Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack

Takami Sato*, Junjie Shen*, Ningfei Wang, Yunhan Jack Jia, Xue Lin, and Qi Alfred Chen

* co-first authors
Automated Lane Centering (ALC) systems

- **Level-2 driving automation** technology that automatically steers a vehicle to *keep it centered in the traffic lane (lateral control)*
Target of our study: OpenPilot

- Open-sourced production ALC with representative design: DNN-based camera lane detection
- Close performance to Tesla AutoPilot and GM Super Cruise*

*https://www.caranddriver.com/features/a30341053/self-driving-technology-comparison/

Driver can hand off steering wheel

OpenPilot dashcam device
- Detect Lane by camera
- Override cruise mode
- Control vehicle via OBD-II
Target of our study: OpenPilot

- Open-sourced production ALC with representative design: DNN-based camera lane detection
- Close performance to Tesla AutoPilot and GM Super Cruise*

*https://www.caranddriver.com/features/a30341053/self-driving-technology-comparison/
Is DNN model in ALC secure?

Widely reported to be vulnerable to physical-world adversarial attacks

Can DNN-level vuln lead to whole ALC system-level attack effect?
Our study

First to systematically study security of DNN-based ALC in designed operational domains (i.e., road w/ lane lines) under physical-world adversarial attacks.
Our study

First to systematically study security of DNN-based ALC in designed operational domains (i.e., road w/ lane lines) under physical-world adversarial attacks

Challenge 1: Lack of domain-specific & deployable attack vector
Our study

First to systematically study security of DNN-based ALC in designed operational domains (i.e., road w/ lane lines) under physical-world adversarial attacks

Challenge 1: Lack of domain-specific & deployable attack vector

Road-side objects do not include ROI (input).
Our study

First to systematically study security of DNN-based ALC in designed operational domains (i.e., road w/ lane lines) under physical-world adversarial attacks

Challenge 1: Lack of domain-specific & deployable attack vector

Road-side objects do not include ROI (input).
Our study

First to systematically study security of DNN-based ALC in designed operational domains (i.e., road w/ lane lines) under physical-world adversarial attacks

Challenge 1: Lack of domain-specific & deployable attack vector

Road-side objects do not include ROI (input).

Challenge 2: Camera frame content inter-dependency due to attack
Our study

First to systematically study security of DNN-based ALC in designed operational domains (i.e., road w/ lane lines) under physical-world adversarial attacks

Challenge 1: Lack of domain-specific & deployable attack vector

Challenge 2: Camera frame content inter-dependency due to attack

Challenge 3: Lack of differentiable attack objective function design for ALC
Challenges

• **Lack of domain-specific & deployable attack vector**
  - How to handle semantic gap from perturbations in physical-world driving environment to those in model inputs?

• **Camera frame content inter-dependency due to attack**
  - Successful attack on a single frame can only cause \(<0.3 \text{ mm}\) at 45 mph.
  - How can such attack be continuously effective on sequential camera frames?

• **Lack of differentiable attack objective func design for ALC**
  - How to change the shape of detected lane lines?
    - Existing ones concentrate on changing object classes or bounding boxes
    - Popular lateral control (e.g., MPC) is not differentiable
Challenge 1: Lack of domain-specific & deployable attack vector

What on the road surface can be both seemingly benign & possible for attack?
Challenge 1: Lack of domain-specific & deployable attack vector

Can dirty road patterns attack ALC?
Novel attack vector: Dirty Road Patch (DRP)

- DRP attack pretends to be **benign road patch** but the surface patterns are designed for **adversarial attack**
- Attacker can print malicious perturbation on patch and quickly deploy it

- Grayscale perturbation
- Brightness limits
- Preserving original lane line information
- Perturbation area restriction

http://www.americanroadpatch.com/
Attack demos
Attack demo 1: Miniature-scale physical-world setup

Studio lights

Road texture

Official OpenPilot dashcam device
Attack Demo 2

Software-in-the-Loop Simulation with LGSVL

Target ALC: OpenPilot v0.6.6
Scenario: Local Road at 45 mph (72 km/h)
Attack demo 3: Safety impact on real vehicle

- We inject attack trace into real-world driving to see if other driving assistance features (e.g., AEB) can prevent crash.

Replace model output with the one obtained in the simulator.
Pre-collision alert starts 0.74 sec before the crash
*Alert Only.* Pre-collision braking is enabled but not applied.
Challenges

- **Lack of domain-specific & deployable attack vector**
  - How to handle semantic gap from perturbations in physical-world driving environment to those in model inputs?

- **Camera frame content inter-dependency due to attack**
  - *Successful attack on a single frame can only cause <0.3 mm at 45 mph.*
  - How can such attack be continuously effective on sequential camera frames?

