ASSURANCE MONITORING IN LEARNING-ENABLED CYBER-PHYSICAL SYSTEMS

XENOFON KOUTSOUKOS

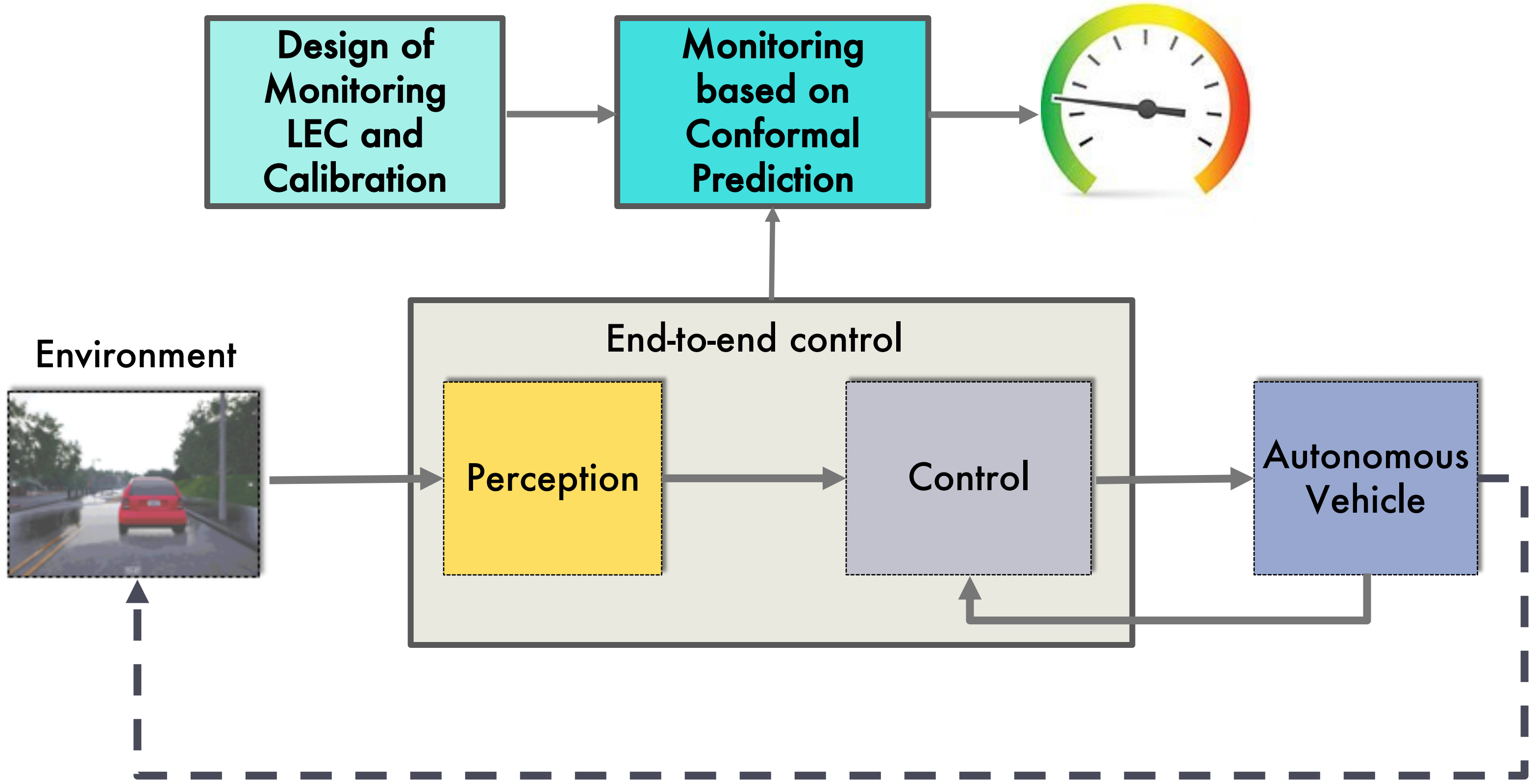
INSTITUTE FOR SOFTWARE INTEGRATED SYSTEMS

ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

VANDERBILT UNIVERSITY

VANDERBILT UNIVERSITY

ASSURANCE MONITORING OF LEARNING-ENABLED CYBER-PHYSICAL SYSTEMS



Assurance monitoring based on inductive conformal anomaly detection

- Variational autoencoder (VAE)
- VAE for regression
- Adversarial Autoencoder (AAE)
- Deep support vector description (SVDD)

Evaluation

- Airport image dataset
- Self-driving simulator and open datasets
- Autonomous underwater vehicle

NOVELTY DETECTION IN HIGH-DIMENSIONAL TIME SERIES

- In autonomous systems, inputs are high-dimensional sensor measurements (e.g., camera, LiDAR) and arrive one by one based on the sampling rate of the sensors
- After observing each input, inductive conformal anomaly detection is used to quantify the degree to which the input disagrees with the training data
- Main idea: Train an appropriate neural network architecture which can be used for detection in real-time
  - Use multiple examples sampled from a learn representation from the input distribution
  - A nonconformity measure (NCM) to evaluate the degree to which a new example disagrees from the distribution of the training data
  - Compute empirical p-values used for statistical significance testing
  - Perform a randomness test to compute an assurance measure using a martingale process of the p-values

VAE-BASED NONCONFORMITY MEASURE

Original Image

Reconstructed Image

Nonconformity measure

$$\boxed{\alpha'_k = A_{\mathrm{VAE}}(z_t, z'_k) = ||z_t - z'_k||^2}$$

Given an input example at time *t*, the encoder portion of the VAE is used to approximate the posterior distribution of the latent space

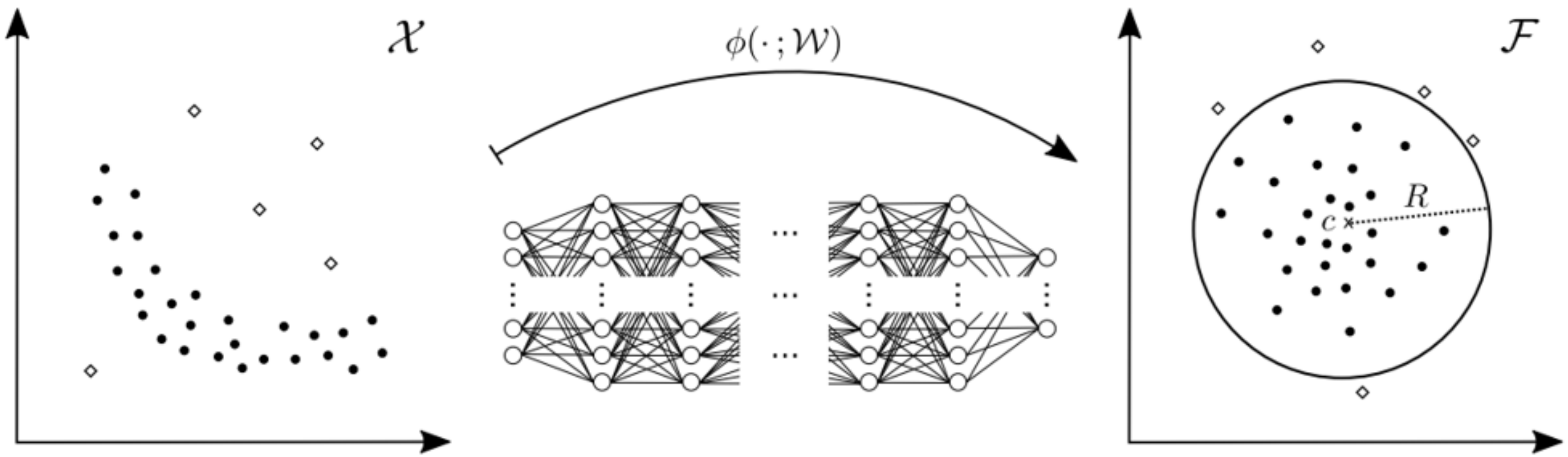
- Typically, the posterior of the latent space is approximated by a Gaussian distribution

Sampling from the posterior generates multiple encodings so that the decoder is exposed to a range of variations of the input example

- An in-distribution input should be reconstructed with a relatively small reconstruction error.
- Conversely, an out-of-distribution input will likely have a larger error.

