2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 1-5, 2023. Detroit, USA

Lateral-Direction Localization Attack in High-Level Autonomous Driving: Domain-Specific Defense Opportunity via Lane Detection

Junjie Shen Yunpeng Luo Ziwen Wan Qi Alfred Chen
University of California, Irvine {junjies1, yunpel3, ziwenw8, alfchen}@uci.edu

2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-9190-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IROS55552.2023.10342017

***Abstract*— Localization in high-level Autonomous Driving (AD) systems is highly security critical. Recently, researchers found that state-of-the-art Multi-Sensor Fusion (MSF) based localization is vulnerable to GPS spoofing, which can cause road hazards such as driving off road or onto the wrong way. In this work, we perform the first exploration of using Lane Detection (LD) to detect and correct deviations caused by such attacks and design a novel LD-based system-level defense, LD^3. We evaluate LD^3 on real-world sensor traces and find that it can achieve effective and timely detection against the state-of-the-art attack with 100% true positive rates and 0% false positive rates. Results show that LD^3 can be highly effective at steering the AD vehicle to safely stop within the current traffic lane. We implement LD^3 on 2 open-source AD systems and validate its end-to-end defense capability using an industry-grade AD simulator and also in the physical world with a real vehicle-sized AD R&D vehicle.**

I. Introduction

Recently, high-level Autonomous Driving (AD) vehicles [1], e.g., Level-4 ones, are gradually becoming part of the transportation system by providing commercial services [2], [3]. To achieve high driving automation, the *high-level AD system* (the "*brain*") in such a vehicle needs to localize itself with *centimeter-level* accuracy on the map [4], [5] to ensure safe and correct driving. Thus, today's industry-grade high-level AD systems predominantly adopt a Multi-Sensor Fusion (MSF) based localization design, which combines sensor inputs, typically GPS, LiDAR, and IMU, for overall higher accuracy and robustness in practice [6], [7].

Due to the reliance on sensor inputs, AD localization is inherently vulnerable to sensor spoofing attacks, in particular GPS spoofing [8], [9], a long-existing security problem that is fundamentally difficult in both prevention and detection in practice [8], [10]. Although the MSF-based design is generally more robust against such single-source sensor attacks, recent work [8] finds that state-of-the-art industry-grade MSF-based AD localization algorithms can still be vulnerable to *strategic* GPS spoofing in practical settings due to non-deterministic and practical factors such as sensor noises and algorithm inaccuracies. To understand the real-world exploitability of such non-deterministic vulnerabilities, the authors devise a lateral-direction localization attack, dubbed *FusionRipper*, to *opportunistically* inject large lateral deviations (e.g., can be 10 meters) in the MSF localization outputs that are sufficient to cause various levels of safety damages such as unintended lane departure, driving off road, or to the wrong way [11].

Fortunately, we find that the AD context may have a unique opportunity to defend against such lateral-direction

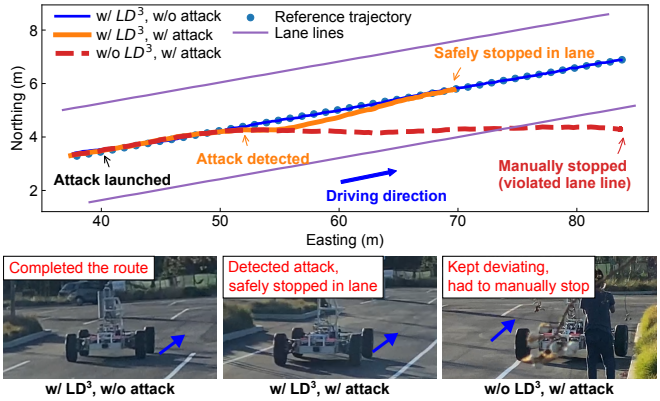

Fig. 1: Physical-world end-to-end demonstrations of LD^3 effectiveness using a Level-4 AD R&D vehicle with closed-loop control. (Top) Vehicle driving trajectories in bird's eye view. (Bottom) Final stopping positions under the three experimental settings. The driving direction and vehicle heading are annotated with blue arrows.

localization attacks – *Lane Detection (LD)* [12], which is directly related to the attack goals since it can measure the vehicle's physical lateral deviation in the ego lane in real time. Today, LD is already widely used in low-level AD localization (e.g., for lane centering in Level-2 AD). However, to the best of our knowledge, *LD is currently not generally used for high-level AD localization (e.g., Level-4) in industry settings*. This is because what LD provides by nature is only *local* positioning (i.e., relative positioning within ego lane), while high-level AD requires *global* positioning (i.e., in world coordinates on a map) for safe and correct driving without human drivers.

In this work, we perform the first study to explore the potential of such domain-specific defense opportunities in AD settings, by designing and prototyping the first LD-based system-level defense approach, called LD^3 (Lane Detection based Lateral-Direction Localization attack Defense), which is capable of both attack *detection* and *response*. Recognizing that existing attack cannot deterministically predict when and where will large deviations occur in MSF [8], LD^3 detects attack at the MSF output level to take advantage of such non-determinism. In the attack response (AR) stage, LD^3 is designed to safely stop the vehicle in the ego lane, which can minimize the attackable duration after detection and thus fundamentally bounds the attack-achievable deviation. To account for the inherent LD-side adaptive attack surface introduced by LD^3, we further design a novel safety-driven

fusion between LD and MSF that systematically penalizes the source that is more aggressive in causing lateral deviations, which can fundamentally reduce the attacker's capability in causing safety damages in AR period even in adaptive attack settings.

