

Too Good to be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations

Jing, Pengfei, et al. (2021)

Background and Motivation

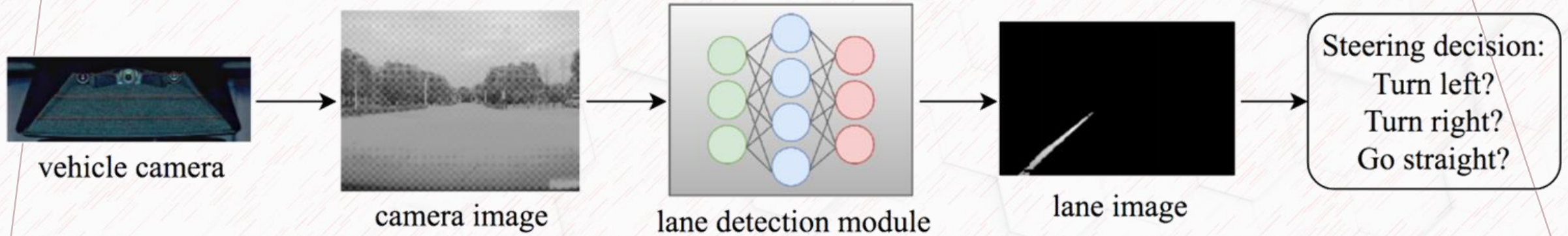
What's the importance of this study?

Autonomous driving is safety-critical!



- Camera technology together with deep learning algorithms are used to identify lane markings.
- Lane markings are important for autonomous vehicles to navigate safely and reliably on the highway.
- The over-sensitivity of the lane detection module in AVs is a weakness in autonomous driving.

How does lane detection work?



1

Images are collected by a camera.

2

Based on the camera images, lane detection module generates the corresponding lanes

3

Autonomous vehicle behaves based on the lane detection result

Are these questions answered in the current systems?



How does the AV
behave on a
snow-covered
road?

Are these questions answered in the current systems?

How about faded lanes?



Even on a sunny day, the white lines are almost invisible between the lanes on this busy, curving section of East North Street in Rapid City. City officials say they repaint the lines each year, but tire wear, weather and plows wear the markings down each winter. Photo: Bart Pfankuch

Are these questions answered in the current systems?




What of
construction
errors?

Are these questions answered in the current systems?

Bear in mind, all these are nature/human induced.

What then happens when an attacker who understands the system intentionally designed crafty situations to mislead a victim's vehicle?



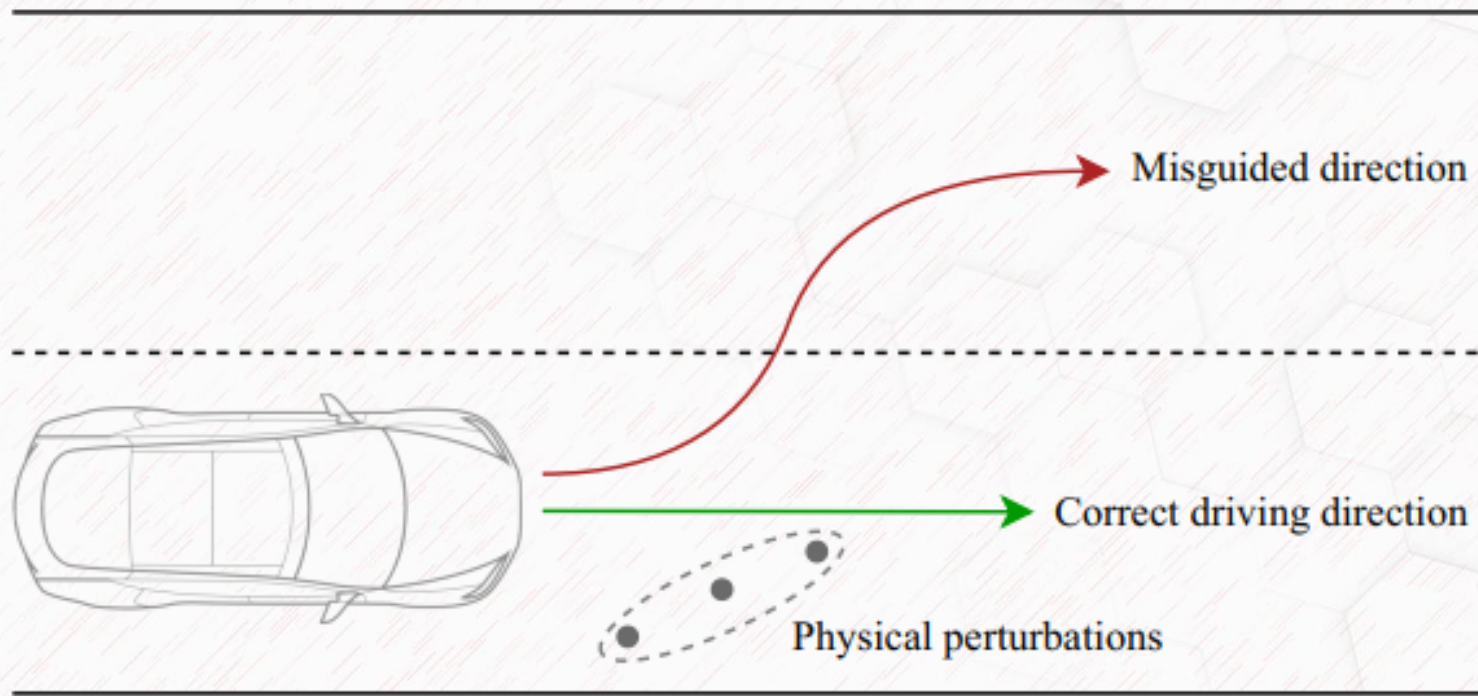
Clearly visible road markings support the human driver and the machine in navigating through the roadway



Failsafe detection of lane markings is critical to guarantee safety in autonomous driving.