The reconstruction error is a good measure of the strangeness of the input relative to the training set and it is used as the nonconformity measure

DEEP SUPPORT VECTOR DESCRIPTION (SVDD)



Loss function: $\min_{R,\mathcal{W}} \quad R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\max\{0, \|\phi(\boldsymbol{x}_i;\mathcal{W}) - \boldsymbol{c}\|^2 - R^2\} \; + \frac{\lambda}{2}\sum_{\ell=1}^{L}\|\boldsymbol{W}^{\ell}\|_F^2$

SVDD maps the training data into a hypersphere characterized by center c and radius R of minimum volume

- Training should avoid hypersphere collapse: c must be selected appropriately, no bias terms or bounded activation functions

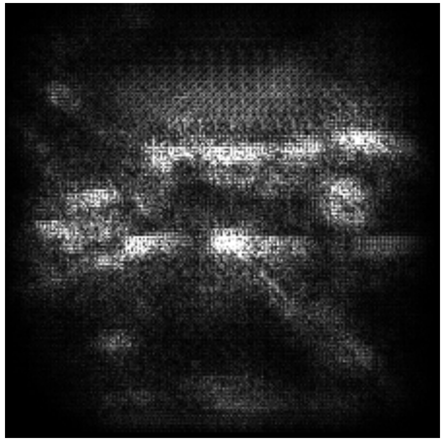
Mappings of normal examples fall within, whereas mappings of anomalies fall outside the hypersphere

The distance from the center can be used as the NCM

$$\boxed{\alpha'_t = A_{\text{SVDD}}(z_t) = ||\phi(z_t;\mathcal{W}^*) - \boldsymbol{c}||^2}$$

IMPROVING ROBUSTNESS OF DETECTION USING SALIENCY MAPS

Original Image



Saliency Map

VAEs have difficulty generating fine-granularity details of the original image

Fine-granularity details and other input features may not affect the LEC output

Saliency Map:

- **Quantify the spatial support of the LEC prediction for a given image input**

Nonconformity Measure:

- **Reconstruction error x saliency map**

AIRPORT IMAGE DATASET (BOEING)

Open set classification

- Individual labeled frames with three classes and bounding boxes around the objects
  - Airplane, Ground Vehicle, and Person
  - Person to be treated as the unknown class

Training and calibration dataset (contain only known classes)

- Training: 23403 images/Calibration: 5841 images

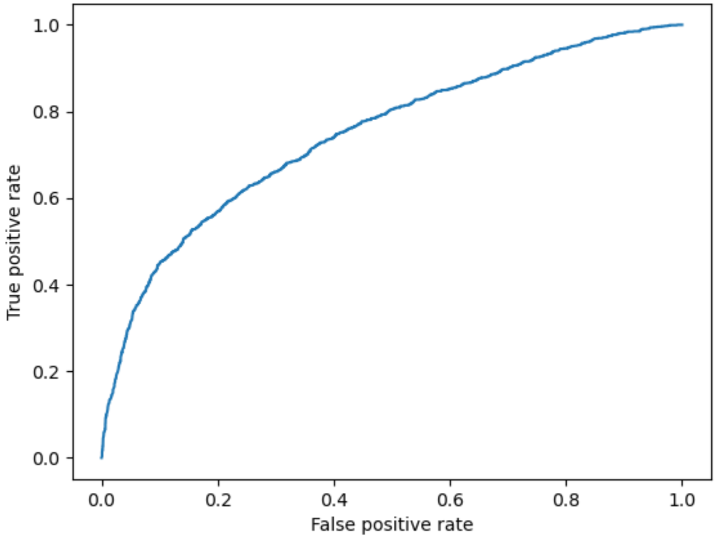
Testing dataset

- Contains both known classes (3249 images) and unknown classes (1135 images)