We evaluate our defense against the latest lateral-direction localization attack on a diverse set of real-world sensor traces with various environmental conditions. Results show that LD^3 is effective and timely in attack detection, with 100% true positive rates (TPR) and 0% false positive rates (FPR). We also evaluate LD^3 in realistic end-to-end closed-loop controlled setting using an industry-grade AD simulator as well as validated in the physical world on a real vehicle-sized AD R&D vehicle. Fig. 1 shows the vehicle driving trajectories and stopping positions in the physical world experiments. Results show that LD^3 can promptly detect the attack and safely stop the vehicle at the center of the lane, while without LD^3, the vehicle drives out of lane boundary, and we have to manually stop the vehicle to prevent the collision. The demo videos of these experiments and source code are available at **https://sites.google.com/view/cav-sec/LD3**.

II. Background and Threat Model

A. Lateral-Direction Localization Attack via GPS Spoofing

For high-level (e.g., Level-4 [1]) AD localization, a direct threat is the attacks targeting the localization sensors, especially GPS spoofing [13], [14], which has been practically shown on various end systems including AD vehicles [9]. MSF is often considered as a promising defense strategy for GPS spoofing [15], [16]. Contrary to the common belief, prior work [8] proposes an *opportunistic* lateral-direction localization attack method, called *FusionRipper*, which shows the practical exploitability of GPS spoofing alone to inject sufficiently-large *lateral deviations* in the MSF outputs that can cause the AD vehicle to drive off-road or onto the wrong way. *FusionRipper* has shown high attack effectiveness on the representative MSF algorithms, including the one in the industry-grade Baidu Apollo AD system [17]. To our best knowledge, *FusionRipper* [8] is the so far only localization attack that is able to defeat the industry-grade MSF based localization algorithm in practical high-level AD systems.

B. Threat Model

In this work, we assume the attacker can launch practical lateral-direction localization attacks through external means such as GPS spoofing, which can cause lateral deviations in the localization outputs. Specifically, we focus on the lateral-direction attacks since such attacks (1) can cause the AD vehicle to violate the traffic norm that a vehicle should be driving within its designated lane boundaries and should not have unexpected lane straddling behaviors, and (2) pose a direct threat to the AD vehicle and road safety [11].

III. LD^3: Novel LD-based System-Level Defense

Motivation and novelty. Currently, no software-based defense solutions have been proposed to address latest GPS spoofing-based lateral-direction localization attack in high-level AD system (§II-A). The closest ones are the physical-invariants based detectors used for small robotic vehicles (e.g., drones) [18], [19], which estimate the physical dynamics to validate the GPS signal. Although they show high effectiveness for small robotic vehicles with large deviation goals (e.g., 5-10 meters [18]), the effectiveness in AD is fundamentally more limited since (1) vehicle driving in real world is more diverse and complex (e.g., commonly have high-speed or curvy-road), and thus much harder to model accurately [20], [21], and (2) the attack deviation goals in AD context can be much smaller while still being highly safety-critical, e.g., ∼1-2 meters [8]. As we concretely evaluate later in §IV-B, direct adaptation of such existing approach to the AD context can actually suffer from high false positives and the performance is close to random guessing.

As described in §I, we observe a novel and unique defense opportunity in AD context as opposed to small robotic vehicles — *Lane Detection (LD)* [12]. Specifically, we find LD has various levels of advantages if leveraged for such defense purposes: (1) *General defense capability*, since it can provide real-time information directly related to the *attack goal* of lateral-direction localization attacks (i.e., lane departure); (2) *Technology maturity*, as it is already widely adopted in commercial Level-2 AD such as Tesla Autopilot, GM Cadillac, etc.; (3) *Defense deployability*, since today's high-level AD vehicles are all equipped with cameras for object detection, and thus using them for defenses is readily deployable without the need to install new hardware; and (4) *Defense coverage*, since we analyzed all attack traces in the *FusionRipper* paper [8] and found that among all attack starting points, only *0.8% (15/1813)* achieved the attack goal in road regions without lane lines. This means that an LD-based defense, if effective, can already effectively provide protection for 99.2% of the possible attack attempts. However, as discussed in §I, due to its inherent incompatibility (local versus global localization) and the practical accuracy limitation for global localization, LD is currently not generally used for high-level AD localization (e.g., Level-4) in industry settings. In this paper, we propose to be the first to explore novel use of it for *defense purposes*, by designing and prototyping the first LD-based system-level defense approach, called LD^3, which is designed to support both attack *detection* and *response*. An overview of the LD^3 design is in Fig. 2.

A. Attack Detection Design

As shown in Fig. 2, LD^3 performs the attack detection at the MSF output level. We choose to detect at this level instead of at the GPS output level since (1) in normal conditions, GPS positions can naturally have large noises while MSF outputs are at centimeter-level accuracy [6], which means performing the detection at the MSF level can better reduce false positives; and (2) detecting at the MSF output level also allows LD^3 taking advantage of the *opportunistic* property of lateral-direction attacks in MSF settings [8], for which the attacker cannot predict where and

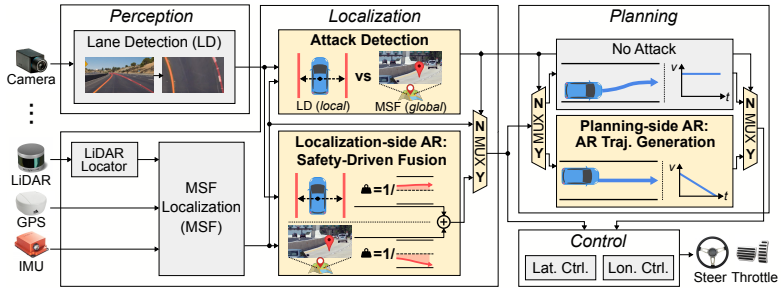

Fig. 2: Overview of the LD^3 design integrated in a typical high-level AD system. New components are highlighted in yellow.

