WIP: Deployability Improvement, Stealthiness User Study, and Safety Impact Assessment on Real Vehicle for Dirty Road Patch Attack

Takami Sato*, Junjie Shen*, Ningfei Wang, Yunhan Jack Jia†, Xue Lin‡, and Qi Alfred Chen
Univesity of California, Irvine; †ByteDance; ‡Northeastern University

***Abstract*—Automated Lane Centering (ALC) systems are convenient and widely deployed today, but also highly security and safety critical. Recently, Dirty Road Patch (DRP) attack is proposed as a state-of-the-art adversarial attack against ALC systems. In this work, we report our recent progress of improving the DRP attack on attack deployability, attack stealthiness, and effectiveness on real vehicle. We also discuss future directions.**

I. Introduction

Automated Lane Centering (ALC) is a Level-2 driving automation technology that automatically steers a vehicle to keep it centered in the traffic lane. Due to its high convenience for human drivers, today it is widely available on various vehicle models such as Tesla, GM Cadillac, etc. While convenient, such system is highly security and safety critical: When the ALC system starts to make wrong steering decisions, the human driver may not have enough reaction time to prevent safety hazards such as driving off road or colliding into vehicles in adjacent lanes. Thus, it is imperative and urgent to understand the security property of ALC systems.

Dirty Road Patch (DRP) attack [1] is recently proposed as a domain-specific adversarial attack to Deep Neural Network (DNN) based ALC systems. This attack identifies *dirty road patches* as an attack vector for physical-world adversarial attacks on ALC systems due to 2 unique advantages: (1) Road patches can appear to be legitimately deployed on traffic lanes in the physical world, e.g., for fixing road cracks; and (2) Since it is common for real-world roads to have dirt or white stains, using similar dirty patterns as the input permutations can allow the malicious road patch to appear more normal and thus stealthier. With the attack vector, the DRP attack systematically generates a malicious surface pattern mimicking the normal dirty patterns and achieves high attack effectiveness against a production ALC system. However, it is known that the DRP attack has major limitations on deployability, attack stealth, and effectiveness against real vehicles.

In this work, we report our recent progress to address such limitations of the DRP attack. In §II, we design a multi-piece mode that can significantly reduce the attacker's effort in deploying attacks while maintaining a high attack success rate. In §III, we conduct a user study to more rigorously evaluate the stealthiness of the DRP attack. In §IV, we assess safety impact on real vehicle by injecting attack traces into an ALC system installed in a real vehicle to simulate other driver assistance features that are commonly used with ALC at the same time in real-world driving, for example, Lane Departure Warning (LDW), Adaptive Cruise Control (ACC), Forward Collision Warning (FCW), and Automatic Emergency Braking (AEB). Finally, we discuss the future research directions in §V.

*The first two authors contributed equally.

Workshop on Automotive and Autonomous Vehicle Security (AutoSec) 2021
21 February 2021
ISBN 1-891562-68-1
https://dx.doi.org/10.14722/autosec.2021.23027
www.ndss-symposium.org

II. Attack Deployability Improvement.

To deploy the attack, the direct approach is to print the malicious dirty patterns on one single road patch and deploy it at once. However, if the required patch size is large, it may not be easy to be quickly deployed at once, which may increase the risk of being noticed by police officers or road guards. To improve this, we design an optional *multi-piece patch attack* mode, which allows the attackers to deploy the DRP attack with multiple small pieces of road patches. In this mode, the attacker can specify (1) the size of a small piece that they can quickly deploy at once, denoted as $size_p$, and (2) the total number of such pieces based on their affordable deployment efforts and risks, denoted as N_p. Fig. 1 shows an example multi-piece patch attack from the driver's view, which includes 8 pieces of quickly-deployable small road patches. With this, the attacker can deploy one small piece at a time to avoid drawing too much attention, and can also parallelize the deployment of different pieces to further accelerate the process.

Evaluation methodology and setup. First, we consider a road patch of size 1.6-1.8 m wide and 6-8 m long as a *deployment unit* for 1-2 people at a time, based on normal arm span and available adhesive road patch lengths online.Different from a legitimate road patch deployment, the attacker does not need to fill asphalt and stamp the road, which are the most time-consuming steps. Instead, she only needs to place the patch on the road surface, and this step only takes 5-10 seconds for the defined deployment unit above based on demo videos of the adhesive patch deployment. With this, we can then estimate the deployment effort for 1-2 people by calculating the number of deployment units for a given attack road patch. We start from an attack patch size of 5.4m×36m (width×length), the size used in previous sections, and then vary the allowed number of deployment units in the patch area using multi-piece patch attack mode to evaluate the trade-off between deployability and attack effectiveness.