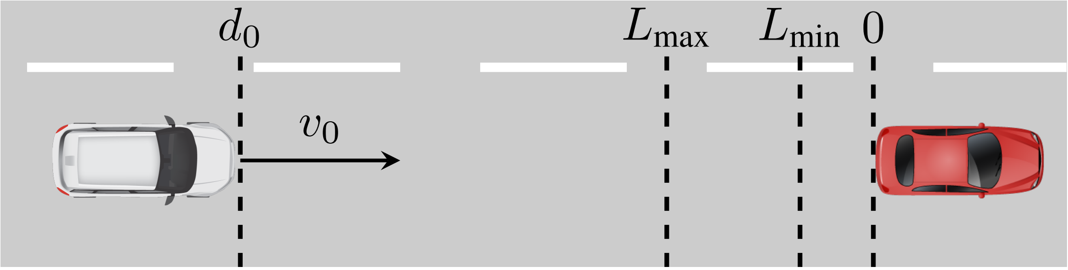
VAE for Classification + Deep SVDD

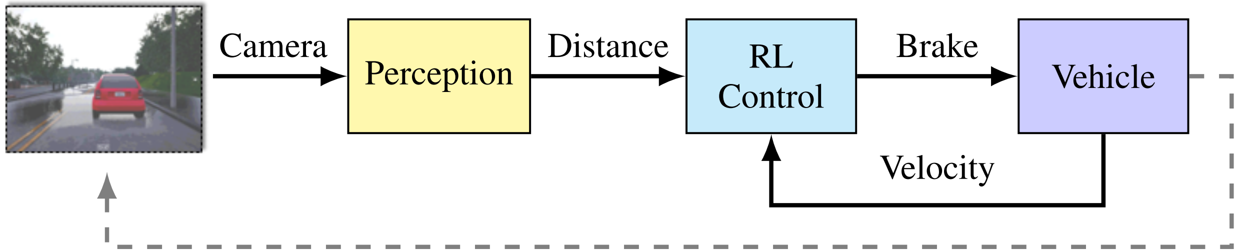
- Sample N examples using VAE for classification model
- Feed N reconstructed examples into deep SVDD
  - Nonconformity measure: Distance of the representation to the center of the hypersphere
- Compute p-values and assurance measure (martingale M) for each test example
- If $\log M > \varepsilon$, the test example is a considered a novelty

Area Under ROC Curve ≈ 0.85



ADVANCED EMERGENCY BRAKING SYSTEM (AEBS)