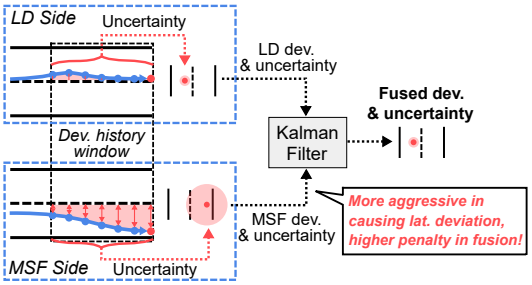

Fig. 3: Illustration of our safety-driven fusion design in the attack response stage (§III-B).

when MSF will exhibit large deviations.

Since the MSF and LD outputs are in different coordinate systems, to make them comparable we first convert them to a unified lateral deviation representation w.r.t. the *lane centerline* since that's directly related to the lateral-direction attack goal. We use semantic map [22], which is a standard utility in high-level AD systems consisting of coordinates about the road surface elements such as lane lines and road signs, to calculate this on the MSF side, and use the detected left and right lane line polynomial functions to calculate this on the LD side. Next, we use their deviation consistency to detect potential attacks. We apply the widely-used CUSUM anomaly detector [18], [23], by calculating a statistic

$$S_i = max(0, S_{i-1} + |r_i| - b), \tag{1}$$

where $S_0 = 0$, $r_i = D_i^{\text{MSF}} - D_i^{\text{LD}}$ is the residual between the MSF and LD lateral deviations at timestamp i, and b is a weight to prevent the CUSUM statistic from monotonically increasing in the benign scenarios. We consider an attack detected if S_i is over a certain threshold τ; after that, we switch to the Attack Response (AR) stage.

B. Attack Response (AR) Design

Since high-level AD vehicles are traveling at high speed and by design do not assume onboard human drivers are ready for take-over at any time (already the case in some commercial AD services [2], [24]), it is necessary to further design an attack response step that can (1) minimize the safety risks during response, and (2) assume no dependence on human assistance. There are several possible design choices, e.g., maintaining driving in the current lane waiting for the system to recover, or pulling over to the roadside. However, these cannot apply to the context of AD localization attacks, since without knowing the accurate real-time location, we cannot even know how to safely and correctly drive in the current lane or to the roadside. To this end, we choose *safe in-lane stopping* as our AR objective, since (1) it has minimal reliance on the attack-time localization accuracy for maximizing safety in the AR period, and (2) this minimizes the attackable duration after detection, it can fundamentally bound the attack-achievable deviation in AR period. Even though stopping in the ego lane is not ideal, it is commonly recognized [25] as one of the fallback strategies to transition to a minimal risk condition when the AD vehicle cannot operate safely. In most driving scenarios, stopping in the ego lane shall not cause a collision as long as the tailgating vehicle is driving with a safe following distance and speed, which is much safer than driving out of the ego lane. As shown in Fig. 2, we need a system-level AR design involving both the planning and localization modules:

Algorithm 1 Safety-driven fusion for attack response

Notations: D: deviation to lane centerline; P: uncertainty from MSF or LD outputs; MSF: MSF position output; kf: 1-dimensional Kalman Filter; R: uncertainty for KF update

function FUSEDPOSE(D^{MSF}, D^{LD}, P^{MSF}, P^{LD}, MSF)
  $R^{\text{MSF}}, R^{\text{LD}} \leftarrow$ UNCERTAINTY(D^{MSF}, D^{LD}, P^{MSF}, P^{LD})
  $kf.update(D^{\text{MSF}}, R^{\text{MSF}})$; $d \leftarrow kf.predict()$
  $kf.update(D^{\text{LD}}, R^{\text{LD}})$; $d \leftarrow kf.predict()$
  $pose_{\text{center}}$, $heading_{\text{center}} \leftarrow$ MAPLANEPOINT(MSF)
  $pose_{\text{fusion}} \leftarrow$ ADDDEVTOPOINT($pose_{\text{center}}$, $heading_{\text{center}}$, d)
  return $pose_{\text{fusion}}$
end function

Planning-side AR: AR trajectory generation. To enforce the AR goal, the planning module needs to generate an AR trajectory with a stopping motion. Since our AR goal is to stop in the ego lane, we design the AR trajectory to be aligned with the lane centerline. To reduce the speed, we set a slowing-down speed profile on the AR trajectory based on a safe deceleration used in high-level AD systems. Generally, a deceleration $<4.6\ \text{m/s}^2$ is considered as safe for maintaining steady control [26]. Thus, to calculate the speed profile of the AR trajectory, we apply 4 m/s^2 as the deceleration, which is also generally used in practical high-level AD systems as the maximum allowed deceleration to ensure safety [17].

Localization-side AR: safety-driven fusion. On the localization module side, one direct AR design is to simply fall back to an LD-based automatic lane centering design as in low-level AD systems. However, this easily exposes the AR period to adaptive attacks on the LD side, which does have concrete examples discovered in recent years [27]. Thus, to account for such inherent LD-side adaptive attack surface introduced by LD^3, we further design a novel safety-driven fusion between LD and MSF that systematically penalizes the source that is more aggressive in causing lateral deviations, which can fundamentally limit the attacker's capability in causing safety damages in AR period in adaptive settings.

