CARAMEL

Autonomous Vehicles: Cyber-attacks Detection & Mitigation

Petros Kapsalas (Panasonic Automotive)

Caramel Workshop

15th November 2021

Objectives

Cyberthreat Detection and Response Techniques for Autonomous Automated Vehicles

❑ GOAL:

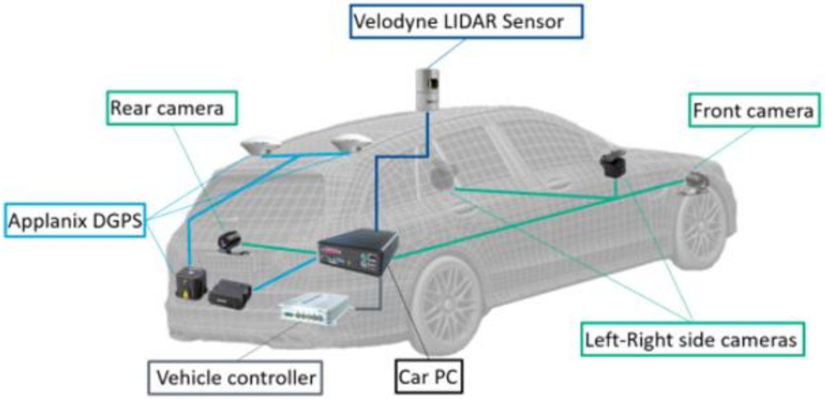
- Design and Implementation of detection and response techniques focused on Deep Learning the multimodal fusion of the available sensor data using both sparse and deep priors.
- Evaluation of the impact of noise mitigation on two different layers: 1) viewing layer, 2) perception layer.
- Deep learning technologies can be used for cyber threat intelligence in anticipation of cyberattacks to identify malicious activity trends and correlating them with attackers' information, tools and techniques.

❑ Data features:

- Multi-modal data acquired by different systems such as LiDAR, mono and stereo-vision cameras, USS, GPS, etc.
- Heterogeneous data: different sensors of the same modality may have significantly different specifications, e.g., different resolutions, different sampling rate etc.

❑ Data treatment:

- Handling missing, noisy or inconsistent data
- Identifying and/or removing outlier/false injected data
- Dealing with contaminated/noisy data segments

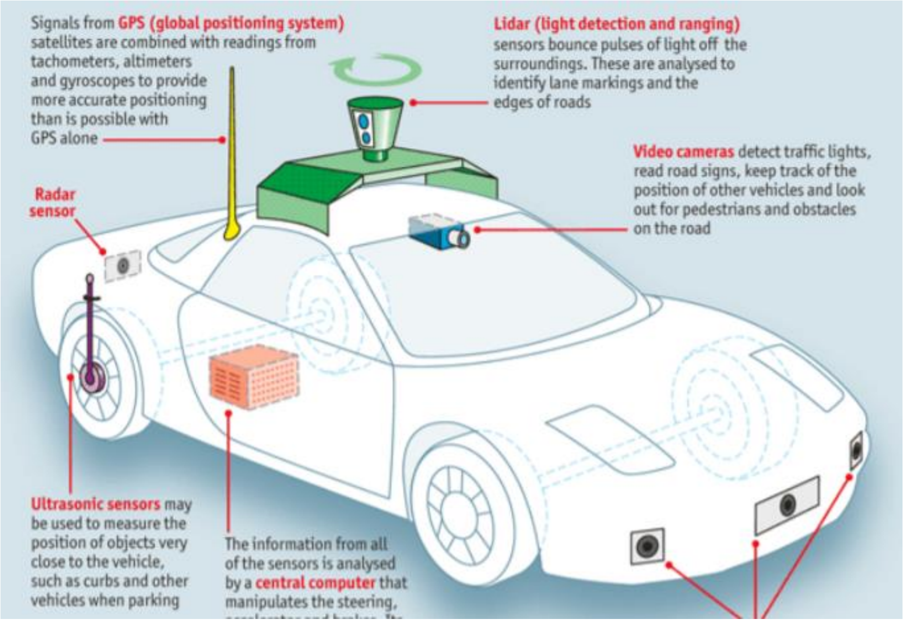


Expected Outcomes

- PKI-enabled Vehicle Identity Management System
- Solution against cyberattacks on machine vision systems
- Situational Awareness Solution
- Fallback actions for Minimal Risk

Basic Physical Ecosystem of Autonomous Car

- Global Positioning System (GPS).
- Light Detection and Ranging (LIDAR)
- Cameras (Video).
- Ultrasonic Sensors
- Central Computer
- Dedicated Short-Range Communications-Based Receiver V2X.
- Differential GPS (DGPS)



Signals from **GPS (global positioning system)** satellites are combined with readings from tachometers, altimeters and gyroscopes to provide more accurate positioning than is possible with GPS alone

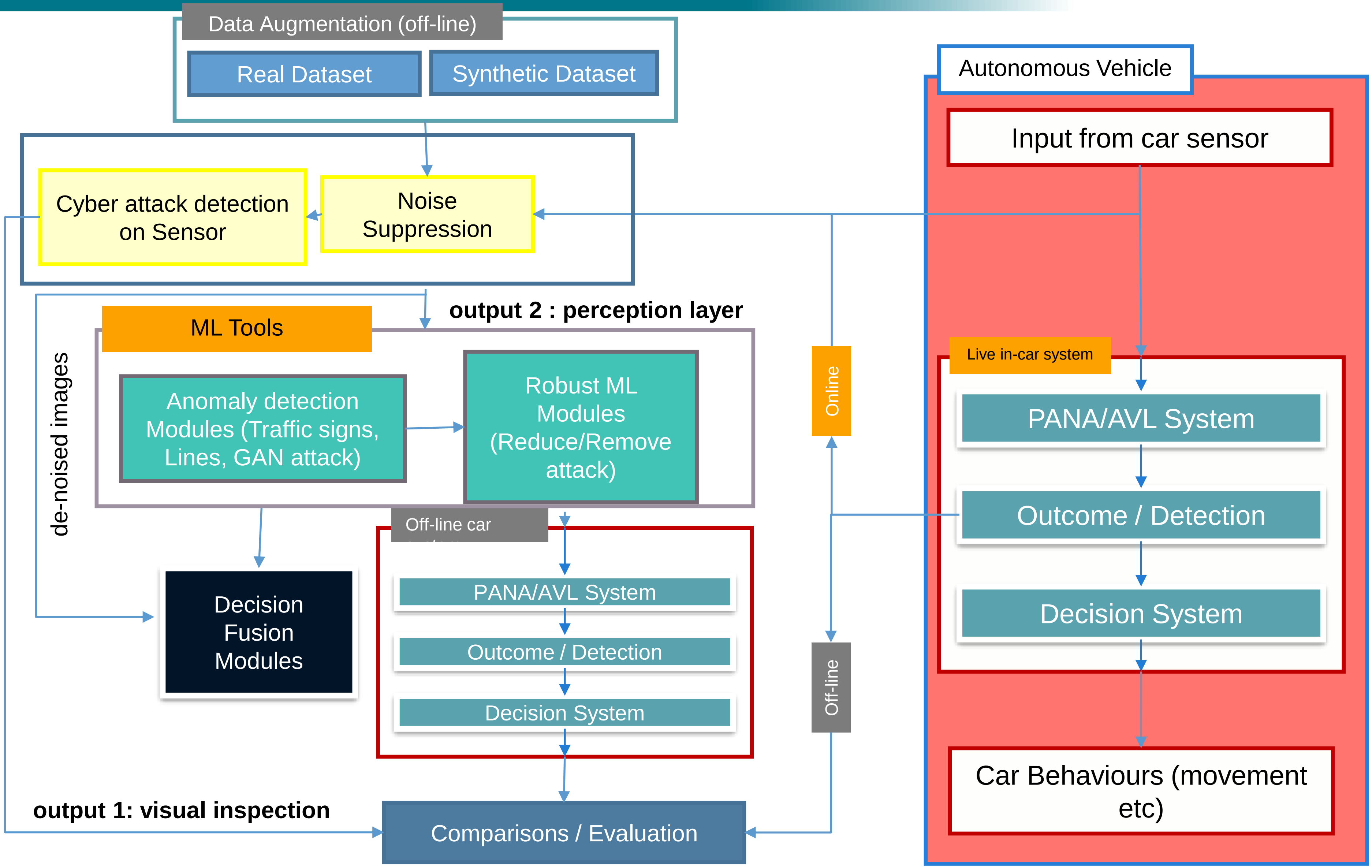
Lidar (light detection and ranging) sensors bounce pulses of light off the surroundings. These are analysed to identify lane markings and the edges of roads