Attack Model

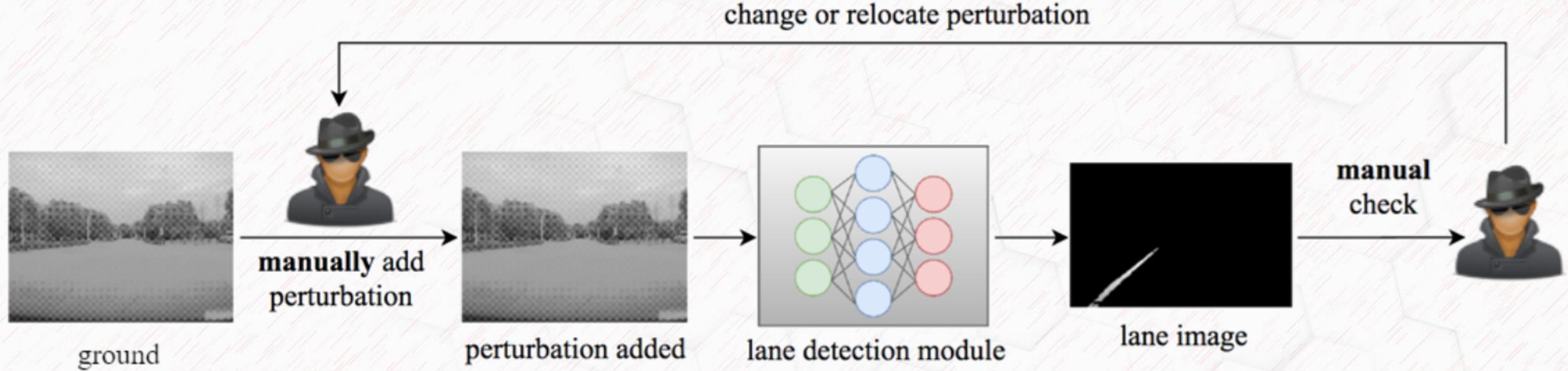
What is the attacker's goal?



Tricking the autonomous vehicle to steer into the reverse traffic lane

If the physical perturbations added by an adversary are recognized as a lane, the vehicle is likely to follow the fake lane and swerve into the wrong direction

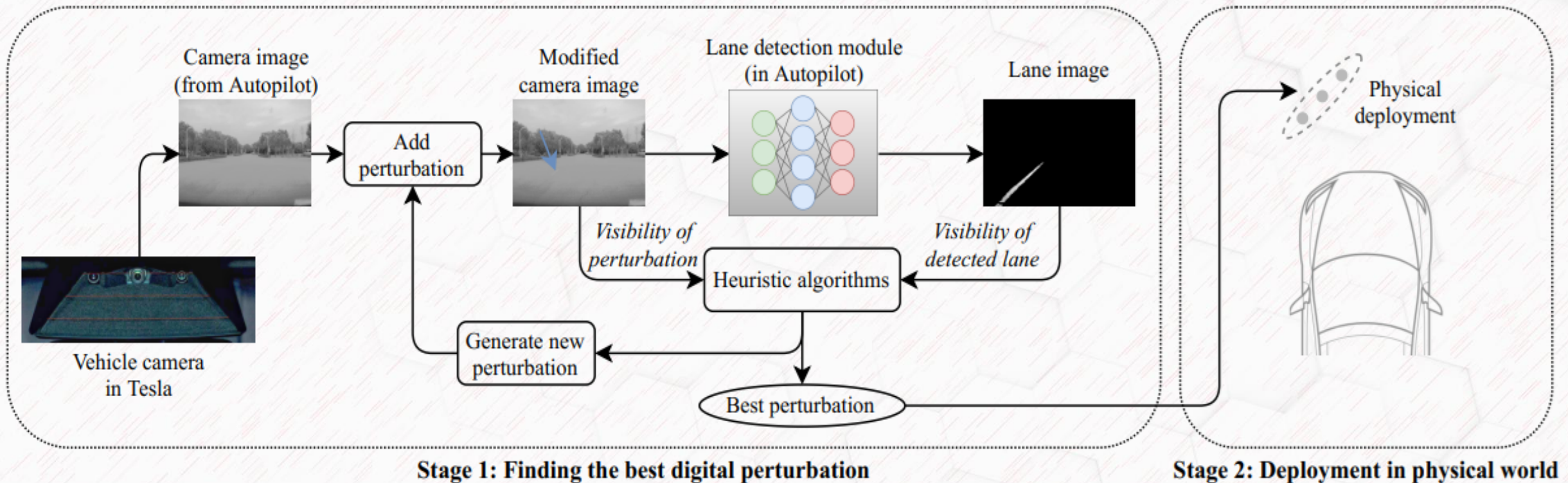
Creating a fake lane – An intuitive approach



📌 Add perturbations and check whether the module will be affected **manually**. If not, the perturbation should be changed or relocated.

