Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Supplementary materials for

Shuai ZHAO, Boyuan ZHANG, Yucheng SHI, Yang ZHAI, Yahong HAN, Qinghua HU, 2025. A comprehensive survey of physical adversarial vulnerabilities in autonomous driving systems. *Front Inform Technol Electron Eng*, 26(4):510-533. https://doi.org/10.1631/FITEE.2300867

1 Related works

This section will briefly review related works, i.e., the adversarial attacks and defenses, the perception tasks and models in ADS, related traffic prediction datasets, and the traffic simulators related to ADS.

1.1 Adversarial Attack and Defenses

Based on the attackers' knowledge of target ADS, the adversarial attacks in the physical world can be divided into white-box attacks and black-box attacks. White-box attacks assume that the attackers have complete access to the victim ADS. Black-box attacks assume no knowledge of the target ADS. More strictly, the adversary is unaware of its training process and parameters. One mainstream black-box attack is the transfer-based attack, which generates adversarial examples against a substitute model and transfers them to the target black-box model. The other one is the query-based attack, where the attacker can query the target ADS and exploit its output to optimize adversarial examples. Although black-box attacks are more realistic in the physical world, the white-box attack could demonstrate the adversarial vulnerability of autonomous driving systems.

Adversarial machine learning was first proposed since Szegedy et al. (2014) first detected that adversarial examples could fool DNNs with imperceptive perturbations. After that, the adversarial vulnerability of ADS has attracted attention in related communities. Goodfellow et al. (2015) designed the FGSM attack, the first powerful attack against image classifiers. The FGSM attack still cannot attack a black-box ADS in the real world. To narrow the gap between the digital world and the physical world, Athalye et al. (2018) designed the Expectation Over Transformation (EOT) algorithm, which presents a structure to generate adversarial examples that maintain their attacking nature across a specified transformation distribution. The fundamental concept underlying EOT is incorporating these perturbations into the optimization process. This strategy has demonstrated robustness in executing adversarial attacks within the physical world.

The adversarial defense strategies on ADS can be categorized as input preprocessing, adversarial example detection, and model enhancement. The input preprocessing strategies aim at cleaning inputs to make them benign for the target model. For example, JPEG-based compression (Guo et al., 2018) could remove adversarial perturbations from images. The preprocessing strategies can be easily used in conjunction with other defense mechanisms. Adversarial example detection could detect whether the input data is an adversarial example. Grosse et al. (2017) utilized an extra class in the classifier to detect adversarial examples. Metzen et al. (2017) trained a detection neural network to detect adversarial examples. Model enhancement represents researchers enhancing the robustness of the perception module without any preprocessing. The most common model enhancement strategy is adversarial training (Goodfellow et al., 2015). Adversarial training creates and then incorporates adversarial examples into the training process. However, the defense capacity of adversarial training depends on the corresponding attack method, which is difficult to generalize the defense capacity against other attacks. Therefore, certified defenses were proposed to provide a mathematical guarantee of the model's robustness against such attacks within a certain range of perturbations (Chiang et al., 2020).

1.2 Models and Tasks related to the Autonomous Driving Systems

This section will introduce commonly used perception models in autonomous driving systems, including classification models, detection models, and infrared detection systems. These models are widely applied in tasks such as Traffic Sign Recognition, Traffic Light Recognition, Vehicle Detection, Road lane detection, Monocular Depth Estimation, and Person Detection.

For Traffic Sign Recognition and Traffic Light Recognition tasks, the most commonly used classification algorithm is Convolutional Neural Network (CNN) (LeCun et al., 1998). CNN (LeCun et al., 1998) is a deep learning algorithm that extracts image features through multiple layers of convolution and pooling operations, and then classification using fully connected layers. ResNet (He et al., 2016), a representative network of CNN, addresses the gradient vanishing and exploding problems in training deep networks by introducing residual blocks and shortcut connections, enabling the training of very deep models and achieving significant success in computer vision tasks, including image recognition. It exhibits outstanding performance in image recognition tasks and demonstrates high accuracy and robustness in traffic sign recognition, making it widely employed in this field.