Data Generation using CARLA

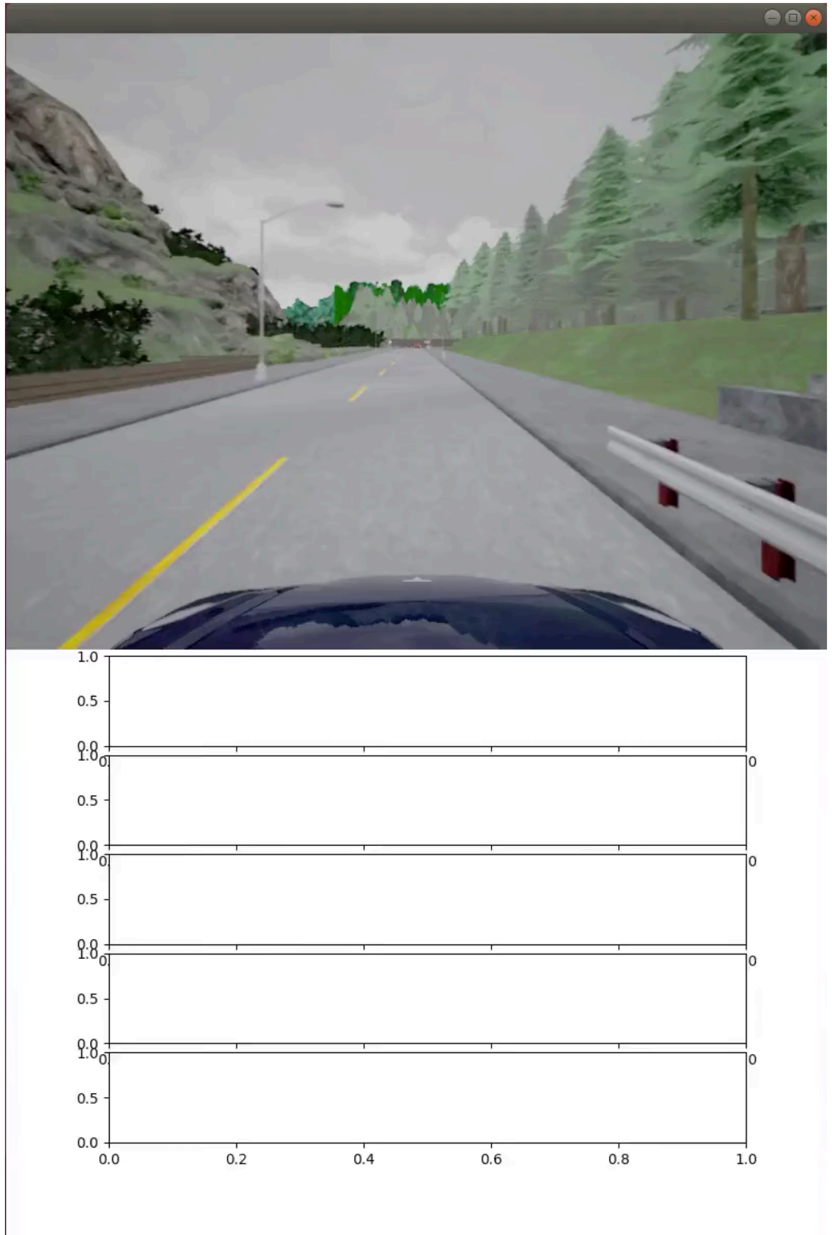
d_0	100 m approximately
v_0	Randomly sampled between 90 and 100 km/h
L_{min}	1 m
L_{max}	3 m
CARLA precipitation parameter r	Randomly sampled between 0 and 20
Sampling period	1/20 sec = 50 ms

