Securing Autonomy
Resiliency of Perception-Based Control

Miroslav Pajic

CPSL@Duke

Department of Electrical and Computer Engineering

Department of Computer Science

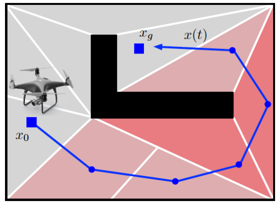
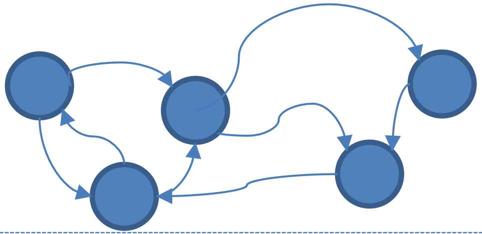
Duke University

Duke PRATT SCHOOL of ENGINEERING

DUKE ROBOTICS

Security-Aware Autonomy

Duke
PRATT SCHOOL of ENGINEERING

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

Adding Resiliency

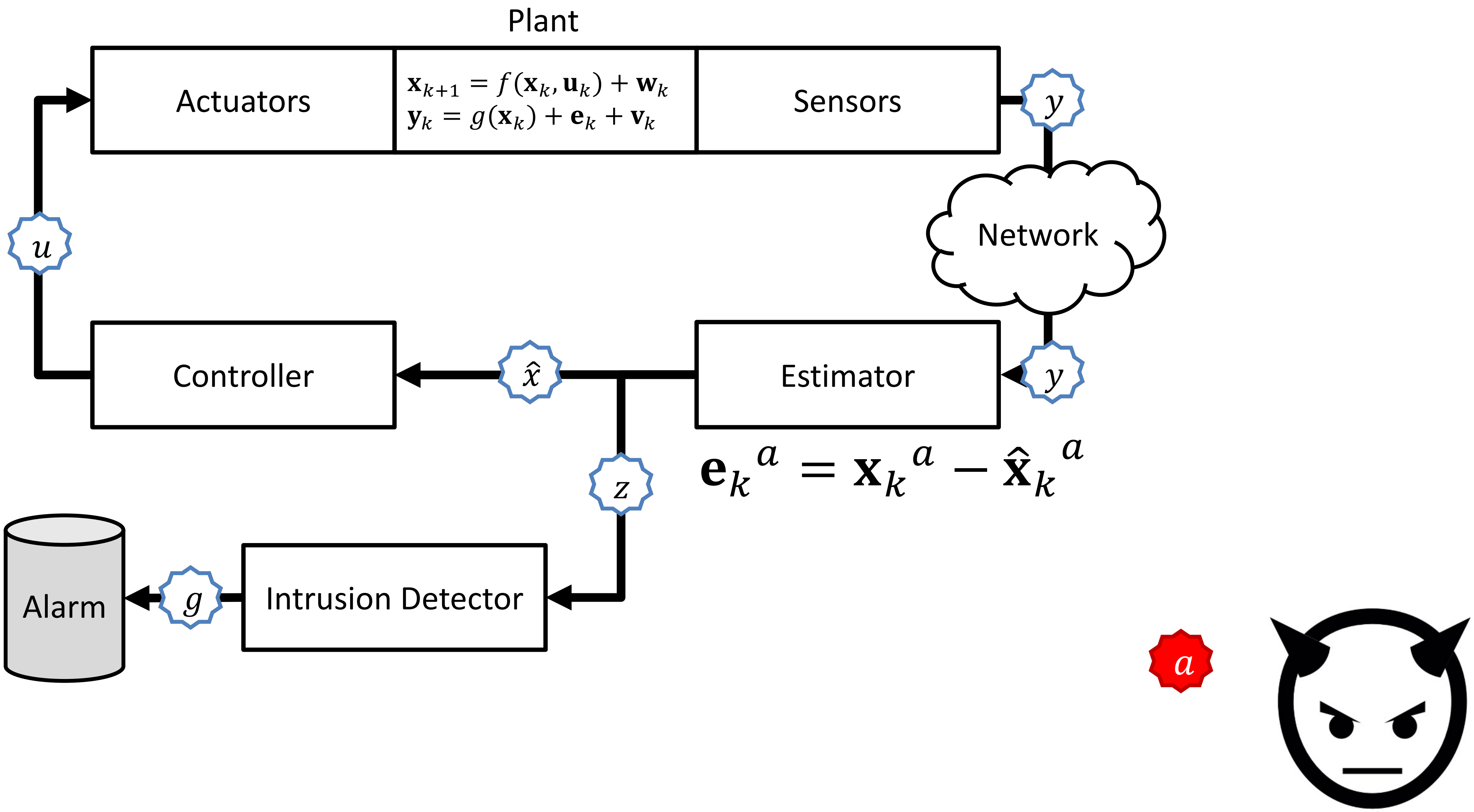
[USENIX Sec'22, TAC22*, CDC21, ICRA21a, ICRA21b, ICRA20, ICRA19, CAV'19a, THMS19]

[Aut22*, TII21, TASE21, CDC19a, CDC19b, IoTDI19]

[L4DC, ICCPS22a, AUT22, AUT21a, TCPS20, ACC20, AUT18, TECS17, RTSS17, TCNS17a, TCNS17b, CSM17, CDC17, CDC18,...]

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

Low-Level Control in the Presence of Attacks



Can Attacker Reach Any State?

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k) + \mathbf{w}_k$$
$$\mathbf{y}_k = C\mathbf{x}_k + \mathbf{a}_k + \mathbf{v}_k$$

$$supp(\mathbf{a}_k) = \mathcal{K}$$
$$\mathbf{a}_{k,i} = 0, \forall i \in \mathcal{K}^C$$

Theorem 1 [1,2,3]:
A system presented above is perfectly attackable if and only if it is **unstable**, and at least one eigenvector **v** corresponding to an unstable mode satisfies $supp(\mathbf{Cv}) \subseteq \mathcal{K}$ and **v** is a reachable state of the dynamic system.

