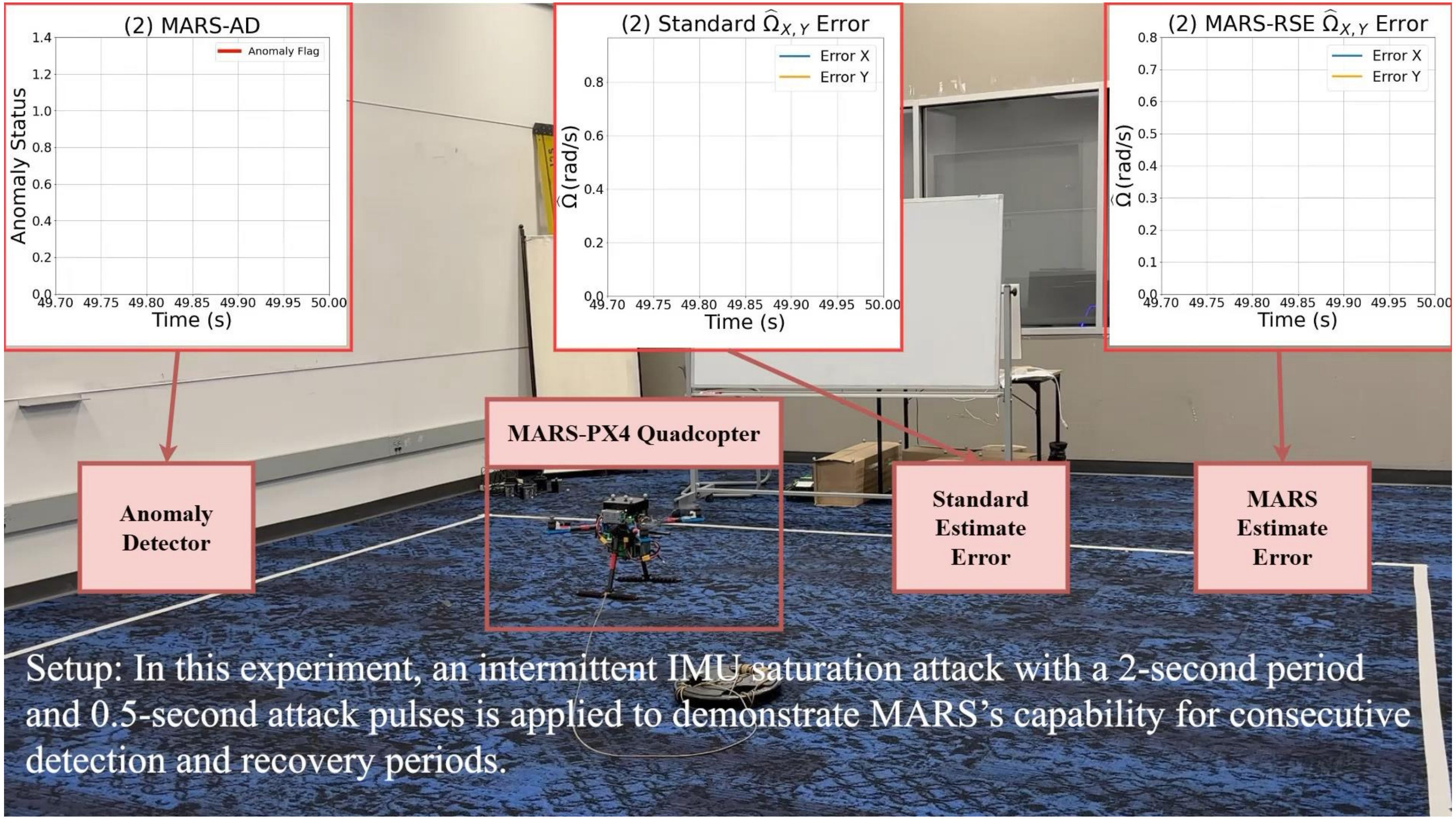
MARS-PX4 Autopilot Architecture
ROS
MAVLink
PX4-uORB
Drone
ROS-MARS High-level Controllers
ROS Position Controller
MAVROS
Attitude setpoint
Standard estimate
Resilient estimate
*
MAVLink Streamer/ Receiver
PX4 Controllers
Attitude Controller
Attitude Rate Controller
Mixer
3
4
PX4 Standard State Estimator
PX4-MARS Anomaly Detector
PX4-MARS Resilient State Estimator
Or
6
7
5
8
Serial Connection
Tachometers
GPS
Compass
IMU
ESC
Rotors
1
2
1 MARS high level commands
2 Position estimate/setpoint
3 Attitude rate setpoint
6 Standard estimate
4 Torque and thrust setpoint
7 MARS resilient estimate
5 Rotor speed command
8 Sensor measurements
* Note: only the controller related MAVLink messages are shown.