- **Lack of differentiable attack objective func design for ALC**
  - How to change the *shape* of detected lane lines?
    - Existing ones concentrate on changing object classes or bounding boxes
    - Popular lateral control (e.g., MPC) is not differentiable
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are **dynamically changed due to attack**
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are *dynamically changed due to attack*

| When collecting frames |
|------------------------|
| ![Frame 1](image1)    |
| ![Frame 2](image2)    |
| ![Frame 3](image3)    |
| ![Frame 4](image4)    |

| When attacking        |
|-----------------------|
| ![Frame 5](image5)    |
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are **dynamically changed due to attack**
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are **dynamically changed due to attack**

*How to obtain attack-influenced camera frame contents?*
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are **dynamically changed due to attack**

When collecting frames

When attacking
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are **dynamically changed due to attack**

When collecting frames

When attacking

Model input only uses part of camera frame (ROI)
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are **dynamically changed due to attack**
Challenge 2: Camera frame content inter-dependency due to attack

- Challenge: Frame contents are **dynamically changed due to attack**
Motion model-based input generation

- Calculate attack-influenced vehicle positions & heading with **vehicle motion model**

- Obtain attack-influenced steering angle under attack.
Motion model-based input generation

- Calculate attack-influenced vehicle positions & heading with **vehicle motion model**
Motion model-based input generation

- Calculate attack-influenced vehicle positions & heading with vehicle motion model.
- Use **perspective transformation** to dynamically synthesize the content inside ROI based on position changes.
Motion model-based input generation

- Calculate attack-influenced vehicle positions & heading with **vehicle motion model**
- Use **perspective transformation** to dynamically synthesize the content inside ROI based on position changes
- ≥46% better than possible alternative methods such as **single-frame EoT**
- Also make it possible to judge attack success directly at **lateral deviation** level during optimization
Challenges

• Lack of domain-specific & deployable attack vector
  - How to handle semantic gap from perturbations in physical-world driving environment to those in model inputs?

• Camera frame content inter-dependency due to attack
  - Successful attack on a single frame can only cause $<0.3 \text{ mm}$ at 45 mph.
  - How can such attack be continuously effective on sequential camera frames?

• Lack of differentiable attack objective func design for ALC
  - How to change the shape of detected lane lines?
    - Existing ones concentrate on changing object classes or bounding boxes
  - Popular lateral control (e.g., MPC) is not differentiable
Challenge 3: Lack of differentiable attack objective func design for ALC

**Key idea:** maximize/minimize the derivative at each waypoint

- Can be a **differentiable surrogate to steering angle** at lateral control design level
- Named “*lane-bending objective function*”
DRP attack generation framework

- Alternatively update patch and vehicle trajectory
  - Update patch with gradient information of current frames
  - Update vehicle trajectory with current frames

Dirty Road Patch (DRP)
- Grayscale Perturbation
- Preserve Lane Line
- Brightness Limit
- Perturbable Area

Lane-bending Objective Function

Motion model-based input generation
Evaluations

• **Real-world driving trace-based evaluation**
  - ≥97.5% attack success rate w/ < 0.903 sec avg success time (avg driver reaction time is 2.5 sec)

• **Physical-world miniature-scale evaluation**
  - >20° steering angle under all 12 lighting conditions & 45 different viewing angles

• **Software-in-the-loop simulation**
  - 100% success rates from all 18 starting positions

• **Comparison with baseline attacks**
  - ≥46% better than possible alternative methods such as single-frame EoT

• **Attack stealthiness user study**
  - 100 human subjects on Amazon MTurk (IRB exempt)
  - As innocent as the benign road patch at 2.5 sec before attack succeeds

More evaluations in the paper...
Defense evaluation & discussion

• DNN model level defenses
  • Evaluated 5 popular defense methods that are directly applicable (e.g., Bit-depth reduction)
  • None of them can defend against our attack without harming normal driving
    • E.g., Bit-Depth reduction can defend 46% attacks but cannot handle 10% benign driving

• Sensor/data fusion-based defenses
  • Fusion with High Definition (HD) map
    • Create & maintain it is time-consuming, costly, & hard to scale
    • Tesla explicitly claims that it is a “non-scalable approach”*
  • Maybe necessary for security purposes

• Short-term mitigation: At least put dirty road & dirty road patches into the list of unhandled scenarios so users can be aware
  • Checked ALC manuals from 11 companies (e.g., Tesla, GM Cruise, OpenPilot, Honda Sensing, and Toyota LTA) but none of them list them today

*https://electrek.co/2020/06/18/tesla-approach-self-driving-harder-only-way-to-scale/
Responsible vulnerability disclosure

- As of 7/7/21, informed 13 companies developing ALC systems
  - 10 companies (77%) have replied and have started investigation
  - Some companies already had meetings with us to facilitate such investigations
Conclusion

First to systematically study security of DNN-based ALC in designed operational domains under physical-world adversarial attacks

• Adopt an optimization-based approach with 2 novel designs: motion model-based input generation and lane-bending objective function
• Evaluate our attack on a production ALC with real-world driving traces, physical-world miniature-scale setup, a production-grade simulator, and also stealthiness, deployability, and robustness to different viewing angles & lighting conditions
• Evaluate safety impact on real vehicle by injecting attack traces
• Evaluate 5 DNN model-level defenses, discuss sensor/data fusion-based defenses, propose short-term mitigation suggestions
• Informed 13 companies developing ALC systems
Thank you!

For demos, data/source code, FAQ & other details, Please visit our project website:
https://sites.google.com/view/cav-sec/drp-attack