Applied Computer Systems
ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
December 2021, vol. 26, no. 2, pp. 96–106
https://doi.org/10.2478/acss-2021-0012
https://content.sciendo.com

Adversarial Attacks and Defense Technologies on Autonomous Vehicles: A Review

K. T. Y. Mahima[1*], Mohamed Ayoob[2], Guhanathan Poravi[3]
[1–3]*Department of Software Engineering, Informatics Institute of Technology, Colombo, Sri Lanka*

***Abstract* – In recent years, various domains have been influenced by the rapid growth of machine learning. Autonomous driving is an area that has tremendously developed in parallel with the advancement of machine learning. In autonomous vehicles, various machine learning components are used such as traffic lights recognition, traffic sign recognition, limiting speed and pathfinding. For most of these components, computer vision technologies with deep learning such as object detection, semantic segmentation and image classification are used. However, these machine learning models are vulnerable to targeted tensor perturbations called adversarial attacks, which limit the performance of the applications. Therefore, implementing defense models against adversarial attacks has become an increasingly critical research area. The paper aims at summarising the latest adversarial attacks and defense models introduced in the field of autonomous driving with machine learning technologies up until mid-2021.**

***Keywords* – Adversarial attacks, adversarial robustness, autonomous driving, computer vision, machine learning.**

I. Introduction

With the rapid growth of machine learning and artificial intelligence, various fields such as Autonomous Vehicles (AVs), computer vision, and natural language processing are developing rapidly [1]. Developing an autonomous vehicle is a core goal of conceptualizing the future vision of smart and autonomous cities [2]. Over the past decade, the development of autonomous cars has drawn much attention and lots of simulations and prototypes have been introduced. At present, a variety of blue-chip companies like Google, Apple (rumored), and Uber are in the process of productionalizing commercialized AVs. Companies such as Tesla have already commercialized vehicles on offer [3].

In the development process of AVs, there are various artificial intelligence and machine learning components. Due to the high availability of the data and the high performance, most of the time the research community employs deep learning models for machine learning-related tasks [4]. At present, they have been deployed in various domains such as image, text, and audio classifications, time-series predictions and many more [5]. Some AV capabilities are integrated into modern vehicles, which also use deep learning models to provide functionality such as object detection, image classification, and semantic segmentation to improve the security of the passengers. In the case of AVs, these models are used in more complicated tasks such as decision making, steering and pathfinding of the vehicles [6], [7]. Furthermore, to improve the effectiveness of AVs, a communication schema between stakeholders named Vehicle-to-Everyone (V2X) technique, which is a composition of communications like Vehicle-to-Infrastructure (V2I) and vehicle-to-vehicle (V2V), has been introduced to transmit information such as traffic conditions and resource allocation [8]. Fig. 1 depicts a general architecture of an autonomous driving machine and its assorted data ingressing components.

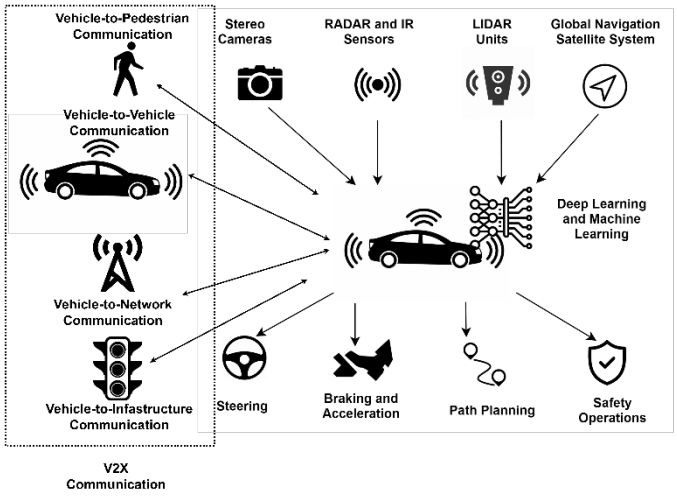


Fig. 1. Overview of the autonomous driving machines [9], [10]. Inputs gathered from the cameras, other sensors and V2X communication is processed by the deep learning models and other controlling units in the AVs to perform various tasks.

A. Adversarial Attacks

Present research has discovered that deep learning models are not fully secure because an attacker can input adversarial examples, which are specifically designed perturbations to cause these deep learning models to predict erroneously [11]. These perturbations are undetectable by humans, but they are strong enough to reliably fool the model. There are considerable research works that have been done on defensive strategies against adversarial attacks as well.

An attacker is able to execute attacks on a machine learning model in two main stages in a machine learning pipeline. The first category is the machine learning model-training phase

* Corresponding author's e-mail: yasas.2018362@iit.ac.lk

attacks [12]. This can be further divided into three main types. The first type of attack is known as the data injection attack where the attacker injects adversarial samples into the training dataset to change the distribution by poisoning but without any knowledge about the training dataset. The second type is data modification attacks where the adversary is modifying or contaminating the training dataset with the knowledge of it. In these two types of attacks, it assumes that the adversary has no knowledge about the target model. The latter type of attack is known as the logic corruption attack where the attacker tries to modify the target model. It is assumed here that the attacker is fully aware of the model [12], [13].

The second type of attack is testing phase attacks [12]. During this stage, the attacker primarily causes misclassifications in the model output by generating adversarial perturbations. From the above two main categories (i.e., training phase and testing phase attacks), testing phase attacks get higher attention because there are many studies conducted for attacks and defense methods at this stage [14], [15].

Attacks during the testing phase are further classified into three categories: black-box, white-box, and grey-box attacks [5]. In black-box attacks, it is assumed that the attacker does not know the model, in white-box attacks, it is assumed that the attacker has full knowledge of the model, including architecture and defense methods, and in the grey-box attacks, the attacker has some knowledge about the model such as the structure of the model and training data but no knowledge about the weights of the model [13], [16]. Considering the adversary's knowledge about the network, the white-box attacks could be stated as the strongest and the black box attacks are the weakest ones. Fig. 2 depicts the summary of the above classification of the adversarial attacks.

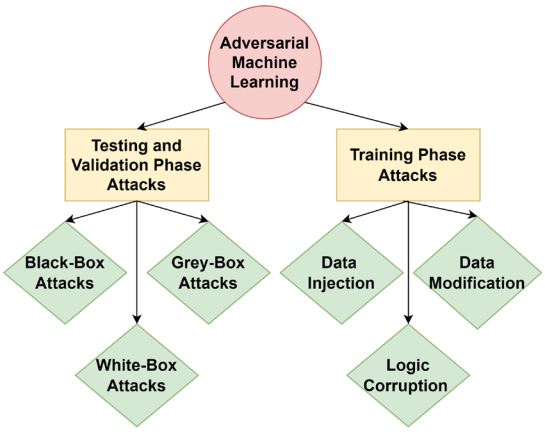


Fig. 2. Concept map of the adversarial attacks.