Real-time flight recovery

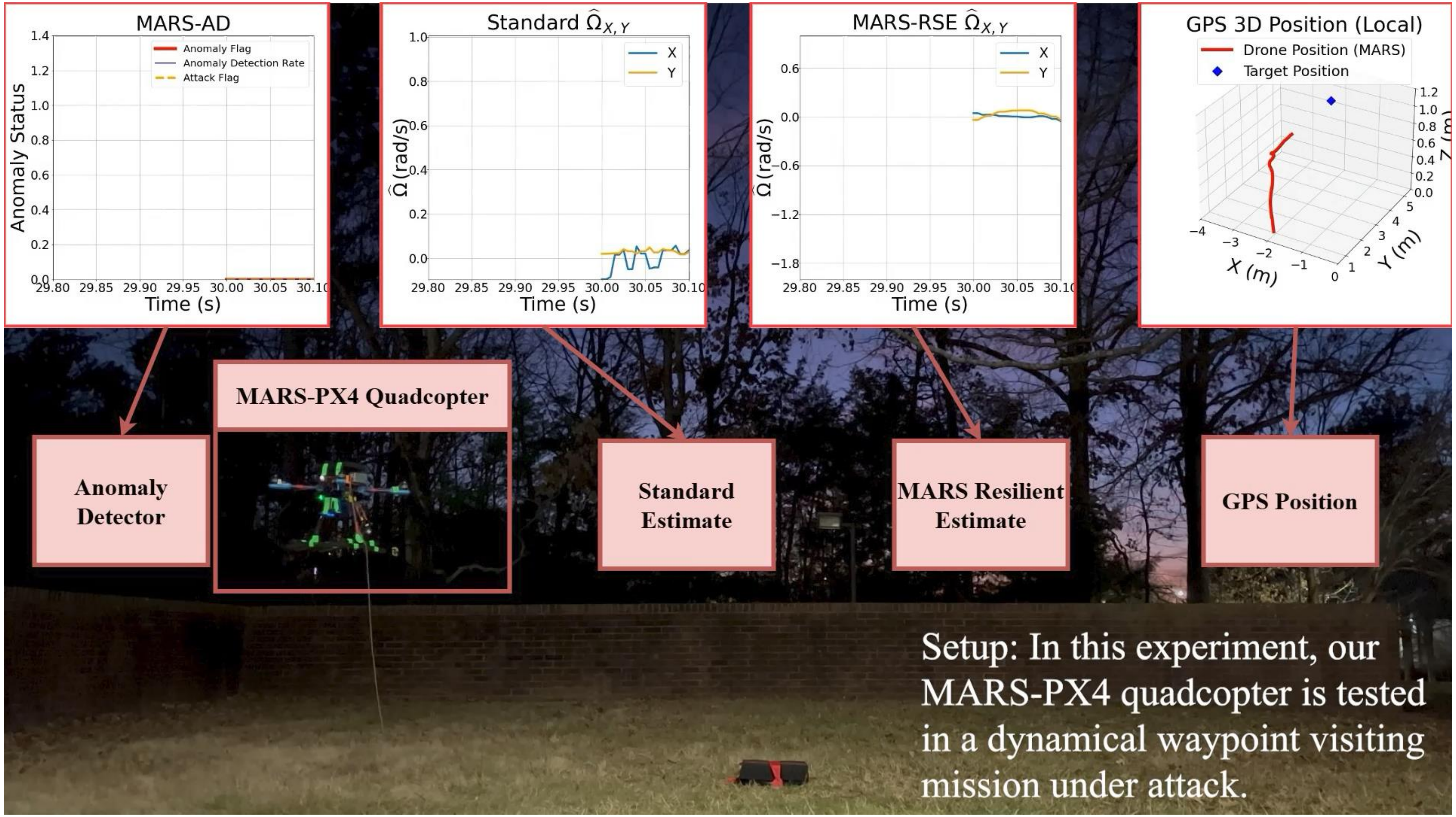
- An intermittent IMU saturation attack at 2s period with 0.5s attack pulses. It shows MARS's capability in consecutive detection and recovery periods.

- **Monitor-Left**: MARS anomaly detector status.
- **Monitor-Middle**: PX4 IMU-based standard angular velocity estimate error.
- **Monitor-Right**: MARS angular velocity estimate error. With suboptimal performance, It is robust to attacks on IMU.

