Security Analysis of Camera-LiDAR Fusion Against Black-Box Attacks on Autonomous Vehicles

R. Spencer Hallyburton, Yupei Liu and Miroslav Pajic, *Duke University*
Yulong Cao and Z. Morley Mao, *University of Michigan*

Sudhanshu Madhukar Tarale

1. Summary of the paper's findings

1.1. Introduction and context of the study

This paper performs an analysis of camera-LiDAR fusion in autonomous vehicles (AVs) where the AV's are under LiDAR spoofing attacks. A novel attack, called the *frustm attack* is developed to show that all 8 widely used perception algorithms under different architectures are vulnerable to this attack. We also see how this attack preserves consistencies between camera and LiDAR semantics and is thus stealthy.

The idea of perception; trying to find where an object/obstacle is situated is critical to the safety of AV's and its surroundings. AV's decision making is hinged on the fact that it can or cannot detect an obstacle. Thus because of how important the safety factor is, it is necessary to conduct research on the limitations of the AV's sensors and algorithms that it uses to make critical decisions (in this case it's mainly camera and LiDAR sensor).

1.2. Research questions

In existing security analysis of LiDAR based perception, the perception model is known to the attacker and those perception models mainly consisted of single sensors. This paper attempts to perform a security analysis where multiple sensor perception is used and the attacker doesn't know which perception model is being used. They call this as black box testing, and this is one of the first papers to perform this kind of testing.

1.3. Main papers contributions

The paper has following 4 main contributions:

1. The sensor fusion algorithm is quite secure. Less than 1% of the algorithms are affected by naïve LiDAR spoofing attacks.
2. Thus the paper works on attacking the sensors with the *frustum attack* and noting down how the attack affects the existing hardware in different real life scenarios.
3. Performing analysis on LiDAR-only and camera-LiDAR perception and also show that the frustum attack can compromise 8 high-performing perception algorithms across 3 LiDAR-only and 3 camera-LiDAR fusion architectures. The analysis also shows that the attack is stealthy.
4. Showing that the attacker has high level of attack success when targeting the short and long range of the AV's software. The paper also talks about doing a longitudinal study of security against perception attacks. Longitudinal studies means that the researchers observe and collect data on a number of variables without trying to influence those variables.

1.4. Threat model

Here we look at how AV's systems are attacked, what knowledge do we assume to have and all the background we need to know in order to attack and draw inferences from it. The goal of the attack is to make the driver perform dangerous maneuvers to avoid the fake object that the *frustum attack* manufactures for the AV's software's.

1.4.1 Attacker Capability

Assumption is made that that the attacker has no access to AV's internal mechanism. A threat model is followed wherein about 200 spoofing points are injected. The spoofing attack that is used is from this paper [1]. The attack in the referred paper shows that how a deceiving physical signal into a victim sensor can cause problems. Meaning that the attack will spoof a vehicle in from the of the victims AV. An assumption was made with these attacks that the spoofing points must be in a very high precision shape (eg: shape of a car) but these restrictions are relaxed by allowing the attacker to randomly placing the points. This also takes into consideration attacks caused by a noisy laser beam as in those cases, the laser will be in random spots and not forming any precise shapes.

1.4.2 Attack Strategy

This paper looks at 2 types of attacks:

Naive attack: This kind of attack compromises a single sensor without considering the consistency between multiple sensors or the environment.

Frustum attack: This attack retains consistency across multiple sensors. It stems from the fact that a 2D camera cannot see how far an object is and thus there is 3D uncertainty in the 2D camera. Attacking withing the range of a target vehicle retains consistency with semantic and feature information between camera and LiDAR data.

1.4.3 Attacker Knowledge

The attacker only needs to know what kind of data is required by the relay system. They do not need to know the perception model or the architecture used in order to perform attacks. There is an assumption that the attacker knows the

approximate position of the target vehicle so that they know where to place the spoofing points with regards to the target vehicle.

1.5. Summary of the methodology

The paper is trying to implement a spoofing attack talked about in these papers [1] and [2] but with some relaxations on the assumption that the spoofed points have to form something meaningful like an outline of a car or any other obstacle on the road. They have defined *frustum* as the 3D uncertainty of a 2D camera and are using this to retain consistency with the data that the senor is supposed to get and thus keeping the attack stealthy. This paper takes more real life scenarios into consideration as compared to previous papers and studies how the AV's sensors will behave to those attacks.