To achieve this, we leverage the classic Kalman Filter (KF) fusion algorithm, which can systematically determine the contributions of each fusion source based on their uncertainties. In the original design, the uncertainty score calculation is based on the noise measurements reported by

the sources themselves, which are thus no longer suitable in attack settings since such measurements are also fundamentally under the attacker's control. To systematically realize our safety-driven fusion design, we thus still leverage such uncertainties-based fusion framework but design novel uncertainty score calculation based on *their tendencies to cause lane departure*. Specifically, we calculate MSF and LD uncertainties R^{MSF} and R^{LD} as

$$R^{MSF} = \lambda\sqrt{\frac{\sum D_w^{MSF}}{\sum D_w^{LD}}} + (1-\lambda)P^{MSF}$$
$$R^{LD} = \lambda\sqrt{\frac{\sum D_w^{LD}}{\sum D_w^{MSF}}} + (1-\lambda)P^{LD}, \quad (2)$$

where w is a fixed-sized historical deviation window for MSF and LD. The left components are equivalent to taking the ratio of MSF or LD deviation to their geometric mean. This can greatly penalize the source with a larger cumulative deviation. To increase the design flexibility, we include both our cumulative lateral deviation based uncertainty and the uncertainty from MSF/LD algorithms (the right components) in the final uncertainty and use λ to adjust their fractions. With the uncertainties, we apply standard KF update/prediction to fuse the MSF and LD lateral deviations (lines 3 and 4 in Alg. 1). We then add the fused lateral deviation to the closest centerline point along the lateral direction based on the lane heading to instantiate a fused localization in the global coordinate system (lines 5 and 6 in Alg. 1). Fig. 3 illustrates such safety-driven fusion process.

IV. Trace-based Evaluations

A. Evaluation Methodology

We prototype LD^3 on the industry-grade full-stack Baidu Apollo AD system [17]. For targeted attacks, we evaluate against the *FusionRipper* attack [8]. We follow the same evaluation methodology as in the *FusionRipper* paper [8] using real-world sensor traces from the KAIST complex urban dataset [28]. As listed in Table I, we include 562 attack traces covering diverse driving scenarios, e.g., different road types, driving speeds, time-of-day, and road conditions. For each trace, we find the most effective attack parameters for *FusionRipper* attack. The attack success rates are >98%. In our evaluation, we adopt the LD model in OpenPilot [29], which is already used commercially for Automated Lane Centering. Similar to *FusionRipper* paper, we assume the lateral deviations in the MSF localization will be directly reflected as physical world deviations to the *opposite* direction. We then model the attack-influenced LD outputs by adding the physical world deviations to the lateral deviations calculated from the benign LD outputs.

Baselines. We compare the detection effectiveness of LD^3 to the latest physical-invariant based defense, SAVIOR [18]. For the AR stage, we additionally include a naive AR design, denoted as *NaiveAR*, which tries to reach our safe in-lane stopping goal by only applying the maximum deceleration (the most common strategy for emergency stop [30]) instead of our safety-driven fusion design (§III-B).

TABLE I: Details of the 562 total attack traces used in our evaluation and the *FusionRipper* attack effectiveness. The attack goal deviation is for achieving the off-road attack goal defined in [8], which is smaller (1.3 m) on local-road traces due to the narrower lane widths.

	Attack Trace #	Road Type	Avg. Speed	FusionRipper Attack			
				Attack Goal Dev	Best d	Best f	Success Rate
ka-local31	174	Local	10.9 m/s	1.3 m	0.5	1.2	99.4%
ka-local33	170	Local	9.5 m/s	1.3 m	0.3	1.3	98.3%
ka-highway36	182	Highway	26.3 m/s	1.9 m	0.3	1.3	100%
ka-highway18	36	Highway	24.8 m/s	1.9 m	0.3	1.3	100%

Evaluation metrics. We separate the evaluation into attack detection and response evaluations. For the former, we use the *ROC curves* to systematically show the TPRs and FPRs under different CUSUM parameters b and τ (§III-A). We also report the maximum lateral deviation before the attack is detected by LD^3, denoted as *DetectDev*, to indicate the detection timeliness. A *DetectDev* smaller than the lane straddling deviation (i.e., touching the lane lines) means that the detection is early enough. For AR evaluation, we focus on the lateral deviations since our AR goal is to steer the vehicle to stop within the lane boundaries. Specifically, we report two metrics, *MaxDev*, which measures the maximum lateral deviation before the vehicle fully stops, and *StopDev*, which is the final lateral deviation when the vehicle stops.

B. Attack Detection Effectiveness

Attack detection rates. As shown in the top figures in Fig. 4 LD^3 can achieve effective detection with 100% TPRs and 0% FPRs on all traces. In contrast, SAVIOR's detection performance is only slightly better than random guessing. As discussed in §III, we suspect that this might due to more challenging to accurately model the AD vehicle kinematics in real-world driving scenarios, especially when the deviation it needs to differentiate is quite small (1-2m) in AD settings (Table I). On the other hand, LD is shown to be a much more reliable source in terms of lateral deviation measurement: in benign drivings the differences between MSF and LD lateral deviations are always bounded within <0.6 m, which makes the attack reliably detectable by LD^3.

Attack detection deviation. The bottom figures in Fig. 4 show the distributions of DetectDevs (box plots with pink background). As shown, LD^3 can *always* promptly detect the attack before the vehicle has lane straddling, and the average DetectDevs are all <0.5 m, which is far away from reaching the attack goal deviations (Table I).