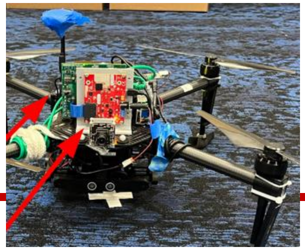
MARS real-time attack detection and recovery experiment



Real-World Experiments

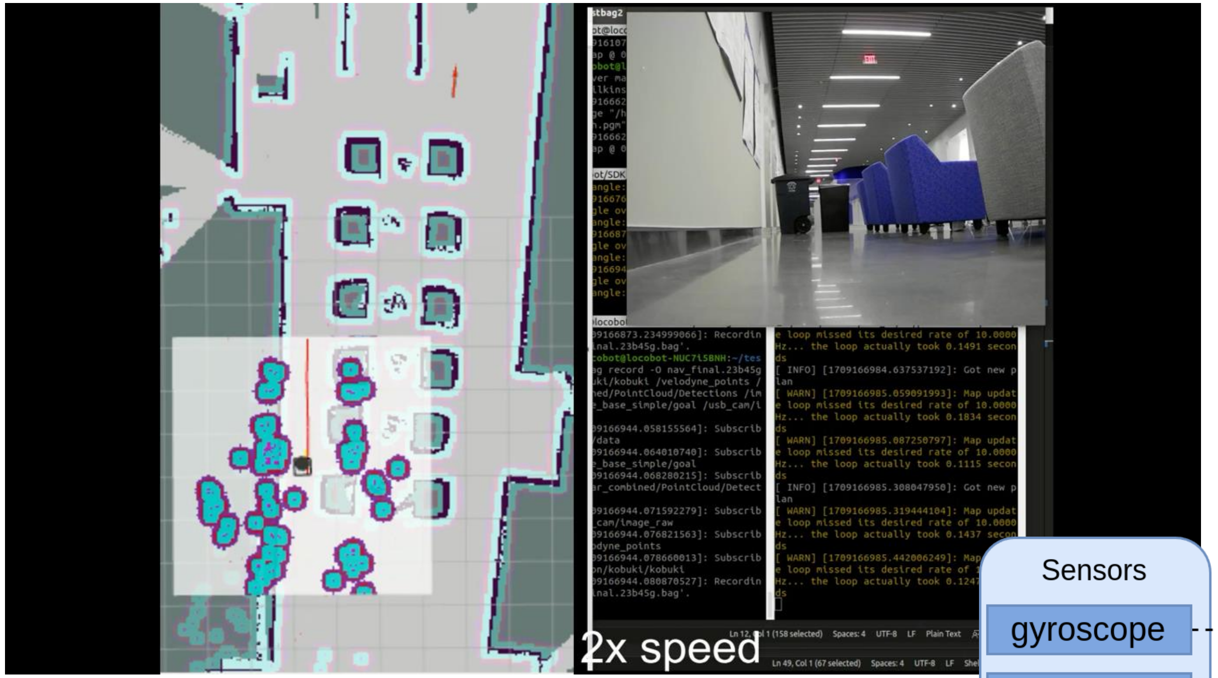


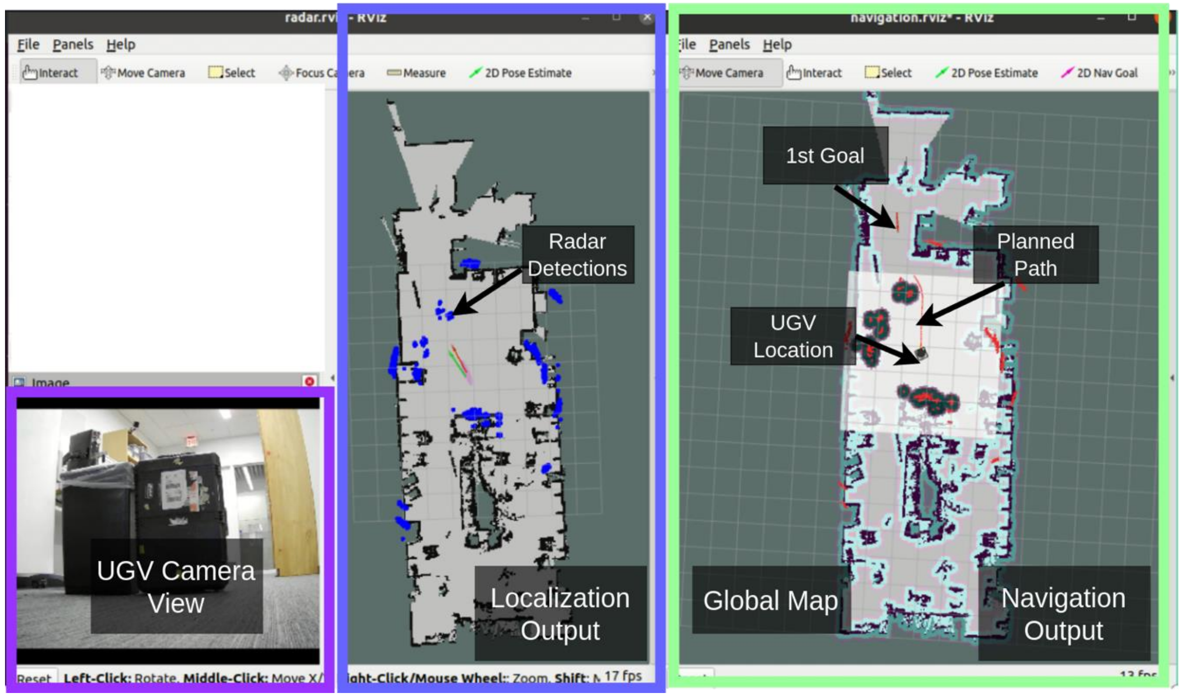
mmWave Radars for Resilient Autonomy (ICRA’24, IROS’25*)