Video cameras detect traffic lights, read road signs, keep track of the position of other vehicles and look out for pedestrians and obstacles on the road

Ultrasonic sensors may be used to measure the position of objects very close to the vehicle, such as curbs and other vehicles when parking

The information from all of the sensors is analysed by a **central computer** that manipulates the steering, accelerator and brakes. Its

CARAMEL Architecture



Data Augmentation (off-line)

Real Dataset

Synthetic Dataset

Cyber attack detection on Sensor

Noise Suppression

ML Tools

output 2 : perception layer

Anomaly detection Modules (Traffic signs, Lines, GAN attack)

Robust ML Modules (Reduce/Remove attack)

de-noised images

Decision Fusion Modules

Off-line car

PANA/AVL System

Outcome / Detection

Decision System

output 1: visual inspection

Comparisons / Evaluation

Online

Off-line

Autonomous Vehicle

Input from car sensor

Live in-car system

PANA/AVL System

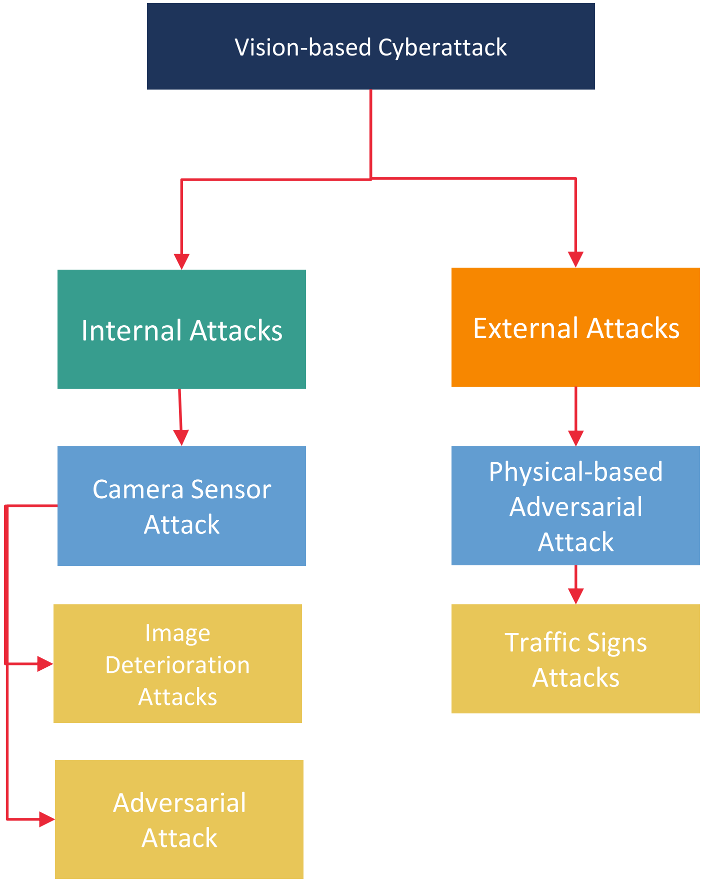
Outcome / Detection

Decision System

Car Behaviours (movement etc)

Overview

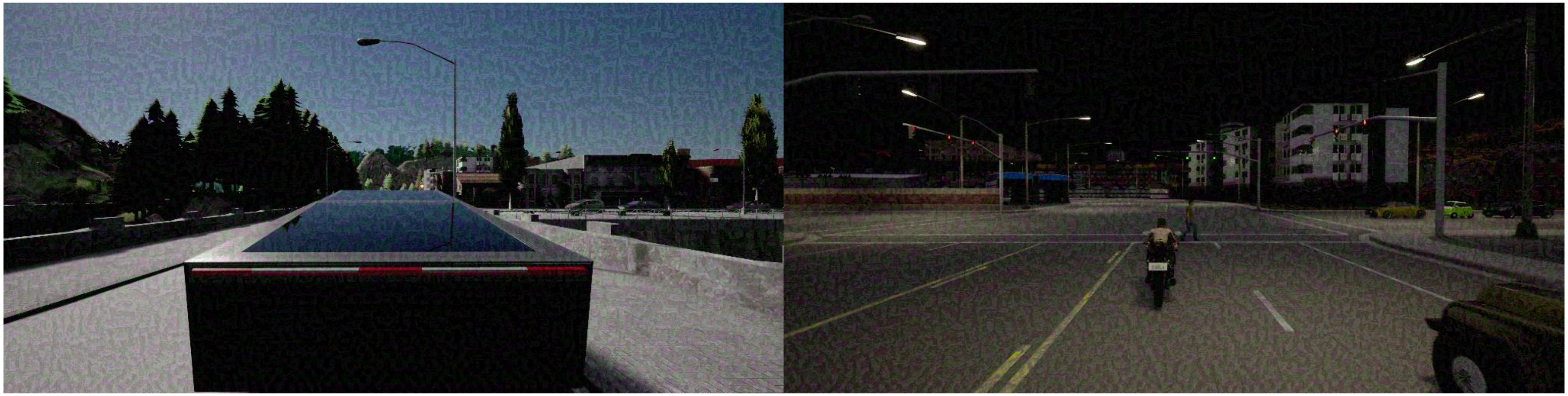
- ❑ **Vison-based Cyberattacks**
  - **Types of Attacks**
    - ➢ Internal Attacks
    - ➢ External Attacks
- ❑ **Includes 3 use case scenarios:**
  - Physical-based Adversarial Attack
    - ➢ Attacks on the external environment
    - ➢ Examined the attacks on Traffic signs.
  - Camera Sensor Attack
    - ➢ Attacks on the internal component of the autonomous vehicle
    - ➢ Focused on camera image deterioration attack
  - Camera Sensor Attack – Adversarial Attack
    - ➢ Investigated the attacks on the internal component
    - ➢ Examined Deep Learning adversarial attack on camera sensor