📌 Unfortunately, such an approach is very **labor-intensive** and **error-prone**

Creating a fake lane – The proposed approach



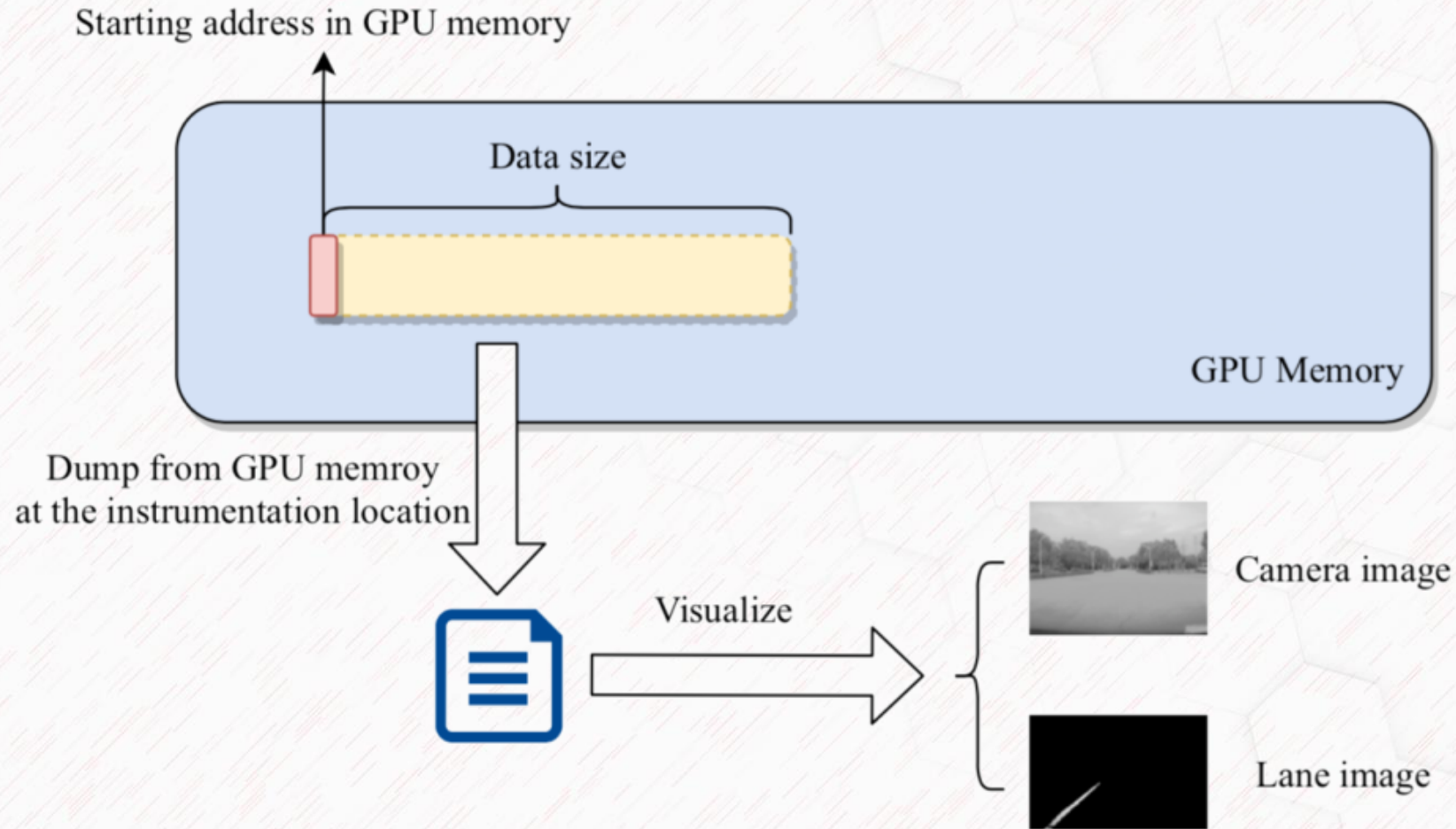
1 Add the perturbation to the **camera image** to trigger the lane detection module to generate the corresponding **lane image**.

2 Formulate an optimization problem based on the **visibility** of perturbation and that of detected lane and adopt **heuristic algorithms** to find the best perturbation.