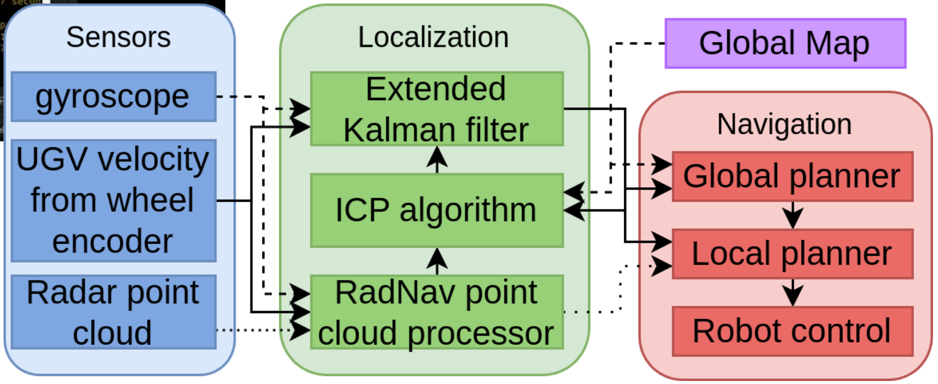


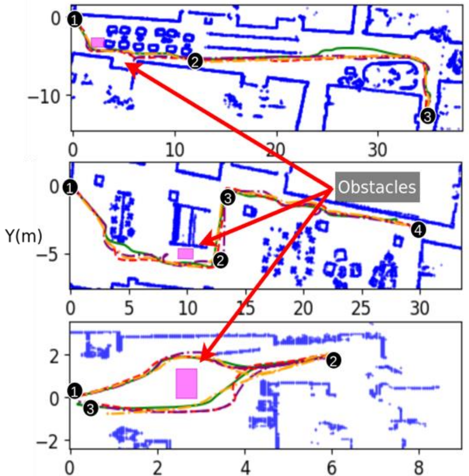
Duke PRATT SCHOOL of ENGINEERING

Goal: ***Low-cost*** (~$100), ***low-weight*** solution for resilient real-world autonomy on ***computationally constrained*** systems





