TOWARDS ARTIFICIAL-INTELLIGENCE-BASED CYBERSECURITY FOR ROBUSTIFYING AUTOMATED DRIVING SYSTEMS AGAINST CAMERA SENSOR ATTACKS

THIS PAPER IS A PREPRINT OF A PAPER SUBMITTED TO AND ACCEPTED FOR PUBLICATION IN PROCEEDINGS OF IEEE COMPUTER SOCIETY VLSI SYMPOSIUM - ISVLSI2020

Christos Kyrkou[1]*, **Andreas Papachristodoulou**[1], **Theocharis Theocharides**[1,2]
[1]*KIOS Research and Innovation Center of Excellence*
[2]*Department of Electrical and Computer Engineering*
University of Cyprus
1 Panepistimiou Avenue, Nicosia Cyprus
{kyrkou.christos,apapac03,ttheocharides}@ucy.ac.cy

Andreas Kloukiniotis[3], **Andreas Papandreou**[3],**Aris S. Lalos**[3,4], **Konstantinos Moustakas**[3]
[3]*ECE Department, University of Patras*
[4]*Industrial Systems Institute, Athena Research Center*
Patras, Greece
{kloukiniotisandreas,apapandreou,aris.lalos,moustakas}@ece.upatras.gr

July 27, 2020

ABSTRACT

CARAMEL is a European project that aims amongst others to improve and extend cyberthreat detection and mitigation techniques for automotive driving systems. This paper highlights the important role that advanced artificial intelligence and machine learning techniques can have in proactively addressing modern autonomous vehicle cybersecurity challenges and on mitigating associated safety risks when dealing with targetted attacks on a vehicle's camera sensors. The cybersecurity solutions developed by CARAMEL are based on powerful AI tools and algorithms to combat security risks in automated driving systems and will be hosted on embedded processors and platforms. As such, it will be possible to have a specialized anti-hacking device that addresses newly introduced technological dimensions for increased robustness and cybersecurity in addition to industry needs for high speed, low latency, functional safety, light weight, low power consumption.

***K*eywords** Autonomous Driving, Cyber-security, Computer Vision, Deep Learning, Machine Learning

1 Introduction

Autonomous vehicles, are already technologically feasible and continuously being developed and enhanced with increasing levels of connectivity and automation in an effort to increase transport safety and to reduce casualties. The list of vehicles that are expected to become autonomous include, apart from private cars, taxis, buses, and trucks. Automotive manufacturers have already solved complex problems like collision detection and avoidance, and navigation. There is still more work needed on defending against a full spectrum of malicious attackers, wielding both traditional cyberattacks and a new generation of attacks [15]. The damaging effects of cyberattacks include among others the damage in the reputation of vehicle manufacturers, the increased denial of customers to adopt cooperative, connected and automated mobility (CCAM), the loss of working hours (having direct impact on the European GDP), material

*ckyrkou@gmail.com,www.christoskyrkou.com

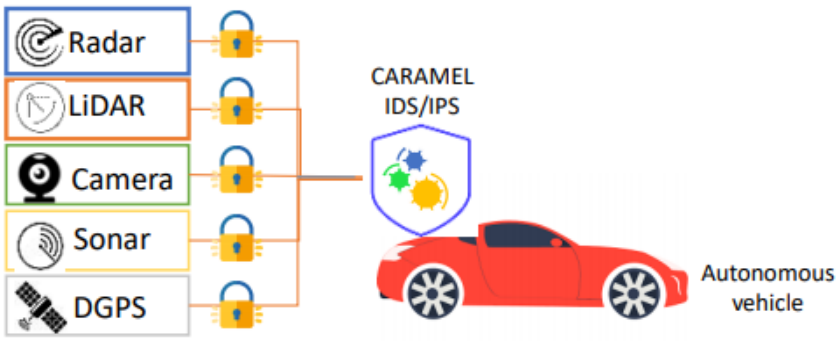

Figure 1: Sensors used by an autonomous vehicle to sense/perceive its environment.

damages, increased environmental pollution due e.g., to traffic jams or malicious modifications in sensors' firmware, and ultimately, the great danger for human lives, either they are drivers, passengers or pedestrians.