Results. Table I shows the experiment results with 4 to

TABLE I: Evaluation results for attack deployability using multi-piece patch attack mode. Piece #: Allowed number of 1.8m×7.2m patches for the attack; 15 pieces cover the entire deployable area.

Piece #	Succ. rate	Succ. time (s)	Deploy. time improvement
15 (full)	100%	0.89	-
12	98.8%	0.94	-20%
8	93.8%	1.28	-47%
6	76.3%	1.63	-60%

Fig. 1: Driver's view in multi-piece patch attack mode (§II) at 2.5 seconds before the attack succeeds. The attack in the figure has 8 pieces of 1.8m×7.2m road patches, each requiring only 5-10 sec to deploy for 1-2 people.

15 deployment units. To fully cover the 5.4m wide and 36m long patch placement area, 15 deployment units, each for a piece of 1.8m×7.2m patch, are required for 1-2 people. When the allowed piece number decreases, the attack success rate decreases accordingly since the perturbable area size becomes smaller. Interestingly, the success rate is still as high as 93.8% when only 8 pieces are allowed, which is able to significantly reduce the original deployment efforts by nearly 50%. This suggests that it is not necessary to cover the entire attack deployable area to achieve a high success rate for the DRP attack, which thus concretely shows the benefit of our multi-piece patch attack mode design in improving deployability. Fig. 1 shows an example DRP attack with such 8 pieces of 1.8m×7.2m road patches from the driver's view. To deploy 8 such pieces, the attacker only needs to find opportunities to block the road for 1-2 min in total. To maximize the stealthiness, such an effort can be spread up to 8 different times, each for a single piece or more. Each time, a single piece only requires a 5-10 seconds vacant period of the target road, which is common in the U.S. since the majority of the driving scenarios in the U.S. are *free-flow* scenarios, where a vehicle has at least 5-9 seconds headway. In the late night, such opportunities can be even more frequent.

III. Attack Stealthiness User Study

We conduct a user study to more directly evaluate the stealthiness of the DRP attack. We have gone through the IRB process and our study is determined as in the IRB Exempt category since it does not involve the collection of any Personally Identifiable Information (PII) or target any sensitive population.

Evaluation methodology. We use the generated attacks on real-world driving traces to perform the user study. For an attack scenario, we ask the participants to imagine that they are driving with the ALC system taking control, and then show a sequence of image frames with the malicious road patch from the driver's view at 3, 2.5, 2, 1.5, and 1 second(s) before the attack succeeds. Here, 1 second before the attack succeeds is right before the attack starts to take effect. At this point, the patch is placed at 7 m away from the vehicle. For each image frame, we ask whether they will decide to take over the driving to avoid danger or potential safety risks. These questions are also asked for the image frames with a benign road patch that only has the base color without the malicious dirty patterns as a control group.

Since our attack is designed for drivers who are in favor of using ALC system in normal cases, the same set of questions are asked at the beginning for the original image frames without attack, and we only accept a participant if she does not choose to take over the driving for these cases. This process also helps filter out ill-behaved participants who just provide random answers. Since DRP is a new form of attack vectors on the road, we do not tell the participants that the study is related to security attacks. Instead, we only tell them that our focus is on surveying driver's decisions under ALC systems for different road surface patterns such as road patches and scratches. At the beginning of the study, we also provide an introduction of ALC systems with demo videos to ensure that the participants fully understand what driving technology we are surveying about. To understand the distribution of the participant background, we also ask demographic information and background information related to driving and ALC usage. None of the questions in our study involve PII or target any sensitive population; our study is thus determined as in the IRB Exempt category.

Evaluation setup. We use Amazon Mechanical Turk to perform this study, and in total collected 100 participants. All of them have driving experience, which is confirmed by asking them the age when first licensed and the weekly driving mileage. A local-road driving trace is used in this study, and for the scenarios with attack, we evaluate 3 stealthiness levels (i.e., $\lambda = 10^{-2}, 10^{-3}, 10^{-4}$ as defined in [1]). The survey is available at [2]. Among the 100 participants, 56% are male and 44% are female. The average age is 32.3 years old. 79% of them have experienced at least one ALC system, among which Tesla Autopilot has the largest share (28%). Statistics of ALC experiment and demographic information are shown in Fig. 3.