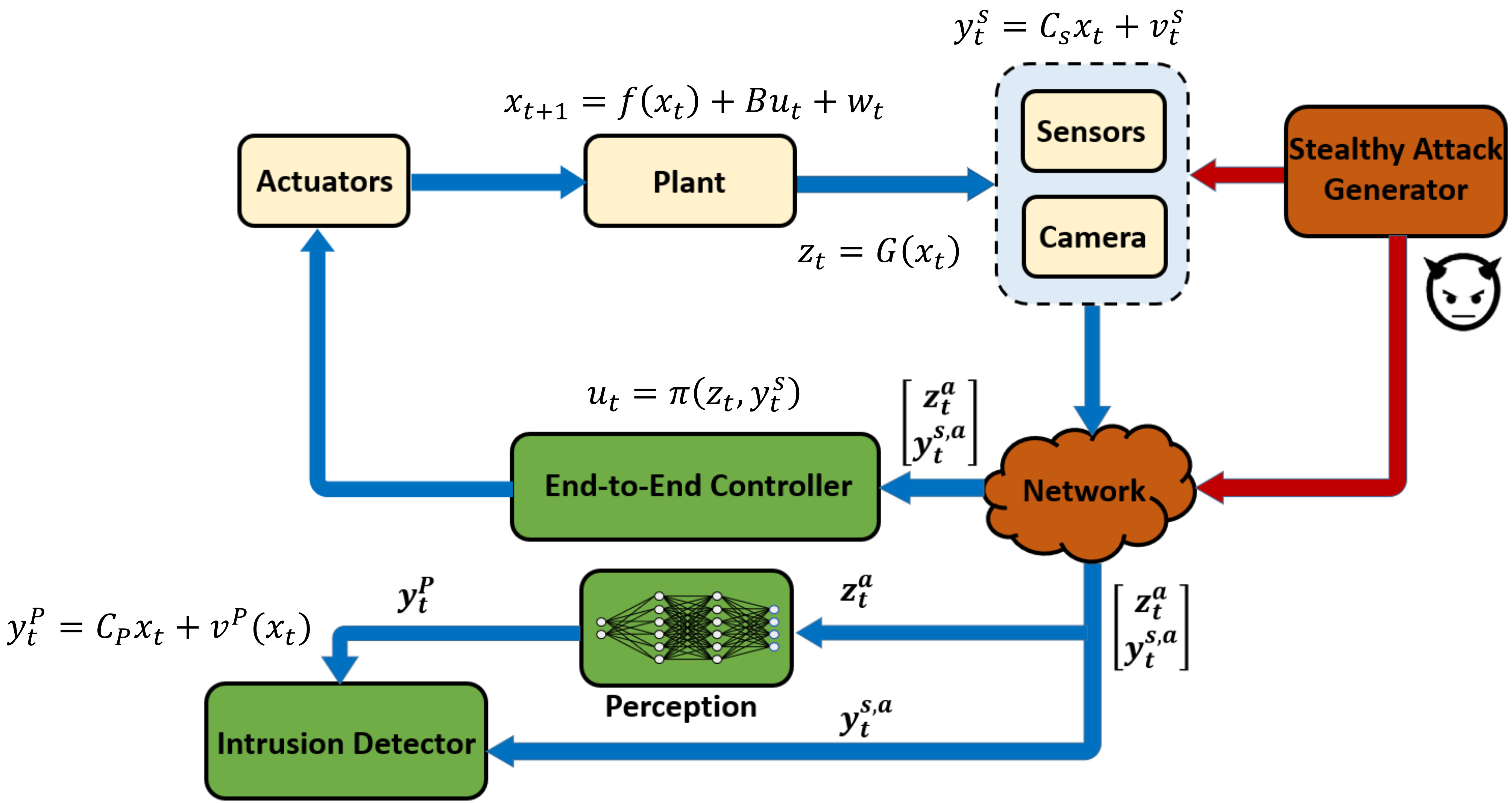
Physics-based detectors cannot always protect us from an intelligent attacker

[1] I. Jovanov and M. Pajic, "*Relaxing Integrity Requirements for Attack-Resilient Cyber-Physical Systems*", IEEE Trans. on Automatic Control, 2019
[2] A. Khazraei and M. Pajic, "*Perfect Attackability of Linear Dynamical Systems with Bounded Noise,*" ACC 2020.
[3] A. Khazraei and M. Pajic, "*Attack-Resilient State Estimation with Intermittent Data Authentication,*" Automatica, 2022.

What happens when we include perception?

A. Khazraei, H. Pfister, and M. Pajic, “Resiliency of Perception-Based Controllers Against Attacks”, *Learning for Dynamics and Control* (**L4DC**), 2022, **accepted, spotlight paper**.

System Model/Architecture



Intrusion Detector

- H_0: Normal condition (the ID receives $\mathrm{Y} = y_0 : y_t$ with distribution $\boldsymbol{P}$)