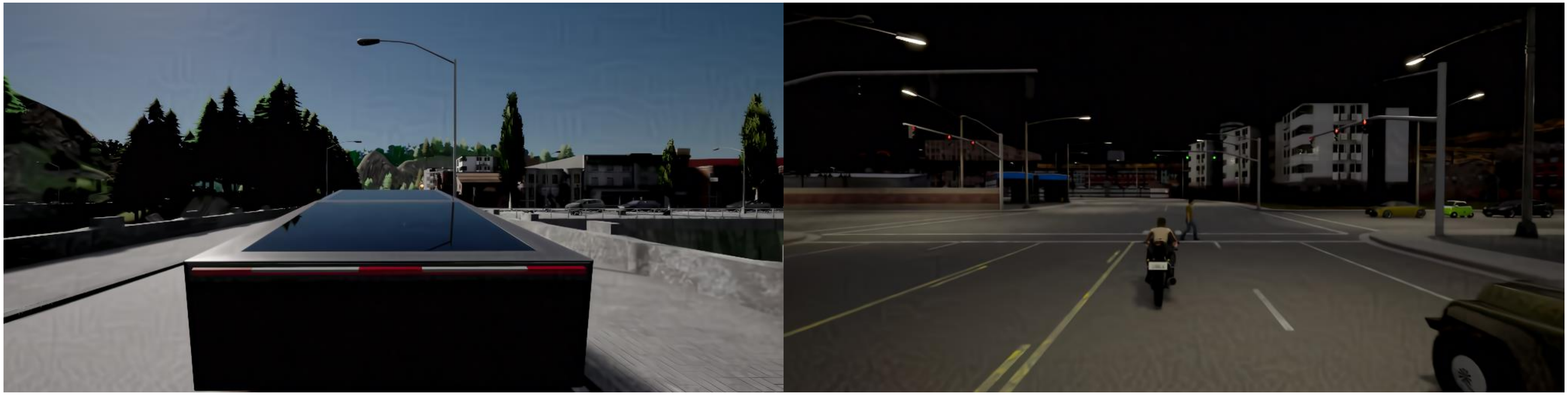
Inspecting Results on the Viewing layer

Adversarial Noise Suppression using Deep Restoration Network

Attacked images



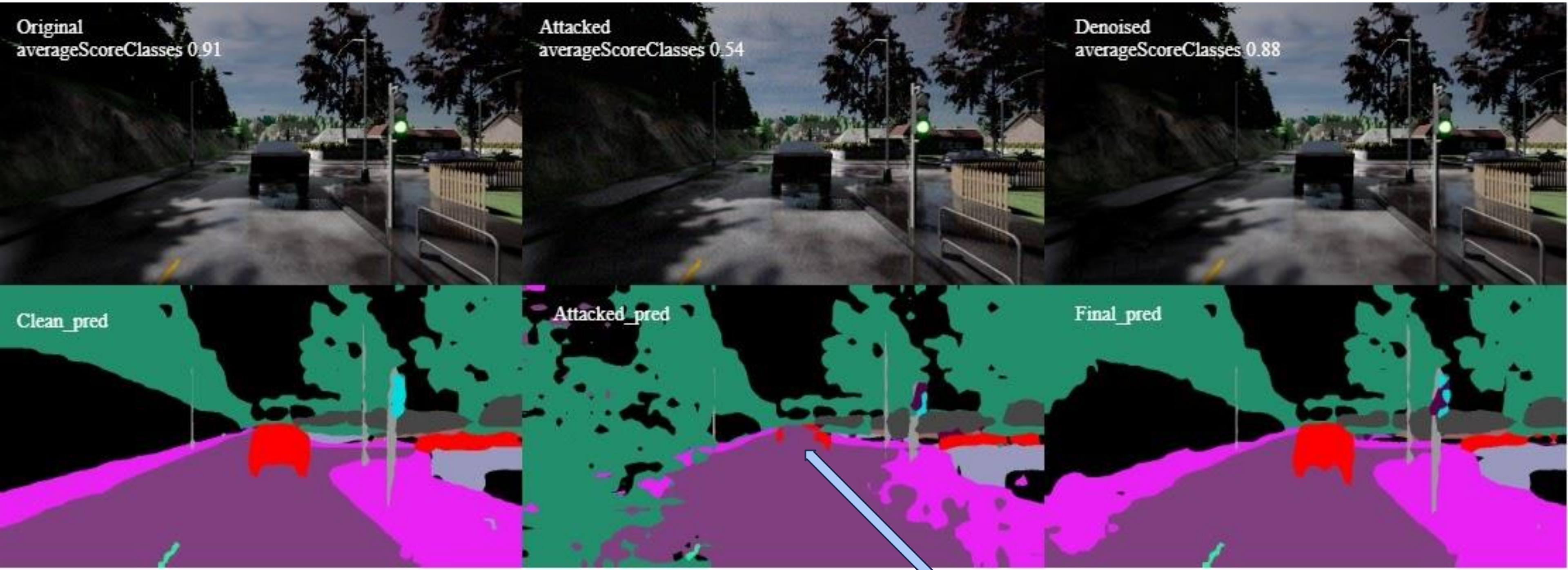
Denoised images



Inspecting Results on the Perception layer

Adversarial Noise Suppression using Deep Restoration network – Examples

- **Effect of FFGSM Adversarial Attack on the segmentation module**



Original rgb image /
Original segmentation output

Attacked rgb image /
Attacked segmentation output

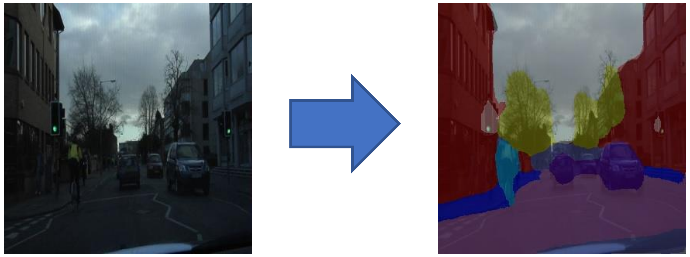
Denoised rgb image /
Denoised segmentation output

The car is hidden from the perception engine

Inspecting Results on the Perception layer

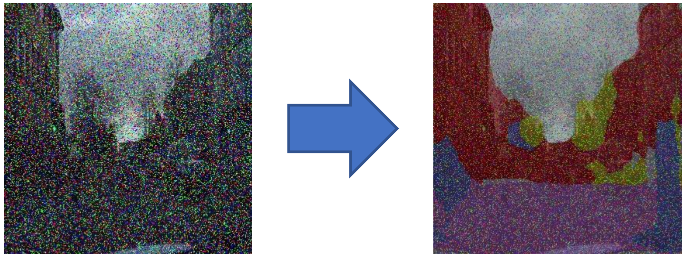
Mitigating Image Noise Attacks with Deep Learning

- **Problem:** Images attacked with noise can have detrimental effect on perception modules such as semantic segmentation