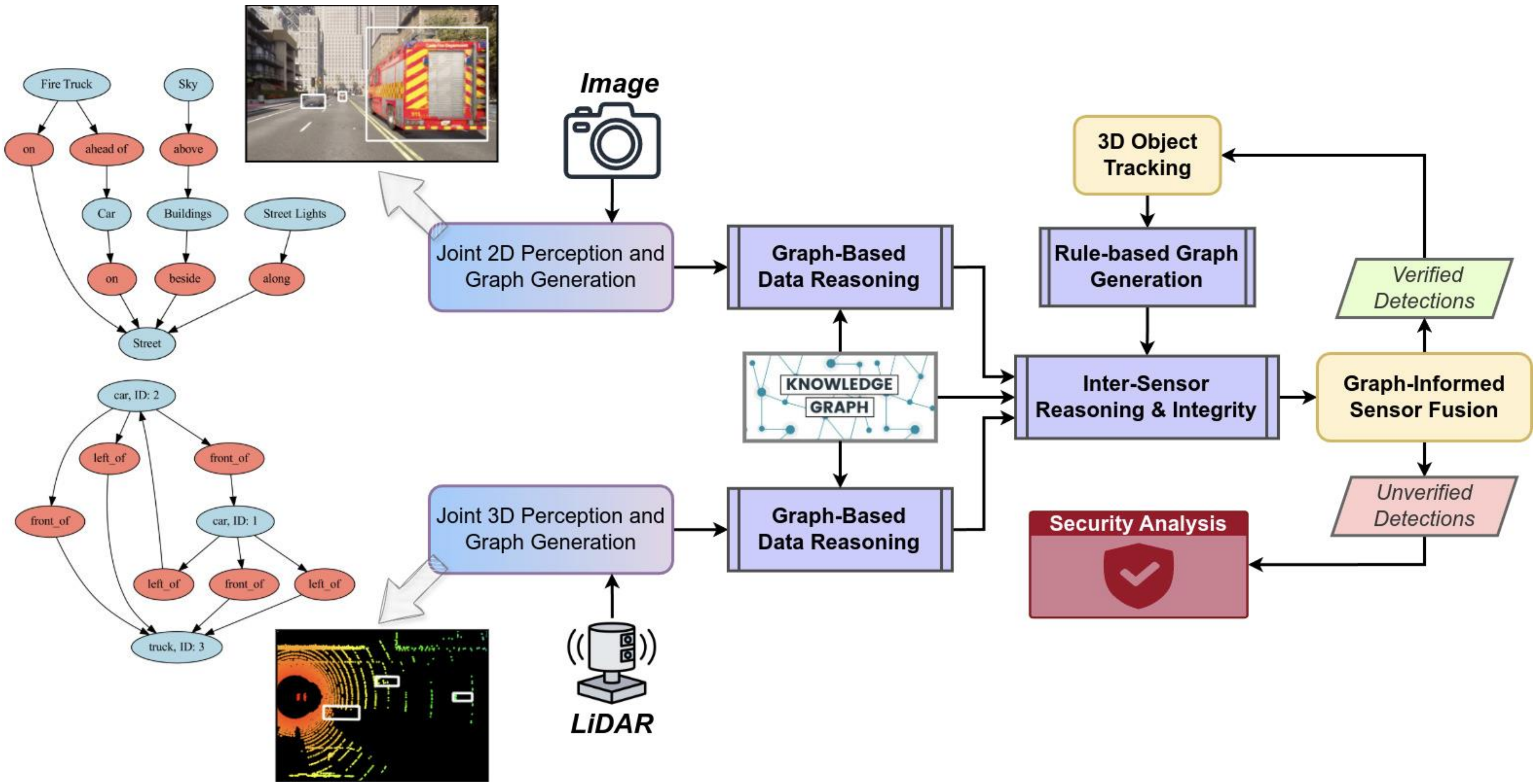




Assured Autonomy with Neuro-Symbolic Perception (NeuS'25*)