$$y_t = \begin{bmatrix} y_t^P \\ y_t^s \end{bmatrix}$$

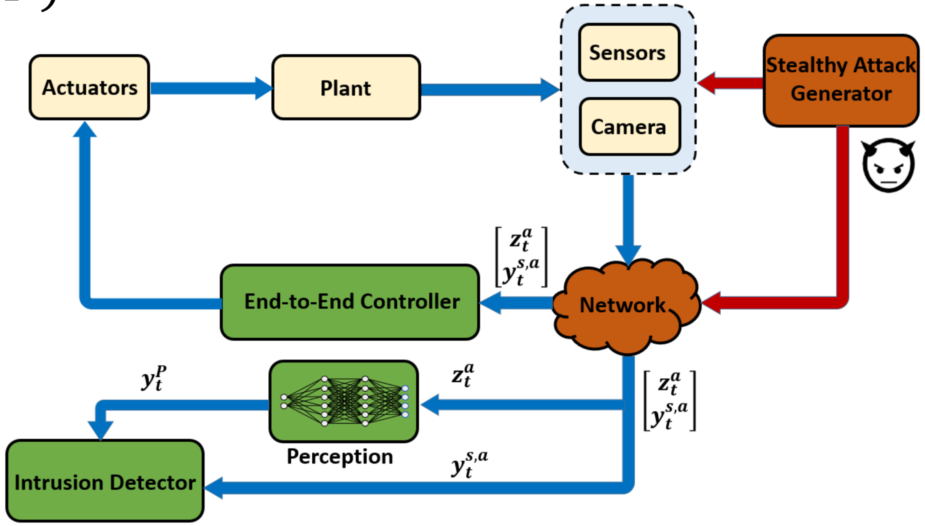
- H_1: Abnormal behavior (the ID receives $Y^a = y_0^a : y_t^a$ with distribution $\boldsymbol{Q}$)

$$y_t^a = \begin{bmatrix} y_t^{P,a} \\ y_t^{s,a} \end{bmatrix}$$

Intrusion Detector: $\mathcal{D}(\bar{Y}) \rightarrow \{0,1\}$

$$p_t^e = \mathbb{P}(\mathcal{D}(\bar{Y}) = 0 | \bar{Y} \sim \boldsymbol{Q}) + \mathbb{P}(\mathcal{D}(\bar{Y}) = 1 | \bar{Y} \sim \boldsymbol{P})$$

Random Guess: $p_t^e = \mathbb{P}(\mathcal{D}(\bar{Y}) = 0) + \mathbb{P}(\mathcal{D}(\bar{Y}) = 1) = 1$



Assumptions

Assumption 1: There exists a safe set $\mathcal{S}$ around the operating point such that for all $x \in \mathcal{S}$, it holds that $\|P(z) - C_P x\| \leq \gamma_e$, where $z = G(x)$– i.e., for all $x \in \mathcal{S}$, $\|v^P(x)\| < \gamma_e$. Without loss of generality, in this work we consider the origin as the operating point – i.e., $x_o = 0$.

Assumption 2: We assume that for the closed-loop system (4) is exponentially stable on a set $\mathcal{D} = B_d$.

Using the converse Lyapunov theorem, there exists a Lyapunov function that satisfies the following inequalities hold with constants c_1, c_2, c_3 and c_4 on a set $\mathcal{D} = B_d$

$$c_1\|x_t\|^2 \leq V(x_t) \leq c_2\|x_t\|^2 \qquad V(x_{t+1}) - V(x_t) \leq -c_3\|x_t\|^2 \qquad \left\|\frac{\partial V}{\partial x}\right\| \leq c_4\|x\|$$

Class of unstable functions

Definition: The class of functions $\mathcal{U}_\rho$ contains all functions f such that the dynamics $x_{t+1} = f(x_t) + d_t$, where d_t satisfies $\|d_t\| \leq \rho$, becomes arbitrarily large for some nonzero initial state x_0. Also, for a function f from $\mathcal{U}_\rho$ and initial condition x_0, we define $T_f(\alpha, x_0) = min\{t | \|x_t\| > \alpha\}$.

Proposition: Let $V\colon \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function satisfying $V(0) = 0$ and define

$$U_{r_1} = \left\{x \in B_{r_1} \middle| V(x) > 0\right\}.$$

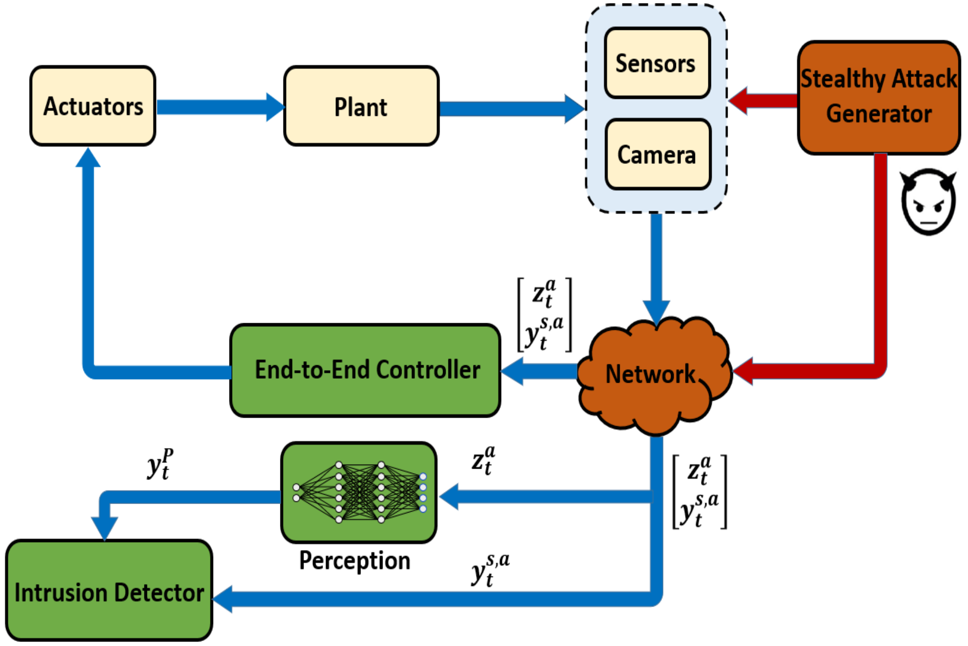
Assume that $\left\|\frac{\partial V(x)}{\partial x}\right\| \leq \beta(\|x\|)$ and for any $x \in U_{r_1}$ it holds that $V\big(f(x)\big) - V(x) \geq \alpha(\|x\|)$ where $\beta(\|x\|)$ and $\alpha(\|x\|)$ are in class $\mathcal{K}$ functions. Further, assume that r_1 can be chosen arbitrarily large.