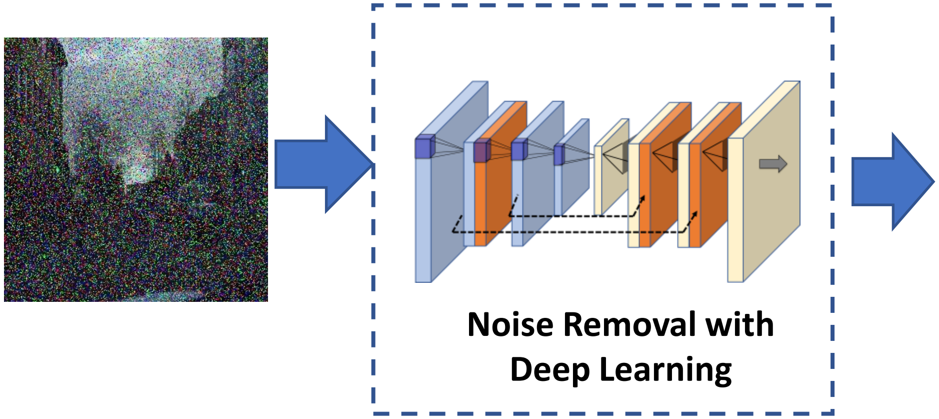
Normal segmentation output from original image

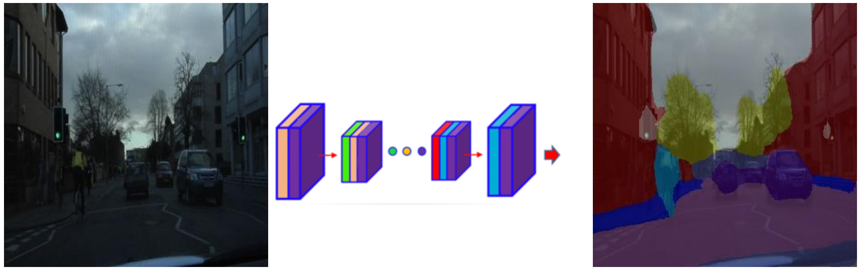
Vs.



Effect of attacked image on segmentation output

- **Solution:** Train deep convolutional neural network to reconstruct images attacked with noise/artifacts
  - The convolutional autoencoder learns to restore the image and remove any noise and artifacts
  - As a result, perception modules such as image segmentation can be robustified.





Improved Segmentation Outcome

Inspecting Results on the Perception layer

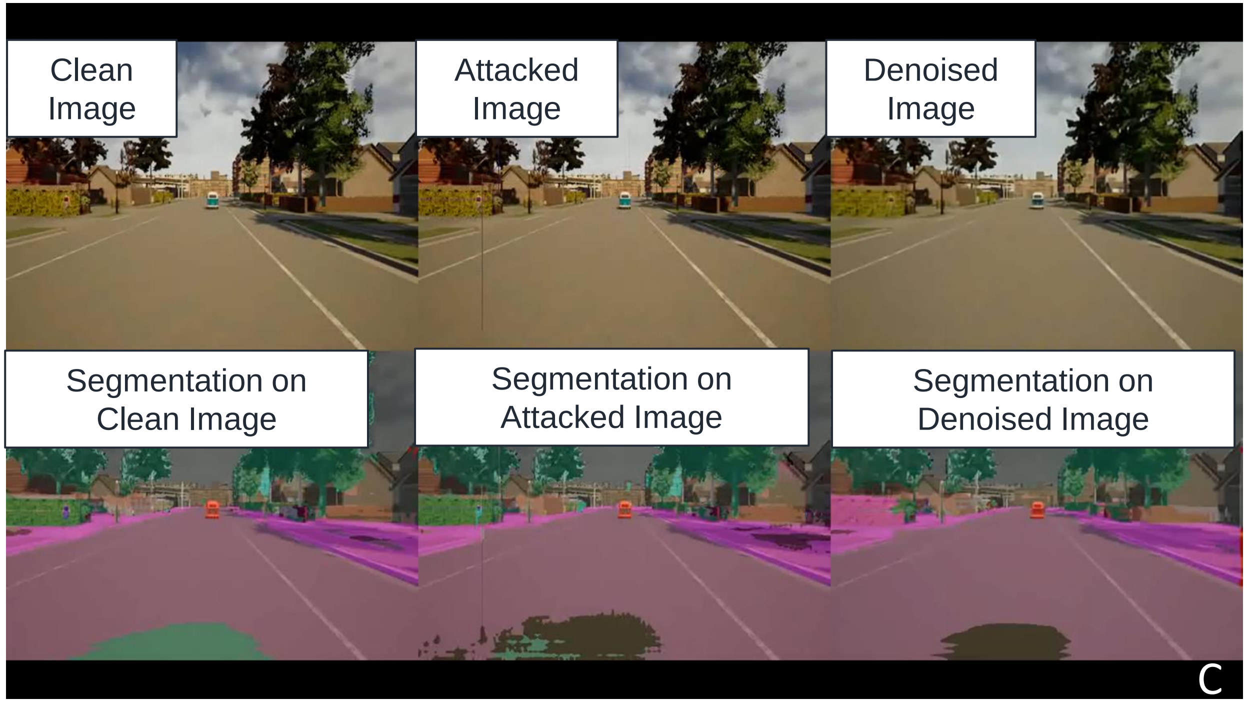
Mitigating Image Noise Attacks with Deep Learning

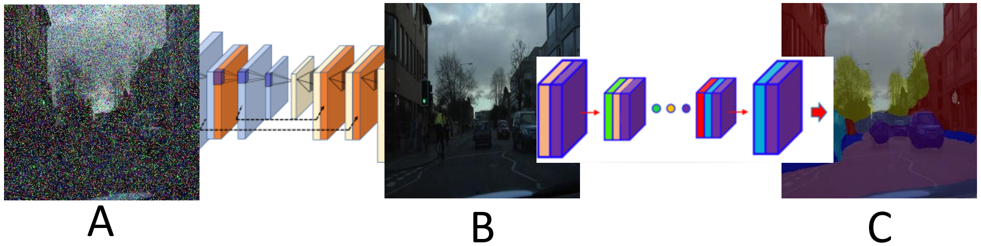
A. The attack is performed stochastically

B. The image is passed through the convolutional autoencoder for image quality restoration.

C. As a result, the segmentation output is improved compared to the attacked version

Visual Results on Synthetic Data from CARLA simulator


Clean Image
Attacked Image
Denoised Image
Segmentation on Clean Image
Segmentation on Attacked Image
Segmentation on Denoised Image
C