They have also performed black-box testing which means that the attacker will not have any information about the perception sensors or the algorithm that is being used and thus simulating a more real life test case. This kind of testing is important as it tests the sensors against more practical/likely attacks rather than attacks that are performed just so that the sensors fail (or in this case give wrong output).

1.6. Summary of the experiments and results

The paper exposed the vulnerability of LiDAR-only perception and camera-LiDAR fusion to the frustum attack. The attack was used on three distinct LiDAR-only architectures and five models within three different architectures of camera-LiDAR fusion, including fusion at the semantic, feature, and tracking levels. It was demonstrated that a singular black-box model attack where no knowledge of perception algorithm is required is capable of compromising each class perception in AV's. Having achieved such success with black-box attack testing, this points to the fact that the existing LiDAR-only and camera-LiDAR perception algorithms have many vulnerabilities that need to be looked at as these issues are safety critical.

Frustum attack has been tested in against a few different scenarios. The first case is when vehicles are at intersection. Here the vehicles are initially static. The goal here was to make a vehicle speed up towards the victim by successfully spoofing cluster of points behind the victim. It was found that 8 out 10 injections by the attacker were falsely detected by perception and resulted in creating an adversarial track.

The second scenario is when vehicles are in cruise control on a highway. AV's use perception to monitor objects and keep up with traffic. As vehicles are already in motion, the sensors already have a track decided based on it's surroundings. Thus it becomes difficult for the attacker to disrupt this track as major deviations in it will cause doubt. So here, 5 out of 10 spoofs were successful in deviating the victim dangerously. The researchers have also used the frustum attack on Baidu Apollo algorithm with success.

Thus, we can see that shown that frustum attacks can have high-impact on tracking, decision, and control levels of a vehicle. An attacker can use mere seconds of real-time to create false scenarios of predicted collision or accelerate the flow of traffic.

2. Main takeaways and limitations of the work

2.1. Major takeaways

The major takeaway would be that it was demonstrated that the current camera-only and camera-LiDAR perception sensors on the AV's are vulnerable to a particular attack model. The important part to note here that these experiments are conducted in black box environment; meaning the attacker doesn't know about the perception algorithm used by the AV. What we can understand from this is that the attacker has a success rate in a real-world type situation which results in safety critical situations. Although as mentioned in the paper, they showed that sensor fusion is intrinsically more robust to naive attacks. They have demonstrated that the perception models and defenses perform poorly under a new class of attacks: the frustum attacks. This attack is effective against both LiDAR-only and camera-LiDAR fusion and is stealthy to the defenses.

2.2. Major limitations

Dataset: The dataset used was not fully representative of a real-world scenario; more data points are necessary for that.

Attack on Baidu Apollo algorithm: The algorithm is only tested when the attacker spoofs points in front of the target vehicle. Testing needs to be done with points behind the target (i.e, shadow region). But the problem is that the algorithm does not take any evasive measure when points are spoofed in the back of the target vehicle. This is shown to be consistent with findings from this paper [3].

Dynamic Spoofing: The frustum attack has shown great success when the victim is static. The current experiments have not shown attack feasibility for spoofing with different angles and speeds of the victim. This would require the spoofing device to dynamically track and aim at the victim, and this engineering feat has not yet been fully demonstrated.

3. Fundamental previous work

3.1. Previous paper title 1

Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures by J.Sun, Y. Cao, Q. A. Chen, and Z. Morley Mao.

3.2. Research problem

This paper performs the first study to explore the general vulnerability of current LiDAR-based perception

architectures and discover that some patterns in LiDAR point clouds make self-driving cars vulnerable to spoofing attacks. The existing attacks (at the time this paper was published;2020) weren't as effective and this paper explore how LiDAR sensors can be exploited even more. A defense study is also performed to mitigate LiDAR spoofing attacks.

3.3. Main contributions of the work

This work performs the first study about the general vulnerabilities existing among three state-of-the-art LiDAR-based 3D object detection model designs: bird's-eye view-based, voxel-based, and point-wise. They have referred to previous work [4] which found that found that an attack trace with merely 60 points is sufficient to spoof a front-near vehicle in Apollo 2.5, while a valid trace should have ∼2000 points [5] (the paper referred is the one that created the KITTI Dataset hence we can look at it and see how many points are required to spoof a vehicle).

