CERIAS
The Center for Education and Research in Information Assurance and Security

Anomaly Detection Against GPS Spoofing Attacks on Connected and Autonomous Vehicles Using Learning from Demonstration

Zhen Yang[1], Jun Ying[2], Junjie Shen[3], Yiheng Feng[2], Qi Alfred Chen[3], Z. Morley Mao[4], and Henry X. Liu[1]

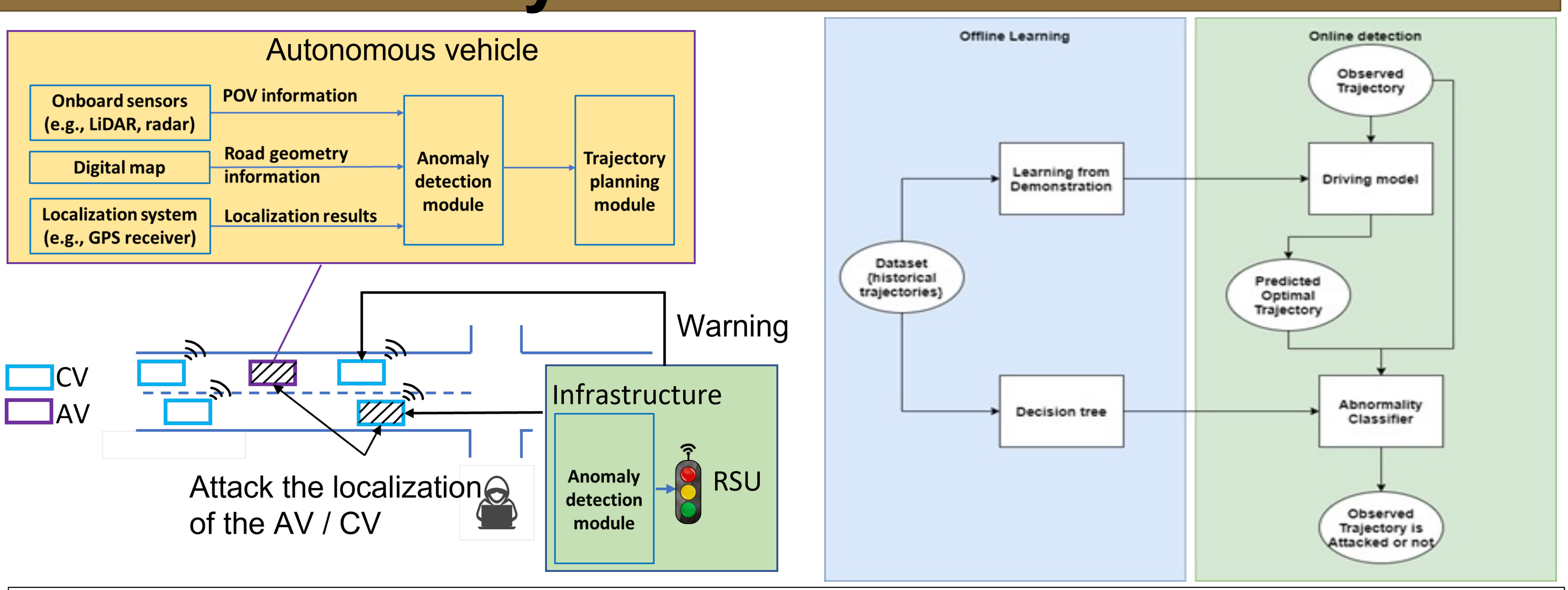
1. Dept. of Civil & Environmental Engineering, University of Michigan; 2. Lyles School of Civil Engineering, Purdue University

3. Department of Computer Science, University of California, Irvine 4. Department of Electrical Engineering and Computer Science, University of Michigan

Introduction

- In both autonomous Vehicles (AVs) and connected vehicles (CVs), the localization module, which provides accurate local and global positions, plays a critical role in vehicle navigation and ITS applications.
- GPS spoofing attacks pose great challenges to safety applications of connected vehicles (CVs) and localization of autonomous vehicles (AVs).
- This study proposes a generic detection framework to detect anomalies in the localization module of AV/CV using learning from demonstration.

Anomaly Detection Framework



- The anomaly detection framework consists of two steps: offline learning and online detection
- Learning from demonstration is applied to learn the normal driving policy via maximum entropy inverse reinforcement learning using historical trajectories.
- An anomaly classifier (i.e., a decision tree) is trained with both historical trajectories and known attack trajectories.
- Observed trajectories are compared with predicted optimal trajectories from the learned driving policy to detect anomaly.