Results. Fig. 2 shows the study results. As shown, the closer it is to the attack success time, the more participants choose to take over the driving in the attacked scenarios since the dirty patterns become increasingly larger and clearer. Among the 3 stealthiness levels, the driver decisions are consistent with our design: the lowest stealthiness level ($\lambda = 10^{-4}$) has the highest take-over rate, while the highest level ($\lambda = 10^{-2}$) has the lowest. In particular, we find that even for the lowest stealthiness level ($\lambda = 10^{-4}$), only *less than 25%* of the participants decide to take over before the attack starts to take effect. At this stealthiness level the white dirty patterns are quite dense and prominent. Thus, these results suggest that the majority of human drivers today do not treat dirty road patches as road conditions where ALC systems cannot handle.

As discussed in [3], 2.5 seconds is commonly used as the average driver reaction time to road hazard. Thus, at 2.5

seconds or more before the attack succeeds, the human driver still has a chance to take over the driving to prevent the damage in common cases, as long as she can realize that it is a road hazard. However, our results show that only less than 20% of the participants decide to take over at 2.5 and 3 seconds before our attack succeeds even for the lowest stealthiness level. In particular, when the stealthiness levels are $\lambda = 10^{-2}$ and $\lambda = 10^{-3}$, the take-over rates at these 2 time points are similar to the rates for the benign road patch with only the base color. This suggests that at the time when there is still a chance to prevent the damage in common cases, our attack patches at $\lambda = 10^{-2}$ and 10^{-3} *appear to be as innocent as normal clean road patches to human drivers*. In these cases, the take-over rates are only *less than 15%*, which are from participants who will take over even for normal clean road patches. Note that the take-over rates in practice are likely to be lower than this since (1) this study is performed for a local-road scenario, while the road patches in highway scenarios are much farther and thus much less noticeable as shown in Fig. 7 in [1], and (2) the road patches in this study are digitally synthesized into the image frames, which may appear less natural and thus may more easily alert the participants.

Discussion. While we try our best to evaluate the stealthiness of the DRP attack, our user study has 3 potential limitations: (1) The participants may not represent the normal drivers. Unfortunately, we cannot fully address this limitation as long as we use cloud-sourcing. To the best of our ability, we use the responses for normal cases to filter out people who are not in favor of using ALC systems and those who answer random responses. However, this filtering may introduce a bias to the participants. While we intentionally target people who are supportive to use ALC systems, the results of the user study may change in the future along with the spread of ALC systems. (2) Our patch may increase the attention level of drivers from a distance due to its noisy appearance of our patch, and it could reduce the reaction time to less than 2.5 sec. We agree that such a case may occur, but we consider that it only has a minor impact. The driver's reaction time (2.5 sec) consists of 1.75 sec of perception time and 0.75 sec of physical reaction time [3]. In the target scenario, our attack starts to take effect at 1 second before the attack succeeds. In this case, the driver needs to react within 0.25 sec to avoid the DRP attack, which requires a significant reduction from 1.75 sec perception time since the attention level can only improve the perception time. (3) The multi-piece patch mode may draw more driver's attention. In the user study, we only evaluate the stealthiness of the single-piece patch. If the multi-piece patch mode harms the stealthiness, we need to find the best balance between the stealthiness and deployability.

IV. Safety Impact on Real Vehicle

While the simulation-based evaluation is able to show severe safety impacts of our attack, it does not simulate other driver assistance features that are commonly used with ALC at the same time in real-world driving, for example Lane Departure Warning (LDW), Adaptive Cruise Control (ACC), Forward Collision Warning (FCW), and Automatic Emergency Braking (AEB). This makes it unclear whether the safety damages discussed in [1] are still possible when these features are used, especially the safety-protection ones such as AEB. In this section, we thus use a real vehicle to more directly understand the safety impact.