In order to achieve a secure vehicle development process with explicit cybersecurity considerations, well established methodologies coming from the information and communication (ICT) sector are required, allowing to assess vulnerabilities and potential cyberattack impacts and considering the entire supply chain of operations. Although past initiatives and cybersecurity endeveours related to the automotive industry have reached to security assurance frameworks for networked vehicles, several newly introduced technological dimensions like 5G, autopilots, and smart charging of Electric Vehicles (EVs) introduce cybersecurity gaps, not addressed satisfactorily yet. Recently, these systems have attracted increased attention within academia, and the academic community has begun to investigate the systems' robustness to various attacks. Recent studies [1],[11], [16] showed that advanced driver assistant systems (ADAS) can be compromised by applying adversarial machine learning techniques targeting the camera sensors and the underlying computer vision algorithms. Thus causing scene structural elements (e.g. traffic signs, pedestrians, etc.) to not be detected.

One of the objectives of CARAMEL is to demonstrate the use of AI/ML-based techniques in detection and possibly mitigation of dynamic cyber-attacks on the camera system/data in the context of automated driving systems. In many applications in mobility, particularly those addressed in CARAMEL, due to the diversity of the data streams the patterns of interest cannot be reliably classified by explicit programming. In such cases, if sufficiently large amounts of example (training) data are available or feasible to obtain, the most usual approach is to employ ML (typically using Neural Networks - NNs). This paper highlights the relevant cybersecurity challenges with regards to vision-based perception and functionality of autopilots and ADAS. Some ideas are then discussed on how they can be addressed with machine learning (ML) methods for detecting attacks and anomalies and how this can be realized through a dedicated anti-hacking device.

2 Attacks on Vision Sensors

The autopilots required for the CCAM are heavily dependent on computer vision and Artificial Intelligence (AI) techniques because a vehicle perceives visible data quite differently from a human (Fig. 1). Deep Learning (DL) teaches vehicles how to act in particular situations based on what is being detected. The potential for cyberattacks here arises when images are intentionally manipulated in a way that a human recognises them correctly, but a neural network or computer vision-based system misclassifies them. The use of such manipulated images opens the door to new attacks on autonomous vehicles and Advanced Driver-Assistance Systems (ADAS), requiring non-conventional cybersecurity techniques to detect and possibly to mitigate them. Notably, these attacks may sometimes not require any physical access to the vehicle or tampering with the communication system and their consequences can be critical. Examples of such attacks are shown in Fig. 2.

Adversarial attacks seek small perturbations of the input causing large errors in the estimation by the perception modality. Attacking perception functions using adversarial examples (AE) is a popular way to examine the reliability of learning approaches for data classification [14]. The key to all such attacks is that the change to the image should be minor yet have a large influence on the output. Adversarial examples typically involve small perturbations to the image that are not noticeable by the human eye. The adversaries are shown to work even when a single pixel is perturbed in the image [1]. Although these attacks reveal limitations of deep networks, they are hardly replicated in real-world settings. For instance, it is rather difficult to change a scene such that one pixel captured by a camera is perturbed in a specific

Figure 2: Attack scenarios on vehicle autopilot: (a) Lane lines removal can cause the car to steer in the wrong direction. (b) Adversarial patches can be use to mislead the visual detection module so that it does not detect pedestrians. (c) Adversarial attack on input image that causes the perception module to not detect the pedestrians.

way to fool the network. However, recent work on [10],[18] demonstrate that adversarial examples can also work when printed out and shown to the network under different illumination conditions. [2] shows that adversarial examples can be 3D printed and are misclassified by networks at different scales and orientations. [19] constructs adversarial glasses to fool facial recognition systems. [7] show that stop signs can be misclassified by placing various stickers on top of them.

Apart from the adversarial attacks, which involve scene modifications on the physical layer, within the Autonomous driving, the vulnerability of the Perception Engine is also an important issue to address. The Perception engine has to be secured against a variety of cyber-attacks at the sensors layer with the help of proper approaches for detecting the attacks and mitigating them.

Recently, it has been shown that it is possible to generate robust AEs to attack the state-of-the-art object detectors [21] and driving models [3] used in the real world. Attackers can succeed with these tactics even if they don't know the details of how the target neural net was constructed or trained. Adversarial tampering can be extremely subtle and hard to detect, even all the way down to pixel-level. Typically, two forms of attacks are considered. Hiding Attack (HA) which makes the perception module failing to recognize an entity in the environment, and Appearing Attack (AA), which makes the perception module misrecognize the AE as a different element specified by the attacker.