Duke
PRATT SCHOOL of ENGINEERING

Enforcing spatial and temporal consistency with neurosymbolic architectures

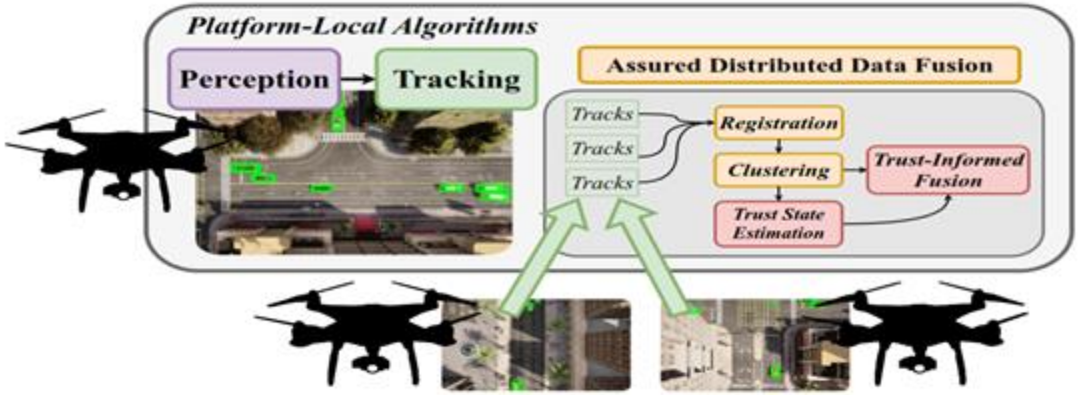


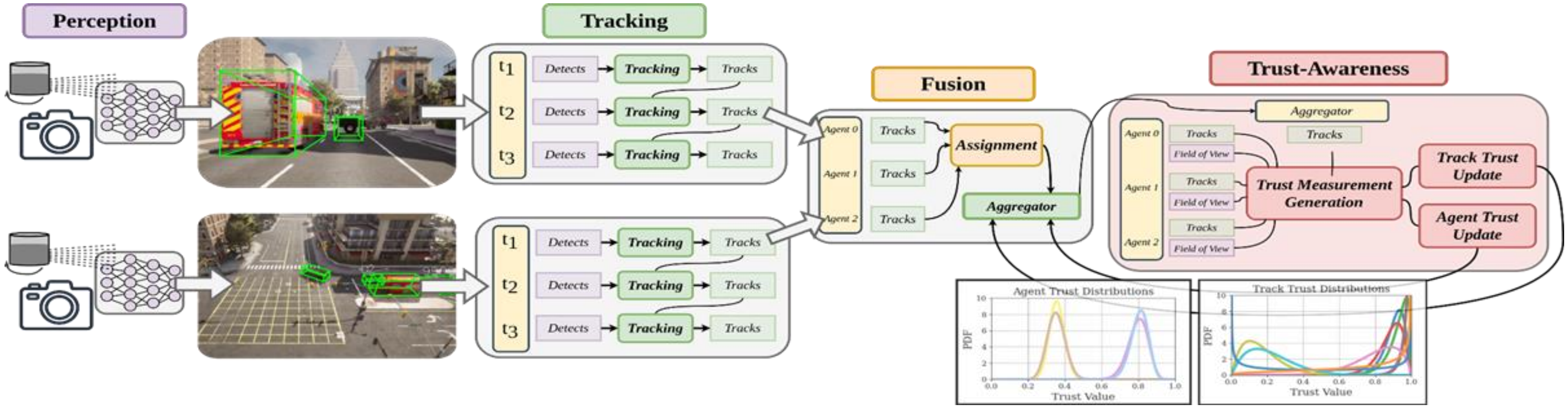
Trust-Informed Data Fusion (CDC24, Usenix Sec'25*, ICCPS25)

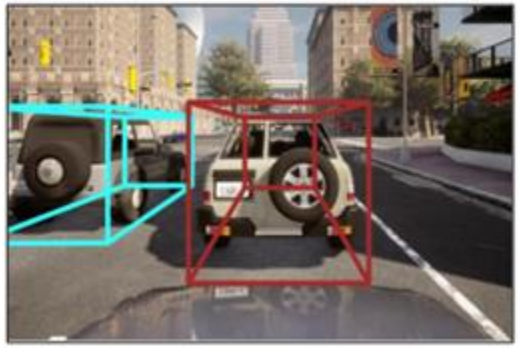
Duke
PRATT SCHOOL of ENGINEERING

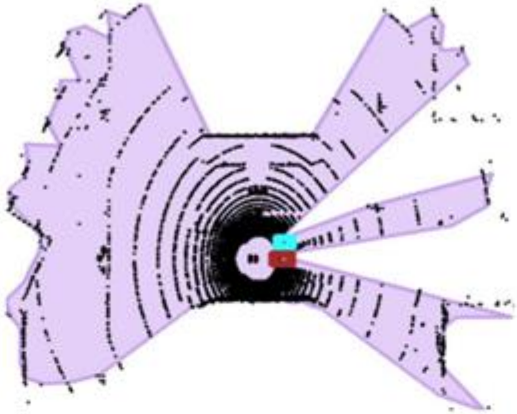
Trust-Informed Fusion

- *Agent trust* used to weight sensor fusion updates
  - (centralized) weighted Kalman updates
  - (distributed) weighted covariance intersection
- *Track trust* used to filter/single-out peculiar tracks
  - Low trust → further investigation
  - Track trust can inform motion planning