3 Deploy the best perturbation in physical world for evaluation.

Attack Methodology

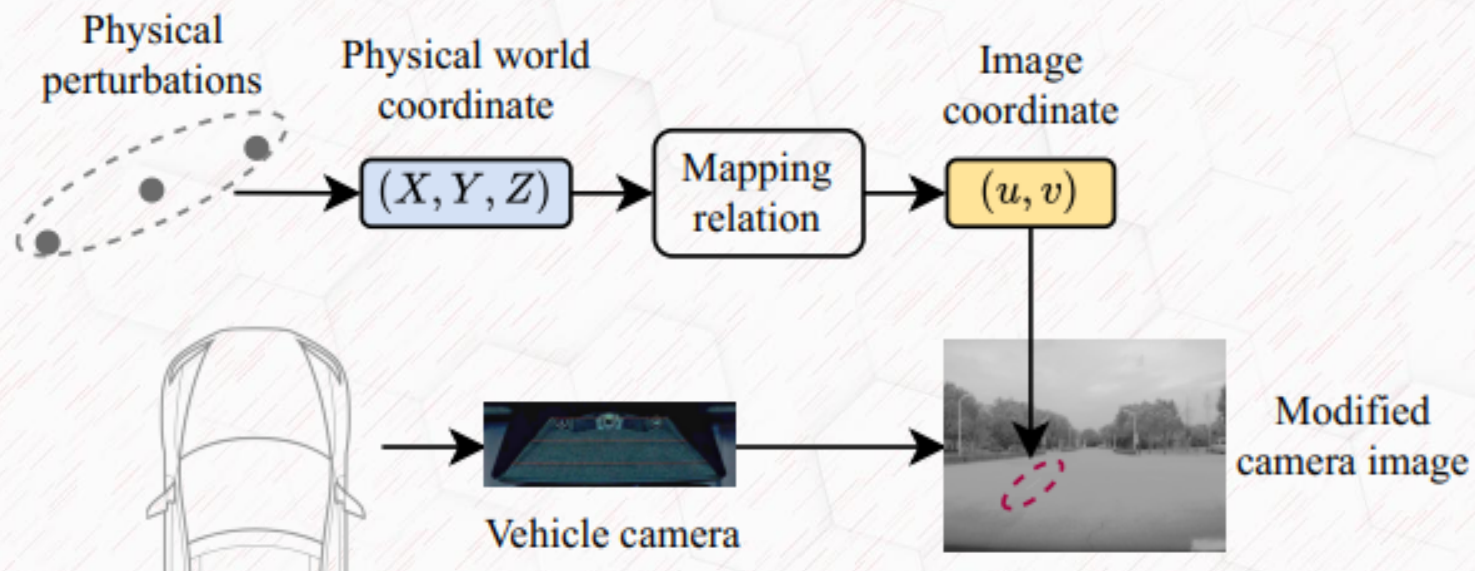
How to extract the data from the real vehicle, which is not exposed to users?



Conduct static and dynamic analysis on the firmware responsible for lane detection to collect the data (camera image and lane image) from the vehicle

Dumping and visualizing the target data from the GPU on Autopilot

How to add perturbations to input camera image?

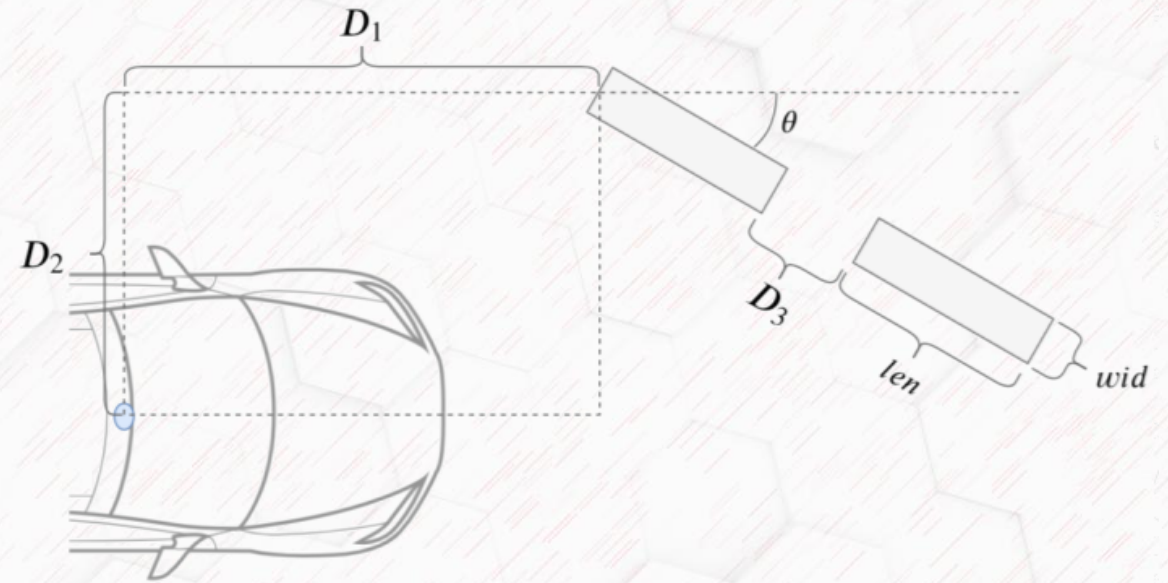


Mapping the coordinate of $(X;Y;Z)$ on markings in physical world to the coordinate of $(u;v)$ on perturbations in digital world

How to add perturbations to input camera image?