The commonly used object detection algorithms in Person Detection and Vehicle Detection tasks include Faster R-CNN (Ren et al., 2015), YOLO (You Only Look Once) (Redmon et al., 2016), Mask R-CNN (He et al., 2017), and SSD (Single Shot Multibox Detector) (Liu et al., 2016). YOLO (Redmon et al., 2016) is a fast real-time object detection algorithm that transforms the detection task into a single neural network prediction problem, achieving simultaneous object localization and classification in a single forward pass, making it well-suited for pedestrian and vehicle detection due to its real-time capabilities. SSD (Liu et al., 2016) is an efficient real-time object detection algorithm that simultaneously predicts the object's location and class, completing the entire detection process in a single forward pass by predicting multiple boxes at different scales to handle objects of various sizes. This enables SSD to excel in speed and accuracy, making it particularly suitable for real-time applications in object detection tasks.

Faster R-CNN (Ren et al., 2015) is a classical two-stage object detection algorithm that differs from one-stage methods by introducing the Region Proposal Network (RPN) to achieve end-to-end object detection, thereby improving detection accuracy. Mask R-CNN (He et al., 2017) is an extension of Faster R-CNN that not only accurately detects object positions and categories but also generates precise instance segmentation masks for each object, accomplishing the tasks of object detection and semantic segmentation simultaneously. Cascade R-CNN (Cai and Vasconcelos, 2018) is an improved version of object detection algorithms, which employs cascaded training of multi-level detectors to gradually improve the detection performance for small objects, leading to better precision and recall in object detection. Tan et al. (2020), using EfficientNet as the backbone network, incorporated Bi-directional Feature Pyramid Network (BiFPN) and specific scaling coefficients to build a multi-scale feature pyramid for more efficient detection of objects of different sizes. EfficientDet D0 (Tan et al., 2020) is the smallest version of the EfficientDet series, suitable for resource-constrained scenarios, such as real-time object detection on mobile devices. When pedestrians or vehicles are in motion, the detection difficulty significantly increases.

Besides image recognition models, LiDAR-based perception models support camera sensors to recognize complex environments in the real world. PointRCNN (Shi et al., 2019) adopts PointNet++ as its backbone and consists of two stages: stage 1 performs proposal generation by processing each foreground point, while stage 2 refines the proposals in the canonical coordinate. On the other hand, PointPillar (Lang et al., 2019) introduces a rapid point cloud encoder using a pseudo-image representation. It partitions the point cloud into bins and employs PointNet to extract features for each pillar. Baidu Apollo 5.0 model (Peng et al., 2020), as an industry-level BEV-based model, has 6 hard-coded feature maps in the BEV and outputs the grid-level confidence score. Stereo R-CNN (Li et al., 2019) and DSGN (Chen et al., 2020) are two 3D object detection modules. PIXOR (Yang et al., 2018) is a detection network that processes input point clouds into occupancy voxels and generates bounding boxes in a bird eye view.

With the help of several sensors and the above perception models, ADS has achieved satisfied perfor-

mance in recognizing the surrounding environments. The adversarial vulnerability needs to be investigated.

1.3 Datasets and Traffic Simulators Related to Autonomous Driving Systems

In this section, we introduce related scenario datasets and traffic simulators in autonomous driving systems. Based on different perception tasks, researchers have designed various datasets and simulators to evaluate the performance of ADS.

For traffic sign recognition, there are mainly four datasets: LISA Dataset (Mogelmose et al., 2012), GTSRB Dataset (Stallkamp et al., 2011), SBU Shadow Dataset, and Mapillary Traffic Sign Dataset (Ertler et al., 2020). The LISA Dataset (Mogelmose et al., 2012) was designed for autonomous driving and computer vision research. It includes multiple subsets covering traffic signs, traffic lights, weather conditions, and various real-world driving scenarios. The GTSRB Dataset (Stallkamp et al., 2011) stands for the German Traffic Sign Recognition Benchmark containing over 50,000 images of different traffic signs. The SBU Shadow Dataset is utilized for shadow detection and removal research. It comprises outdoor scene images collected from social media websites, each accompanied by the corresponding shadow mask annotations. The Mapillary Traffic Sign Dataset (Ertler et al., 2020) is a large-scale traffic sign dataset includes high-resolution images of traffic signs from various locations worldwide. It provides detailed annotation information, such as sign categories, locations, and orientations. The Inria Person Dataset (Watanabe et al., 2009) is a commonly used dataset for pedestrian detection tasks, which is primarily used for training and evaluating the performance of pedestrian detection algorithms.

For vehicle trajectory prediction tasks, the NGSIM dataset (Alexiadis et al., 2004) was developed by the National Highway Traffic Safety Administration (NHTSA), which is a highway traffic dataset containing real traffic flow and vehicle trajectory data. The nuScenes dataset (Caesar et al., 2020), on the other hand, is a large-scale autonomous driving dataset created by nuTonomy (now part of Aptiv). It includes high-quality LiDAR point clouds, camera images, radar data, vehicle trajectories, and scene annotations captured from urban streets in Boston and Singapore.