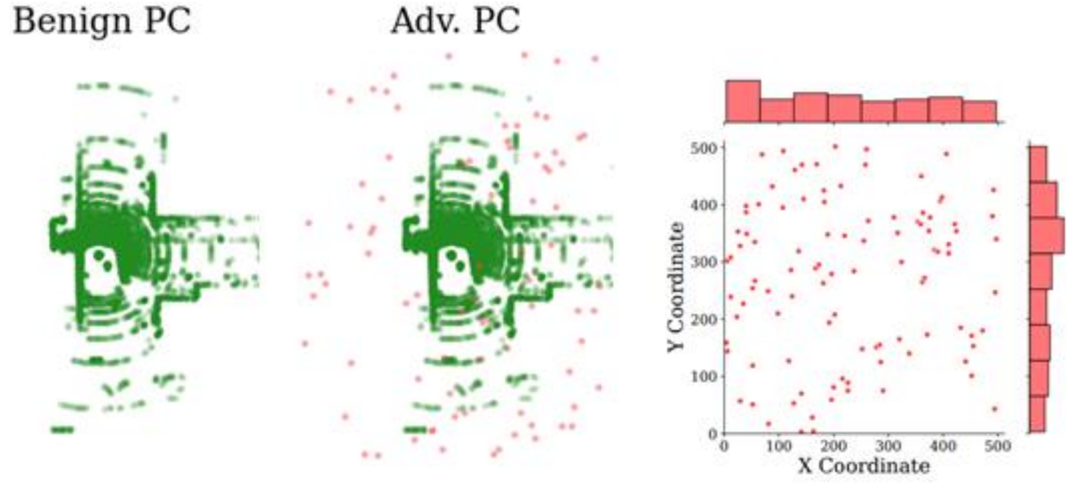


Analysis of Field-of-View Components

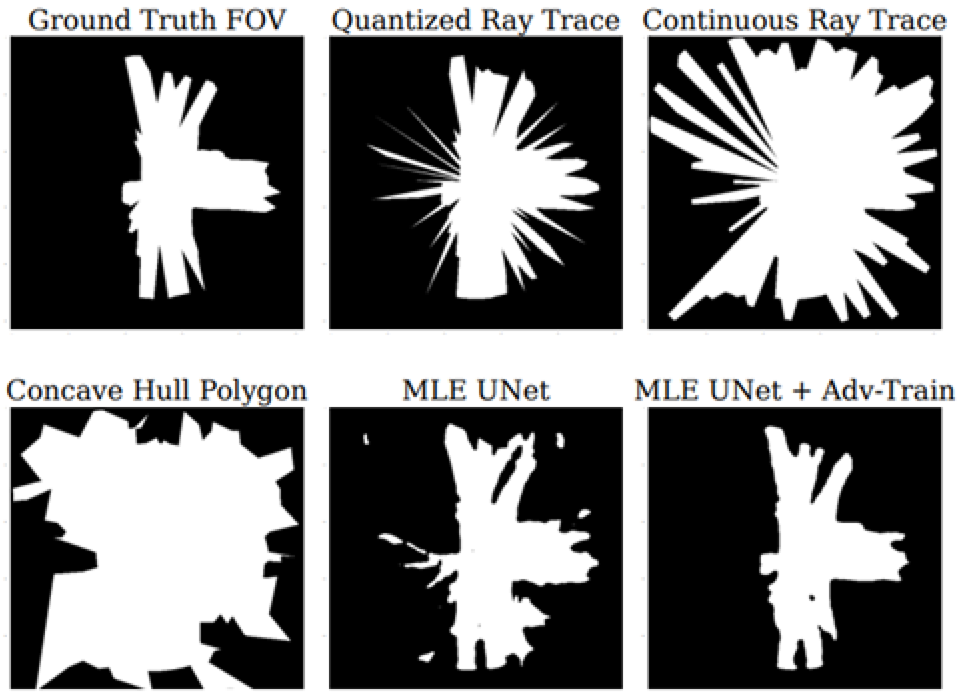
Probabilistic Segmentation for Robust Field of View Estimation

Uniform spoofing compromising traditional models

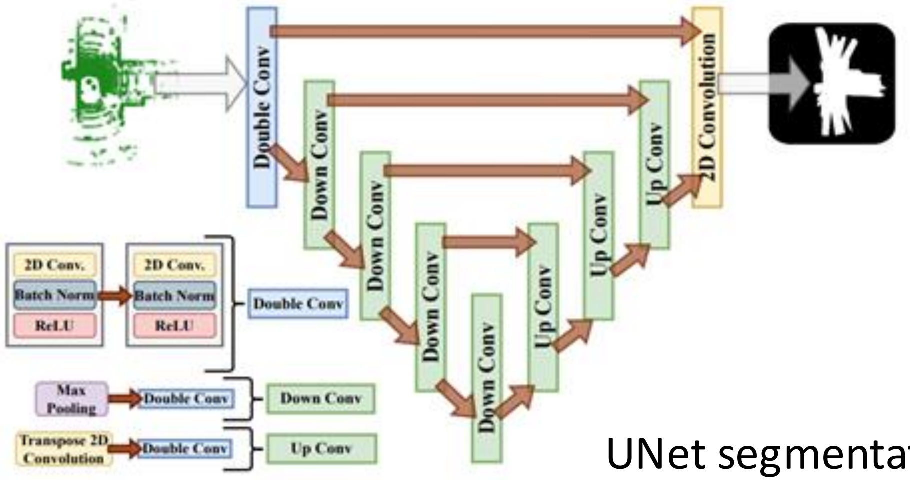
Duke
PRATT SCHOOL of ENGINEERING



(a) Point cloud in BEV frame. (b) Spoof point spread.