Parameters	Explanation
len	Length of a single perturbation
wid	Width of a single perturbation
D_1	Longitudinal distance from the vehicle camera to the edge of the first perturbation
D_2	Lateral distance from the vehicle camera to the edge of the first perturbation
D_3	Distance between adjacent perturbations
ΔG	Increment of grayscale value of the perturbed pixels
θ	Rotation angle of the perturbation
n	Number of the perturbations

Parameters determining the added perturbation



$$x = (len, wid, D_1, D_2, D_3, \Delta G, \theta, n) \in X$$

Illustration of the parameters

For the ease of deployment, 8 parameters, which form a vector x were used to represent the attributes of the perturbations. With pinhole camera model and undistortion techniques, these perturbations can be accurately mapped to digital images

How to find the best perturbations?

$$V_{lane}(x) = \sum_{p \in lane_o(x)} G_p$$

$$V_{perturb}(x) = \sum_{p \in perturb_i(x)} \Delta G$$

$$S(x) = \frac{V_{lane}(x)}{V_{perturb}(x)}$$

Parameters	Explanation
p	One single pixel in the image
$lane_o(x)$	Lane pixels in the output image
$perturb_i(x)$	Pixels on the added perturbations
G_p	Grayscale value of pixel p
$V_{lane}(x)$	Visibility of the fake lane created by x
$V_{perturb}(x)$	Visibility of the perturbations added by x
$S(x)$	Overall score of the parameter x

Explanations of parameters

$S(x)$ represents the overall score, based on which we use heuristic algorithms to find the perturbation with the highest score:

$$x^* = \max_{x \in X} S(x)$$

Formulate an optimization problem to find the best perturbations. Specifically, quantify the quality by the visibility of lane and visibility of perturbation. The visibility of lane should be high (to make the attack effective), and the visibility of perturbation should be low (to make the perturbation unobtrusive)

Evaluation/ Results

Does the research answer the following important questions?

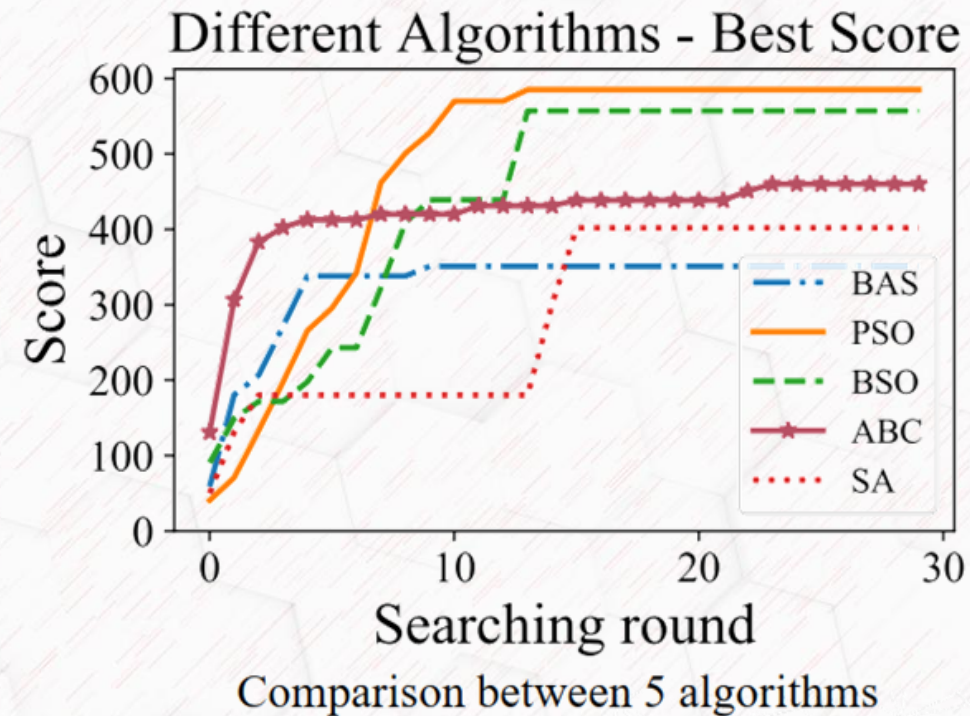
- 1** How efficient are the heuristic algorithms to find the best perturbations?
- 2** Can the vehicle be misguided in physical world?
- 3** How do the perturbation number n and the rotation angle θ affect the best perturbation?
- 4** How is the performance of the attack given different input camera images?
- 5** What are the common characteristics of the best perturbations?
- 6** How effective is the attack in the physical world?

How efficient are the heuristic algorithms to find the best perturbation?

Approach: We use 5 heuristic algorithms to find the best perturbations:

- *Beetle Antennae Search (BAS)*
- *Particle Swarm Optimization (PSO)*
- *Beetle Swarm Optimization (BSO)*
- *Artificial Bee Colony (ABC)*
- *Simulated Annealing (SA)*

Answer: *PSO* is the most efficient one and thus we use it in other experiments.



How efficient are the heuristic algorithms to find the best perturbation?

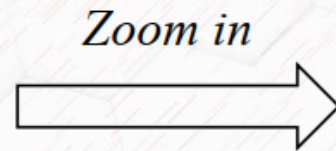
A best perturbation is shown below. The added perturbation is only 1cm wide in physical world, but it causes the lane detection module to generate a fake lane.



Original camera image



Normal output (no lane)



Modified camera image



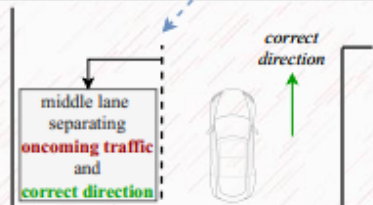
Fake lane detected



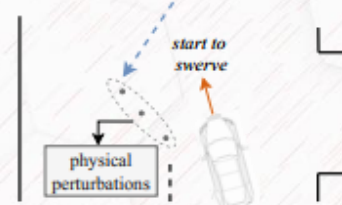
The perturbation can be hardly noticed

Effect of a best perturbation

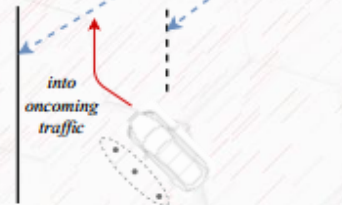
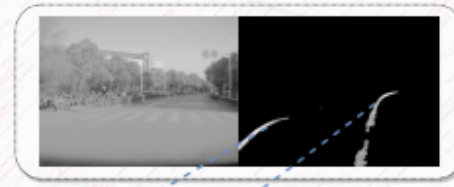
Can the vehicle be misguided in the physical world?