Thus, having learnt this indifference in the model, this work furthers the previous work with an attempt to generalize to other state-of-the-art 3D object detection model designs.

3.4. Correlation with the presented work

The presented work referred to this paper to use it's LiDAR spoofing threat model to show that camera-LiDAR fusion give additional robustness against general naive LiDAR attacks; this is because the naive spoofing does not retain consistency between camera data. In this work, it also point to the fact that attack success could be greatly reduced with sensor fusion when not all sensors are compromised. The presented work picks up on this and performs a systematic study to understand more about sensor fusion and how to attack it. The presented paper then comes up with an attack of their own (frustum attack) to check vulnerabilities on sensor fusion.

3.5. Previous paper title 2

Vision meets Robotics: The KITTI Dataset by Andreas Geiger, Philip Lenz, Christoph Stiller and Raquel Urtasun

3.6. Research problem

As the authors of this paper put it, the main purpose of this paper that gives us the KITTI dataset is to is to push forward the development of computer vision and robotic algorithms targeted to autonomous driving. There hasn't been any high quality raw data available for researchers previously, but this dataset gives us raw real time data with annotations so that a model can learn about the what object it sees on the road.

3.7. Main contributions of the work

This is a very important dataset. This seems to be one of the first papers to capture very high-quality data which helps with computer vision algorithms. It's a very vast dataset which is broken down into different categories like 'Road', 'City', 'Residential' etc. This is very important as it gives future researchers the opportunity to test their attacks or defenses in different scenarios leading to better inferences. The authors have also mentioned that the data gathering took place over a 5 day period and only in daytime.

They have pointed out that there is a need to collect more data in different scenarios and different lighting conditions. They plan on expanding the dataset by adding additional 3D object labels for currently unlabeled sequences and recording new sequences. This paper and the work done on it further seems to very important for the AV community as it gives them something to work with in terms of data. This is a goldmine of a dataset to start researching about the sensors and how AV's react to being spoofed due to how vast it is.

3.8. Correlation with the presented work

This dataset is used by the presented paper. It's interesting to note that this paper does say that they have only put about 25% of their raw data online but one of the limitations given by the presented paper is that the dataset isn't big enough. The presented paper reproduces the LiDAR spoofing attack by using patterns of stopped vehicle from KITTI and then use these vehicles position as attack points for the target vehicle by starting with 10 points and going upto 200 in increments of 10.

The presented paper found that they don't exactly need to spoof points in the shape of stopped cars extracted for the KITTI dataset but rather that spoofing using a normal distribution of points with moments can achieve performance on-par.

4. Proposed defenses

4.1. Defense 1 description

A defense referred to by the presented paper is Shadow-Catcher [6]. As pointed out in [6], it is easy to spoof small objects like pedestrians, cycle etc and thus no existing defense system could sufficiently detect such spoofing. This defense technique also catches shadows of cars which is quite important as these things are easy to spoof and there are no other defenses that detects shadows and classifies them.

4.2. Advantages of the defense if implemented.

From [6] we can see that this defense is quite effective as it achieves 94% and 96% average accuracy in identifying anomalous shadows and classifying them as either ghost or invalidation attacks. Shadow-Catcher can analyze objects in real time (0.003s–0.021s on average, a 2.17x speedup compared to other algorithms). This mechanism is agnostic to the classification model targeted: any detected object, either genuine or fake, will be verified.

Although it has been pointed out in the presented paper that ShadowCatcher does not perform well in detecting the frustum attack. The reason that a high success rate for the defense is not possible because the original work made

several assumptions that are unrealistic, including tuning parameters on the test set (tuning parameters show us the complexity of the model and also affect any variance-base trade-off that can be made), hand labeling the objects instead of outputs of a perception algorithm, which significantly alters the region to be tested, and not enough testing..

4.3. Defense 2 description

For the second defense we look at PyCRA [7]. Traditional methods of security include a query-response check or encryption algorithms, but in PyCRA, the authors propose security to sensors at a point prior to the digitization of the sensor response. It turns off the active sensing signal at random times, called challenge periods. PyCRA assumes that an attacker cannot detect a challenge immediately due to its hardware and signal processing latency.