Attack scenarios on Autonomous Mobility can be classified into two big categories: physical adversarial attacks and attacks on the camera sensor data. In this section, we will present them in more detail. When referring to physical adversarial attacks we will consider attack scenarios where changes in the physical world will cause the cyber system in the autonomous vehicle to misbehave (as in TESLA accident). Such is the example of physically manipulating structural elements such as road lane line markings, or printable patterns designed to alter the behaviour of a perception algorithm. Camera sensor attacks refer to the scenario where an attacker manages to access critical vehicle systems and manipulate directly the camera image via modification through internal malicious software.

2.1 Scene Perception Model

Typically attacks target a scene perception model that is executed on the vehicle hardware. This can be a model that utilizes AI/ML to detect a specific environment object (road area, parking markings, vehicle, pedestrian, lane line, etc). Such models need to be trained on relevant data to then be able to recognize instances of objects in images. This is a tedious process and hence, usually exploits the process of transfer learning where some generic feature extractor neural networks are used as the backbone for building more complex perception systems. Two main scene perception models will be studied in CARAMEL:

- **Detection Models:** Methods that predict object bounding boxes and class probabilities at the same time.
- **Segmentation Models:** Semantic segmentation is a form of dense prediction task. With dense prediction we mean that we want to assign a label or in general make a prediction for each pixel in the image rather than a single label or some local predictions.

2.2 External Attack on Camera Sensor

Deep learning solutions are used in several autonomous vehicle subsystems in order to perform perception, sensor fusion, scene analysis, and path planning. State-of-the-art and human-competitive performance have been achieved by ML on many computer vision tasks related to autonomous vehicles [9].Over the last years it was demonstrated that ML solutions are vulnerable to certain visual attacks [20] that can cause the autonomous vehicles to misbehave in unexpected and potentially dangerous ways, for example on physical modification of the environment and especially traffic signs [17],[13]. It is considered that in these attacks, modifications are physically added to the objects themselves aiming to make the ML system fail but most humans would not consider it suspicious [8]. This is an attack on the physical layer. It assumes disturbance of the visual appearance of structural elements of the scene like the road lane lines and traffic signs. According to this attack, minor changes might be introduced, e.g.: printed patterns attached on an enviromental element in such a way that they might disturb the scene perception output (Fig. 2). The change could also involve deterioration in a multitude of appearance characteristics with regards to the appearance of lane/parking markings, e.g.: shape/length/colour. This attack should introduce minor changes in such a way that they should not be immediately noticeable to the human eye but that could end up altering the output of the scene perception engine.

2.3 Direct attack on Camera Sensor Data

Besides physical attacks that can induce erroneous cyber-system behaviour there is also the possibility that the camera data can be manipulated directly thus eliciting false algorithmic inferences. This can cause an AI-based perception module/controller to make incorrect decisions, such as when an autonomous vehicle fails to detect a lane/ parking marking and results to a departure from the correct lane. Camera sensor attacks refer to the scenario where an attacker manages to access the vehicles' critical systems and installs and activates some malicious software that distorts the captures camera data.

2.3.1 Adversarial attacks

The last few years there has been a growing concern on the cyber-security of perception modules for object localization such as object detectors and object segmentation [12], because deep neural networks are known to be vulnerable to adversarial examples (AEs) Adversarial attacks can be performed on the machine vision algorithm and video/image processing algorithm used for object detection (road, obstacles, road signs, etc.) by altering the image captured by the camera [12]. Most of them rely on the propagation of gradient signals in order to find the right perturbations that can result in an error in the output from the perception model. Autonomous vehicles developed nowadays lack robustness to adversarial conditions.