If $\lim\limits_{\|x\| \to \infty} \frac{\alpha(\|x\|)}{\beta(\|x\|)} \to \infty$, then f $\in \mathcal{U}_\rho$ for any $\rho > 0$.

If $\lim\limits_{\|x\| \to \infty} \frac{\alpha(\|x\|)}{\beta(\|x\|)} = \gamma$ then f $\in \mathcal{U}_\rho$ for any $\rho < \gamma$.

Attack Model

- The attacker has full knowledge of the system, its dynamics and employed architecture
- The attacker has the required computation power to calculate suitable attack signals to inject a subset of sensors, while planning ahead as needed
- The attacker has the ability to compromise camera images by z_t^a
- The attacker has the ability to compromise the sensor measurements
- Attack objective: ***effective*** and ***stealthy***!



Stealthiness Definition

Definition: An attack sequence is **strictly stealthy** if there exists no detector such that the sum of conditional error probabilities p_t^e satisfies $\boldsymbol{p_t^e < 1}$, for any $t \geq 0$. An attack is **ϵ-stealthy** if for a given $\epsilon > 0$, there exists no detector such that $\boldsymbol{p_t^e < 1 - \epsilon}$ for any $t \geq 0$.

$$p_t^e = \mathbb{P}(\mathcal{D}(\bar{Y}) = 0 | \bar{Y} \sim \boldsymbol{Q}) + \mathbb{P}(\mathcal{D}(\bar{Y}) = 1 | \bar{Y} \sim \boldsymbol{P})$$

Theorem: An attack sequence is **strictly stealthy** if and only if

$$\boldsymbol{KL}\big(\boldsymbol{Q}(y_0^a : y_t^a) || \boldsymbol{P}(y_0 : y_t)\big) = \boldsymbol{0} \text{ for any } t \geq 0,$$

(KL represents the Kullback-Leibler divergence operator).

An attack sequence is **ϵ-stealthy** if the corresponding observation sequence satisfies

$$\boldsymbol{KL}\big(\boldsymbol{Q}(y_0^a : y_t^a) || \boldsymbol{P}(y_0 : y_t)\big) \leq \log\left(\frac{1}{1-\epsilon^2}\right)$$

Attacker's Goal

Definition 2: Attack sequence, denoted as $\{z_0^a, y_0^{s,a}, z_1^a, y_1^{s,a}, \dots\}$ is an ***$(\boldsymbol{\epsilon}, \boldsymbol{\alpha})$-successful attack*** if there exists $t' \geq 0$ such that $\|x_{t'}\| \geq \alpha$ and the attack is ϵ-stealthy for all $t \geq 0$.

When such a sequence exists for a system, the system is called **$(\boldsymbol{\epsilon}, \boldsymbol{\alpha})$-attackable**. Finally, when the system is (ϵ, α)-attackable for arbitrarily large α the system is referred to as ***perfectly attackable***.

Definition 3: For an attack-free state trajectory $x_0 : x_t$, and for any $T \geq 0$ $b_v > 0$ and $b_x > 0$, $\delta(T, b_x, b_v)$ is the probability that the system state and physical sensor noise v^s remain in the ball with radius b_x and b_v, respectively, during $0 \leq t \leq T$– i.e.,

$$\delta(T, b_x, b_v) = \mathbb{P}\left(\sup_{0 \leq t \leq T} \|x_t\| < b_x, \sup_{0 \leq t \leq T} \|v_t\| < b_v\right)$$

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 e $= x_t^a - s_t$,

Theorem 2: 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 $(\boldsymbol{\epsilon}, \boldsymbol{\alpha})$**-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 $b = \frac{c_4}{c_3-(L_1+L_3\|B\|)c_4}\sqrt{\frac{c_2}{c_1}\frac{L_2 b_\zeta}{\theta}}$.

Attack Strategy II

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: $s_{t+1} = f(s_t)$

Theorem: Assume that the functions 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 $L_1 = L_f'(\alpha + d)$, $L_2 = L_f'(\alpha + b_x)$ and $L_3 = L_\Pi'(b_x + d + b_v)$. Moreover, assume that b_x has the maximum value such that the inequalities $L_1 + L_3\|B\| < \frac{c_3}{c_4}$ and $L_2 b_x < \frac{c_3-(L_1+L_3\|B\|)c_4}{c_4}\sqrt{\frac{c_1}{c_2}}\theta r$ for some $0 < \theta < 1$, 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 $f \in \mathcal{U}_0$ and $b = \frac{c_4}{c_3-(L_1+L_3\|B\|)c_4}\sqrt{\frac{c_2}{c_1}}\frac{L_2 b_x}{\theta}$.

Attack on LTI Systems

Corollary 1: Consider an LTI perception-based control system with $f(x_t) = Ax_t$.
If $L_3\|B\| < \frac{c_3}{c_4}$ with $L_3 = L'_\Pi(b_x + d + b_v)$ and the matrix A is ***unstable***, the system is $(\boldsymbol{\epsilon}, \boldsymbol{\alpha})$ **-attackable** with probability $\delta(T(\alpha + b_x, s_0), b_x, b_v)$ for arbitrarily large α and $\epsilon = \sqrt{1 - e^{-b_\epsilon}}$, where

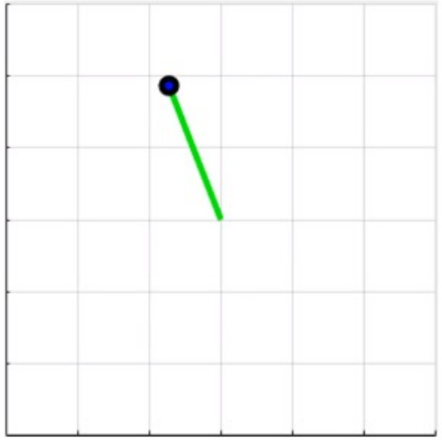
$$b_\epsilon = \left(\lambda_{max}(\Sigma_w^{-1}) + \lambda_{max}(C_s^T \Sigma_v^{-1} C_s + \Sigma_w^{-1}) \min\left\{T(\alpha + b_x, s_0), \sqrt{\frac{c_2}{c_1} \frac{e^{-\beta}}{1 - e^{-\beta}}}\right\}\right) \|s_0\|$$

and $e^{-\beta}$ is the largest eigenvalue of the closed-loop system.