Learning-Enabled Components

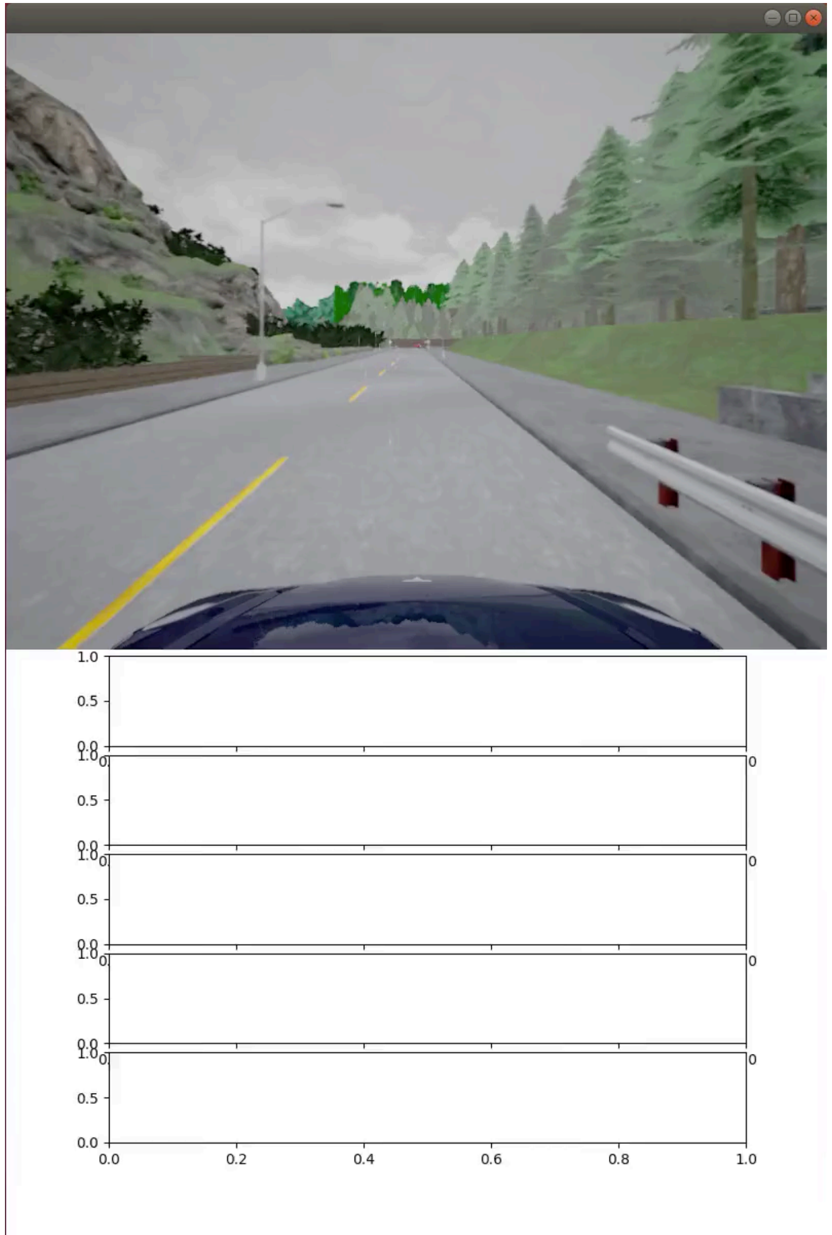
- Perception: CNN with 11 layers
- Control: Reinforcement learning controller trained using DDPG
- VAE: CNN encoder with 4 layers, 1024 FC layer, and symmetric decoder
- SVDD: 4 convolution layers and 1568 FC layer

SIMULATION RESULTS

In-distribution



Out-of-distribution



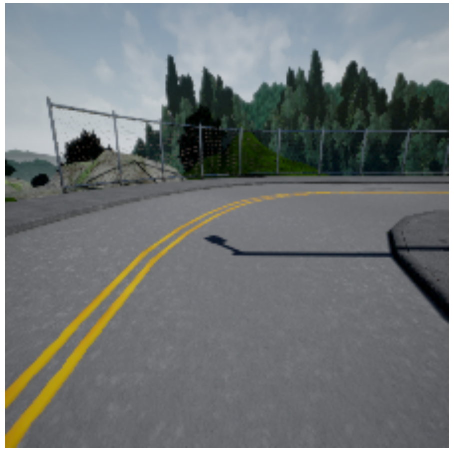
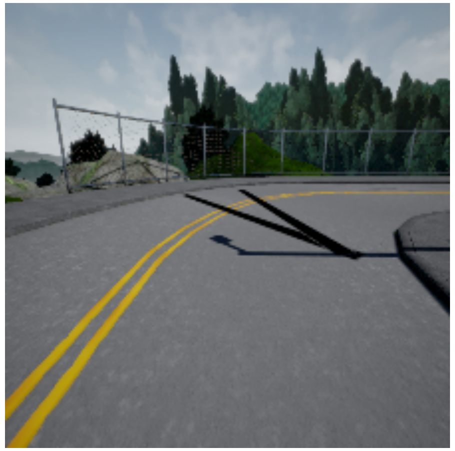
SELF-DRIVING END-TO-END CONTROLLER (SDEC)

CARLA provides an SDEC trained using imitation learning

- Uses camera images as inputs and computes steering, acceleration, and brake actuation signals
- Implemented using a CNN trained using 14 hours of driving data recorded by human drivers
- The sampling period is Δt = 100 ms

A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017.