There are two classes of adversarial examples, non-targeted and targeted ones. In non-targeted attacks the goal is to generate adversarial examples for which the model's prediction is any label other than the ground-truth label. A common approach for this is the fast gradient sign method (FGSM) which produces adversarial examples by increasing the loss function, which is often the cross-entropy loss, of the network on the input x as shown in Eq. 1. In targeted adversarial attacks, we seek adversarial images that can change the prediction of a model to a specific target label. This can be reformulated in an alternative approach as to maximize probability of some specific target class which is unlikely to be the true class for a given image as shown in Eq. 2. Iterative methods apply the FGSM attack multiple times over the image.

$$\hat{x} = x + \epsilon \bullet sign(\nabla_x J(\theta, x, y_{true})) \tag{1}$$

$$\hat{x} = x - \epsilon \bullet sign(\nabla_x J(\theta, x, y_{target})) \tag{2}$$

where $\nabla_x J$ the gradient of the models loss function with respect to the original input pixel vector x,y_{true} is the true label vector for x, y_{target} is the target attack label, and θ is the model parameter vector.This can also be generalized to semantic segmentation problems where networks are trained with an independent cross-entropy loss at each pixel.

2.3.2 Image deterioration attacks

Image deterioration attacks aim at altering the input image in order to cause the vehicles perception modules to fail. In constrast to adversarial examples these attacks are not guided by a target label but the objective is to degrade the quality of the input image so that the perception module's output becomes erroneous. These attacks are rather simple and do

not require from an attacker to have knowledge or access to any model or infrastructure and thus can be considered as more common. Deterioration techniques include adding artifacts such as lines, or adding noise to the measured data (gaussian, salt and pepper, etc.), or blanking out image regions.

3 Impact of Attacks

In this section we attempt to demonstrate the impact that internal attacks on the camera sensor data can have on performance of a perception module. For this a simulation framework has been developed and some preliminary experiments were carried out to assess the robustness of existing approaches. Specifically, we carry out two different experiments, first we examine the impact of the deterioration attacks using some available pretrained models for semantic segmentation with different backbones to asses the general robustness of the models. Then we train a model based on the backbone that had the best trade off between deterioration efficiency and computational complexity and evaluate its performance against adversarial attacks.

3.1 Evaluation Metrics

For deterioration attacks we use the overall pixel accuracy (Eq. 3), which is the percentage of pixels in the image which were correctly classified, since noise is added randomly across the image. For the adversarial attacks we use the intersction-over-union (IoU) metric (Eq. 4) which is the per class mean of the intersection of the *target* and *predicted* binary segmentation maps for a given class, divided by their union.

$$PixelAccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{4}$$

3.2 Simulation Framework

CARLA [6], an open-source realistic driving simulator, provides a convenient way to train autonomous driving agents with a variety of different sensor suites (cameras, LIDAR, depth, etc.), and test them in realistic traffic scenarios. We employ this simulator to craft different perturbation attacks and test them for the problem of image segmentation with deep learning. Specifically, we measure how different noise models applied at different strength levels affect the overall pixel-level classification accuracy.

We use CARLA to collect data from camera sensors attached to specific locations of an autonomous vehicle. The sensor properties are derived with regards to the specifications of each model that we use in the experiments. The reproducibility of the CARLA simulator and the variety of environmental conditions it supports, enables us to better understand the effect of different attacks and accordingly build different defence mechanisms.

Our initial effort in creating these attacks and evaluating them under realistic scenarios is to understand whether adversarial examples and other forms of image quality degradation pose a real threat to autonomous vehicles. Our plan is to also extend the CARLA simulator for simulating adversarial attacks and perturbations as well as any developed mitigation mechanisms.

3.3 Impact of Deterioration Attacks

In this section we demonstrate the effect of attacking on the camera sensor data by adding noise and artifacts. Such attacks are easy to be added as the attacker does not need to have any information regarding the underlying algorithm and models and can lead to erroneous output of a scene perception module. First, we assume that the goal of the attacker is not to completely fail the detection-based system but to lead the detection model to erroneous behaviour with potentially catastrophic results.

To study the effect of noise addition attacks on the image segmentation task we train different segmentation models using [22] with different backbone models. The backbone refers to the network which takes as input the image and extracts the feature map upon which the rest of the network will build to extract the segmentation map. This will also allow us to see if different backbone networks are affected differently from the random noise attack. Specifically, for backbones we use the Resnet50,Resnet101, and MobileNetV2 networks. Figure 3 shows the pixel-level accuracy for