Learn from Demonstration

$minimize_s\ \boldsymbol{\theta}^T \boldsymbol{f}(\boldsymbol{s},\boldsymbol{u})$

$s.t. Vehicle\ dynamic\ constraints$

- Feature $\boldsymbol{f}$ includes different driving behaviors.
- Vehicle dynamic constraints represent the kinematics of vehicle motion, assuming the vehicle follows the bicycle model.
- The weight vector $\boldsymbol{\theta}$ is learned via inverse reinforcement learning.

Maximum entropy inverse reinforcement learning algorithm :

Compute the empirical feature vector over all demonstrations $\tilde{\boldsymbol{f}}_0 = \frac{1}{m}\sum_{s_j \in D} \boldsymbol{f}(\boldsymbol{s}_j,\boldsymbol{u}_j)$. Normalize the feature, denoted as $\tilde{\boldsymbol{f}}$.

Initialize every entry of the weight vector $\boldsymbol{\theta}$ with 1.

While $\frac{1}{m}\sum_{j=1} \boldsymbol{f}(\boldsymbol{s}_j^{\theta},\boldsymbol{u}_j) - \tilde{\boldsymbol{f}} > threshold$

{

For each demonstrated trajectory collected in the dataset

{

fix the initial condition and the environment states and optimize the trajectory. The optimized trajectories are denoted as $\{\boldsymbol{s}_1^{\theta},\dots,\boldsymbol{s}_m^{\theta}\}$.

}

The gradient can be calculated as $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1} \boldsymbol{f}(\boldsymbol{s}_j^{\theta},\boldsymbol{u}_j) - \tilde{\boldsymbol{f}}$.

Update the parameter vector: $\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) + \gamma \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, in which γ is the learning rate.

}

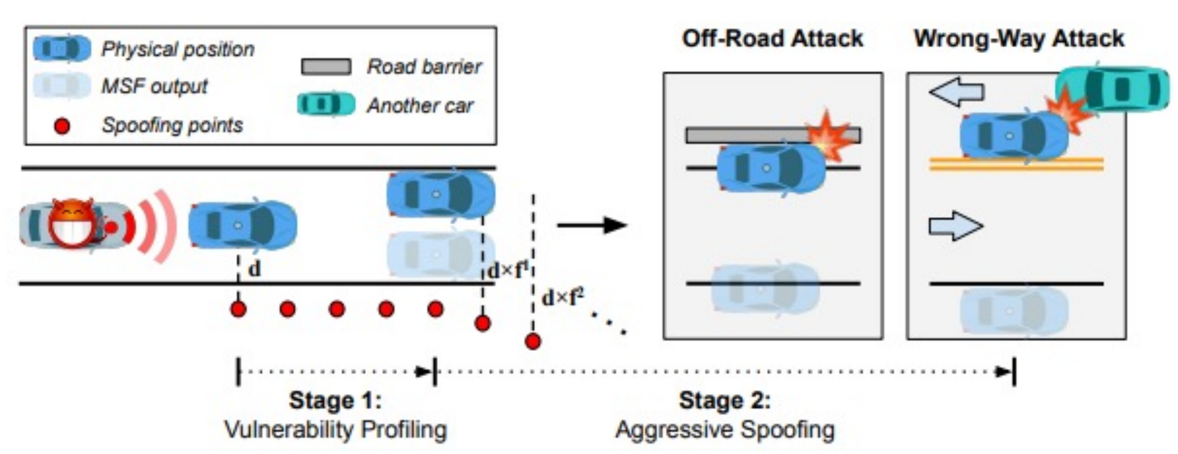
[1] Shen J, Won JY, Chen Z, Chen QA. Drift with Devil: Security of Multi-Sensor Fusion based Localization in {High-Level} Autonomous Driving under GPS Spoofing. In29th USENIX Security Symposium (USENIX Security 20) 2020 (pp. 931-948).