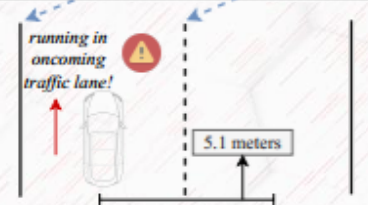
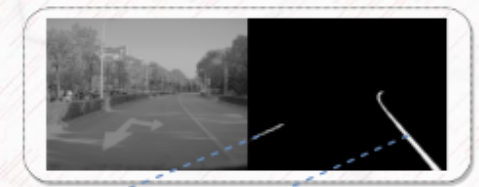
(a) Vehicle is running on the correct direction.



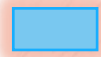
(b) Fake lane is detected and vehicle starts to swerve.



(c) Vehicle follows the fake lane into oncoming traffic.



(d) Vehicle finally runs in the oncoming traffic lane!



Right before the vehicle runs into the crossroads, the perturbations are detected and recognized as the fake lane, and therefore the vehicle starts to swerve along with the detected lane.



The vehicle is in auto-steer mode, and its average speed is above 40km/hr which is already very dangerous in the real world.

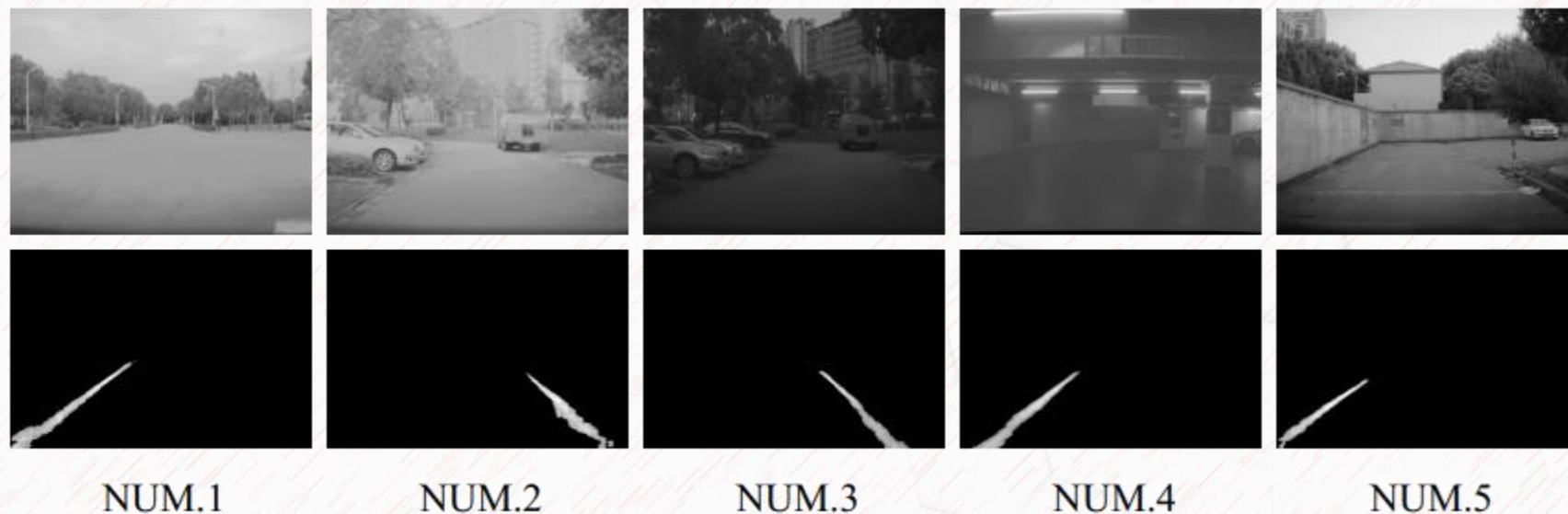
Answer: Fake lane resulting from the unobtrusive perturbations can successfully fool the vehicle in auto-steer mode to swerve, and even misguide the vehicle into oncoming traffic (might hit other cars in the oncoming traffic lane), thus demonstrating the potential severe threats in real world

How is the performance of the attack given different input images?

Num	Environmental Features
1	Clean and bright ground, without other disturbing objects in view
2	Clean and bright ground, with disturbing objects in view
3	Clean and dark ground, with disturbing objects in view
4	Clean and dark ground, without other disturbing objects in view
5	Dirty and bright ground, with disturbing objects in view

Answer: Given different input images, their approach can successfully generate high-score perturbations that can mislead the lane detection module without being noticed by the driver.

Table 3: Environmental features of different input images



(a) Perturbations in different input images and the corresponding outputs

How effective is the attack in the physical world?