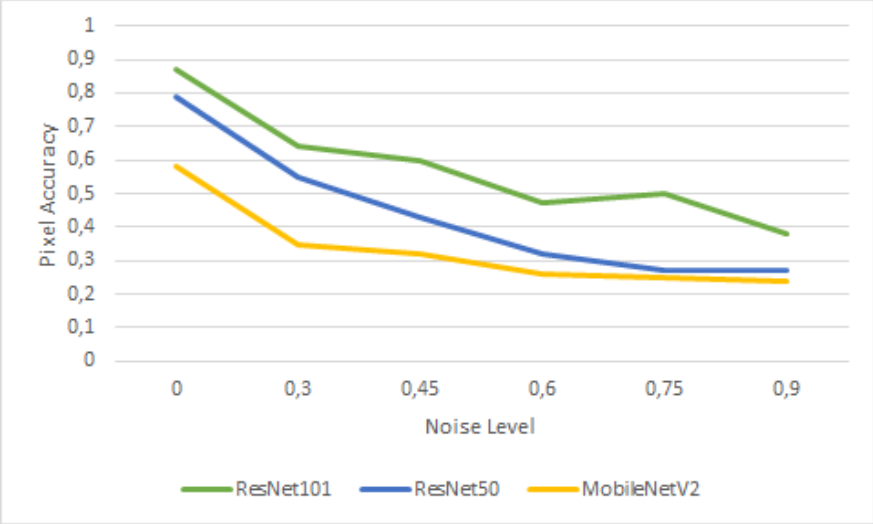

Figure 3: Accuracy drop for different backbone models. Performance drop is observed across all models.

Table 1: Effect of untargeted and iterative targeted FGSM attack on IoU

ϵ	**Untargeted FGSM attack**	**Iterative Targeted FGSM attack**
1	53%	49%
12	35%	18%

each backbone for different levels of noise. The larger ResNet101 model shows higher accuracy with respect to the other two backbone models. However, its accuracy is still almost reduced to half at the highest noise level which makes it unusable for autonomous vehicle applications. Fig. 4 shows an example of how the semantic segmentation map prediction changes with the introduction of noise.

3.4 Impact of Adversarial Attacks

In this section we examine two classes of adversarial examples, Untargeted FGSM attack and terative Targeted FGSM attack. We want to generate adversarial examples for which the model's prediction is changed. In order to demonstrate these vulnerabilities, we trained a state of art deep neural network architecture, Deeplab V3 [4] with Resnet50 backbone based on the previous experiment and on data from CARLA simulator.

In table 1 we provide statistics of how the performance of Deeplab trained in CARLA, got affected by untargeted and iterative targeted FGSM attack. For evaluation we use the IoU score which quantifies the percent overlap between the target mask and the predicted output. It is observed that as the strength of the adversarial attack ϵ increases the IoU metric decreases. In Fig. 5 we can see that this can cause the network to miss important objects in the scene such as vehicles and traffic lights.

It is clear from both forms of attacks that the performance of a computer vision model can be compromised. This provides an attacker with a wide range of possible attacks that can be deployed depending on the capabilities of the adversary. It is critical then to develop solutions to detect such attacks and mitigate their negative effects.

4 Potential Defence/Mitigation Strategies

Existing pre-processing approaches as shown in Fig. 6, such as filtering (bilateral or gaussian) even though effective against specific attack types may fail to completely remove artifacts and will also distort non-attacked images. Hence, the goal of CARAMEL is to improve the robustness of the environment perception module through the use of AI/ML models to first detect and then reduce the impact of the attack.

To make Autonomous Vehicles less vulnerable to any attacker, it is necessary to develop adversarial robust ML/DL solutions. Adversarial training or input reconstruction where adversarial samples will be cleaned to transform them

