# Adversarial Driving: Attacking End-to-End Autonomous Driving Systems

**Han Wu , Wenjie Ruan**[1]

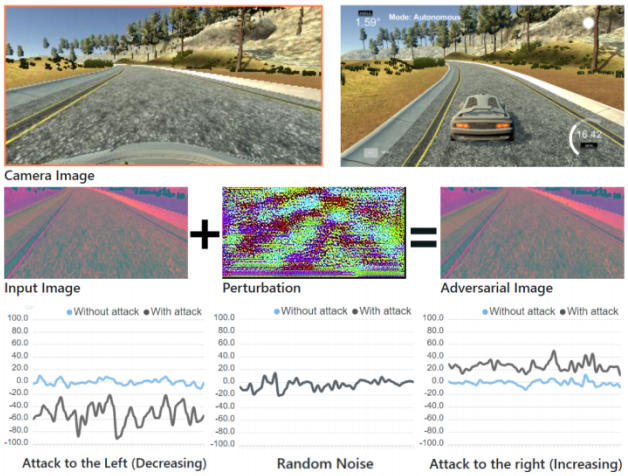
## Abstract

As the research in deep neural networks advances,deep convolutional network become feasible for automated driving tasks. There is an emerging trend of employing end-to-end models in the au-tomation of driving tasks. However,previous re-search unveils that deep neural networks are vul-nerable to adversarial attacks in classification tasks.While for regression tasks such as autonomous driving, the effect of these attacks remains uncer-tain. In this research, we devise two white-box targeted attacks against end-to-end autonomous driving systems. The driving model takes an image as input and outputs the steering angle. Our attacks can manipulate the behavior of the autonomous driv-ing system only by changing the input image.The implementation of both attacks can achieve real-time performance on CPUs. This demo aims to raise concerns over applications of end-to-end models in safety-critical systems.

## 1. Introduction

As computational capacity increases, the end-to-end deep learning model has been employed in autonomous driving systems that may bring new concerns about its safety. Autonomous driving systems are normally divided into sub-tasks: localization and mapping, perception, assessment,planning and decision making, vehicle control, and humanmachine interface [Yurtsever et al., 2020]. Recently, end-to-end driving that maps inputs directly to steering commands started to rise as an alternative to modular systems. The earli-est attempt of end-to-end driving dates back to a 3-layer fully connected network that was trained to produce the direction to be followed [Pomerleau, 1989]. There are also applica-tions of end-to-end systems for off-road driving [Muller et al., 2006]. More recently, researchers build a convolutional neural network to map raw pixels from a single front-facing camera directly to steering commands [Bojarski et al., 2016]. End-to-end learning may lead to better performance and smaller systems, but they can be fooled by adding perturba-tions to inputs without being perceived by humans.

**Demo:**
**https://www.youtube.com/watch?v=?**




Fugure 1: Adversarial Driving: The behavior of end-to-end au-tonomous driving model can be manipulated by adding unperceiv-able perturbations to the input image..

Existing adversarial attacks can be categorized into white-box,gray-box,and black-box attacks.In white-box attacks,the adversaries have full knowledge of their target model, in-cluding model architecture and parameters.In gray-box at-tacks, the adversaries only have access to the structure of the target model. In black-box attacks,the adversaries can only gather information about the model through querying [Ren et al., 2020].Current research on adversarial attacks majorly fo-cuses on classification tasks, while the effect of these attacks against regression tasks remains uncertain.

The contributions of this paper are summarized as follows:

- We introduce adversarial driving: the first online attack against autonomous driving, which is a regression task (Figure 1).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

- We devise two white-box attacks that can be mounted in real-time: one calculates the perturbation for each frame,while the other uses a pre-calculated patch for all frames.
- The online attacking system is open-sourced and is scal-able to include more attacks for further research.

## Adversarial Driving

### 1.1. Adversarial Attacks on Regression Tasks

Adversarial attacks on classification tasks have been widely studied for decades, but we find that relatively few works fo-cus on regression tasks.

For classification tasks, the attack is effective if the prediction differs from the ground truth. While for regression tasks, an admissible prediction could be within a range. For instance, the predicted house price can fluctuate within a reasonable range. Taking the actual value as the ground truth [Nguyen and Raff, 2018], we can use Root Mean Square Er-ror (RMSE) to measure the effectiveness of attacks. An effec-tive attack should produce a higher RMSE loss than random noises [Villar et al., 2019].

For autonomous driving, which is a regression task, current research focuses on asynchronous offline attacks. The driving record is split into static images and corresponding steering angles, then the attack is applied on each static image, and the overall success rate can be concluded [Deng et al., 2020].However, many traffic incidents are caused by minor mis-takes at a critical point. Thus some stealth attacks that have low overall success rates could still be perilous. On the other hand, similar to human drivers, driving models could also re-act to adversarial attacks, some attacks may be neutralized by models' reactions if the attack is applied synchronously.

To investigate those stealth attacks and driving models' reactions to those attacks, we would like to perform online at-tacks, which means the attack will be applied while the vehi-cle is navigating. It's risky to perform online attacks against real-world autonomous driving systems. Thus our adversarial driving system employs a self-driving simulator.