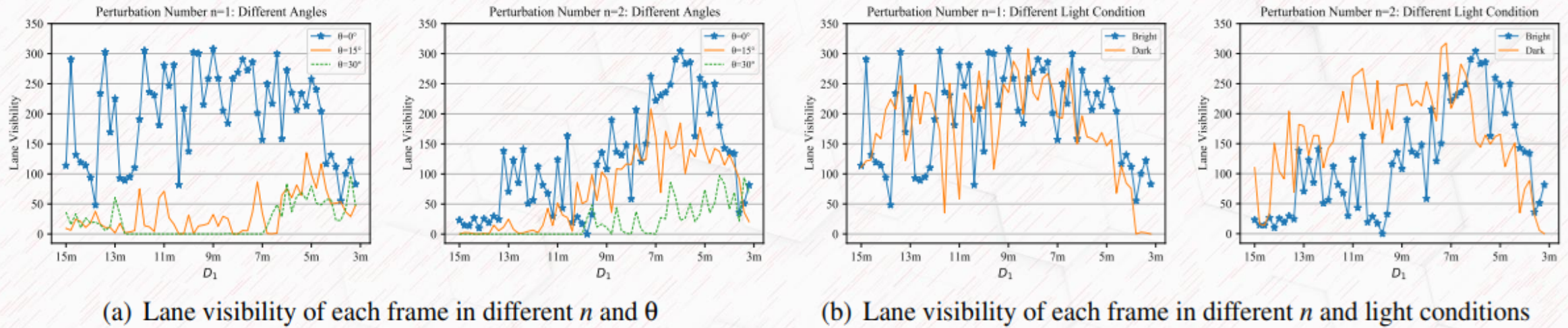


Figure 11: The visibility of lane changes with D_1 . Straight perturbations ($\theta = 0$) have higher lane visibility. Perturbation number n and light condition have little effect on the lane visibility.

Answer: The crafted perturbations can be detected as fake lanes while staying imperceptible to humans.

USENIX Security 2021 Paper

Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations

Video demonstration - timeline:

0:05 ~ 1:05: Research Question 5 - Investigate effectiveness of perturbations in physical world.

1:06 ~ 1:15: Research Question 6 - Misguide the vehicle to the oncoming traffic.

The Hong Kong Polytechnic University & Tencent Security Keen Lab

Defense Mechanisms

Enhancing the lane detection module

1. *Detecting abnormal lane lines by features*

- Since the attackers want to make the perturbations unobtrusive, the size of the perturbations for generating the fake lane should be much smaller than the normal lanes.
- Moreover, as the attackers want to mislead the vehicle to cause safety and/or security consequences, the detected fake lane will be inconsistent with the real lanes (e.g., generating sharp turns).
- As a result, the lane detection module can leverage these features to reject the abnormal lanes in advance.

2. *Including adversarial examples in training data*

- Adding adversarial examples in the training data can make the model more robust to adversarial attacks.
- Images with perturbations can be included in the training data to help the lane detection module distinguish between crafted perturbations and real lane lines.

Enhancing the control policy

1. *Taking into consideration other visual elements*

|| The vehicle is vulnerable to our attacks if the steering control policy just relies on the lane detection result

|| It can be enhanced by involving other visual elements (i.e., coming traffic, pedestrian) to assist the steering control.

2. *Multi-Sensor fusion*

|| The control policy should also take into account the information from sensors like LiDAR, Radar, sonar and GPS

|| Data from GPS and Radar can be used to detect whether the vehicle is deviated or running in the oncoming traffic lane.

3. *Advanced warning*

|| As the security of autonomous driving may not be fully guaranteed, the vehicle should warn the driver in advance when any abnormal lane line is detected (e.g., the size of the lane is too small or the angle of the lane is too sharp, etc.)

|| Moreover, to ensure safety, the vehicle should demand the driver for manual control and quit auto-steer mode.

High quality road marking systems

High-quality road markings based on cold plastic with high visibility

e.g. radar-reflective road markings based on cold plastic

Rain



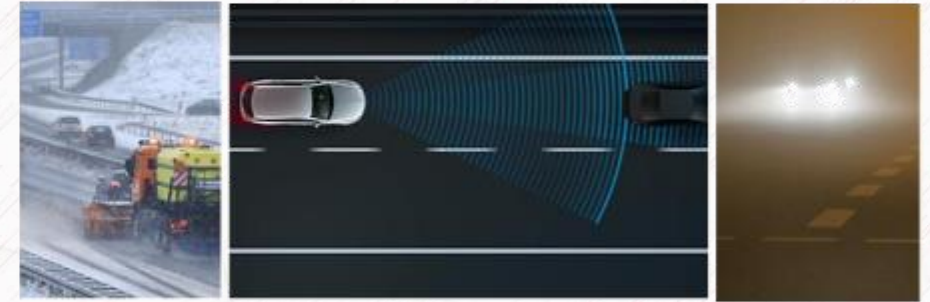
Type II markings
(wet-night visibility)

Dirt



Dirt-resistant markings
(anti-blackening effect)

Snow / Fog



Detection outside the visual range
(redundancy)

High quality road marking systems



Cars' onboard cameras and GPS can map, photograph, and grade lane lines from Ideal to Missing



HONDA VIA YOUTUBE



HONDA VIA YOUTUBE

- 1 The vehicles will capture location coordinates (longitude and latitude), along with images and video clips of the roads.
- 2 The system views and classifies lane lines to the left and right of the vehicle and color-codes them by condition: green (Ideal), yellow (Good), red (Needs Repair), and gray (no lane lines).
- 3 The collected data will be anonymously sent to a secure platform where it can be analysed and then shared with the state DOT.