Considering the intention of the adversary, the attacks can be divided into two main categories. If the attack is performed for a specific feature, it is a targeted attack; if the attack performs all the features, it is an untargeted attack [13]. Apart from the above classifications, there is a set of attacks named exploratory attacks to gain sensitive information about the model and training set without modifications. Some examples of exploratory attacks: 1) Model extraction attacks: where the attacker tries to extract parameters by querying the model in a black-box manner, and 2) Model inversion attacks: where the attacker attempts to reconstruct or extract private and sensitive information related to the training set from the model parameters or using the prediction outputs [12], [13], [17].

B. Defending against Adversarial Attacks

There are many attacks introduced in each of the above-discussed stages. Moreover, when defending against these adversarial attacks, the researchers introduce several defense strategies. Among such strategies, the adversarial training method where the model is re-trained by augmenting adversarial examples to the training dataset with their correct labels is widely used. Here, the adversarial perturbations are generated by selecting one or many attacks [14]. Another popular defense mechanism is the defensive distillation method. The main objective of the defensive distillation method is to make the learning process smooth and remove the volume of gradients around the inputs [18]. Apart from these, Generative Adversarial Networks (GAN) based approaches (Defense GAN) [19] and denoiser-based defense approaches [20] have been introduced. Nevertheless, adversarial training is the most promising adversarial defense approach and several seminal improvements have been carried out and introduced [21].

Qiu et al. divided these adversarial defense methods into three main strategies [1]. The first one is modifying the data-based defense method, which refers to modifying both training and testing phase data and improving the robustness of the models. Defensive techniques such as adversarial training [14], gradient hiding [1], [22], [23] and input transformation [24] belong to this method. The second strategy is modifying the model-based defenses. This strategy includes defense mechanisms like defensive distillation and regularization [18]. The last one is auxiliary tool-based defense models, which use additional tools such as GAN networks in the defense model [1]. This is the summary of the adversarial attacks and the defense models commonly used in present.

The security threat of the adversarial attack has received more attention among the research community with the arrival of AVs, which detect and classify objects, control speed and plan paths via the use of deep neural networks [25], [26]. This is understandable given the growth of computer vision technologies in AVs based on machine learning, and the future of autonomous and unmanned vehicles is likely to be based on machine learning [27], [28]. Thus, these adversarial attacks would be a significant threat. At present, there are several research works devoted to introducing novel adversarial attacks and defense models, pipelines in AVs. However, there is still a technical barrier to making fully robust defense models against these adversarial attacks.

C. Motivation

The importance and the usage of machine learning applications are increasing in the AV industry. However, as mentioned earlier, these machine learning and deep learning models are vulnerable to adversarial attacks, which are limiting the performance of the applications. Moreover, these attacks on AVs could be a great problem for society because if the output

of a machine-learning model in an AV is misclassified due to an adversarial attack it will result in an increase in automobile accidents, traffic delays, and even impairments and death. There are comprehensive surveys and reviews on adversarial attacks, but there is a limited number of works specifically in the domain of AVs. Therefore, the primary motivation of this review paper is to summarise the recently identified adversarial attacks, defense models against those attacks, and to discuss their performance and reliability in the field of AVs. In addition, a comprehensive analysis of open research problems in adversarial machine learning on AVs is given. We hope that it will be a valuable addition for those who are interested in conducting research on adversarial attacks on AVs.

D. Organisation of the Paper

In this review paper, we summarise the research on adversarial attacks and defense models, pipelines in AVs. Section II discusses the adversarial attacks on the machine learning models in AVs. Here we review some of the vulnerabilities and attacks that are introduced on deep learning based object detection, object classification, segmentation, and driving simulation models. Section III considers defense models and pipelines to mitigate the above attacks in AVs. A discussion on their performance and limitations will follow. Section IV addresses open research problems and areas in adversarial attacks and defense methods on AVs. Finally, Section V concludes the paper.

II. Adversarial Attacks Introduced on Autonomous Vehicles

This section focuses on the adversarial attacks, which are specially introduced and simulated on machine learning models in AVs. The section includes a discussion of their capabilities, objectives and structure.

A. Attacks Introduced on Image Classification and Object Detection Models

In [29], the researchers introduced an out-of-distribution attack in the traffic sign recognition model in AV. The proposed out-of-distribution attack allows for the generation of adversarial examples starting from anywhere in the training or testing data and going out of the distribution of training and testing data. The main insight of this attack is that since the network is trained on images of traffic signs it can only effectively classify the inputs of traffic signs, which are in distribution to the classifier. Thus, by using out-of-distribution images such as logos and embedding custom signs, the predictions of the model could be misclassified. This attack is able to fool the traffic sign recognition model in both real-world and virtual settings.

In the same research, the researchers were able to fool the convolutional neural network (CNN) based traffic sign recognition model by using the optical phenomenon of different viewing angles, and this attack was called the lenticular printing attack. It synthesizes adverse images using at least two other images and makes a different sign appear based on the viewing angle. The key insight of this attack is the difference between the viewing angles of the human driver and the camera mounted on top of the vehicle. Moreover, to make the proposed out-of-distribution attack effective under real-world conditions, they included transformations, such as brightness adjustments, perspective transformations and re-sizing at the training and evaluation phases. The researchers examined the effectiveness of the proposed attack pipeline against the adversarial training defense strategy. To support their examination, they trained the CNN model using the adversarial perturbations generated by the Fast Gradient Sign Attack (FGSM) with (Epsilon) $\epsilon = 0.3$ and (step size) $\alpha = 0.5$. The training loss of the neural network is modified according to the equation below [14].

$$J\tilde{}(\theta, x, y) = \alpha J(\theta, x, y) + (1 - \alpha)\mathrm{J}(\theta, x + \epsilon \operatorname{sign}\left(\nabla_x J(\theta, x, y)\right) \quad (1)$$

Here, the $J(\theta, x, y)$ denotes the loss function, α denotes the step size, θ denotes the model weights and biases, ∇_x denotes the gradient and $\epsilon \operatorname{sign}\left(\nabla_x J(\theta, x, y)\right)$ is the adversarial perturbation of the FGSM attack. Moreover, ϵ denotes the epsilon value, which is used to make the adversarial perturbation small [30]. When performing the proposed out-of-distribution and lenticular printing attack on this adversarially trained model, they identified that the model was still ineffective for it. The research used the GTSRB [31] and GTSDB [32] datasets. In conclusion, the study emphasized that adversarial attacks could be introduced using the behaviours of the physical world as black-box adversaries.