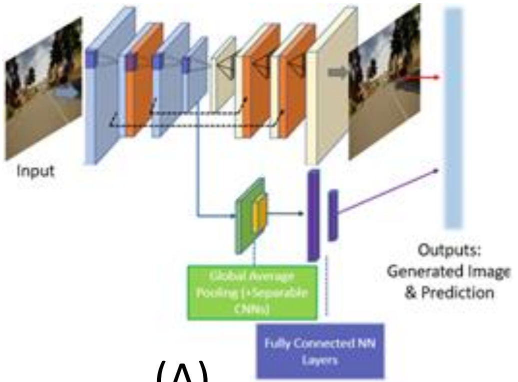


Detection of Attacks on Camera Sensor

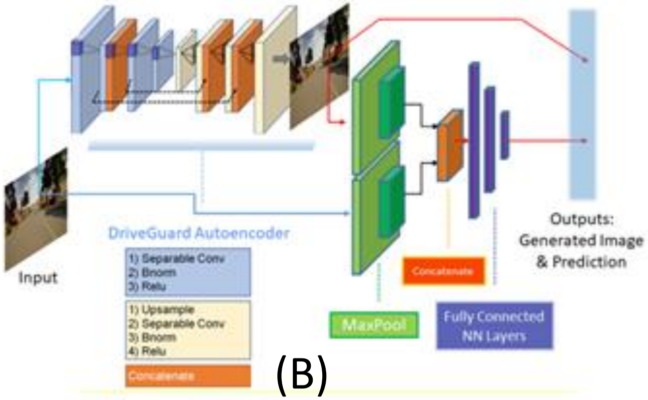
Methods based on Deep Learning

- We aim for a prediction process of whether the input image is attacked, done in parallel to our model's mitigation process for segmentation.
  - The detector uses the output of a convolutional neural network autoencoder, (DriveGuard) that attempts to reconstruct a distorted image
  - When an attack is detected the output of the process is the mitigated reconstructed Image.
  - In the opposite scenario the output is the original input image.
- Different detection approaches:
  A. Parallel Detection Stream (PDS) using bottleneck features
  B. Two-Stage Extension with Fully Connected (TSFC) Neural Networks
  C. Extract similarity measures from the two images and apply machine learning algorithms

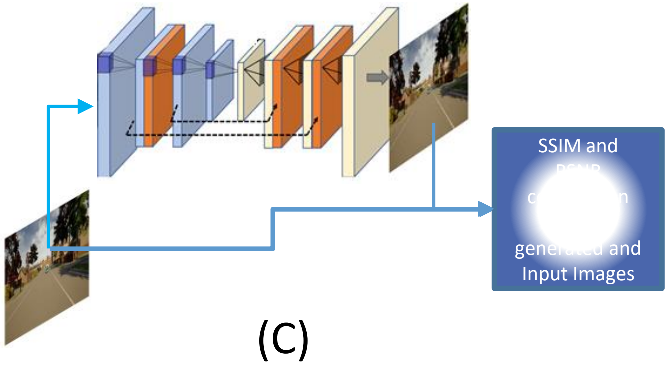
Detector Approach	Total Accuracy	FPS
DriveGuard + Two Stage Fully Connected (TSFC)	0.9859	63.69
DriveGuard + Parallel Detection Stream (PDS)	0.989	72.74
DriveGuard + SVM	0.9992	39.91
DriveGuard + `Linear Classifier	0.9745	40.31


Input
Global Average Pooling (+Seperable CNNs)
Fully Connected NN Layers
Outputs: Generated Image & Prediction

(A)

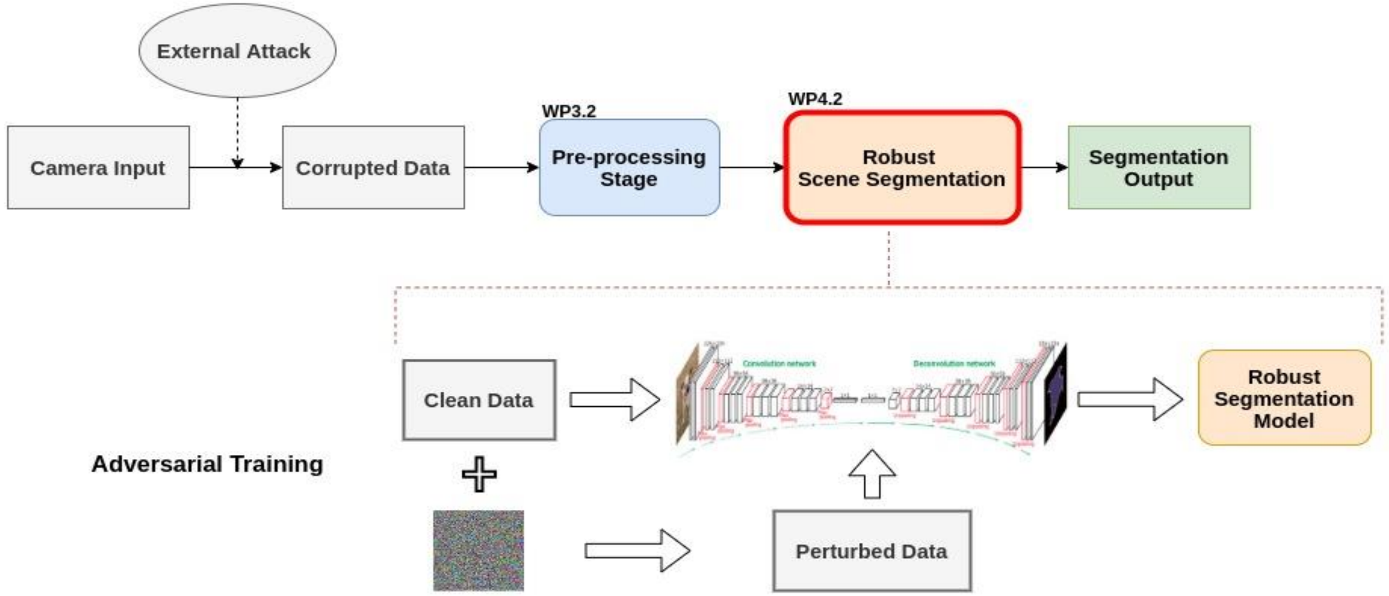


(B)



(C)

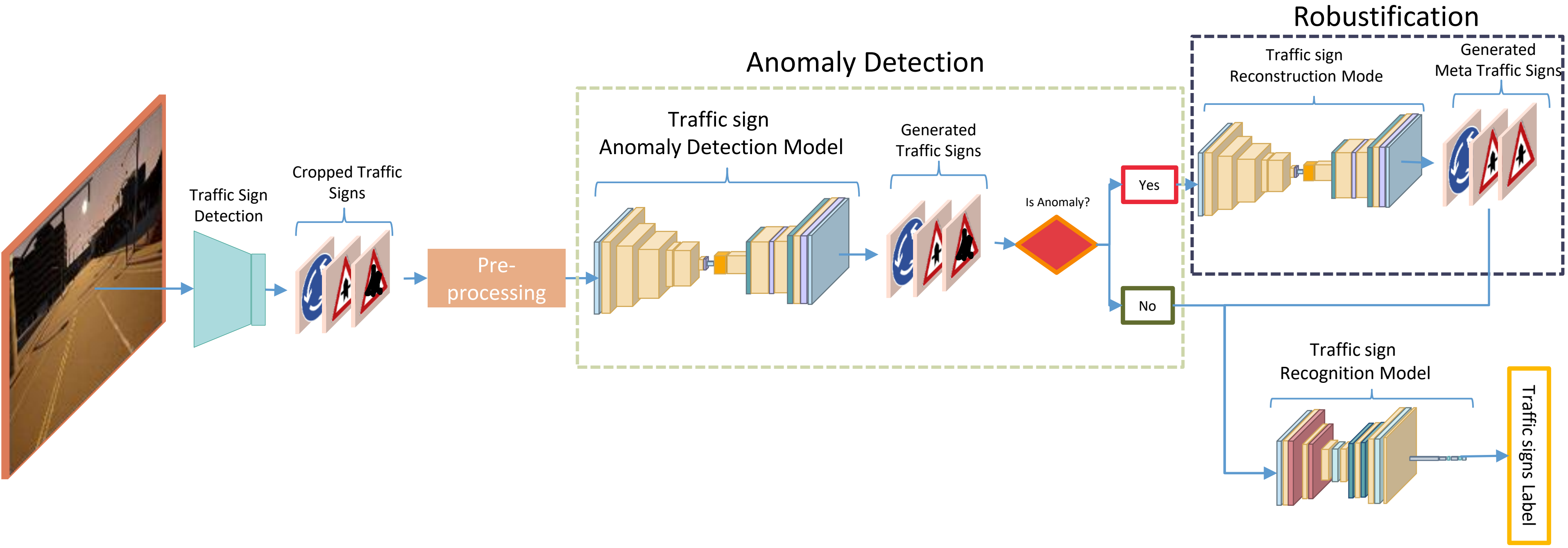
Robust model for adversarial attacks on camera sensors