C. Attack Response Effectiveness

The distributions of the MaxDev and final StopDev are shown in the bottom figures in Fig. 4 (box plots without background colors). During AR, *none* of the attack cases ever reached the attack goal deviation (1.3 m for local and 1.9 m for highway as in Table I), which shows high effectiveness of our AR design. In fact, the victim AD vehicle barely even has any lane straddling when under attack: only 4 (0.7%, all in *ka-highway36*) out of the 562 attack cases have

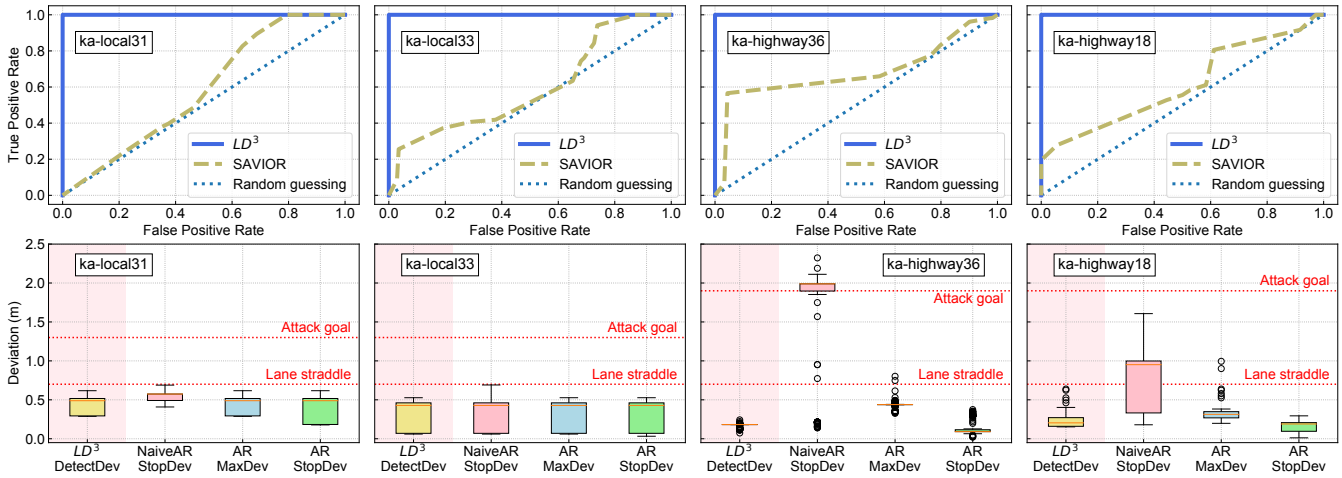

Fig. 4: (Top) Attack detection ROC curves; (Bottom) Detection and Attack Response (AR) deviations in the LD^3 evaluation.

MaxDevs exceeding lane straddling deviation (0.7 m), while at stopping, *all* attack cases are successfully corrected back to be within the lane boundaries, which effectively meets our AR goal of safe in-lane stopping (§III-B).

In comparison, the StopDevs in NaiveAR are significantly higher, especially on the highway traces due to the longer AR periods. Particularly, since the lateral deviations on *ka-highway36* increase very quickly, over 75% of the attacked cases still reach a lateral deviation higher than the attack goal deviation, which consequently leads to $>75\%$ attack success rate despite the correct attack detection. In contrast, with our safety-driven fusion AR design, such attack success rate is effectively reduced to 0%.

Evaluation against adaptive attacks. We take a step further to examine LD^3's capability under potential adaptive attacks, including (1) an idealized stealthy attack that can evade the detection, and (2) the latest LD-side attack, which is the inherent new attack surface introduced by LD^3 approach (§III-B). Our results show that LD^3 can effectively bound the deviations of the stealthy attack from reaching the attack goals and can safely stop the vehicle under the LD-side attack. More details can be seen in our extended version **https://arxiv.org/abs/2307.14540** [31].

V. End-to-End Evaluations

In this section, we implement LD^3 on 2 open-source full-stack AD systems, Baidu Apollo [17] and Autoware [32], and evaluate LD^3 in driving scenarios using both simulation and a Level-4 AD R&D vehicle. The demo videos are at **https://sites.google.com/view/cav-sec/LD3**.

A. Evaluation in Industry-Grade AD Simulator

Experimental setup. We implement LD^3 in Baidu Apollo v5.0.0 [17]. Specifically, we run the complete Baidu Apollo AD system with all functional modules enabled in an industry-grade AD simulator [33]. To simulate the *FusionRipper* attack effect, we add lateral deviations to localization based on the most aggressive *FusionRipper* attack trace, which only takes 10 sec to reach a 2 m lateral deviation. We evaluate LD^3 under benign and attacked drivings in 4 simulation scenarios with different speeds and roads.

Results and demos. Our simulation shows that the attack detection rates for both LD^3 and LD^3-NaiveAR are all 100%, and none of the benign drivings are falsely detected as under attack. With LD^3, the average MaxDev achieved in the simulation are *all* smaller than lane straddling deviation in the 4 scenarios and the vehicle can always safely stop within the ego lane. In comparison, due to the blind trust of the localization outputs, LD^3-NaiveAR has much higher ($\sim$2.40$\times$ on average) MaxDev and the vehicle's stopping positions are always either lane straddling or already crashing into the road curb/barrier. The NoDefense setting is even worse, with $\sim$8.35$\times$ higher MaxDev than LD^3 on average. The video demos for the 4 simulation scenarios and 3 defense settings are all available on our project website.