Tietz et al. introduced a GAN based adversarial perturbation generation mechanism for the traffic sign recognition models in AV [33]. The structure of their GAN model is similar to the AdvGAN model, which was introduced in [34] to generate adversarial perturbations. The experiments were carried out on the GTSRB dataset. The AdvGAN model contains mainly three components: the target classifier f, generator $\mathcal{G}$ and discriminator $\mathcal{D}$. Initially, $\mathcal{G}$ gets the original image and makes the adversarial examples $(x + \mathcal{G}(x))$. Then adversarial examples are pointed to $\mathcal{D}$ to ensure the generated samples are indistinguishable from their original class. At the same time, target network f gets the generated adversarial samples and outputs loss ($\mathcal{L}_{\mathrm{adv}}$), which represents the distance between the original and the predicted class for the adversarial samples [34].

In this study, the researchers checked the effect of the transformations of the input image and other hyperparameters of the GAN model on the success rate of the attack. As the initial step, they checked the performance of the attack for different perturbation threshold values with greyscale and RGB images. Then they verified the performance of the attack with hyper-parameters of the GAN models such as weighted loss, learning rate, and the number of filters in the generator model. After the experiments, results showed that the proposed attack was able to reduce the test accuracy to 17.4 % and obtained a misclassification rate of 82.9 % as the optimal result. Moreover, as a future enhancement of the research, the researchers mentioned making a defense model using adversarial training or defensive distillation methods.

Li et al. proposed a novel black-box attack on traffic sign classification models in AVs [3]. This method was based on the square attack implemented using the random search proposed in [35]. Moreover, this attack was evaluated for both targeted and untargeted approaches and it performed well in both cases. In summary, these are some main attacks introduced for image classification-related machine learning models in AVs.

In AVs, object detection also plays a major role. In terms of object detection, at present RetinaNet, YOLO, and Faster R-CNN are the trending methods [36]. A recent study has introduced an attack by compressing the size of the malicious samples to stickers that can fool the YOLO and Faster R-CNN object detectors [37]. Furthermore, ShapeShifter: Adversarial attacks on Faster R-CNN object detector [38] proposed a new attack method that resulted in object detection failures in AVs. The proposed attack was inspired by two approaches introduced for image classification named change-of-variable attack [39] and Expectation over Transformation (EOT) [40]. For the evaluation, they trained a Faster R-CNN model using the Microsoft Common Objects in Context (MS-COCO) dataset [41] and performed the proposed attack in both indoor and driving conditions. Their experiments showed that the proposed attack was able to fool the object detector with a maximum of 93 % success rate.

Eykholt et al. proposed an adversarial physical perturbation-based attack to fool classification models [42]. The main objective of the work was to introduce a novel algorithm named robust physical perturbations (RP2) that synthesized adversarial samples under different physical conditions such as environmental conditions and spatial constraints. The evaluation of this study was performed using classification models made on GTSRB and LISA [43] datasets and the proposed attacks were robust to a wide range of distances and angles. Based on the RP2 algorithm introduced in [42], Eykholt et al. proposed two novel attacks on YOLO and Faster R-CNN traffic sign object detection models [44]. These attacks make object detectors ignore the traffic signs (disappearance attack) or start to detect objects that are not available in the frame (creation attack) by covering the signs with adversarial posters or attaching stickers. Both object detectors were fooled with an average of 85% by the disappearance attack with poster perturbations in a lab environment.

Lovisotto et al. proposed a novel attack technique named SLAP (Short-Lived Adversarial Perturbations) which used a specific adversarial pattern on traffic signs using a projector [45]. To ensure the strength of the perturbations for physical conditions such as viewing angles, various input transformations were used with EOT method-based optimization. The attack success rate was evaluated on different distances, angles, and indoor/outdoor conditions. The results showed that an adversarial training defense approach could successfully improve the resistance against attack; however, the input randomization defense approach was unable to improve resilience [45].

B. Attacks Introduced on Segmentation and Driving Simulation Models

Apart from classification and object detection, semantic image segmentation is also a key annotation technique used in AVs. Xu et al. introduced an iterative projected gradient-based attack on segmentation models in AVs [46]. For the investigation of the attack, they used the DeepLab-V3+ segmentation model. In the evaluation, the untargeted way of the attack reached approximately 65 % of D-mIoU (Drop-in mean interaction over-union) rate and the attack was improved when the number of iterations increased. Based on the observations, they also proposed an adversarial training method to defend against attacks on segmentation models.

When developing autonomous driving models, simulation engines do a great job. In [47], the researchers investigated the effect of physical adversarial attacks using the CARLA simulator. Here they generated adversarial perturbations such as painting of a black line on the road to investigate the models and they identified that those physical perturbations could easily deceive the steering functionalities of an AV. Moreover, [48] proposed a novel attack to cause failure in the motion planning system of AVs using adversarial billboards. This attack was evaluated in physical conditions such as brightness, weather conditions. According to the results, they were able to mislead the steering angle error of the vehicle by nearly 26 degrees.

Wu et al. devised FGSM on regression (FGSMr) and universal adversarial perturbation on regression (UAPr) attacks for steering tasks in AVs using a simulator [49]. The results showed that the FGSMr attack had the ability to deviate the vehicle within seconds while the UAPr attack caused underperformance of the vehicle at certain critical points.

According to the attacks discussed above, it can be seen that most of the proposed attacks concentrate on image classification and object detection-based machine learning models. Moreover, since adversarial attacks on semantic segmentation and other fields are still ongoing research areas, in the future there will likely be more attacks introduced. Furthermore, we have noticed that the researchers have concentrated on the strength of the proposed adversarial perturbations under various physical world conditions. Table I provides a summary of the state-of-the-art adversarial attacks specially introduced for autonomous driving machines.

III. Defense Methods Introduced on Autonomous Vehicles against Adversarial Attacks

This section focuses on the adversarial defense mechanism introduced on AVs. In addition, we will discuss the advantages and disadvantages of these defense models.

The authors of [50] performed an in-depth analysis of four existing defense methods against five adversarial attacks on CNN-based driving models. They evaluated adversarial training, defensive distillation, anomaly detection, and feature squeezing defense methods against Iterative Targeted Fast Gradient Sign Method (IT-FGSM) attack, optimization based attack, AdvGAN [34] universal adversarial perturbation attack, and AdvGAN universal adversarial perturbation attack [51]. As

a result, adversarial training and defensive distillation methods are only robust to IT-FGSM and optimization-based attacks up to some extent, while other defenses are able to detect more attacks with their own limitations. It was proposed to make a collaborative defense approach.

Wan et al. evaluated adversarial training and defensive distillation for adversarial attacks on traffic light classification [52]. First, they trained the model using a dataset, which contained original and adversarial examples and then implemented defensive distillation. In the investigation process, they verified the robustness of the model using spatial, one-pixel, Carlini & Wagner (C&W) and boundary attacks. As a result, each defense approach was marginally successful or susceptible to those attacks.