There are also comprehensive datasets that can be used for multiple tasks. For example, the MS COCO (Microsoft Common Objects in COntext) dataset (Lin et al., 2014) provides annotations for multiple tasks, including object detection, segmentation, and keypoint detection serving as a crucial benchmark dataset.

The Apolloscape dataset (Huang et al., 2018), developed by the Baidu Apollo autonomous driving team, is a large-scale autonomous driving dataset featuring high-resolution images, LiDAR data, and GPS trajectories from different cities. It offers diverse road scenes and weather conditions, making it applicable for advancing infrared sensor applications in autonomous driving. However, no existing adversarial datasets are available for conducting adversarial robustness evaluations of autonomous driving systems instead of the perception module in ADS.

For simulator, high-quality simulator environments are developed to help researchers test the performance and safety of autonomous driving systems without experimenting on actual roads (Karopoulos et al., 2022). The selection of an autonomous driving simulator needs to be considered from various aspects, such as the development system, application scenarios, and operating efficiency.

The physics engine is the main factor affecting the dynamic and rendering effects of the simulation (Zhong et al., 2021). CARLA and AIRSIM use Unreal Engine 4 (UE4) (Sanders, 2016) as their physics engine, SVL and PARACOSM use Unity engine (Haas, 2014), and CARSIM simulator is considered to be the most realistic vehicle dynamics engine at present. CARLA (Dosovitskiy et al., 2017) is mainly aimed at the urban driving environment from the perspective of a single vehicle. It can basically be used to help train all modules of autonomous driving, including perception system and localization. CARLA has a variety of sensor models that simulate the real world (Deschaud et al., 2021), including cameras, lidar, acoustic radar, and more.

SVL (Rong et al., 2020) is an end-to-end autonomous driving simulator built by LG Electronics US R & D Lab. Users can mark on the basis of the 3D scene in Unity and export it into a high-definition map format that matches the automatic driving system.

The Apollo platform (Xu et al., 2020) is a cloud service platform that can perform simulation tests on the cloud (Peng et al., 2020). It is divided into two scenarios: World-sim and Log-sim, while Log-sim is a scene extracted from road test data, which truly reflects the complex and changeable obstacles and traffic conditions in the actual traffic environment.

Although plenty of datasets and simulators have been proposed, a critical gap remains while none of the existing resources have been specifically designed with a focus on the unique challenges posed by adversarial examples in autonomous driving systems. The lack of datasets and simulators tailored to adversarial scenarios in autonomous driving hinders our ability to fully understand, anticipate, and mitigate the adversarial risks on ADS.

2 Physical adversarial vulnerability in other tasks

In this section, we delve into the adversarial vulnerabilities beyond the commonly perception tasks in autonomous driving systems, exploring under-researched areas such as license plate recognition and adversarial weather conditions. These tasks are crucial for the comprehensive functioning of autonomous vehicles but have received comparatively less attention in the related research of adversarial attacks. By examining these specific domains, we aim to shed light on the broader spectrum of challenges that autonomous driving technologies face in securing against adversarial threats.

2.1 The real-world attacks

In other tasks, we discuss the vulnerability of license plate recognition and others in the real-world scenario. Qian et al. (2020) designed spot evasion attack against License Plate Recognition (LPR) systems. Unlike previous methods that tamper with all pixels in the image, they focused on locally modifying the license plate character by adding simulated spots with specific colors. While the aforementioned adversarial image perturbations against autonomous cars, these perturbed images were generated offline. Yoon et al. (2023) proposed a multi-level stochastic optimization framework to generate adversarial perturbations in real-time scenarios. The method uses a generative adversarial network (GAN) to generate adversarial images, a reinforcement learning agent to misguide the vehicle, and a binary decision-maker to determine when to use image attacks. They tested their method with a small indoor drone in an office environment. Similar to Zolfi et al. (2021), Man et al. (2020) utilized a projector to attack the camera sensor directly. The authors proposed a method called GhostImage, which remotely and unobtrusively exploits the perception domain to create spurious objects or alter existing objects, leading to misperception by camera-based image classification systems. The attack progress is shown in Fig. S1. They investigated the lens flare effects and exposure control in camera sensors to enhance the attack performance.