B. Evaluation on Level-4 AD R&D Vehicle

Experimental setup. We experiment on an AD R&D vehicle as shown in Fig. 1, which is specifically designed for Level-4 AD system testing. The R&D vehicle is of real-vehicle size (2.7m$\times$1.5m), closed-loop controlled, and fully equipped with Level-4 AD sensors including LiDAR, GPS, IMU, cameras, RADARs, and ultrasonic sensors. Since AD vehicle testing is not allowed on public roads by default, we reserve a parking lot in our institute for the experiments. We mark a straight traffic lane with 3.5 m width in the parking lot and create the corresponding semantic map for Autoware.

We implement LD^3 in the Autoware AD system [32]. For the attacked scenario, we directly inject the same *FusionRipper* attack trace used in §V-A to the localization outputs in Autoware. We evaluate 3 defense settings: (1) *w/ LD^3, w/ attack*; (2) *w/o LD^3, w/ attack*; and (3) *w/ LD^3, w/o attack*. For each, we experiment in low driving speeds of 2 m/s (4.5 mph) and 4 m/s (9 mph) to reduce safety risks.

Results and demos. Our results show that LD^3 on average can detect the attack when the vehicle's physical deviation is still small ($\sim$0.5m). Within the AR period, the average maximum deviations are 0.36 m and 0.27 m at speeds of 4 m/s and 2 m/s, respectively, and the final stopping deviations

are always within 0.1 m. In comparison, without LD^3, the vehicle keeps deviating, and we have to manually press the emergency button on the remote to prevent it from crashing into the curb. Such distinctive driving behaviors with and without LD^3 are consistent with our trace-based (§IV) and simulation results (§V-A). Without the attack, the vehicle's trajectories well align with the road centerline and eventually complete the route and stop at the center of the lane. We record demo videos of the vehicle driving behaviors under the three settings (available on our website). As an illustration, Fig. 1 visualizes the driving trajectories in the bird's eye view and shows the snapshots of final stopping positions at driving speed of 4 m/s.

VI. Related Work

Physical-invariant based defenses. Recently, researchers propose physical-invariant based defenses [18], [19], to detect sensor attacks such as GPS spoofing by cross-checking sensor measurements with system state estimations based on the physical invariants, i.e., the relationships between system states and control inputs. However, as shown in §IV-B, the direct adaptation of such existing approaches to the AD context is far from effective in practical scenarios, likely due to the much higher complexity of the physical dynamics and much smaller attack deviation goals in the AD context. Also, none of them has proposed attack response designs, which is especially important for AD systems. Nevertheless, we would like to note that such physical-invariant based attack detection methods are complimentary to LD^3 and can be incorporated into our design for attack detection if the accuracy of state-estimation model can be further improved.

Attack response/recovery. Existing defenses in CPS security area mostly focus on attack detection and very few studied attack responses [34]. Particularly, Choi et al. [35] and Zhang et al. [36] propose *attack recovery* methods, which apply similar state estimations as above to replace attacked sensors in the attack recovery period. However, this means that they are also suffering from the same model accuracy limitations in the AD context. Moreover, they can only maintain normal operations of the system for a short duration until the system is taken-over by the human driver, which does not apply to high-level AD vehicles that aim at driverless deployment [2], [24]. In addition, they all assume an effective attack detection in the first place, which does not yet exist in for high-level AD localization.

AD system security. Prior works studied attacks and defenses of AD system components, such as object detection, localization, lane detection, and planning [27], [37]–[48]. Only *FusionRipper* [8] is able to break the MSF localization on high-level AD systems and cause lateral deviations in MSF outputs. In this paper, we show that LD^3 can effectively detect *FusionRipper* (localization attack) and steer the vehicle to safely stop in ego lane.

VII. Conclusion and Limitation Discussions

In this work, we perform the first exploration of using LD to defend against state-of-the-art lateral-direction AD localization attacks. We design and prototype a novel system-level defense approach, LD^3, shown capable of detecting attacks accurately and timely, and safely stopping the vehicle within the current lane in various evaluations using sensor traces, simulation, and AD R&D vehicle in physical world.

Same as all CPS security research that uses sensor cross-checking/fusion for defense purposes [49]–[54], a fundamental limitation of our current design is simultaneous attacks on both MSF and LD, which can in theory fundamentally bypass our defense. However, LD^3 can sufficiently raise the bar of the attacker in both theory and practices. For example, all the 3 practical LD attacks today targeting AD settings [27], [55], [56] leveraged malicious patterns on the ground (e.g., via road patches or stickers) as the attack vector. Considering the non-deterministic nature of the existing high-level localization attacks [8], it would be fundamentally hard, if not impossible, for the attacker to figure out where to place the attack pattern beforehand, not to mention how to carefully synchronize the malicious pattern with the localization-side attack to effectively bypass LD^3. Thus, we consider such simultaneous attack design still an open research question and leave the systematic exploration of it and the countermeasures as a future direction.

Another limitation is that our detection and response happen after the attack has occurred to some extent (i.e., some deviations have already been caused by the attack). Even though our system can greatly reduce the safety consequences and transition the vehicle into a minimal-risk condition, it is still better if we can detect the attack immediately after the first injection is sent to the system. We thus consider this as another future direction.

Acknowledgment

We would like to thank the anonymous reviewers for their valuable feedback on our work. This research was supported in part by the NSF under grants CNS-1929771, CNS-2145493, and USDOT under grant 69A3552047138.