Robust Environment Perception:

- Implement state-of-the-art adversarial attacks
- Generate image dataset with adversarial attacks
- Robustify 2d segmentation model dedicated to the prevention of attacks to scene analysis

Physical-based Adversarial Attack



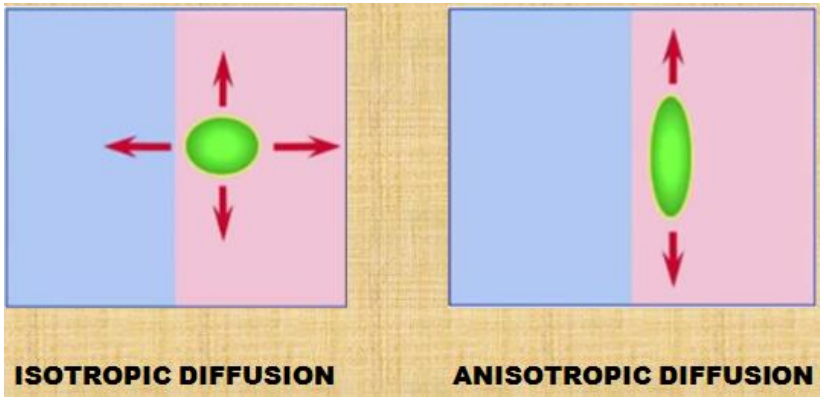
- ❑ Use case considers various attacks on traffic signs, such as pattern, noise, and graffiti attacks
- ❑ Proposed a robust pipeline using Deep Learning approach
  - Developed Deep Learning models
    - ➢ Traffic signs Anomaly Detection models
    - ➢ Traffic signs Reconstruction models
    - ➢ Traffic sign Recognition models
- ❑ Various Synthetic datasets were generated for the purpose

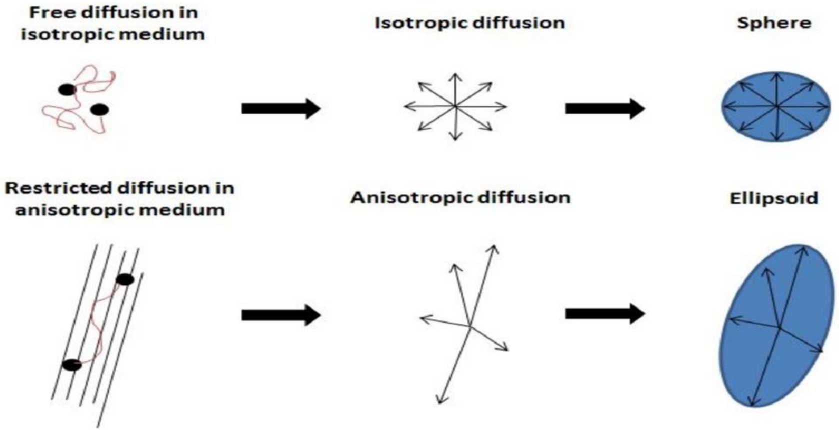
Performance of our models on Nvidia RTX 2080

Type	Model Accuracy	FPS
Anomaly Detection	0.94	222.99
Reconstruction	0.91	251.82
Recognitions	0.99	314.15
Together	-	121.5

Fully Unsupervised (Variational) Techniques for Restoration

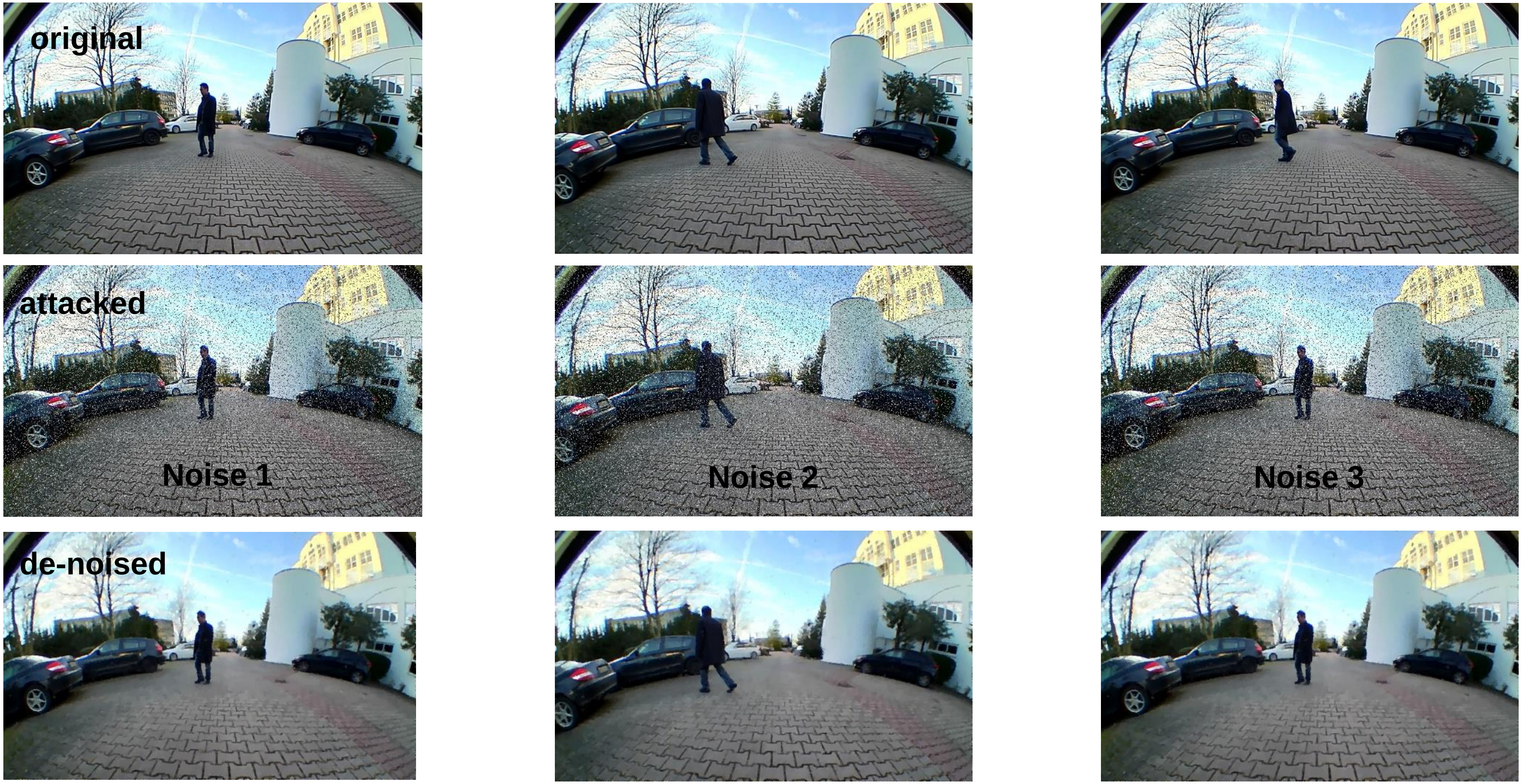
- Introduction of Continuous Instead of Discrete models for Image Analysis.
  - Advantages:
    - More Consistent & intuitive Mathematical Formalization of the Solution.
    - Use of Physical Measures and Physical Phenomena as inspiration for solution modeling.
    - For Theoretical Development, strong background in applied maths is needed
    - Highly Accurate and Stable Algorithms due to consistent Numerical Calculus methods.
  - Mainly associated to the minimization of a functional.