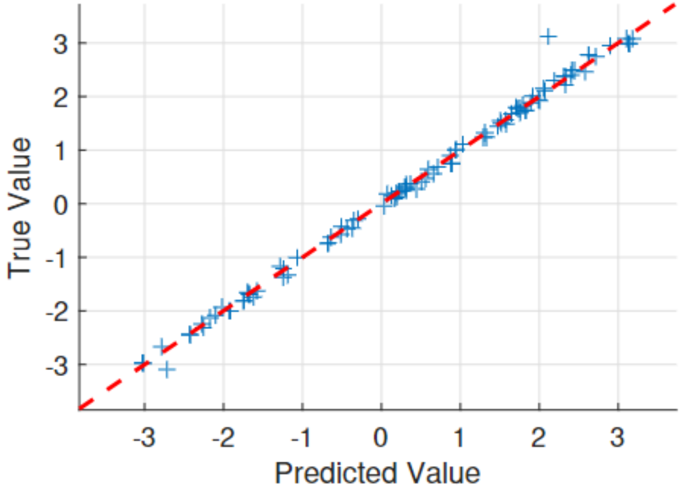
Corollary 2: Consider an LTI perception-based control system with $f(x_t) = Ax_t$ and a ***linear feedback controller***.
If the matrix A is ***unstable***, the system is (ϵ, α) -attackable with **probability 1** for arbitrarily large α and $\epsilon = \sqrt{1 - e^{-b_\epsilon}}$, where $b_\epsilon = \lambda_{max}(\Sigma_w^{-1}) + \lambda_{max}(C_s^T \Sigma_v^{-1} C_s + \Sigma_w^{-1}) \sqrt{\frac{c_2}{c_1} \frac{e^{-\beta}}{1 - e^{-\beta}}}$ and $e^{-\beta}$ is the largest eigenvalue of the closed-loop system.

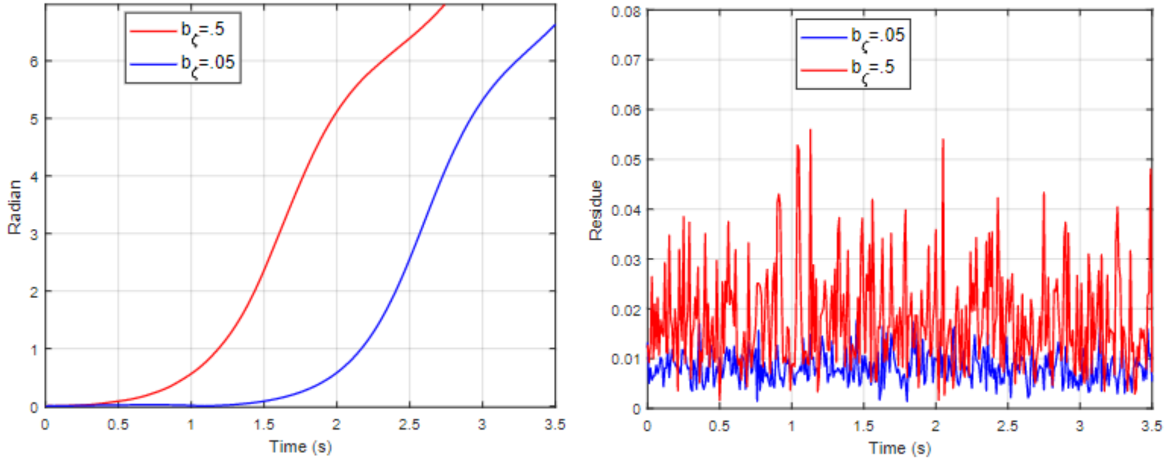
Case Study : Inverted Pendulum

Duke PRATT SCHOOL of ENGINEERING

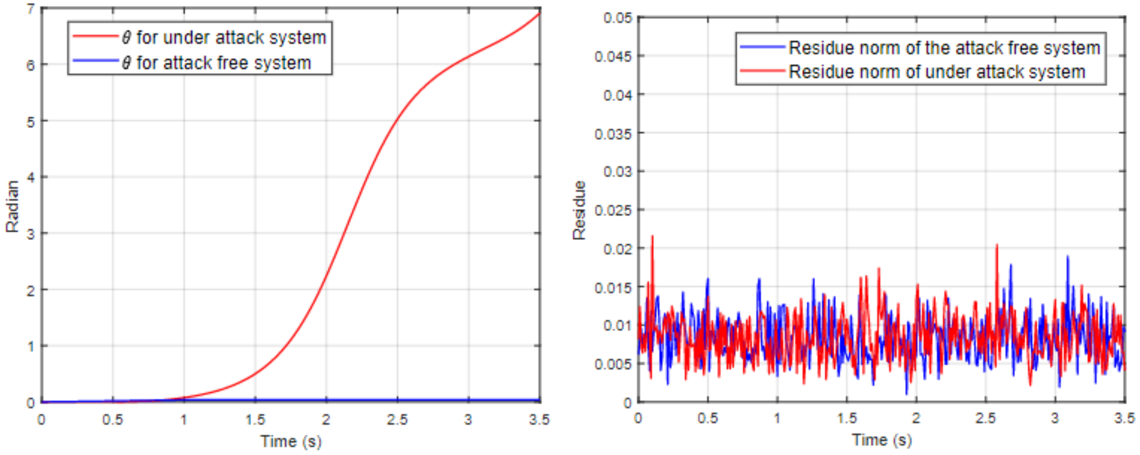


Perception map performance





Evolution of the angle's θ absolute value over time for different levels of b_ζ (left). The norm of the residue over time when the attack starts at time $t = 0$ (right)