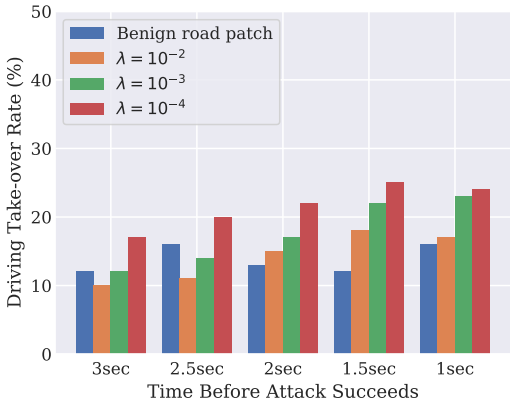

Fig. 2: Results of the attack stealthiness user study. Driving take-over rate is the percentage of participants who choose to take over the driving at a particular time point before the attack succeeds.

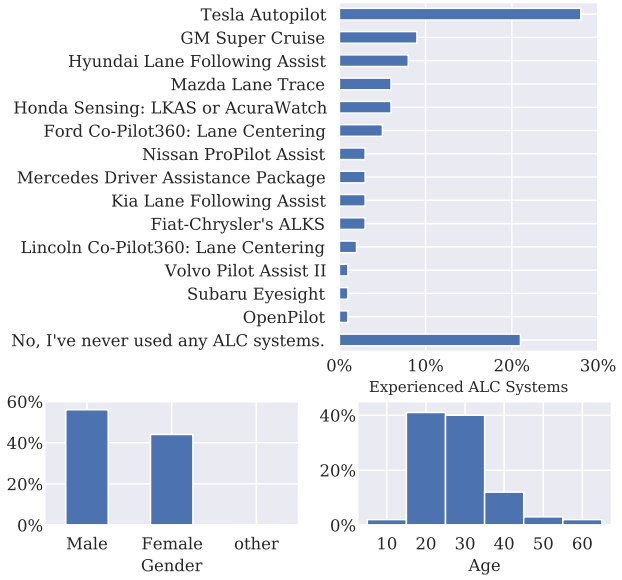

Fig. 3: Statistics of the ALC system experience and demographic information in the attack stealthiness user study.

Evaluation methodology. We install OpenPilot on a Toyota 2019 Camry, in which case OpenPilot provides ALC, LDW, and ACC, and the Camry's stock features provide AEB and FCW [4]. We then use this real-world driving setup to perform experiments on a rarely-used dead-end road, which has a double-yellow line in the middle and can only be used for U-turn. The driver's view of this road is shown on the left of Fig 4. In our miniature-scale experiment in [1], the attack realizability from the physically-printed patch to the lane detection (LD) model output has already been validated under 12 different lighting conditions. Thus, in this experiment we evaluate the safety impact by directly injecting an attack trace at the LD model output level. This can also avoid blocking the road for sticking printed patches to the ground and cleaning them up, which can easily affect other vehicles. Note that attack trace injection is also a common practice in physical-world attack evaluation for cyber-physical systems.

To create safety-critical driving scenarios, we place cardboard boxes adjacent to but outside of the current lane as shown in Fig 4, which can mimic road barriers and obstacles

Fig. 4: Safety impact evaluation for our attack on a Toyota 2019 Camry with OpenPilot engaged. Even with other driver assistance features such as Automatic Emergency Braking (AEB), our attack still causes collisions in all the 10 trials.

in opposite direction while not causing damages to the vehicle and driver safety. Similar setup is also used in today's vehicle crash tests [5], [6]. While the cardboard boxes are smaller than the object used in [5], [6], our setup has enough potential to damage the vehicle, e.g., the attacker can put heavy weights in the boxes. To ensure that we do not affect other vehicles, we place the cardboard boxes only when the entry point of this dead-end road has no other driving vehicles in sight, and quickly remove them right after our vehicle passes them as required by the road code of conduct.

Experiment setup. We perform experiments in day time with and without attack, each 10 times. The driving speed is kept at ∼28 mph, the minimum speed for engaging OpenPilot on our Camry. The injected attack trace is generated from our simulation environment at the same driving speed.

Results. Our experiment results show that our attack causes the vehicle to hit the cardboard boxes in all the 10 attack trials (100% collision rate), including 5 front and 5 side collisions. The collision variations are caused by randomness in the dynamic vehicle control and the timing differences in OpenPilot engaging and attack launching. In contrast, in the trials without attack, OpenPilot can always drive correctly and does not hit or even touch the objects in any of the 10 trials.