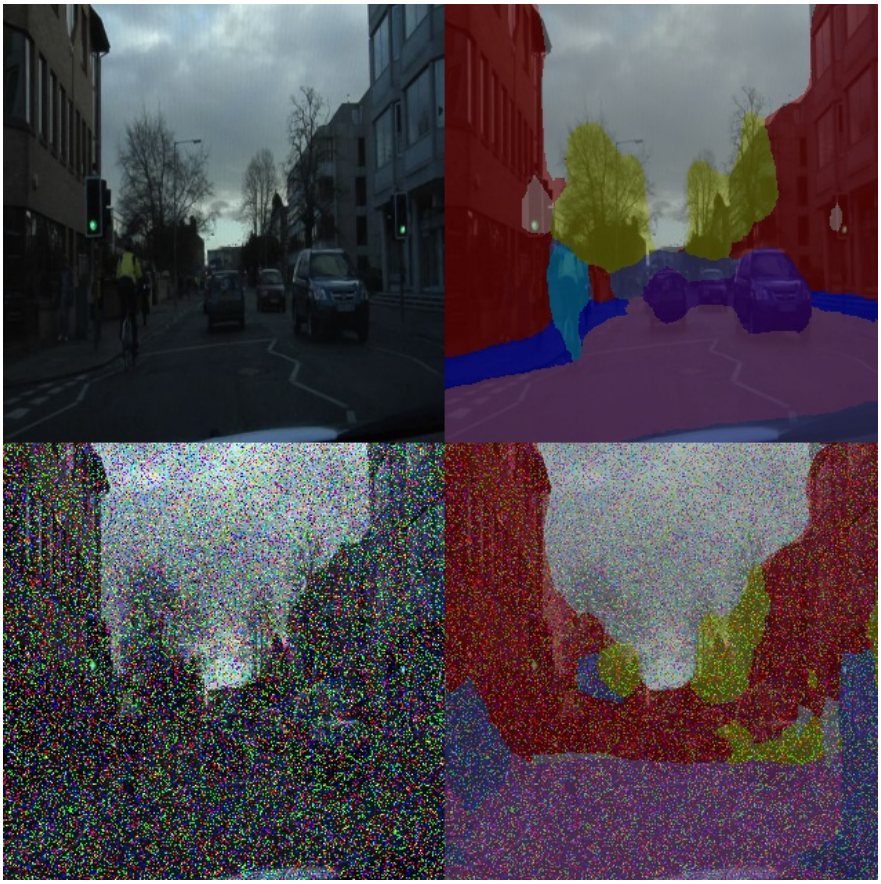

Figure 4: Effect of noise attack on segmentation output. Notice, that the objects are not localized well and objects such as cyclists are completely lost.

back to legitimate ones will lead to more robust methods. Different AI/ML techniques will be investigated to Improve the robustness of the environment perception module and reduce the impact of the attack. Specifically, deep learning techniques such as encoder-decoder networks will be trained to recover image information and mitigate the effect of attacks. Other solutions that can be applied at run-time would be to apply compression techniques such as JPEG to reduce the impact of noise, or transform the image through quantization to lessen the impact of the adversarial attack.

Additionally, combining multiple defence strategies can alleviate some of the adversarial perturbations. Various defences have been proposed to mitigate the effect of adversarial attacks. These defences can be grouped under three different approaches:

- Modifying the training data to make the classifier more robust against attacks, e.g., adversarial training which augments the training data of the classifier with adversarial examples.
- Modifying the training procedure of the classifier to reduce the magnitude of gradients, e.g., defensive distillation.
- Attempting to remove the adversarial noise from the input samples based on the concept that correctly classified examples tend to have greater maximum probabilities than erroneously classified and out-of-distribution examples.

Besides the presented defence mechanisms which rely on ML solutions developed using a single data source, there are possibilities to employ ML solutions leveraging multiple data sources. Multi-sensor data fusion is the process of combining observations from a number of different sensors to provide a robust and complete description of an environment or process of interest. Data fusion finds wide application in many areas of autonomous vehicles such as object recognition, environment perception, road detection, etc. Current approaches for multiple tasks related to autonomous vehicles use either cameras, LIDAR, Radar or other sensors. Cameras can work at high framerates and provide dense information over a long range under good illumination and fair weather. However, being passive sensors,