They leverage an active sensor's ability to emit energy to 1) provide detection of active attackers trying to spoof the sensor, 2) mitigate the effects of active spoofing attacks and 3) detect passive eavesdropping attacks attempting to listen to the information received by the sensor. These properties of the sensor are targeted by this paper to provide defense. For AV's optical and magnetic encoders are some of the active sensors used.

However, there can be a drawback, for example, an automotive radar used for safety-critical applications, e.g., adaptive cruise control and collision warning, will be turned off at random times to detect possible attack. As a result, the availability of the radar system can be decreased, potentially affecting the safety.

4.4. Advantages of the defense if implemented.

This provides a defense when a sensor is challenged physically. PyCRA is able to isolate the response of the attacker from the response of the environment when challenged. It gives the attacker a random challenge code. Sort of like asking an attacker if they have the key to make changes to the sensor data. This level of security is further bolstered by the randomness of the challenge sent to the attacker. The relationship between randomness and security guarantees is a classical relationship that appears in most cryptographic security protocols.

This defense is applicable for very specific sensors. In this case, the defense does not work for passive sensors (sensors that don't send out data, rather detect it without any activity). For a fool proof defence, a fusion of this and other defenses like ShadowCatcher and CARLO could even be useful.

5. Proposed follow-up research idea

5.1. Follow-up research idea description

A follow up research idea would be to find defenses against the frustum attack. There is also a need for white box testing on the sensors along with a fusion of different attacks to further bolster the security of AV's by finding defenses for the attacks that have high success rate (in this case the frustum attack). Black box testing has its advantages, but it cannot be the only parameter to determine if an attack is successful. For now, there isn't complete solution to these attacks and thus complete testing must be done for better understanding. White box testing will target the inner workings of the sensor (since the attacker will know the perception model or the algorithm used by the sensor). This has 2-fold advantages: the first being that maybe better attacks can be developed to test the sensors and how they will behave. The second being that we get a deeper understanding of why the sensors fail in certain scenarios when the attacker knows the algorithms being used.

The attack created by the authors could be researched upon to find a way to avoid it. This is important as the success rate of the frustum attack was quite high in an environment where the attack didn't even know about the algorithms being used.

5.2. Research questions

Conducting a thorough research on the frustum attack with white box and black box technique will ask the question about is the sensor safe. Because we do not know if an attacker has information about the inner workings of AV or not. Hence it makes sense to test using all tools at our disposal. This also opens up to the idea that is there a combination of attacks/defenses that can work rather than just one type of it. This can also lead us to understand how does an attack work against a particular sensor. Is there a group of common sensors that is vulnerable to a particular type of attack?

As cyber physical systems involve a whole lot of computational science, network communication, control theory and other disciplines this technology can maybe translate into nano-level biological robots, fine agriculture, pills electronic endoscopy and other local or micro-system. Finding a defense for such an attack is important as it can translate to many other fields of study where similar sensors are used.

REFERENCES

[1] J.Sun, Y. Cao, Q. A. Chen, and Z. Morley Mao, "To- wards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures," in Proceedings of the 29th USENIX Security Symposium, pp. 877–894, 2020

[2] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," in Proc. of the 2019 ACM SIGSAC Conf. on Computer and Communications Security, pp. 2267–2281, 2019.

[3] A. Piazzoni, J. Cherian, M. Azhar, J. Y. Yap, J. L. W. Shung, and R. Vijay, "ViSTA: a Framework for Virtual Scenario-based Testing of Autonomous Vehicles," arXiv preprint arXiv:2109.02529, 2021.

[4] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pages 2267– 2281. ACM, 2019.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR), 2013.

[6] Z. Hau, S. Demetriou, L. Muñoz-González, and E. C. Lupu, "Shadow-Catcher: Looking Into Shadows to De- tect Ghost Objects in Autonomous Vehicle 3D Sensing," arXiv preprint arXiv:2008.12008, 2020.

[7] Y. Shoukry, P. Martin, Y. Yona, S. Diggavi, and M. Srivastava, "PyCRA: Physical challenge-response authentication for active sensors under spoofing attacks," in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2015, pp. 1004–1015.