In [53], the researchers proposed adversarial training and a defensive distillation-based approach. For the adversarial training, they used perturbations from the FGSM and the Jacobian-based saliency map (JSMA) attacks generated by a separate deep learning network trained on the same dataset. Equation (2) represents the hypothesis of adversarial perturbations from FGSM, and the optimization function of the JSMA attack to generate adversarial perturbations is demonstrated in Eq. (3) [54]:

$$\eta = \epsilon * \text{sign}(\nabla * J(\theta, x, y)) \tag{2}$$

$$\underset{\delta_x}{\text{argmin}} ||\delta_x|| \text{ s.t. } F(X + \delta_x) = Y^*, \tag{3}$$

where Y^* denotes the targeted label and δ_x denotes the perturbation added to the original input X.

The researchers were able to obtain an average of 91 % testing accuracy by using the proposed hybrid defense method. This approach emphasises the importance of the collaborative approach of several defense technologies rather than using a single defense approach.

5G technology is an area that is gaining high focus at present. Wu et al. proposed a 5G network-based approach to make the traffic sign recognition models in AVs robust against adversarial attacks [55]. As the main defense technology, singular value decomposition (SVD) was used where SVD eliminated or filtered out the adversarial perturbation to restore the input image of the neural network model. To address the requirement of real-time combating of the adversaries, they used mobile edge computing (MEC), where the model was deployed in a MEC node with 5G capability, to get the correct (Robust Output) signal for the vehicle. Moreover, to investigate the adversarial robustness, Iterative FGSM (I-FGSM), C&W, Deep Fool and JSMA attacks were used.

In [56], the researchers proposed a restoration method that secured the model by removing the adversarial noise from the adversarial examples and reverting it to the original inputs using an encoder and a decoder (AutoEncoder). According to the output, an average of 97 % restoration rate for the FGSM adversarial perturbations was obtained.

Sun et al. proposed a novel defense approach to counteract the adversarial attacks on object detection models in AVs. The proposed defense strategy was based on adversarial training with a novel regularization term, which considered the local smoothness and stereo information. The evaluation results showed that their approach was more effective than a regular adversarial training approach and it improved the detection performance of the original model as well [57].

Lu et al. showed that the solidity of the adversarial perturbations would degrade with the distances and viewing angles. They empirically examined this phenomenon using a traffic sign detector based on YOLO [58]. However, later researchers used input transformation methods to improve the strength of the proposed attack under these conditions. This is a clear example to show how attackers adopt the particular defense or resilience approaches, and it shows that adversarial machine learning is a boundless research area.

According to the aforementioned adversarial defense methods for AVs, we can clearly understand that researchers are not yet able to implement a fully robust model for the specific components that they paid attention to, such as traffic sign recognition. Besides, most of the research used FGSM or collaborative adversarial training and defensive distillation method as the main techniques. However, each defense approach is not fully robust against adversarial attacks and none of those works concentrated on natural corruptions which appeared as black-box adversaries. Table II summarises the defense methods introduced in AVs. In the next section, we will discuss open research areas in this domain.

IV. Discussion

Based on what has been presented so far, it can be understood that there are various adversarial attacks introduced relevant to machine learning components in AVs. However, the number of defense methods introduced is sparse when compared to the number of attacks introduced. This section will discuss open research areas in adversarial attacks and defense methods in the AV domain.

A. Introducing New Attack Methods

In recent years, most research on adversarial machine learning on AVs has introduced novel attacks. However, we can see that most of those attacks are limited to object detection and classification-related areas. We have identified that research on making adversarial attacks tends to use behaviours of the physical world. As future research, we propose implementing novel attacks on deep learning techniques used in AVs such as semantic segmentation and creating 3D/2D maps. In particular, in AVs photogrammetry and 3D scene, re-construction techniques could be used to accumulate information about objects, geographic properties and other environmental variations [59], [60]. We hope that in the future, cutting-edge 3D scene reconstruction techniques like Neural Radiance Field [61] will be applied in the AVs. Thus, trying out attack methods in an ethical manner is critical for identifying vulnerabilities and improving those technologies. Furthermore, since AVs gather training data for the algorithms, future attacks could use data modifying or poisoning approaches. To sum up, we hope that introducing attacks will help domain experts identify the weak components in the AV architectures.

TABLE I
SUMMARY OF THE ADVERSARIAL ATTACKS INTRODUCED ON AUTONOMOUS VEHICLES

Year	Work	Objective	Summary of the Attack	Used Dataset	Results
2018	[29]	To misclassify the traffic sign recognition models. To investigate the attack against an FSGM adversarially trained network.	Introduced attack named out-of distribution attack effectively synthesizes adversarial perturbations using logos and custom signs. Introduced viewing angle-based attack named a lenticular printing attack.	GTSRB GTSDB	The proposed adversarial attacks are able to fool the adversarially trained robust model.
2018	[62]	To misclassify the traffic sign recognition models.	Modifies non-toxic signs and advertisements to be classified as traffic signs.	GTSRB	Able to fool the traffic sign model with an average of 95 % success rate.
2019	[33]	To misclassify the traffic sign recognition models using adversarial examples generated by a GAN model.	Introduced generative adversarial network (GANs) based approach generates adversarial examples, and the effect for the model is evaluated when adding transformations for the input image and changing the hyperparameters of the GAN model.	GTSRB	The test accuracy was reduced to 17.4 % and an 82.9 % misclassification rate was obtained.
2021	[3]	To misclassify the traffic sign recognition model.	Introduced random search-based adaptive square attack can be performed in both targeted and untargeted ways.	GTSRB	Able to get more than 80 % success rate for both targeted and untargeted ways.
2017	[37]	To fool the YOLO and Faster R-CNN traffic sign object detection model.	Adversarial stickers are attached to the traffic signs.	NA	Successfully fooled the object detection models.
2019	[38]	To fool the Faster R-CNN based traffic sign object detection models.	Introduced attack was inspired by the change-of-variable attack and EOT method originally introduced for image classification.	MS-COCO	Successfully fooled the object detection model with a maximum of 93 % success rate.
2018	[42]	To fool the traffic-sign classification model.	Introduced robust physical perturbation algorithm (RP2) generates adversarial perturbations under different physical conditions, such as environmental conditions and spatial constraints.	LISA GTSRB	Successfully fooled the object detectors, and adversarial samples were robust to a wide range of angles and distances.
2018	[44]	To fool the Faster R-CNN and Yolo object detection models.	Introduced two attacks by adversarial stickers or posters fool the object detection models based on the RP2 algorithm.	NA	Both traffic sign detection models were fooled at a nearly 85 % success rate by ignoring the traffic signs by adversarial posters. Object detector starts recognising objects that are not present in the frame.
2021	[46]	To fool the semantic segmentation models in AVs.	An iterative projected gradient-based attack is introduced on the DeepLab-V3+ segmentation model.	Cityscapes Dataset [63]	Approximately 65 % D-mIoU rate was obtained for the untargeted way of the attack.
2019	[47]	To attack the AV controlling system.	Testing AVs controlling and the robustness for the physical adversaries.	CARLA Simulator	Physical perturbations can fool autonomous cars and make failures in their controlling system.
2020	[48]	To attack the steering tasks of the AV.	Introduced method generates adversarial billboards, which have dynamic behaviours under the viewing angles, illumination changes, and other driving conditions to maximise the error of steering tasks.	Udacity self-driving dataset [64] Dave dataset [65] KITTI [66]	Able to mislead the steering angle error of the vehicle up to 26.44 degrees.
2021	[45]	To fool object detection and traffic sign recognition components.	A specific adversarial pattern was projected to the objects. To improve the physical reliability of adversarial examples, input transformations and EOT-based optimization were used.	LISA GTSRB	The attack was a success on Mask RCNN, YOLO models, and CNN models used for traffic sign recognition. Adversarial training could improve the resistance against the attack.
2021	[49]	To evaluate adversarial attacks on steering tasks on AVs.	It was devised and evaluated how FGSMr and UAPr adversarial perturbations affected the steering tasks in AVs using a simulator.	NVIDIA end-to-end self-driving Unity3D simulator	FGSMr attack can deviate the vehicle within seconds. UPAr attack causes incidents under certain conditions.