References

[1] SAE On-Road Automated Vehicle Standards Committee and others, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," *SAE International*, 2021.

[2] "Baidu To Operate 3,000 Driverless Robotaxis." https://tinyurl.com/3hru7734.

[3] "Waymo's commercial self-driving service." https://tinyurl.com/5n8jaw24.

[4] J. Levinson, M. Montemerlo, and S. Thrun, "Map-Based Precision Vehicle Localization in Urban Environments," in *Robotics: Science and Systems*, vol. 4, p. 1, Citeseer, 2007.

[5] T. G. Reid, S. E. Houts, R. Cammarata, G. Mills, S. Agarwal, A. Vora, and G. Pandey, "Localization Requirements for Autonomous Vehicles," *arXiv preprint arXiv:1906.01061*, 2019.

[6] G. Wan, X. Yang, R. Cai, H. Li, Y. Zhou, H. Wang, and S. Song, "Robust and Precise Vehicle Localization based on Multi-Sensor Fusion in Diverse City Scenes," in *ICRA*, 2018.

[7] Y. Gao, S. Liu, M. Atia, and A. Noureldin, "INS/GPS/LiDAR Integrated Navigation System for Urban and Indoor Environments Using Hybrid Scan Matching Algorithm," *Sensors*, vol. 15, no. 9, 2015.

[8] J. Shen, J. Y. Won, Z. Chen, and Q. A. Chen, "Drift with Devil: Security of Multi-Sensor Fusion based Localization in High-Level Autonomous Driving under GPS Spoofing," in *USENIX Security*, 2020.

[9] "Tesla Model S and Model 3 Vulnerable to GNSS Spoofing Attacks." https://tinyurl.com/3fxv9hpa.

[10] M. L. Psiaki and T. E. Humphreys, "GNSS Spoofing and Detection," *Proceedings of the IEEE*, vol. 104, no. 6, pp. 1258–1270, 2016.

[11] Federal Highway Administration, "Roadway Departure Safety." https://safety.fhwa.dot.gov/roadway_dept/.

[12] A. B. Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: a survey," *Machine Vision and Applications*, vol. 25, no. 3, pp. 727–745, 2014.

[13] C4ADS, "Above Us Only Stars - Exposing GPS Spoofing in Russia and Syria." www.c4reports.org/aboveusonlystars.

[14] A. J. Kerns, D. P. Shepard, J. A. Bhatti, and T. E. Humphreys, "Unmanned Aircraft Capture and Control via GPS Spoofing," *Journal of Field Robotics*, 2014.

[15] D. Davidson, H. Wu, R. Jellinek, V. Singh, and T. Ristenpart, "Controlling UAVs with Sensor Input Spoofing Attacks," in *USENIX Workshop on Offensive Technologies (WOOT)*, 2016.

[16] A. Cardenas, "Cyber-Physical Systems Security Knowledge Area," *The Cyber Security Body Of Knowledge (cybok)*, 2019.

[17] Baidu, "Baidu Apollo." github.com/ApolloAuto/apollo.

[18] R. Quinonez, J. Giraldo, L. Salazar, E. Bauman, A. Cardenas, and Z. Lin, "SAVIOR: Securing Autonomous Vehicles with Robust Physical Invariants," in *USENIX Security*, 2020.

[19] H. Choi, W.-C. Lee, Y. Aafer, F. Fei, Z. Tu, X. Zhang, D. Xu, and X. Deng, "Detecting Attacks Against Robotic Vehicles: A Control Invariant Approach," in *ACM CCS*, 2018.

[20] J. Kong, M. Pfeiffer, G. Schildbach, and F. Borrelli, "Kinematic and Dynamic Vehicle Models for Autonomous Driving Control Design," in *IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2015.

[21] P. Polack, F. Altché, B. d'Andréa Novel, and A. de La Fortelle, "The Kinematic Bicycle Model: a Consistent Model for Planning Feasible Trajectories for Autonomous Vehicles?," in *IV*, IEEE, 2017.

[22] "Semantic Maps." https://tinyurl.com/y65tmaev.

[23] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the Impact of Stealthy Attacks on Industrial Control Systems," in *ACM CCS*, 2016.

[24] Kirsten Korosec, "Waymo's driverless taxi service can now be accessed on Google Maps." https://tinyurl.com/tzrussmp.

[25] "NHTSA Automated Driving Systems 2.0 Voluntary Guidance." https://tinyurl.com/mr4zmkrm.

[26] "Acceleration and Braking Parameters." https://tinyurl.com/bdeth9tf.

[27] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack," in *USENIX Security Symposium*, pp. 3309–3326, 2021.

[28] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex Urban Dataset with Multi-Level Sensors from Highly Diverse Urban Environments," *IJRR*, vol. 38, no. 6, pp. 642–657, 2019.

[29] "Openpilot." github.com/commaai/openpilot.

[30] "Apollo Planning." https://tinyurl.com/3rk5mje4.

[31] "Extended Version." https://arxiv.org/abs/2307.14540.

[32] S. Kato, S. Tokunaga, Y. Maruyama, S. Maeda, M. Hirabayashi, Y. Kitsukawa, A. Monrroy, T. Ando, Y. Fujii, and T. Azumi, "Autoware On Board: Enabling Autonomous Vehicles with Embedded Systems," in *ICCPS*, pp. 287–296, IEEE Press, 2018.

[33] LG, "LGSVL Simulator: An Autonomous Vehicle Simulator." https://github.com/lgsvl/simulator.