These results thus show that driver assistance features such as LDW, ACC, FCW, and AEB are not able to effectively prevent the safety damages caused by our attack on ALC. We examine the attack process and find that LDW is not triggered since it relies on the same lane detection module as ALC and thus are affected simultaneously by our attack. ACC does not take any action since it does not detect a front vehicle to follow and adjust speed in these experiments. FCW is triggered 5 times out of the 10 collisions, but it is only a warning and thus cannot prevent the collision by itself. Moreover, in our experiments FCW is triggered only *0.46 sec* before the collision on average, which is far too short to allow human drivers to react considering the 2.5-second average driver reaction time to road hazard.

In our Camry model, FCW and AEB are turned on together as a bundled safety feature.However, while we have observed some triggering of FCW, we were not able to observe any triggering of AEB among the 10 attack trials, leading to a *100% false negative rate*. We check the vehicle manual and find that this may be because the AEB feature (called *pre-collision braking* for Toyota) is used very conservatively: it is triggered only when the possibility of a collision is *extremely high*. This observation is also consistent with the previously-reported high failure rate (60%) for AEB features on popular car models today [7]. Such conservative use of AEB can reduce false alarms and thus avoid mistaken sudden emergency brakes in normal driving, but also makes it difficult to effectively preventing the safety damages caused by our attack — in our experiments, it was not able to prevent any of the 10 collisions. The video recordings for these real-vehicle experiments are available at **https://youtu.be/OT8seN6pZj4** and **https://youtu.be/Ph42FIaadFo**.

V. Conclusion and Future Directions

In this paper, we address the limitations of the DRP attack from 3 improvements and further evaluation: the multi-piece patch mode, conduct attack stealthiness user study, and safety impact on real vehicle. The multi-piece patch mode can improve the attack deployability significantly. The attack with 8 pieces can reduce the original deployment time by nearly 50%. The stealthiness user study reveals that the DRP attack can appear to be as innocent as normal clean road patches to human drivers and the take-over rates by human drivers can be only less than 15%. The safety impact on real vehicle demonstrates that that driver assistance features such as LDW, ACC, FCW, and AEB are not able to effectively prevent the safety damages caused by the DRP attack on ALC. These results indicate that DNN-based ALC systems have security risks against adversarial attacks in the real world.

To improve the robustness of ALC systems, we plan to explore 2 domain-specific defense strategies in future directions: (1) We will design an attack detection and warning system, which detects cracks or dirty patterns on the road and warn the drivers if such patterns can mislead ALC systems. The challenge in this defense is that the system needs to warn the driver before the driver's reaction time (2.5 sec). (2) We think leveraging HD maps could be a feasible solution for Level-2 AD systems as Level-4 AD systems today heavily utilize HD maps [8]. If such a map can be available, a follow-up research question is how to effectively detect our attack without raising too many false alarms, since mismatched lane information can also occur in benign cases We will conduct a systematic exploration of these directions in future work.

Acknowledgements

This research was supported in part by the National Science Foundation under grants CNS-1850533, CNS-1929771, CNS-1932464, and USDOT UTC Grant 69A3552047138.

References

[1] 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:2009.06701*, 2020.

[2] "Driver Take-Over Decision Survey with Automated Lane Centering System in our Attack Stealthiness User Study," https://storage.googleapis.com/driving-decision-survey/driving_decision_survey.pdf, 2020.

[3] S. of California Department of Motor Vehicles, *California Commercial Driver Handbook: Section 2 – Driving Safely*, 2019. [Online]. Available: https://www.dmv.ca.gov/portal/uploads/2020/06/comlhdbk.pdf

[4] "OpenPilot," https://github.com/commaai/openpilot, 2018.

[5] "Toyota Safety Sense Pre-Collision System (PCS) Settings and Controls," https://youtu.be/IY4g_zG1Qj0, 2017.

[6] C. Miller and C. Valasek, "A Survey of Remote Automotive Attack Surfaces," https://youtu.be/tnYO4U0h_wY?t=1840, 2015.

[7] "Does Your Car Have Automated Emergency Braking? It's a Big Fail for Pedestrians," https://zd.net/2MoUqpd, 2019.

[8] "Building Maps for a Self-Driving Car," https://link.medium.com/Bo5pCOov95, 2016.