TABLE II
SUMMARY OF THE ADVERSARIAL ATTACK DEFENSE METHODS INTRODUCED ON AUTONOMOUS VEHICLES

Year	Work	Objective	Summary of the Defense Method	Defense Category	Dataset	Result
2020	[50]	To evaluate the four existing defense models against five adversarial attacks.	Adversarial training Defensive distillation Anomaly detection Feature squeezing methods	Updating the data Updating the model Auxiliary tool-based defense	Udacity dataset Epoch, Nvidia DAVE-2 and VGG16 Driving Models	None of the defense models can completely defend against the investigated attacks
2020	[52]	To implement a robust model for traffic light classification.	The model was trained using a dataset with adversarial and non-adversarial data. Defensive distillation.	Updating the data Updating the model	Traffic light dataset collected from CARLA simulator	The model was marginally successful against spatial attack, but not robust to the C&W attack.
2017	[53]	To implement an adversarially robust traffic sign classification model.	A collaborative approach of FGSM and the JSMA attacks-based adversarial training with the Defensive distillation method.	Updating the data Updating the model	GTSRB	Able to get an average of 91 % testing accuracy for adversarial samples.
2020	[55]	To implement an adversarially robust traffic sign classification model.	Proposed singular value decomposition, 5G, and MEC-based approach. For the investigation, I-FGSM, JSMA, C&W, and Deep fool attacks were used.	Auxiliary tool-based defense	GTSRB	Able to defend against the attacks at a sufficient level. Allows for real-time combating of adversarial attacks.
2020	[56]	To implement an adversarially robust image classification model (Targeted Model – Traffic sign recognition)	Proposed autoencoder-based adversarial noise removing method.	Auxiliary tool-based defense	GTSRB	An average of 97 % restoration rate was obtained for the FGSM adversarial perturbations
2020	[57]	To implement a robust object detection model.	Proposed defense approach based on adversarial training with a novel regularization term.	Updating the model	KITTI	Robustness to PGD and FGSM attacks is higher than the traditional adversarial training approach.
2017	[58]	To evaluate the robustness of the adversarial perturbations to traffic sign object detectors under the physical world constraints.	Assessed the reliability of the adversarial perturbations under the viewing angles and distances using the YOLO object detection network.	Updating Data	MS-COCO GTSRB	Demonstrates that the attack success rate was degrading under the changes in the viewing distance and viewing angles.

B. Improving the Robustness of the Machine Learning Models

For the defense methods discussed above, most works used FGSM based or collaborative adversarial training approaches and defensive distillation strategies. It is clear that none of those works is fully robust against adversarial attacks. Research by Madry et al. shows that FGSM based adversarial training does not increase the robustness against the adversarial attacks for large epsilon ϵ values and networks can over-fit on the adversarial examples. Additionally, they demonstrate that adversarial training using projected gradient descent (PGD) adversarial perturbations: Equation (4) is more efficient [15].

$$x_{\text{adv}}^{t} = \Pi_{\epsilon}\left(x^{t-1} + \alpha \,.\, \text{sign}(\nabla_x \ell(h(x^{t-1}), y))\right), \qquad (4)$$

where x_{adv}^{t} the adversarial sample in step t and Π the project to the ball of interest (Clipping values between values $[-\epsilon, \epsilon]$).

Furthermore, the information contained in image representations can be utilised as means to generate more transformations and to have a deep semantic understanding of the scene during the training phase defenses. A study by Mahendran et al. [67] has demonstrated that several layers of convolutional neural networks have retained representations of an image to a high level of geometric invariances. These transformations can further augment the training process of the models.

Recent research has identified that deep learning models in AVs are vulnerable not only to adversarial attacks but also to physical world adversaries such as transposing brightness, noise and blur [68]. In [69], the researchers show that these physical conditions cause a decrease in the performance of the traffic sign recognition models in AVs. Moreover, these natural corruptions may manifest single instances or compositions of multiple corruptions. As a result, it is essential to evaluate DL models on both types of corruption [69].

Therefore, when implementing the deep learning models and defense models for AVs, the researchers have to concentrate on these natural corruptions as well. To overcome this problem, several data augmentation/transformation and sensor fusion

approaches are proposed [70], [71]. However, using sensor-based approaches improves the complexity of the system and the performance vs the cost of those approaches have to be assessed. Implementing hybrid defense models using existing defense approaches to improve the robustness of the machine learning models would be an ideal solution. Moreover, we hope that using cutting-edge technologies like reinforcement learning and explainable AI would be a novel solution to improve robustness.

C. Introducing General Adversarial Defense Solutions

Implementing a general defense framework that addresses the vulnerability of both man-made adversarial attacks and physical world adversarial corruptions at the same time would be a promising open research problem, because it would save the cost and the complexity of the system. Gurel et al. proposed an approach named "*Knowledge enhanced machine learning pipeline against diverse adversarial attacks*" (KEMLP), which performs the main task integrated with several auxiliary tasks to improve domain knowledge of the prediction via different factors. This could be identified as recent research on general adversarial robustness [72]. However, since this method uses multiple auxiliary networks during the inference phase, resource consumption of this approach has to be evaluated. This is due to the fact when it comes to the AVs improving the general resilience without using any supporting tools in the inference is essential due to the resource constraints [73].

We hope that a composite approach of adversarial training and data transformation with proper optimizations will be a possible solution for the general adversarial robustness because it will lead to a better-trained model resilient to both man-made and physical world adversaries. This ensures the plug-and-play doctrine, which does not necessitate any other dependencies during the inference phase of the model when deployed. Thus, no resource overheads would appear in the inference. Besides, general robustness to adversarial attacks and physical corruption is not a limited research area for AVs [74].