Conclusion and related works

Related works

Author(date)	Title	Summary of contribution	Implication
Zhou H. et al. (2018)	Deepbillboard: Systematic physical-world testing of autonomous driving systems	They proposed a method called DeepBillboard to generate physical adversarial examples to make the DNN-based autonomous driving system steer to the wrong direction	The attack model is different from reviewed work.
Nassi B. et al. (2020)	Phantom of the ADAS: Phantom attacks on driver-assistance systems	Authors utilized projection to make the vehicle believe that the projection is the real object (phantom attack), and they also tested the lane detection module of Tesla Autopilot. However, the phantom attack only works at night, and it can be easily noticed.	Our review paper's attack can be launched during the day and is more stealthy.
Sato T. (2021)	Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack	Authors systematically study the security of state-of-the-art deep learning based ALC systems under physical-world adversarial attacks in the form of a novel and domain-specific attack vector: dirty road patches.	The attack model and methodology slightly differs from the reviewed paper.

Conclusion

- The authors conducted the first investigation on the security of the lane detection module in a real vehicle and reveal that its sensitivity can be exploited to generate fake lanes and consequently mislead the vehicle.
- They proposed a novel two-stage approach to generate the optimal perturbations against the lane detection module.
- They conducted extensive experiments on a Tesla vehicle to evaluate our approach. The experimental results show that the lane detection module in Tesla Autopilot is vulnerable to our attack.

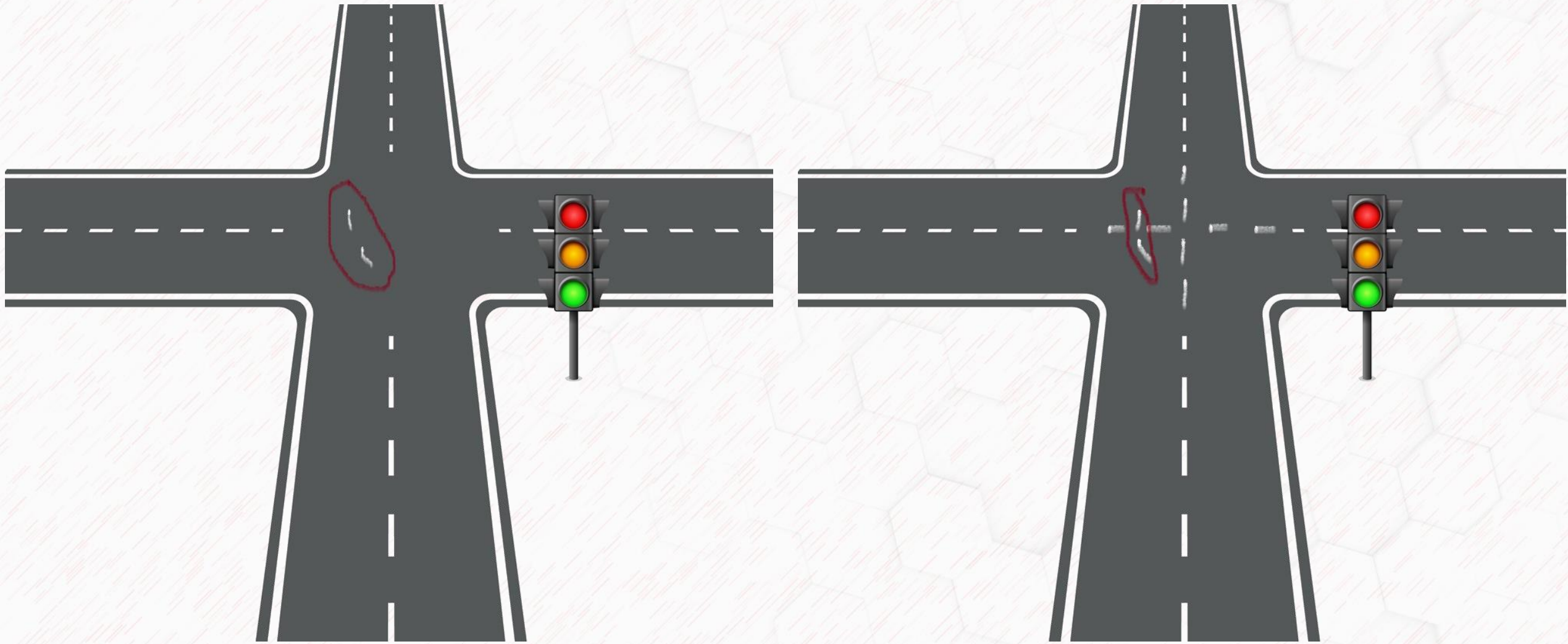


Paper Critique/Discussions

Can we answer the following questions based on our understanding?

- 1** There is a critical flaw in the evaluation of the effectiveness of this attack. Can you identify it?
- 2 Does this flaw completely undermine the gravity of this attack or does it provide a quick intuition on how to nullify the attack?
- 3** What other limitations can we identify from the paper?
- 4 Which other attack methods are possible here?
- 5 Are there other means of defense we can think of?
- 6** What future/interesting research can be done in this area?

The effectiveness of the attack in the real world was based on a single scenario!



A

B