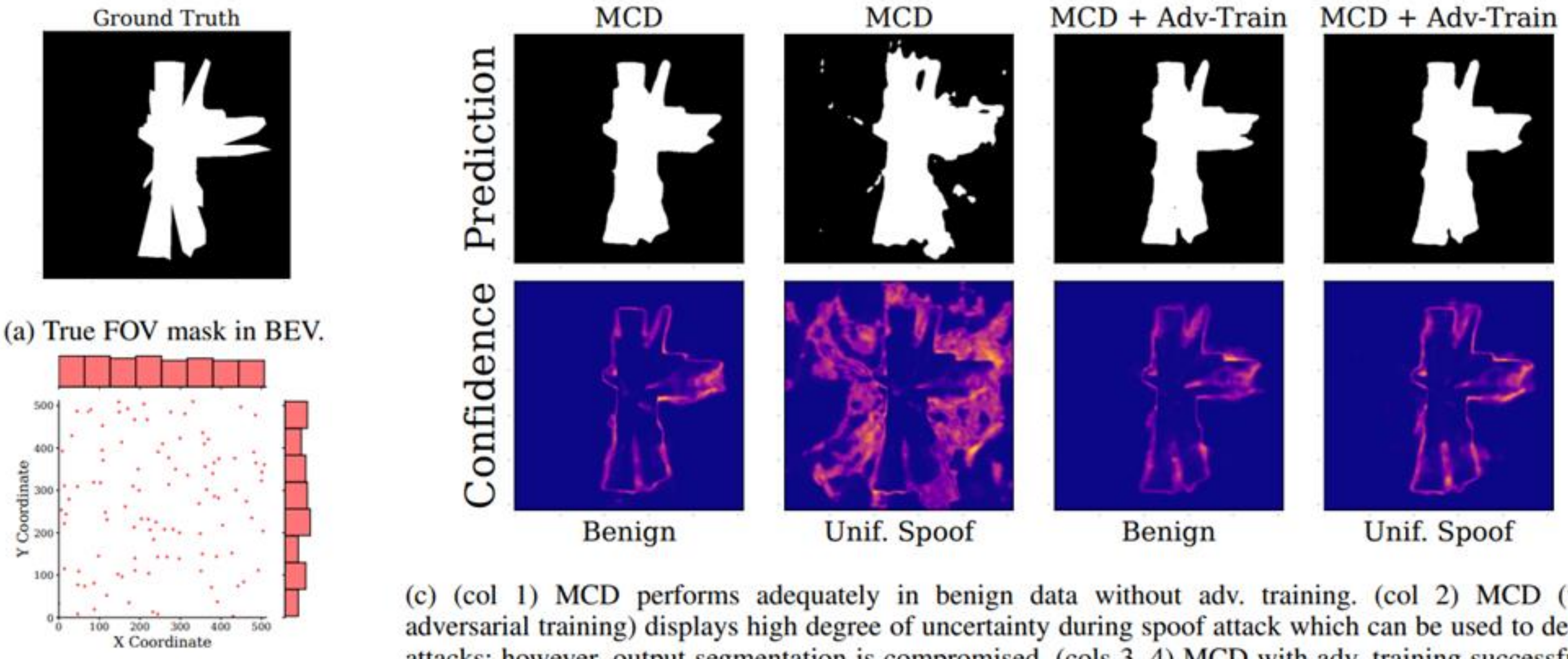


(c) Traditional FOV algorithms vulnerable to uniform spoof, adv. training improves resiliency of UNet segmentation models.



UNet segmentation model for FOV estimation

Defending Against Attacks



(a) True FOV mask in BEV.

(b) Distribution of small # "uniform spoof" points.

(c) (col 1) MCD performs adequately in benign data without adv. training. (col 2) MCD (w/o adversarial training) displays high degree of uncertainty during spoof attack which can be used to detect attacks; however, output segmentation is compromised. (cols 3, 4) MCD with adv. training successfully determines FOV from both benign and adversarial inputs. (row 2, confidence) Brighter colors (red) represent less confidence/more uncertainty.

Fig. 6: A small number of spoofed points can compromise MCD UNet without adv. training. However, confidence map obtained from MC dropout is useful in detecting attacks due to large uncertainty. MCD with adv. training defends attack.

So, what did we learn?

1. Security/resilience as first-class citizens!
2. Real-world is messy – this is both good and bad news!
3. Platform-aware constraints/capabilities must be taken into account (long platform lifetime)
4. Trust but verify!

Thank you

Duke UNIVERSITY | PRATT SCHOOL *of* ENGINEERING