D. Improving the Security of the Machine Learning Model Data

The AV domain is one of the main big data-generating sectors. Deep learning models are data-hungry and their performance relies on the size of the training dataset. Owing to that fact, deep learning models in AVs are trained for a huge amount of data. Besides, reinforcement learning is used to improve the dynamic decision-making capabilities of the AVs where the model is learned by failures and data gathered by the AV itself [75]. Therefore, improving the security of those datasets is necessary because the performance, security and control of AVs rely on these data. At present, the data generated by AVs are stored in a distributed manner, which raises a question about the faultlessness of those data.

Moreover, according to the Eliot Framework [76] and contemporary literature works [77], the ecosystem of AVs will be based on federated machine learning and the trained machine learning and deep learning models can be updated via the cloud infrastructure. This means that there is a security threat on machine learning artifacts and their data when storing them in the cloud as well. For this reason, finding a secure way to store machine learning models and datasets is an essential open research problem.

E. Increasing the Speed and Efficiency of the Defense Models

The ecosystem of the AVs is connected to cloud computing technology. As discussed in defense methods introduced in AVs, some defense methods get considerable time to perform against adversarial attacks. However, real-time combating of adversarial perturbations is essential in a safety-critical domain like AVs. To improve the efficiency of executing adversarially robust machine learning models, this paper has proposed using innovative technologies like 5G. Moreover, since AVs have complicated intelligent components, building lightweight adversarial defense methods and improving the resilience of the existing models naturally without changing the network architecture or without using any supporting tool in the inference would be some promising research areas.

V. Conclusion

Recent research on adversarial attacks raises a question on the security and the effectiveness of AVs, which purely run on machine learning and deep learning technologies, such as object classification and detection, semantic segmentation, etc. This paper has presented a comprehensive analysis of the adversarial attacks and a survey of adversarial defense models specially introduced to machine learning components in AVs in the literature. Previous analysis shows that most of the research to date has focused on implementing new attacks and that the defense models introduced are not fully robust against adversarial attacks. Finally, the authors have discussed the open research areas in AVs and adversarial machine learning related to improving the robustness, securing the machine learning and training data of AVs, and improving the efficiency of executing defense models. The authors have particularly highlighted the importance of a general defense approach for improving the resistance against both adversarial attacks and physical corruptions as a unified solution. Moreover, several methods, which help launch new adversarial attacks on the ecosystem of autonomous driving, have also been discussed.

The authors hope that this paper will help the community identify research gaps to be addressed by those who are going to undertake further research on adversarial attacks and defense technologies for autonomous driving machines.

Acknowledgments

The first author would like to thank Prof. Arosha K Bandara (ScholarX mentor at Sustainable Education Foundation and Professor of Software Engineering at the Open University UK), Ms. Suresha Perera (Lecturer at the Open University Sri Lanka) and Ms. K.T. Senuri De Silva (Senior Software Engineer at Enactor Sri Lanka) for proofreading and guiding the revisions to this paper.

REFERENCES

[1] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, Mar. 2019. https://doi.org/10.3390/app9050909

[2] A. Manfreda, K. Ljubi, and A. Groznik, "Autonomous vehicles in the smart city era: An empirical study of adoption factors important for millennials," *International Journal of Information Management*, vol. 58, Art no. 102050, 2021. https://doi.org/10.1016/j.ijinfomgt.2019.102050

[3] Y. Li, X. Xu, J. Xiao, S. Li, and H. T. Shen, "Adaptive square attack: Fooling autonomous cars with adversarial traffic signs," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6337–6347, Apr. 2021. https://doi.org/10.1109/JIOT.2020.3016145

[4] B. Jason, "What is deep learning?," *Machine Learning Mastery*, 2019. [Online]. Available: https://machinelearningmastery.com/what-is-deep-learning/. Accessed Apr. 05, 2021.

[5] H. Xu *et al.*, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, pp. 151–178, Mar. 2020. https://doi.org/10.1007/s11633-019-1211-x

[6] A. Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," *Array*, vol. 10, Art no. 100057, Jul. 2021. https://doi.org/10.1016/j.array.2021.100057

[7] N. Morgulis, A. Kreines, S. Mendelowitz, and Y. Weisglass, "Fooling a real car with adversarial traffic signs," *ArXiv*, Art no. 1907.00374, 2019.

[8] J. Gao, M. R. A. Khandaker, F. Tariq, K.-K. Wong, and R. T. Khan, "Deep neural network based resource allocation for V2X communications," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Honolulu, HI, USA, Sep. 2019, pp. 1–5. https://doi.org/10.1109/VTCFall.2019.8891446

[9] Y. Tian, K. Pei, S. S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, May 2018, pp. 303–314. https://doi.org/10.1145/3180155.3180220

[10] P. J. Leiss, "The functional components of autonomous vehicles – Expert article," *Robson Forensic*, Sep. 2018. [Online]. Available: https://www.robsonforensic.com/articles/autonomous-vehicles-sensors-expert/

[11] G. Sun, Y. Su, C. Qin, W. Xu, X. Lu, and A. Ceglowski, "Complete defense framework to protect deep neural networks against adversarial examples," *Mathematical Problems in Engineering*, vol. 2020, Art no. 8319249, May 2020. https://doi.org/10.1155/2020/8319249

[12] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *ArXiv*, Art no. 1810.00069, 2018.

[13] X. Liu *et al.*, "Privacy and security issues in deep learning: A survey," *IEEE Access*, vol. 9, pp. 4566–4593, 2021. https://doi.org/10.1109/ACCESS.2020.3045078

[14] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *ArXiv*, Art no.1412.6572, 2015.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *ArXiv*, Art no. 1706.06083, 2018.

[16] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020. https://doi.org/10.1016/j.eng.2019.12.012

[17] M. Rigaki and S. García, "A survey of privacy attacks in machine learning," *ArXiv*, Art no. 2007.07646, 2020.

[18] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA, May 2016, pp. 582–597. https://doi.org/10.1109/SP.2016.41

[19] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," *ArXiv*, Art no. 1805.06605, 2018.

[20] F. Liao, M. Liang, Y. Dong, T. Pang, J. Zhu, and X. Hu, "Defense against adversarial attacks using high-level representation guided denoiser," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018, pp. 1778–1787. https://doi.org/10.1109/CVPR.2018.00191

[21] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," *ArXiv*, Art no. 2102.01356, 2021.

[22] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *ArXiv*, Art no. 1705.07204, 2018.

[23] N. Papernot, P. Mcdaniel, I. Goodfellow, S. Jha, Z. Y. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, Apr. 2017, pp. 506–519. https://doi.org/10.1145/3052973.3053009

[24] E. Raff, J. Sylvester, S. Forsyth, and M. McLean, "Barrage of random transforms for adversarially robust defense," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019, pp. 6521–6530. https://doi.org/10.1109/CVPR.2019.00669

[25] Y. Huang and Y. Chen, "Autonomous driving with deep learning: A survey of state-of-art technologies," *ArXiv*, Art no. 2006.06091, 2020.