Evolution of the angle's θ absolute value over time for attack strategy II. The norm of the residue over time when the attack starts at time $t = 0$ (right)

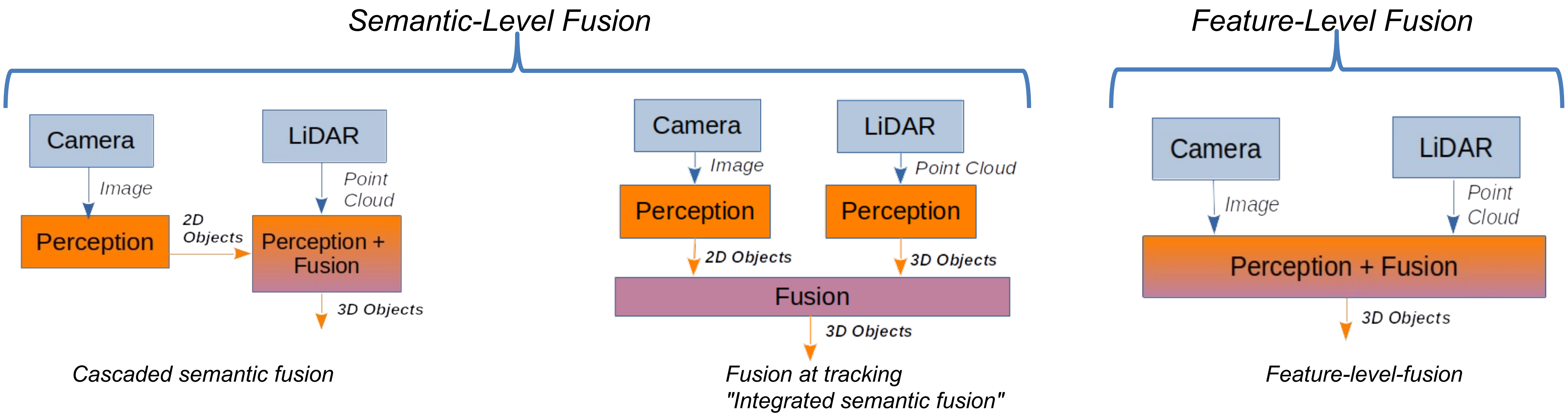
Camera-LiDAR Fusion

Multiple Architectures for Sensor Fusion

- *Semantic fusion* popular across industry due to:
  - Reduce of "curse of dimensionality" of input space
  - Greater flexibility in industry for "plug-and-play"/swap-ability of components
- *Feature-level-fusion* high-performing due to fusion of low-level, machine-learned features
- Fusion touted to improve resiliency and performance compared to single-sensor perception alone

Most common sensors:

- LiDAR data is sparse in R4
  - X-Y-Z-intensity
  - Full 3D resolution
- Camera data is dense in R3
  - R-G-B channels
  - 2D (angles-only) resolution



Beyond Naïve Attack: Novel Frustum Attack Is Feasible

Duke
PRATT SCHOOL *of* ENGINEERING

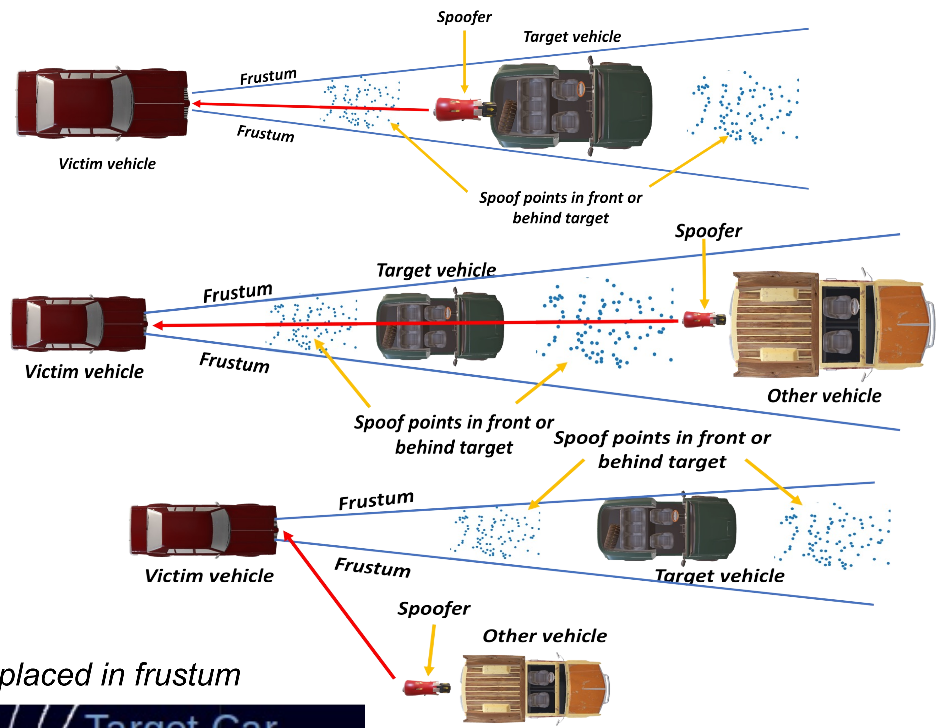
Compromise Fusion (and LiDAR-only)

- Fusion robust against naïve attack because naïve attack is not consistent between sensor modalities
- Ensure consistency by spoofing *within the frustum (i.e. in-view, as seen by camera)* of existing vehicles
- This does not require any knowledge of the camera data

Feasibility

- We validated attack feasibility with limited additional knowledge required over original, naïve black-box spoofing
- Only additional requirement is attack orientation

