

Adversarial Attacks on Autonomous Vehicles

Using Universal Adversarial Network (UAN)

Student: Chia Yi You

Supervisor: Assoc. Prof. Tan Rui

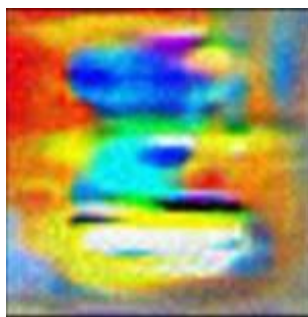



OBJECTIVE

Universal Adversarial Network (UAN)

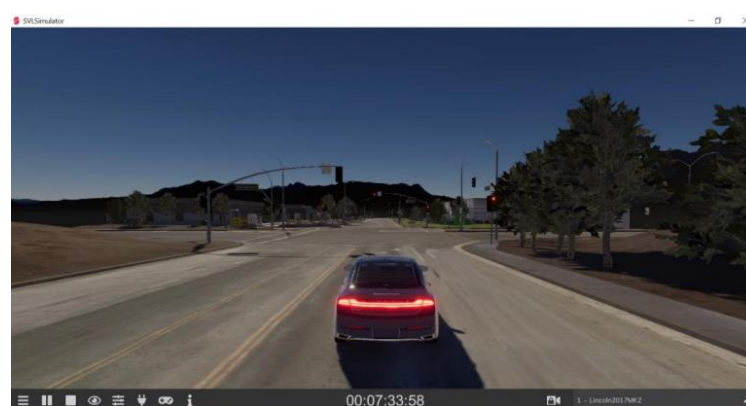
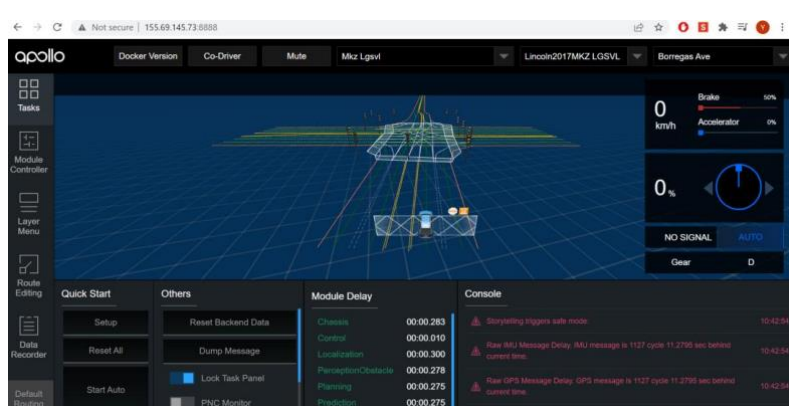
This project aims to study the safety of Autonomous Vehicles, specifically its Traffic Light Recognition systems. These use Convolutional Neural Networks (CNN) to classify traffic lights to determine their colours. Hence, they are prone to adversarial attacks. This paper uses generative models to build Universal Adversarial Perturbations, known as Universal Adversarial Network (UAN) attack, to fool the Traffic Light Recognition model to misclassify these traffic lights. This study uses Apollo Autonomous Vehicle system and its Caffe model.



RESULTS

| Size of Perturbation | UAN Perturbation | Adversarial Image | Success Rate of Attack on Surrogate DenseNet model | Success Rate of Attack on Caffe model |
|----------------------|---|---|--|---------------------------------------|
| 0.2 |  |  | 29.4% | 0% |
| 0.45 |  |  | 88.6% | 30.3% |

UAN is effective in attacking Apollo's Caffe model, resulting in misclassifications



Apollo was still able to classify traffic lights correctly in runtime environment. These were due to Apollo defences against debugging and reverse engineering, as well as sequence and time constraints.

Defences of Apollo was robust to prevent insertion of UAN in runtime environment