[26] K. Ren, Q. Wang, C. Wang, Z. Qin, and X. Lin, "The security of autonomous driving: Threats, defenses, and future directions," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 357–372, Nov. 2020. https://doi.org/10.1109/JPROC.2019.2948775

[27] "Future of driving," Tesla. [Online]. Available: https://www.tesla.com/autopilot. Accessed on: Jun. 04, 2021.

[28] A. Osman Ors, "The role of machine learning in autonomous vehicles," Endeavor Business Media, LLC, 2020. [Online]. Available: https://www.electronicdesign.com/markets/automotive/article/21147200/nxp-semiconductors-the-role-of-machine-learning-in-autonomous-vehicles. Accessed on: Jun. 04, 2021.

[29] C. Sitawarin, A. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "DARTS: Deceiving autonomous cars with toxic signs," *ArXiv*, Art no. 1802.06430, 2018.

[30] A. Madry and Z. Kolter, "Adversarial robustness – theory and practice," 2018. [Online]. Available: https://adversarial-ml-tutorial.org/. Accessed on: Oct. 04, 2021.

[31] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, vol. 32, pp. 323–332, 2012. https://doi.org/10.1016/j.neunet.2012.02.016

[32] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Dallas, TX, USA, Aug. 2013, pp. 1–8. https://doi.org/10.1109/IJCNN.2013.6706807

[33] S. Tietz and K. Nassiri Nazif, "Attacking autonomous driving machine learning algorithms with adversarial examples," Standford University, 2019. [Online]. Available: http://cs230.stanford.edu/projects_spring_2019/reports/18681219.pdf

[34] C. Xiao, B. Li, J. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2018, pp. 3905–3911. https://doi.org/10.24963/ijcai.2018/543

[35] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Computer Vision – ECCV 2020*, *LNCS*, vol. 12368, 2020, pp. 484–501. https://doi.org/10.1007/978-3-030-58592-1_29

[36] N.-D. Nguyen, T. Do, T. D. Ngo, and D.-D. Le, "An evaluation of deep learning methods for small object detection," *Journal of Electrical and Computer Engineering*, vol. 2020, Art no. 3189691, Apr. 2020. https://doi.org/10.1155/2020/3189691

[37] K. Eykholt *et al.*, "Note on attacking object detectors with adversarial stickers," *ArXiv*, Art no. 1712.08062, 2017.

[38] S.-T. Chen, C. Cornelius, J. Martin, and D. H. (Polo) Chau, "ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector," in *Machine Learning and Knowledge Discovery in Databases*, *LNCS*, vol. 11051, 2019, pp. 52–68. https://doi.org/10.1007/978-3-030-10925-7_4

[39] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA, May 2017, pp. 39–57. https://doi.org/10.1109/SP.2017.49

[40] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, Jul. 2018, vol. 80, pp. 284–293. [Online]. Available: http://proceedings.mlr.press/v80/athalye18b.html

[41] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014. Lecture Notes in Computer Science*, vol 8693, 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[42] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 1625–1634. https://doi.org/10.1109/CVPR.2018.00175

[43] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, pp. 1484–1497, Oct. 2012. https://doi.org/10.1109/TITS.2012.2209421

[44] K. Eykholt *et al.*, "Physical adversarial examples for object detectors," in *Proceedings of the 12th USENIX Conference on Offensive Technologies*, USA, 2018, p. 1.

[45] G. Lovisotto, H. C. M. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic, "SLAP: Improving physical adversarial examples with short-lived adversarial perturbations," *ArXiv*, Art no. 2007.04137, 2021.

[46] X. Xu, J. Zhang, Y. Li, Y. Wang, Y. Yang, and H. T. Shen, "Adversarial attack against urban scene segmentation for autonomous vehicles," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4117–4126, Jun. 2021. https://doi.org/10.1109/TII.2020.3024643

[47] A. Boloor, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Simple physical adversarial examples against end-to-end autonomous driving models," in *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, Las Vegas, NV, USA, Jun. 2019, pp. 1–7. https://doi.org/10.1109/ICESS.2019.8782514

[48] H. Zhou *et al.*, "DeepBillboard: Systematic physical-world testing of autonomous driving systems," in *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, Oct. 2020, pp. 347–358. https://doi.org/10.1145/3377811.3380422

[49] H. Wu and W. Ruan, "Adversarial driving: Attacking end-to-end autonomous driving systems," *ArXiv*, Art no. 2103.09151, 2021.

[50] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim, "An analysis of adversarial attacks and defenses on autonomous driving models," in *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Austin, TX, USA, Mar. 2020, pp. 1–10. https://doi.org/10.1109/PerCom45495.2020.9127389

[51] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 4422–4431. https://doi.org/10.1109/CVPR.2018.00465

[52] M. Wan, M. Han, L. Li, Z. Li, and S. He, "Effects of and defenses against adversarial attacks on a traffic light classification CNN," in *Proceedings of the 2020 ACM Southeast Conference*, New York, NY, USA, 2020, pp. 94–99. https://doi.org/10.1145/3374135.3385288

[53] A. M. Aung, Y. Fadila, R. Gondokaryono, and L. Gonzalez, "Building robust deep neural networks for road sign detection," *ArXiv*, Art no. 1712.09327, 2017.

[54] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, Saarbruecken, Germany, Mar. 2016, pp. 372–387. https://doi.org/10.1109/EuroSP.2016.36

[55] F. Wu, L. Xiao, W. Yang, and J. Zhu, "Defense against adversarial attacks in traffic sign images identification based on 5G," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, Art no. 173, Sep. 2020. https://doi.org/10.1186/s13638-020-01775-5

[56] H. Gan and C. Liu, "An autoencoder based approach to defend against adversarial attacks for autonomous vehicles," in *2020 International Conference on Connected and Autonomous Driving (MetroCAD)*, Feb. 2020, pp. 43–44. https://doi.org/10.1109/MetroCAD48866.2020.00015

[57] Q. Sun, A. A. Rao, X. Z. Yao, B. Yu, and S. Hu, "Counteracting adversarial attacks in autonomous driving," in *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, Art no. 83, Nov. 2020, pp. 1–7. https://doi.org/10.1145/3400302.3415758

[58] J. Lu, H. Sibai, E. Fabry, and D. A. Forsyth, "No need to worry about adversarial examples in object detection in autonomous vehicles," *ArXiv*, Art no. 1707.03501, 2017.