2.2 The digital-world attacks

Besides considering the angles and distances of camera sensors, Marchisio et al. (2022) introduced a novel method called fakeWeather to address the adversarial vulnerability of autonomous systems. The method aims to generate adversarial examples by emulating the effects of weather conditions on camera lenses, specifically rain, snow, and hail. The authors observed the effects of atmospheric perturbations on camera lenses, modelled the patterns, and created masks that fake the effects of these weather conditions. The examples of patch patterns are shown in Fig. S2. Rossolini et al. (2023) designed a novel loss function to generate stronger attacks against real-time semantic segmentation (SS) models for autonomous driving. This loss function is a modification of the standard pixel-wise cross-entropy (CE) loss, traditionally used for untargeted digital attacks by adding a perturbation to pixel values.

While the aforementioned fakeWeather attack attempts to simulate atmospheric conditions, the resulting patterns often fall short of realism. Schmalfuss et al. (2023) addressed this limitation by introducing Distracting Downpour, the first 3D weather-based attack that employs a differentiable rendering framework

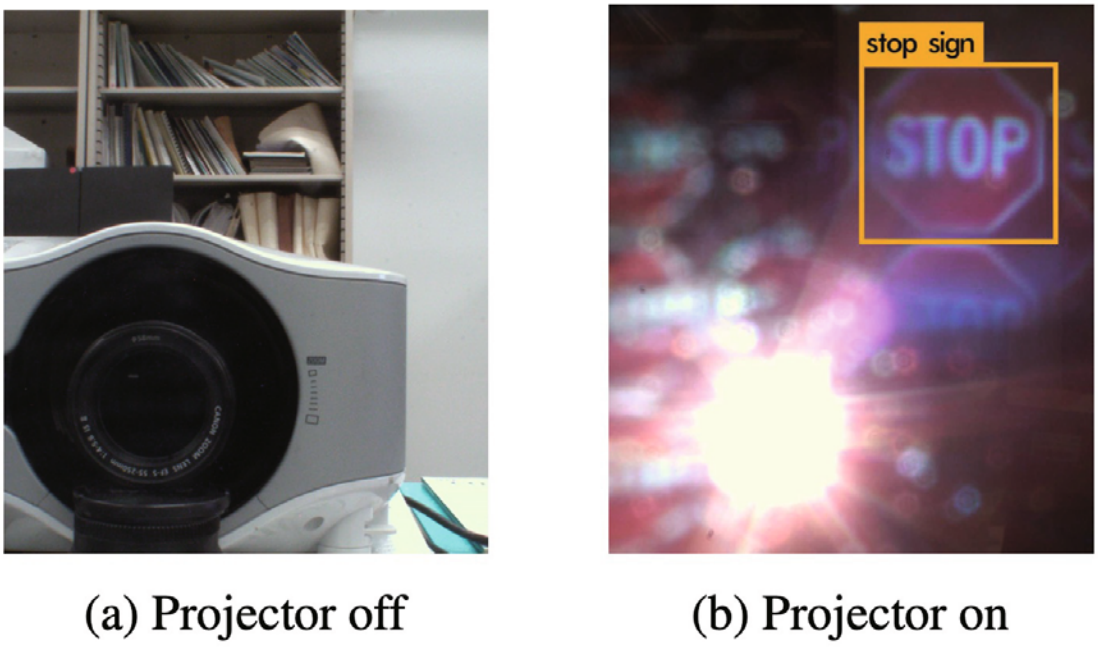

Fig. S1 A stop sign image injected into a camera by a projector, detected by YOLOv3 (Man et al., 2020)