[34] J. Giraldo, E. Sarkar, A. A. Cardenas, M. Maniatakos, and M. Kantarcioglu, "Security and Privacy in Cyber-Physical Systems: A Survey of Surveys," *IEEE Design & Test*, vol. 34, no. 4, pp. 7–17, 2017.

[35] H. Choi, S. Kate, Y. Aafer, X. Zhang, and D. Xu, "Software-based Realtime Recovery from Sensor Attacks on Robotic Vehicles," in *RAID*, pp. 349–364, 2020.

[36] L. Zhang, X. Chen, F. Kong, and A. A. Cardenas, "Real-Time Attack-Recovery for Cyber-Physical Systems Using Linear Approximations," in *RTSS*, 2020.

[37] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "3D Adversarial Object against MSF-based Perception in Autonomous Driving," in *MLSys*, 2020.

[38] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical World Attacks," in *IEEE S&P*, May 2021.

[39] Z. Wan, J. Shen, J. Chuang, X. Xia, J. Garcia, J. Ma, and Q. A. Chen, "Too Afraid to Drive: Systematic Discovery of Semantic DoS Vulnerability in Autonomous Driving Planning under Physical-World Attacks," in *NDSS*, April 2022.

[40] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li, Z. Zhao, C. Qiao, and Q. A. Chen, "SoK: On the Semantic AI Security in Autonomous Driving," *arXiv preprint arXiv:2203.05314*, 2022.

[41] T. Sato, J. Shen, N. Wang, Y. J. Jia, X. Lin, and Q. A. Chen, "WIP: Deployability improvement, stealthiness user study, and safety impact assessment on real vehicle for dirty road patch attack," in *AutoSec*, vol. 2021, p. 25, 2021.

[42] T. Sato, J. Shen, N. Wang, Y. J. Jia, X. Lin, and Q. A. Chen, "Hold Tight and Never Let Go: Security of Deep Learning based Automated Lane Centering under Physical-World Attack," *ArXiv*, 2020.

[43] C. DiPalma, N. Wang, T. Sato, and Q. A. Chen, "Security of Camera-based Perception for Autonomous Driving under Adversarial Attack," in *SPW*, pp. 243–243, IEEE, 2021.

[44] C. Ma, N. Wang, Q. A. Chen, and C. Shen, "WIP: Towards the Practicality of the Adversarial Attack on Object Tracking in Autonomous Driving," in *VehicleSec*, 2023.

[45] N. Wang, Y. Luo, T. Sato, K. Xu, and Q. A. Chen, "Poster: On the System-Level Effectiveness of Physical Object-Hiding Adversarial Attack in Autonomous Driving," in *ACM CCS*, pp. 3479–3481, 2022.

[46] Y. Huai, Y. Chen, S. Almanee, T. Ngo, X. Liao, Z. Wan, Q. A. Chen, and J. Garcia, "Doppelgänger Test Generation for Revealing Bugs in Autonomous Driving Software," in *ICSE*, pp. 2591–2603, IEEE, 2023.

[47] Y. Luo, N. Wang, B. Yu, S. Liu, and Q. A. Chen, "Infrastructure-Aided Defense for Autonomous Driving Systems: Opportunities and Challenges," in *AutoSec Workshop*, 2022.

[48] T. Sato, J. Shen, N. Wang, Y. J. Jia, X. Lin, and Q. A. Chen, "Security of Deep Learning based Automated Lane Centering under Physical-World Attack," in *2021 IEEE Security and Privacy Workshops (SPW)*, pp. 244–244, IEEE, 2021.

[49] Z. Feng, N. Guan, M. Lv, W. Liu, Q. Deng, X. Liu, and W. Yi, "An Efficient UAV Hijacking Detection Method Using Onboard Inertial Measurement Unit," *TECS*, vol. 17, no. 6, pp. 1–19, 2018.

[50] Z. Feng, N. Guan, M. Lv, W. Liu, Q. Deng, X. Liu, and W. Yi, "Efficient Drone Hijacking Detection using Onboard Motion Sensors," in *DATE*, pp. 1414–1419, IEEE, 2017.

[51] W. G. Aguilar, V. S. Salcedo, D. S. Sandoval, and B. Cobeña, "Developing of a Video-Based Model for UAV Autonomous Navigation," in *LAWCN*, pp. 94–105, Springer, 2017.

[52] Ç. Tanıl, S. Khanafseh, and B. Pervan, "Detecting Global Navigation Satellite System Spoofing Using Inertial Sensing of Aircraft Disturbance," *Journal of Guidance, Control, and Dynamics*, 2017.

[53] S. Khanafseh, N. Roshan, S. Langel, F.-C. Chan, M. Joerger, and B. Pervan, "GPS Spoofing Detection using RAIM with INS Coupling," in *PLANS*, pp. 1232–1239, IEEE, 2014.

[54] B.-H. Lee, J.-H. Song, J.-H. Im, S.-H. Im, M.-B. Heo, and G.-I. Jee, "GPS/DR Error Estimation for Autonomous Vehicle Localization," *Sensors*, vol. 15, no. 8, pp. 20779–20798, 2015.

[55] B. Nassi, D. Nassi, R. Ben-Netanel, Y. Mirsky, O. Drokin, and Y. Elovici, "Phantom of the ADAS: Phantom Attacks on Driver-Assistance Systems," *IACR Cryptol. ePrint Arch.*, 2020.

[56] P. Jing, Q. Tang, Y. Du, L. Xue, X. Luo, T. Wang, S. Nie, and S. Wu, "Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations," in *Usenix Security*, 2021.