Decision Tree Classifier

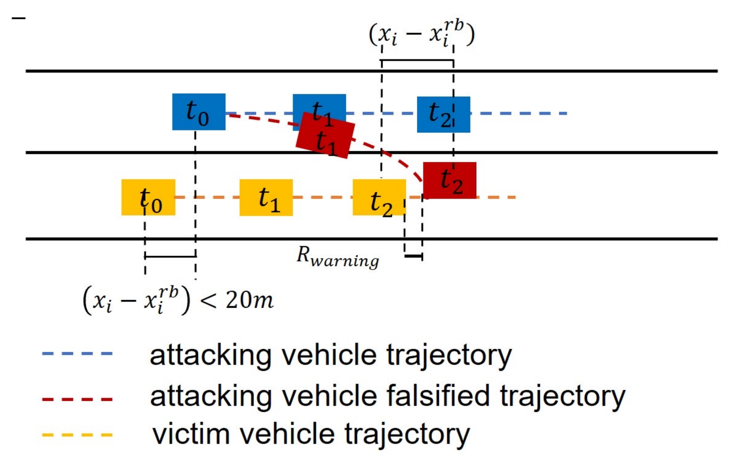
- Objective ratio, normality score and average displacement error are used as classification features.
- Objective ratio: $OR = \max_{1\dots t} OR_t = \frac{\sum_{\tau=1}^{t} observed\ objective_\tau}{\sum_{\tau=1}^{t} optimal\ objective_\tau}$
- Normality score: $NS = \max_{1\dots t} NS_t = \frac{objective_t - objective\ mean_{1\dots t}}{objective\ std_{1\dots t}}$
- Displacement error: $ED = \max_{1\dots t} ED_t = \frac{1}{T}\sum_{i=t}^{t+T}\sqrt{(x_i^{obs}-x_i^{pred})^2+(y_i^{obs}-y_i^{pred})^2}$

Experiment Results

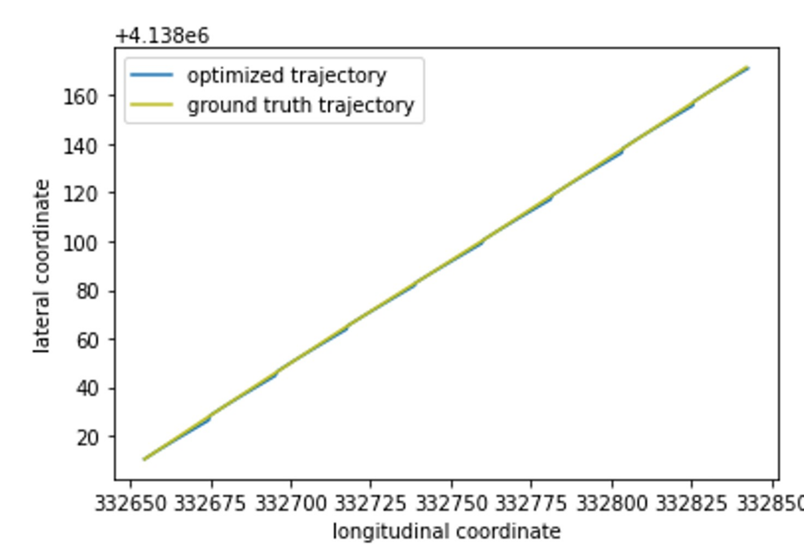
- The anomaly detection algorithm is validated against a Multi-Sensor Fusion attack with the KAIST urban complex dataset and Forward Collision Warning (FCW) attack with the NGSIM Lankershim Blvd. dataset.



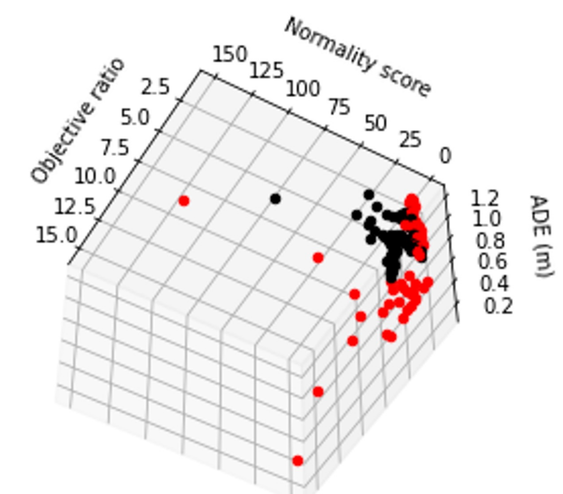
Multi-Sensor Fusion attack[1]



Forward Collision Warning (FCW) attack



Learned Driving Policy (KAIST)