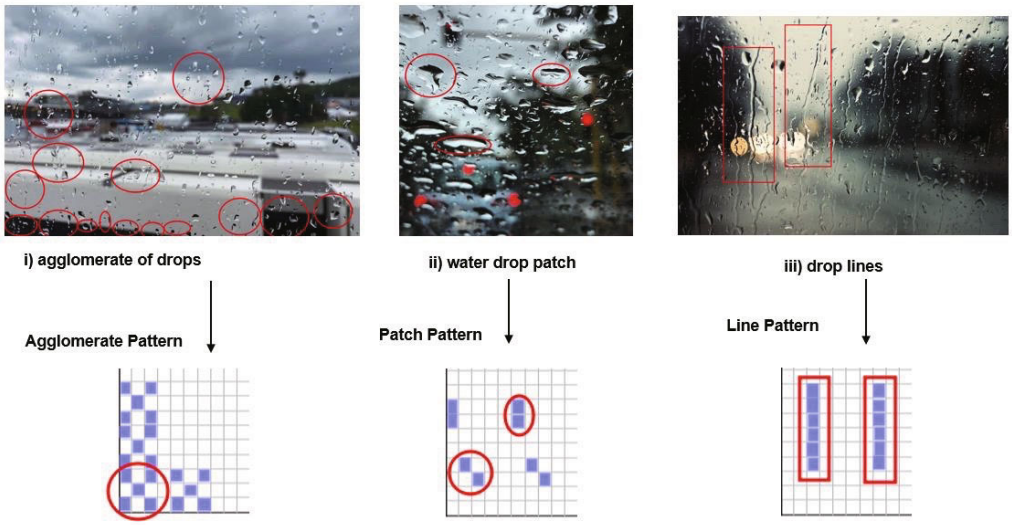

Fig. S2 Several patterns of water drops observed from the real environment (Marchisio et al., 2022)

to authentically recreate rain, fog, and snow. This approach diverges from traditional 2D pixel-by-pixel alterations, focusing on the optimization of 3D spatial positioning and color characteristics of particles within the scene. The outcome is a generation of images that convincingly mimic the motion and visual traits of real weather phenomena. Notably, the technique proves particularly effective against systems that are designed to be robust to minor ℓ_p perturbations.

Kwon and Baek (2021) proposed Adv-plate attack which focuses on adding adversarial noise specifically to the license plate area, which is difficult for humans to discern but can cause misclassification by the LPR system. Im Choi and Tian (2022) focused on the the adversarial vulnerability of YOLO Detectors (Redmon et al., 2016) in autonomous driving scenarios. They designed white-box objectness-oriented attack and tested it on real-world KITTI and COCO traffic datasets. Chen et al. (2022) further explored the adversarial

vulnerability of license plate recognition (LPR) systems to model poisoning attacks and presented a novel attacking strategy. The attack offers insights into the potential security risks associated with LPR systems and highlights the need for effective defense strategies.

References

Alexiadis V, Colyar J, Halkias J, et al., 2004. The next generation simulation program. *Institute of Transportation Engineers ITE Journal*, 74(8):22.

Athalye A, Engstrom L, Ilyas A, et al., 2018. Synthesizing robust adversarial examples. International Conference on Machine Learning, p.284-293.

Caesar H, Bankiti V, Lang AH, et al., 2020. Nuscenes: A multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.11621-11631.

Cai Z, Vasconcelos N, 2018. Cascade r-cnn: Delving into high quality object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p.6154-6162.

Chen J, Gao Y, Liu Y, et al., 2022. Leveraging Model Poisoning Attacks on License Plate Recognition Systems. 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), p.827-834.

Chen Y, Liu S, Shen X, et al., 2020. Dsgn: Deep stereo geometry network for 3D object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.12536-12545.

Chiang Py, Ni R, Abdelkader A, et al., 2020. Certified Defenses for Adversarial Patches. International Conference on Learning Representations.

Deschaud JE, Duque D, Richa JP, et al., 2021. Paris-CARLA-3D: A real and synthetic outdoor point cloud dataset for challenging tasks in 3D mapping. *Remote Sensing*, 13(22):4713.

Dosovitskiy A, Ros G, Codevilla F, et al., 2017. CARLA: An open urban driving simulator. Conference on Robot Learning, p.1-16.

Ertler C, Mislej J, Ollmann T, et al., 2020. The mapillary traffic sign dataset for detection and classification on a global scale. European Conference on Computer Vision, p.68-84.

Goodfellow IJ, Shlens J, Szegedy C, 2015. Explaining and harnessing adversarial examples. 3rd International Conference on Learning Representations, ICLR 2015.

Grosse K, Manoharan P, Papernot N, et al., 2017. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:170206280*, .

Guo C, Rana M, Cisse M, et al., 2018. Countering Adversarial Images using Input Transformations. International Conference on Learning Representations.

Haas JK, 2014. A history of the unity game engine. *Diss Worcester Polytechnic Institute*, 483(2014):484.

He K, Zhang X, Ren S, et al., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p.770-778.

He K, Gkioxari G, Dollár P, et al., 2017. Mask r-cnn. Proceedings of the IEEE International Conference on Computer Vision.

Huang X, Cheng X, Geng Q, et al., 2018. The apolloscape dataset for autonomous driving. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, p.954-960.