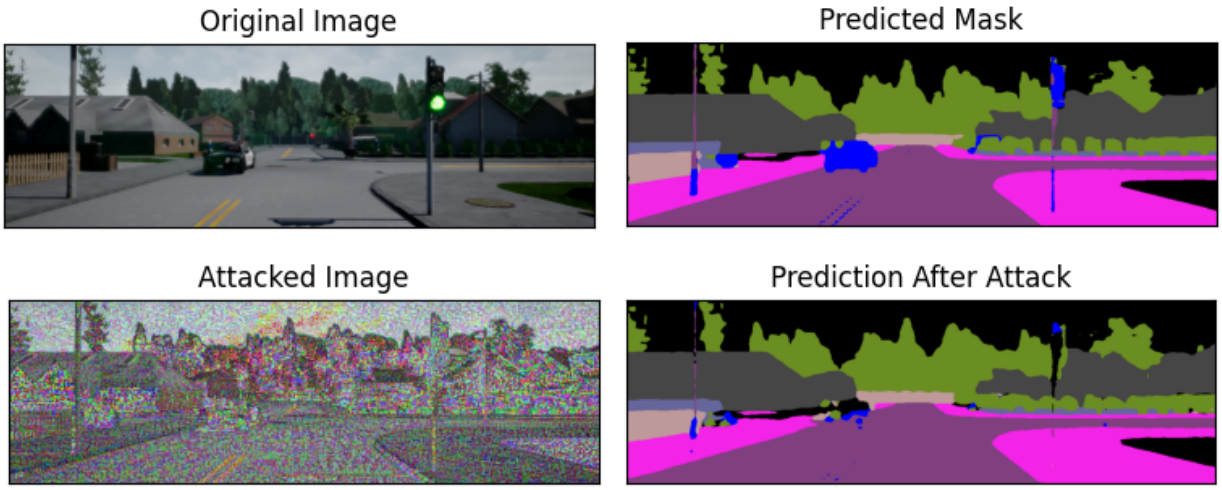

Figure 5: Impact of adversarial attack on segmentation prediction. Notice that for the prediction after the attack, hides important elements such as car and traffic lights are removed.

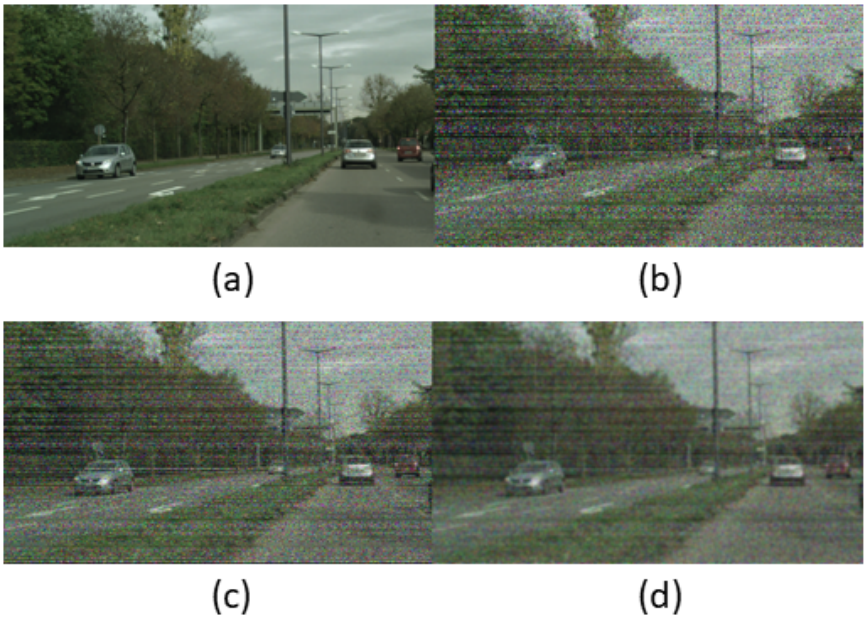

Figure 6: Existing methods for removing noise: (a) Input Image from Cityscapes dataset. (b) Noisy image. (c) Result of applying bilateral filter. (d) Result of applying Gaussian smoothing.

they are strongly affected by the level of illumination. LIDARs sense the environment by using their own emitted pulses of laser light and therefore they are only marginally affected by the external lighting conditions. Furthermore, they provide accurate distance measurements. Based on this description of benefits and drawbacks of these two sensor types, it is easy to see that using multiple sensors might provide an improved overall reliability. An attack can change the confidence with which a model makes a prediction. Having another modality that views the same scene will allow for using that output to detect potential discrepancies [5].