Free diffusion in isotropic medium
Isotropic diffusion
Sphere
Restricted diffusion in anisotropic medium
Anisotropic diffusion
Ellipsoid

Inspecting Results on the viewing layer

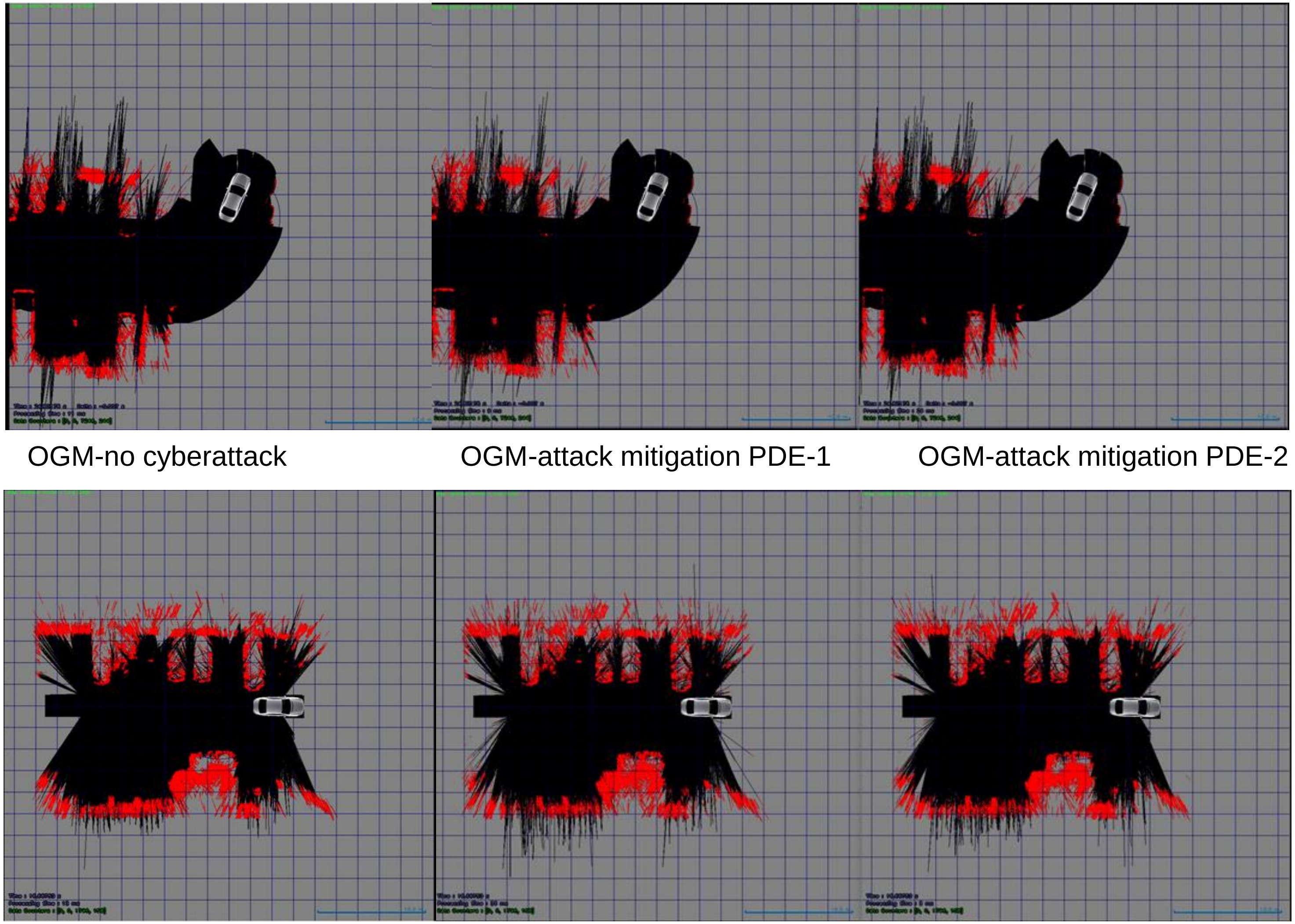
Noise Suppression Using Variational Schemes