Im Choi J, Tian Q, 2022. Adversarial attack and defense of YOLO detectors in autonomous driving scenarios. 2022 IEEE Intelligent Vehicles Symposium (IV), p.1011-1017.

Karopoulos G, Kambourakis G, Chatzoglou E, et al., 2022. Demystifying in-vehicle intrusion detection systems: A survey of surveys and a meta-taxonomy. *Electronics*, 11(7):1072.

Kwon H, Baek JW, 2021. Adv-plate attack: Adversarially perturbed plate for license plate recognition system. *Journal of Sensors*, 2021:1-10.

Lang AH, Vora S, Caesar H, et al., 2019. Pointpillars: Fast encoders for object detection from point clouds. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.12697-12705.

LeCun Y, Bottou L, Bengio Y, et al., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278-2324.

Li P, Chen X, Shen S, 2019. Stereo r-cnn based 3D object detection for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.7644-7652.

Lin TY, Maire M, Belongie S, et al., 2014. Microsoft coco: Common objects in context. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, p.740-755.

Liu W, Anguelov D, Erhan D, et al., 2016. Ssd: Single shot multibox detector. European Conference on Computer Vision, p.21-37.

Man Y, Li M, Gerdes R, 2020. {GhostImage}: Remote perception attacks against camera-based image classification systems. 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020), p.317-332.

Marchisio A, Caramia G, Martina M, et al., 2022. fakeWeather: Adversarial attacks for deep neural networks emulating weather conditions on the camera lens of autonomous systems. 2022 International Joint Conference on Neural Networks (IJCNN), p.1-9.

Metzen JH, Genewein T, Fischer V, et al., 2017. On detecting adversarial perturbations. *arXiv preprint arXiv:170204267*, .

Mogelmose A, Trivedi MM, Moeslund TB, 2012. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE transactions on intelligent transportation systems*, 13(4):1484-1497.

Peng Z, Yang J, Chen TH, et al., 2020. A first look at the integration of machine learning models in complex autonomous driving systems: A case study on apollo. Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, p.1240-1250.

Qian Y, Ma D, Wang B, et al., 2020. Spot evasion attacks: Adversarial examples for license plate recognition systems with convolutional neural networks. *Computers & Security*, 95:101826.

Redmon J, Divvala S, Girshick R, et al., 2016. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p.779-788.

Ren S, He K, Girshick R, et al., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91-99.

Rong G, Shin BH, Tabatabaee H, et al., 2020. Lgsvl simulator: A high fidelity simulator for autonomous driving. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), p.1-6.

Rossolini G, Nesti F, D'Amico G, et al., 2023. On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*, .

Sanders A, 2016. An Introduction to Unreal Engine 4. CRC Press.

Schmalfuss J, Mehl L, Bruhn A, 2023. Distracting downpour: Adversarial weather attacks for motion estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), p.10106-10116.

Shi S, Wang X, Li H, 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p.770-779.

Stallkamp J, Schlipsing M, Salmen J, et al., 2011. The german traffic sign recognition benchmark: A multi-class classification competition. The 2011 International Joint Conference on Neural Networks, p.1453-1460.

Szegedy C, Zaremba W, Sutskever I, et al., 2014. Intriguing properties of neural networks. 2nd International Conference on Learning Representations, ICLR 2014.

Tan M, Pang R, Le QV, 2020. Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.10781-10790.

Watanabe T, Ito S, Yokoi K, 2009. Co-occurrence histograms of oriented gradients for pedestrian detection. Advances in Image and Video Technology: Third Pacific Rim Symposium, PSIVT 2009, Tokyo, Japan, January 13-16, 2009 Proceedings 3, p.37-47.

Xu K, Xiao X, Miao J, et al., 2020. Data driven prediction architecture for autonomous driving and its application on apollo platform. 2020 IEEE Intelligent Vehicles Symposium (IV), p.175-181.

Yang B, Luo W, Urtasun R, 2018. Pixor: Real-time 3D object detection from point clouds. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p.7652-7660.

Yoon HJ, Jafarnejadsani H, Voulgaris P, 2023. Learning When to Use Adaptive Adversarial Image Perturbations against Autonomous Vehicles. *IEEE Robotics and Automation Letters*, .

Zhong Z, Tang Y, Zhou Y, et al., 2021. A survey on scenario-based testing for automated driving systems in high-fidelity simulation. *arXiv preprint arXiv:211200964*, .

Zolfi A, Kravchik M, Elovici Y, et al., 2021. The translucent patch: A physical and universal attack on object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.15232-15241.