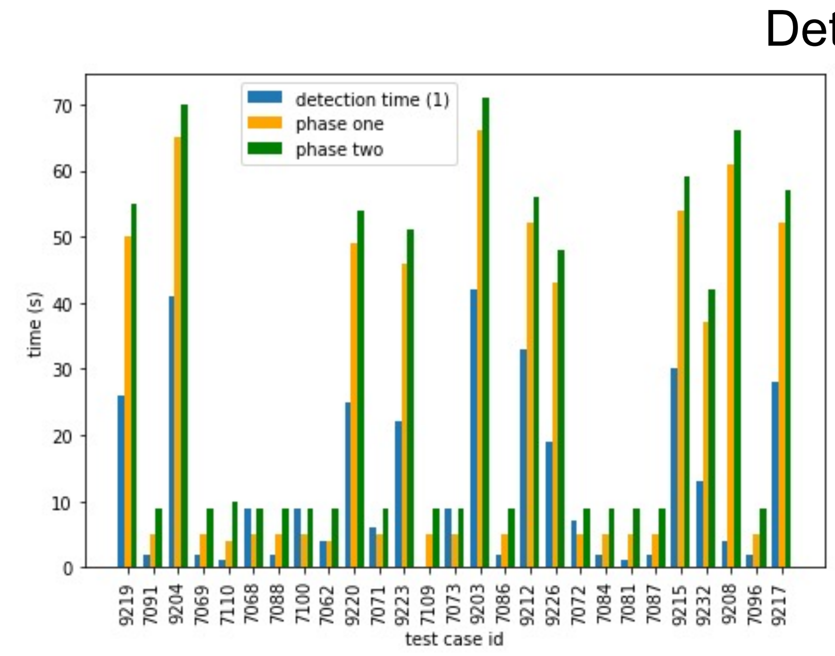


Classification Results (KAIST)

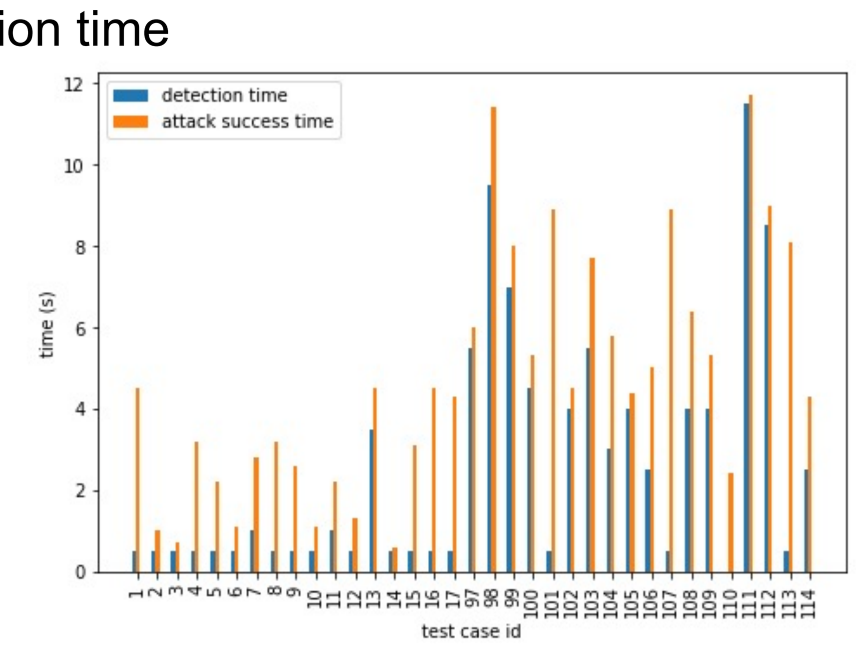
- Performance of online detection on AV/CV threat model

AV threat model				
FP rate	**FN rate**	**Mean attack success time (s)**	**Mean detection time (s)**	**Mean time to attack success (s)**
2/23	1/23	28.7	12.7	16.0
CV threat model				
FP rate	**FN rate**	**Mean attack success time (s)**	**Mean detection time (s)**	**Mean time to attack success (s)**
2/49	1/35	4.7	2.6	2.1

Detection time



KAIST dataset (AV threat model)



NGSIM dataset (CV threat model)

- AV threat model: 94% (47/50) trajectories can be identified no later than the success time of the attack.
- CV threat model: 96% (81/84) trajectories can be identified no later than the success time of the attack.

PURDUE UNIVERSITY Discovery Park