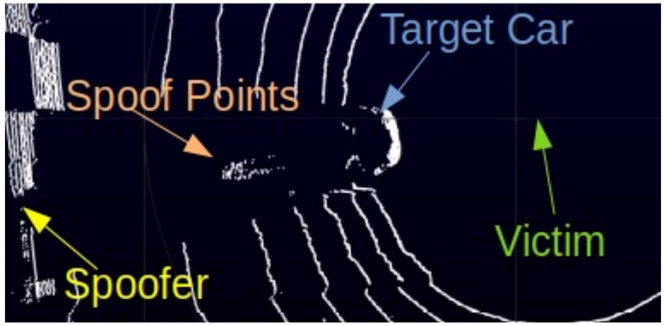
Three candidate realizations of the frustum attack. Additional configurations shown later



Target car in front of victim

Spoofer set behind target car

Stable spoof points placed in frustum



Demonstrated controlling (i.e. moving to attacker's specified location) spoof points stably over time with moving vehicles

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 Symposium* (**USENIX SECURITY**), 2022, **accepted.**

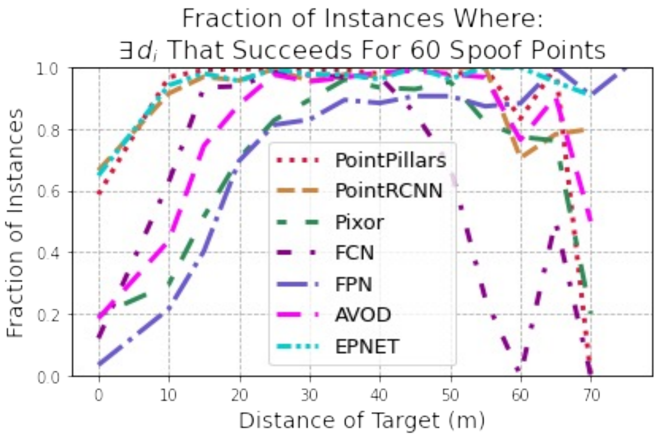
Frustum Attack is Widely Successful

Compromise Fusion (and LiDAR-only)

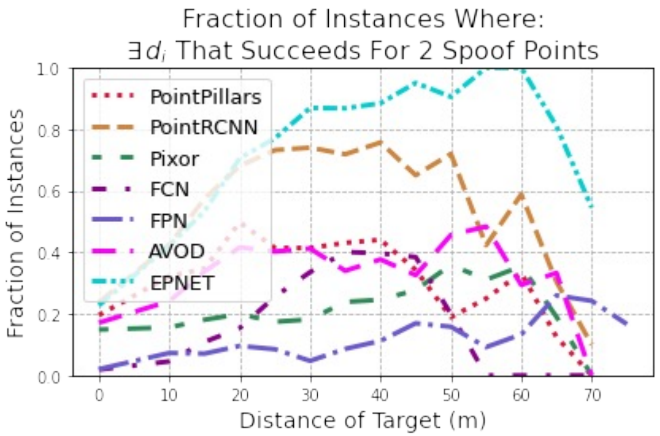
- Frustum attack demonstrated to compromise BOTH LiDAR-only AND camera-LiDAR fusion
- Frustum attack shown indefensible by state-of-the-art defenses (CARLO, SVF, ShadowCatcher, LIFE)

Extensive Evaluations

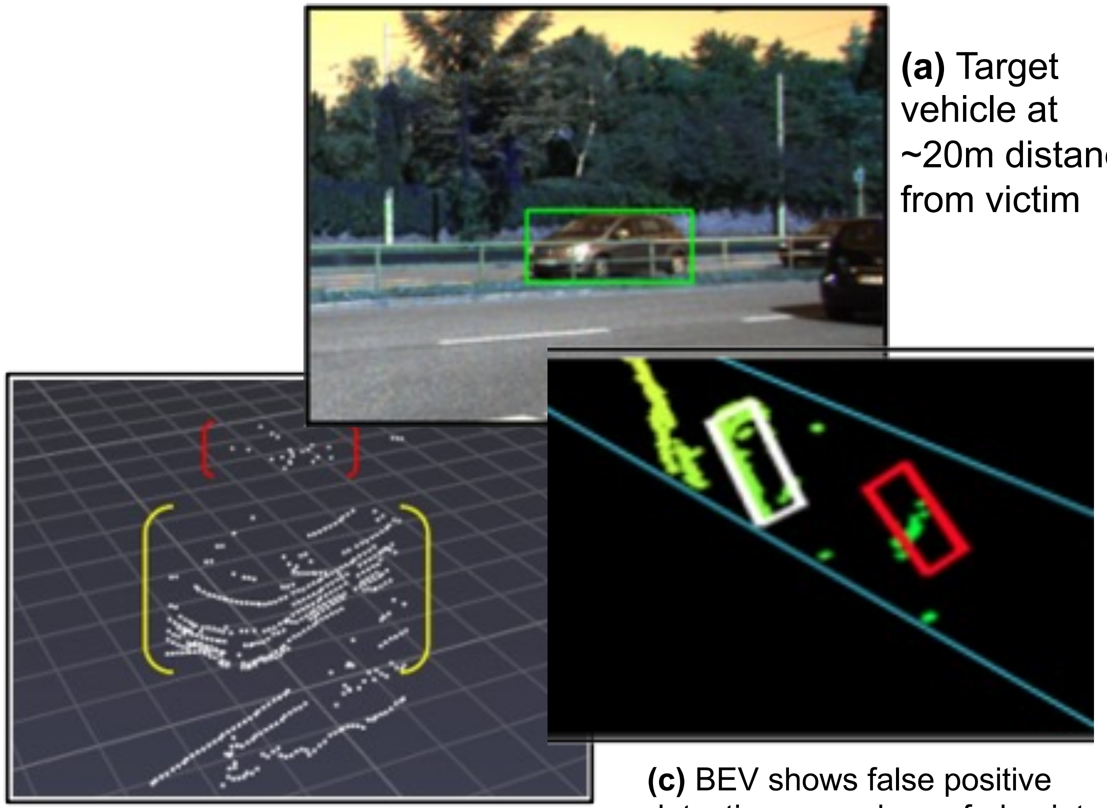
- We perform the most extensive evaluation of attacks on perception to-date with 8 algorithms and 4 defenses (7 and 3 for large-scale evaluation)
- > 75 million attack traces evaluated --> number of spoof points, distance of spoof point placement, each object, each frame of data



Frustum attack widely successful with 60 spoof points



Frustum attack successful even with just 2 spoof points!