5 Anti-Hacking Device

CARAMEL will bring all the algorithmic detection solutions into an embedded anti-hacking device that will be capable for passive detection of attacks on the vehicles visual perception modules. The anti-hacking device will be a physical

controller that will be able to be integrated into an autonomous vehicle and will be an integral part of the project innovation. Its task is to run dedicated ML models that work on the sensor data to detect anomalies that might point to malicious attacks and disable higher level functions in case of a cyberattack. The anti-hacking device will be connected to the busses of the autonomous vehicle carrying the sensor data. It will then passively monitor the bus traffic (e.g. CAN bus frames) and extract the raw sensor data. From there the data will be pre-filtered and aggregated to make it suitable for the machine learning stage to detect threats and attacks. Any security-relevant events are then forwarded to the visualization and mitigation components in the car. To provide added value the device should be able to deal with the data stream(s) at short time-scales with respect to typical use-cases for timely warnings. As such, the anti-hacking device will be based machine learning (ML) hardware such as the Coral Dev Board the Tensorflow Lite Processing Unit (TPU). To enable timely detection of attacks and efficient deployment the project will explore techniques for compression and acceleration of the attack detection/robustification models.

6 Conclusions

CARAMEL will address major challenges and cyber-security concerns affecting the wide adoption of autonomous vehicles. Specifically, through the use of AI/ML, new algorithms and methodologies will be developed capable of assessing whether a vision sensor is under attack or its data cannot be trusted and will also then mitigate this effect. The integration of these algorithms on an anti-hacking device can potentially facilitate the adoption of such defence technologies by OEMs paving the way for the broader adoption of autonomous vehicle technology.

Acknowledgment

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 833611 (CARAMEL). The work of University of Cyprus was also supported by the European Unions Horizon 2020 research and innovation programme under grant agreement No 739551 (KIOS CoE) and from the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development. The work of University of Patras has also received funding from the European Union's Horizon 2020 research and innovation programme CONCORDIA under grant agreement No 830927

References

[1] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[2] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *ArXiv*, abs/1802.00420, 2018.

[3] Adith Boloor, Xin He, Christopher D. Gill, Yevgeniy Vorobeychik, and Xuan Zhang. Simple physical adversarial examples against end-to-end autonomous driving models. *2019 IEEE International Conference on Embedded Software and Systems (ICESS)*, pages 1–7, 2019.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[5] Cory Cornelius, Shang-Tse Chen, Jason Martin, and Duen Horng Chau. Talk proposal: Towards the realistic evaluation of evasion attacks using CARLA. *CoRR*, abs/1904.12622, 2019.

[6] Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. CARLA: an open urban driving simulator. *CoRR*, abs/1711.03938, 2017.

[7] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *CoRR*, abs/1707.08945, 2017.

[8] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *CoRR*, abs/1707.08945, 2017.

[9] R. Hussain and S. Zeadally. Autonomous cars: Research results, issues, and future challenges. *IEEE Communications Surveys Tutorials*, 21(2):1275–1313, 2019.

[10] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Using depth for pixel-wise detection of adversarial attacks in crowd counting. *ArXiv*, abs/1911.11484, 2019.

[11] Jiajun Lu, Hussein Sibai, Evan Fabry, and David A. Forsyth. Standard detectors aren't (currently) fooled by physical adversarial stop signs. *CoRR*, abs/1710.03337, 2017.

[12] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2774–2783, 2017.

[13] Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, and Yuval Weisglass. Fooling a real car with adversarial traffic signs. *CoRR*, abs/1907.00374, 2019.

[14] Dudi Nassi, Raz Ben-Netanel, Yuval Elovici, and Ben Nassi. Mobilbye: Attacking ADAS with camera spoofing. *CoRR*, abs/1906.09765, 2019.

[15] S. Parkinson, P. Ward, K. Wilson, and J. Miller. Cyber threats facing autonomous and connected vehicles: Future challenges. *IEEE Transactions on Intelligent Transportation Systems*, 18(11):2898–2915, Nov 2017.

[16] Jonathan Petit, Bas Stottelaar, and Michael Feiri. Remote attacks on automated vehicles sensors : Experiments on camera and lidar. 2015.

[17] Adnan Qayyum, Muhammad Usama, Junaid Qadir, and Ala I. Al-Fuqaha. Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward. *CoRR*, abs/1905.12762, 2019.

[18] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J. Black. Attacking optical flow. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2404–2413, 2019.

[19] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *CoRR*, abs/1801.00349, 2018.

[20] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. DARTS: deceiving autonomous cars with toxic signs. *CoRR*, abs/1802.06430, 2018.

[21] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. *CoRR*, abs/1904.08653, 2019.

[22] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018.