Edge content of the image is well preserved in the reconstructed image

Inspecting Results on the perception layer

Noise Suppression Using Total Variation PDE



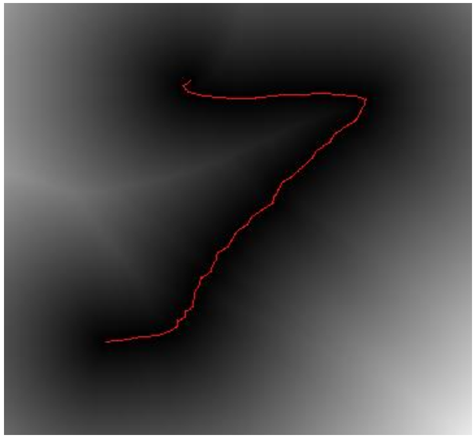
OGM-no cyberattack

OGM-attack mitigation PDE-1

OGM-attack mitigation PDE-2

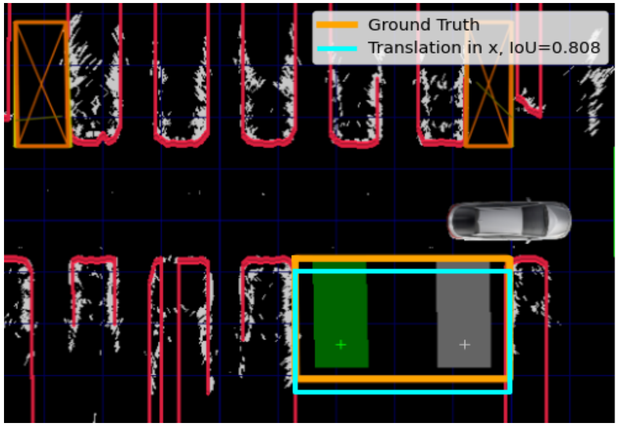
KPIs for Cyber-Attack Mitigation

Distance transform between obstacle-polygons



- $$D(\boldsymbol{p}) = \min_{\boldsymbol{q}\in B} ||\boldsymbol{p}-\boldsymbol{q}||$$
- $$E = \frac{1}{|\Omega_{B_2}|}\sum_{\boldsymbol{p}\in\Omega_{B_2}} D_1(\boldsymbol{p})$$
- Where p_1, p_2 polygons. B_1, B_2 the outlines of polygons as binary images. D_1 distance transform of B_1. A robust measure for the difference between the two polygons can be devised by sampling D_1 along the outline of p_2, e.g.

Intersection Over Units



$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

$$S_{\text{Area}} = \frac{min\{\text{Area}(G), \text{Area}(P)\}}{max\{\text{Area}(G), \text{Area}(P)\}},$$

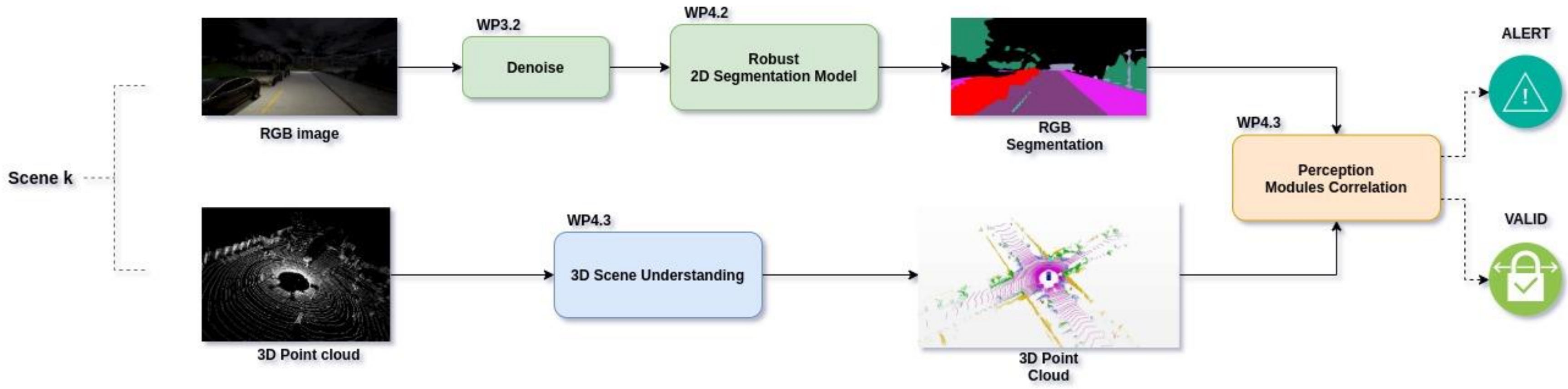
$$S_{\text{Loc}} = \sqrt{\text{Area}(P')/\text{Area}(P)},$$

Demos

- ❑ Physical-based Adversarial Attack can be accessed from here
- ❑ Detection of Attacks on Camera Sensor can be accessed from here

Mitigate the attack on camera sensors using the LiDAR

- ❑ Train Deep CNN on raw point-clouds to detect 3d objects (vehicles, pedestrians, cyclists).
- ❑ Project the detected 3d object to image plane and map them to the output of the segmentation model.
- ❑ Correlate the two outputs to provide improved situational awareness.



Demos

- Mitigation on camera sensor attacks using LiDAR can be accessed from here
- In-vehicle Location Spoofing Attack Detection can be accessed from here

Fallback Actions for Minimal Risk Condition

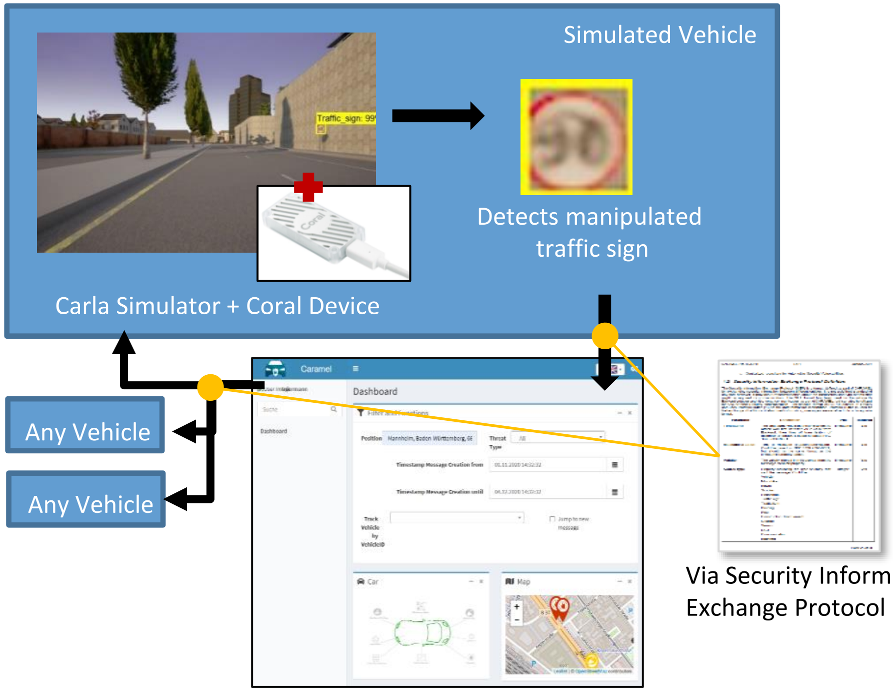
Overview

❑ **Scope:**

- An immutable two-way information exchange protocol between vehicle and infrastructure will be established, which will contain (at least) the level of threat, type, severity, functional specificity, and other information as well as recommendations for the stakeholder.
- Degradation strategies will be developed, which depends on above information about the threat and perceived degree of intrusion.

Scenario 2: Fallback Action through Backend Solution

- ❑ The Backend solution receives and send messages from the Anti-Hacking device or any other In-vehicle devices connected to the Internet based on the Security Information Exchange protocol.
- ❑ The Security Information Exchange protocol (also sometimes referred to as Security Message protocol or SIEP) which defines the format of security messages being shared between Vehicle & Backend.
- ❑ Features of Backend Solution:
  - Dashboard
  - Map Visualization
  - Threat Visualization
  - Localization of Threats on the map
  - …
- ❑ Next Steps Planned:
  - Small Displacement GPS Spoofing
  - Integration of Backend Solution with other CARAMEL Scenarios.
  - …



Via Security Information Exchange Protocol (SIEP)

Backend with Dashboard incl. Threat localization

CARAMEL

Thank you for your attention