[59] Md. T. Hossan *et al.*, "A new vehicle localization scheme based on combined optical camera communication and photogrammetry," *Mobile Information Systems*, vol. 2018, Art no. 8501898, Apr. 2018. https://doi.org/10.1155/2018/8501898

[60] H. Lee, S. Song, and S. Jo, "3D reconstruction using a sparse laser scanner and a single camera for outdoor autonomous vehicle," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, Nov. 2016, pp. 629–634. https://doi.org/10.1109/ITSC.2016.7795619

[61] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the wild: Neural radiance fields for unconstrained photo collections," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, June 2021, pp. 7206–7215. https://doi.org/10.1109/CVPR46437.2021.00713

[62] C. Sitawarin, A. Bhagoji, A. Mosenia, P. Mittal, and M. Chiang, "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos," *ArXiv*, Art no. 1801.02780, 2018.

[63] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[64] "Udacity self-driving car driving data," udacity, 2016. [Online]. Available: https://github.com/udacity/self-driving-car

[65] Y. Zhou, L. Liu, L. Shao, and M. Mellor, "DAVE: A unified framework for fast vehicle detection and annotation," *ArXiv*, Art no. 1607.04564, 2016.

[66] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, Aug. 2013. https://doi.org/10.1177/0278364913491297

[67] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015, pp. 5188–5196. https://doi.org/10.1109/CVPR.2015.7299155

[68] D. Temel, G. Kwon, M. Prabhushankar, and G. Al-Regib, "CURE-TSR: Challenging unreal and real environments for traffic sign recognition," *ArXiv*, Art no. 1712.02463, 2017.

[69] P. Bielik, P. Tsankov, A. Krause, and M. Vechev, "Reliability assessment of traffic sign classifiers," Federal Office for Information Security, Jul. 2020. Accessed: Apr. 07, 2021. [Online]. Available: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Empirical_robustness_testing_of_AI_systems_for_traffic_sign_recognition.pdf?__blob=publicationFile&v=2

[70] M. Shu, Y. Shen, M. C. Lin, and T. Goldstein, "Adversarial differentiable data augmentation for autonomous systems," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, 2021, pp. 14069–14075. https://doi.org/10.1109/ICRA48506.2021.9561205

[71] A. S. Mohammed, A. Amamou, F. K. Ayevide, S. Kelouwani, K. Agbossou, and N. Zioui, "The perception system of intelligent ground vehicles in all weather conditions: A systematic literature review," *Sensors*, vol. 20, no. 22, Art no. 6532, pp. 1–34, Nov. 2020. https://doi.org/10.3390/s20226532

[72] N. M. Gurel, X. Qi, L. Rimanic, C. Zhang, and B. Li, "Knowledge enhanced machine learning pipeline against diverse adversarial attacks," *ArXiv*, Art no. 2106.06235, 2021.

[73] T. Zhang, Y. Deng, G. Lou, X. Zheng, J. Jin, and Q.-L. Han, "Deep learning-based autonomous driving systems: A survey of attacks and defenses," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 7897–7912, Dec. 2021. https://doi.org/10.1109/TII.2021.3071405

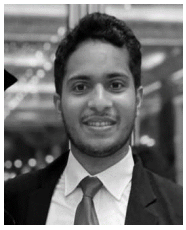
[74] A. Laugros, A. Caplier, and M. Ospici, "Are adversarial robustness and common perturbation robustness independent attributes ?" in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South), Oct. 2019, pp. 1045–1054. https://doi.org/10.1109/ICCVW.2019.00134

[75] B. R. Kiran *et al.*, "Deep reinforcement learning for autonomous driving: A survey," *ArXiv*, Art no. 2002.00444, 2020.

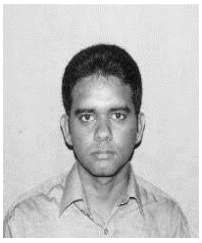
[76] L. Eliot, "Federated machine learning for AI self-driving cars," 2018. [Online]. Available: https://www.aitrends.com/ai-insider/federated-machine-learning-for-ai-self-driving-cars/. Accessed on: Apr. 14, 2021.

[77] A. M. Elbir and S. Coleri, "Federated learning for vehicular networks," *ArXiv*, Art no. 2006.01412, 2020.

K. T. Y. Mahima has a B. Eng. (hons) degree in Software Engineering from the Informatics Institute of Technology, Colombo Sri Lanka affiliated with the University of Westminster, the UK. He is a former web and mobile application Trainee Associate Software Engineer at RevportX Colombo. From 2020 to 2021, he worked as a Trainee Associate Data Engineer at Zone 24×7, Sri Lanka (Zone24×7 Incorporated, 3150 Almaden Expressway, Suite 234, San Jose, California 95118 USA). Currently, he works as an Undergraduate Research Fellow at the Open University, the UK. He was involved in several research projects at ScoreLab, Sri Lanka. His main research interests include the security of intelligent systems, unmanned vehicular systems and data science. Mr. Mahima was a student member at the Institute of Electrical and Electronics Engineers (IEEE) and a member at the Institution of Engineering and Technology (IET) at his university.
E-mail: yasas.2018362@iit.ac.lk or yasasmahima99@gmail.com
ORCID iD: https://orcid.org/0000-0003-4975-9408



Mohamed Ayoob received a B. Eng. (hons) degree in Software Engineering from the University of Westminster, the UK, in 2020. Currently, he is pursuing his MPhil at the University of Nottingham. Currently, he is a Visiting Lecturer at the Informatics Institute of Technology, Colombo, Sri Lanka. From 2018 to 2019, he was a Trainee Associate Data Engineer at Zone 24×7, Sri Lanka (Zone24×7 Incorporated, 3150 Almaden Expressway, Suite 234, San Jose, California 95118 USA). From Oct. 2020 to Jun. 2021, he worked as a Data Scientist at OCTAVE by John Keels, Colombo, Sri Lanka. He was a former visiting student researcher at KAUST (King Abdullah University of Science and Technology), Makka, Saudi Arabia, and a Data Science Mentor at WooTech Singapore. His research interests include computer vision and artificial intelligence. Mr. Ayoob obtained the best final year research project award during his undergraduate studies.
E-mail: nazeem.2016343@iit.ac.lk, hcxmm1@nottingham.edu.my
ORCID iD: https://orcid.org/0000-0003-3585-4383



Guhanathan Poravi received a B. Sc. degree in Information Systems Management from the University of Madras, India in 2004, PgD in Computer Science from the University of Peradeniya, Sri Lanka in 2006, MBA in IT from the University of Moratuwa, Sri Lanka in 2008, and MBCS from BCS, the UK in 2009. From 2004 to 2006, he was a Lecturer and Assistant Manager in education delivery at NIIT, Sri Lanka. From 2007 to 2012, he worked as a Senior Software Engineer at Cambio Healthcare Systems, Sri Lanka. Since 2012, he has been working as a Senior Lecturer (Grade 1) at the Informatics Institute of Technology, Sri Lanka. His main research interests include machine learning, big data & data science, and software engineering. Mr. Poravi was a member of the Institute of Electrical and Electronics Engineers (IEEE), Institution of Engineering and Technology (IET) and British Computer Society (BCS).
E-mail: guhanathan.p@iit.ac.lk