(a) Target vehicle at ~20m distance from victim

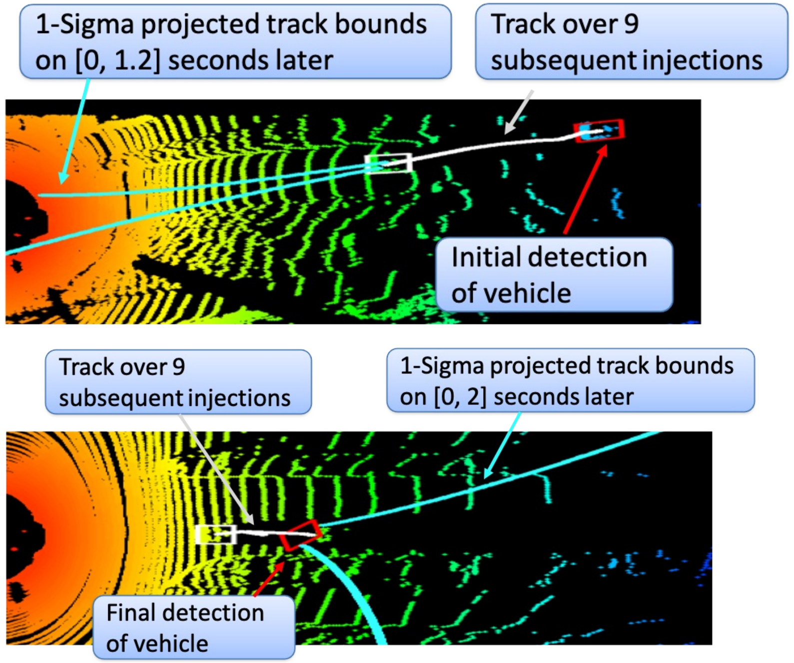
(b) Target victim (yellow, 238 pts) has many more points than the spoof points (red 20pts)

(c) BEV shows false positive detection around spoofed points

Longitudinal Frustum Attacks Are Dangerous

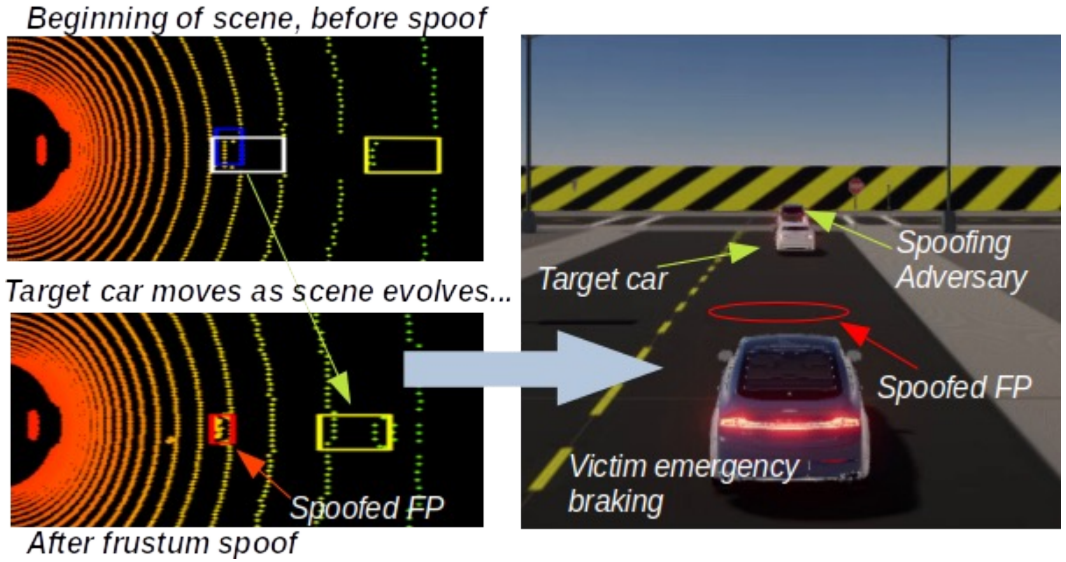
Evaluation of Multi-Frame Tracking

- Use captured KITTI dataset to evaluate impact of frustum attack over multiple frames
- Demonstrated stably executing frustum attack in longitudinally-consistent way to obtain adversarial tracks (white + cyan) that can:
  - 1) project to collide with victim
  - 2) project to accelerate flow of traffic



End-to-End, Industry-Grade AVs

- Preliminary evaluation of the vulnerability of Baidu Apollo perception + control stack to the frustum attack – *emergency braking engaged*
  - Baidu fuses LiDAR and camera detections at the tracking-level
  - Use multi-stage approach since Baidu+SVL combination is still under development
- Physics-based simulations of AV driving with the SVL Simulator



Stealthy Spoofing Frustum-Attacks:
Attacking Baidu's Apollo

Duke PRATT SCHOOL of ENGINEERING

Conclusion

- Perception combined with controls opens new attack surface
- Moving from single instance analysis to longitudinal (i.e., time-series) analysis

Thank you

Duke
UNIVERSITY
PRATT SCHOOL *of*
ENGINEERING