### 1.2. Adversarial Driving: Online Attack

Previous offline attacks usually rely on the ground truth, such as the fast gradient sign method (FGSM) that linearizes the cost function around the current value of , and obtain an optimal max-norm constrained perturbation [Goodfellow et al., 2015]:

However, for a navigating autonomous driving system, the cost J(, x, y) cannot be calculated because there is no ground truth of steering command. Even the same experi-enced human driver can take different operations under the same circumstance. The steering command to drive safely is not unique. Thus we need to define a suitable ground truth for our attacks. Our attack methods follow several assumptions:

Our attacks are online without pre-labeled ground truth.

The driving model is accurate enough so that we can take the model's output as the ground truth. Because if the model itself is fallacious, we do not need to attack.The model itself should fail the driving task.

It's a successful attack if the deviation under attack is greater than the one under random noises. Following these assumptions, we achieve two different kinds of attack that can manipulate the behavior of the end-to-end autonomous driving system. **F**ast Gradient Sign Method on Regression (FGSMr) First, we introduce a white-box attack that calculates the per-turbation at each timestep. This attack can push the vehicle to the desired direction (either left or right). A neural network is denoted as $f(\cdot)$ with input x and pre-diction f(x). Attacking a regression model can be treated as a binary targeted attack. We can either increase the prediction or decrease the prediction.

Instead of linearizing the cost function, we linearize the output of the model directly to manipulate the behavior of the driving model. Linearizing f(x) will increase the output,while linearizing f(x) will decrease the output.

Our steering command ranges from -1 to 1 (from left to right). If we linearize the f(x), the predicted steering command will increase, and thus the attack will push the vehicle to the right side. Similarly, we can attack the vehicle to the left side by linearizing f(x).

**Universal Adversarial Perturbation on Regression (UAPr)** Second, we introduce the other white-box attack that calcu-lates a universal perturbation for all timesteps. The attack consists of two procedures, learning and executing. During the learning procedure, for each frame, we'll gen-erate the universal perturbation by linearizing the output of the model, and find the minimum perturbation that changes the sign of the prediction to the desired direction [Moosavi-Dezfooli et al., 2016]:

The generated universal perturbation can be applied to the input image for all timesteps during the executing procedure.

### 1.3. System Architecture

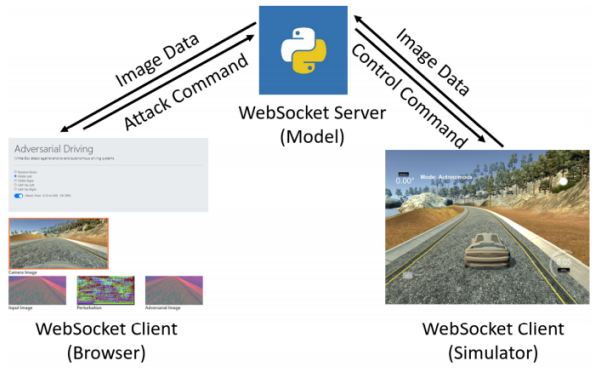



Figure 2: Adversarial Driving: System Architecture

The adversarial driving system consists of three key components: the simulator, the server, and the front-end.

**Simulator**: The self-driving simulator utilizes Unity3D,which is a game engine. The simulator connects to a Web-Socket server, and once connected, it publishes the image of each frame and accepts steering commands from the server.

**Server**: The WebSocket server accepts connections from the simulator and sends back the control command. Besides,it will publish generated adversarial images to the front end,and receive attack commands from the web browser.
**Front-end**: The front end is a website where the attacker can choose different kinds of attacks and monitor the status of the simulator.

### 1.4. Evaluation

We perform several attacks against the NVIDIA end-to-end self-driving model [Bojarski et al., 2016]. The steering command ranges from -1 to 1 (from left to right). The abso-lute and relative deviations of steering angle in average under 1000 attacks are listed in Table 1 respectively.

**FGSMr is a strong attack.**Once under attack, the vehicle will run off the road in several seconds. The absolute deviations are similar for different attack directions.

**UAPr is a stealth attack.**The absolute deviation is larger than random noises, indicating it's effective, while much smaller than the FGSMr. The attack is stealth because the di-gression of the steering command is slight while making the vehicle hard to control. This could lead to incidents at certain critical points. Besides, the same perturbation is applied for all frames, thus it's much faster than the FGSMr.

In conclusion, we devise one strong attack (FGSMr) and one stealth attack (UAPr). A strong attack deviates the vehicle in several seconds, while a stealth attack could cause incidents at certain critical points. As a result, the end-to-end driving model is vulnerable to adversarial attacks.

## Conclusion and Future Work

In this research, we devise two white-box targeted attacks against end-to-end autonomous driving systems. The behavior of the driving model can be manipulated by adding perturbations to the input image. Our research demonstrates that the autonomous driving system, which is a regression task, is also vulnerable to adversarial attacks.

Besides, further research could also be conducted to investigate the effect of black-box attacks against end-to-end autonomous driving systems. It is also possible that modular systems with inputs from multiple sensors are vulnerable to adversarial attacks. This research may raise concerns over applications of end-to-end models in safety-critical systems.

## References

Mariusz Bojarski, Davide Del Testa,Daniel Dworakowski, Bernhard Firner, Beat Flepp, Pra-soon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars, 2016.