Detect physically realizable attacks

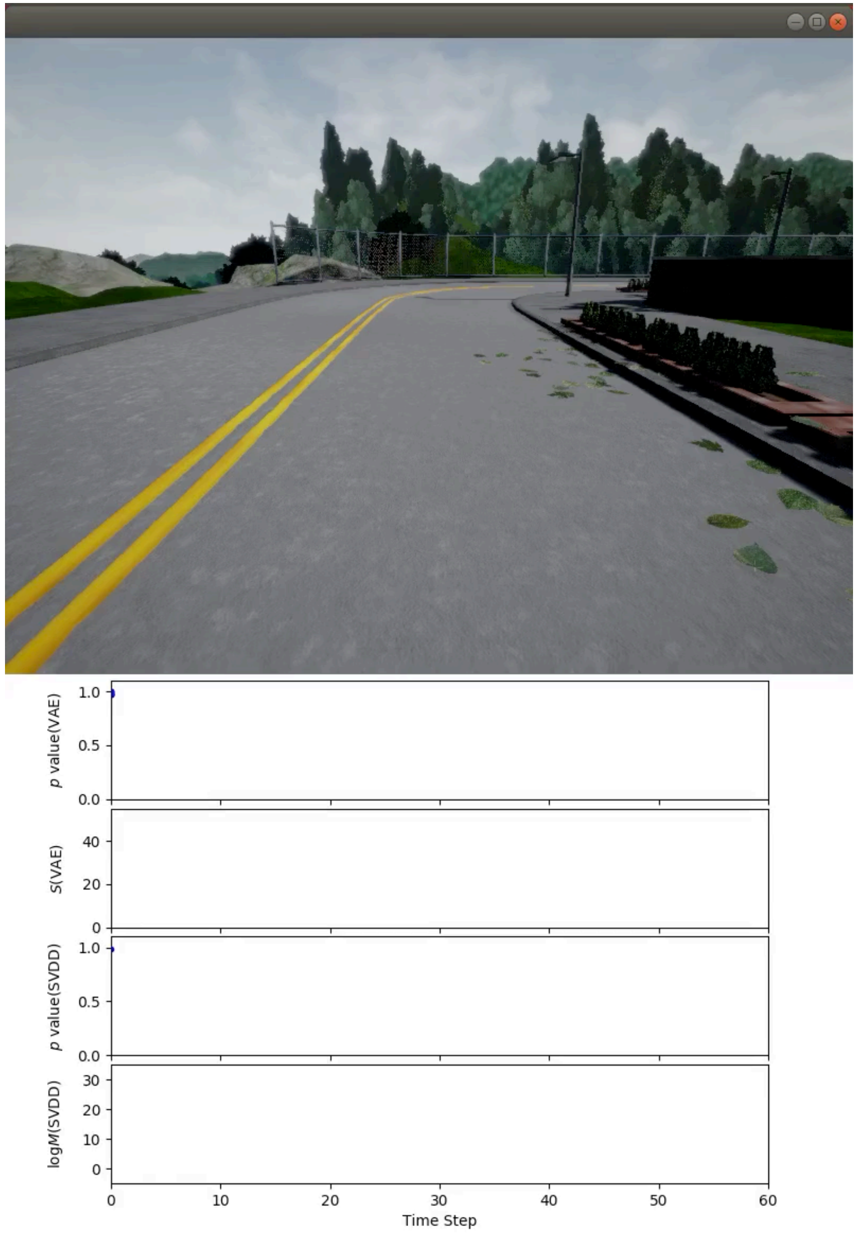
Boloor A, Garimella K, He X, Gill C, Vorobeychik Y, Zhang X. Attacking vision-based perception in end-to-end autonomous driving models. *Journal of Systems Architecture*. 2020 Apr 4:101766.

Data generation for training the VAE and SVDD
- Weather patterns: clear and cloudy noon
- Turning right, left, and going straight

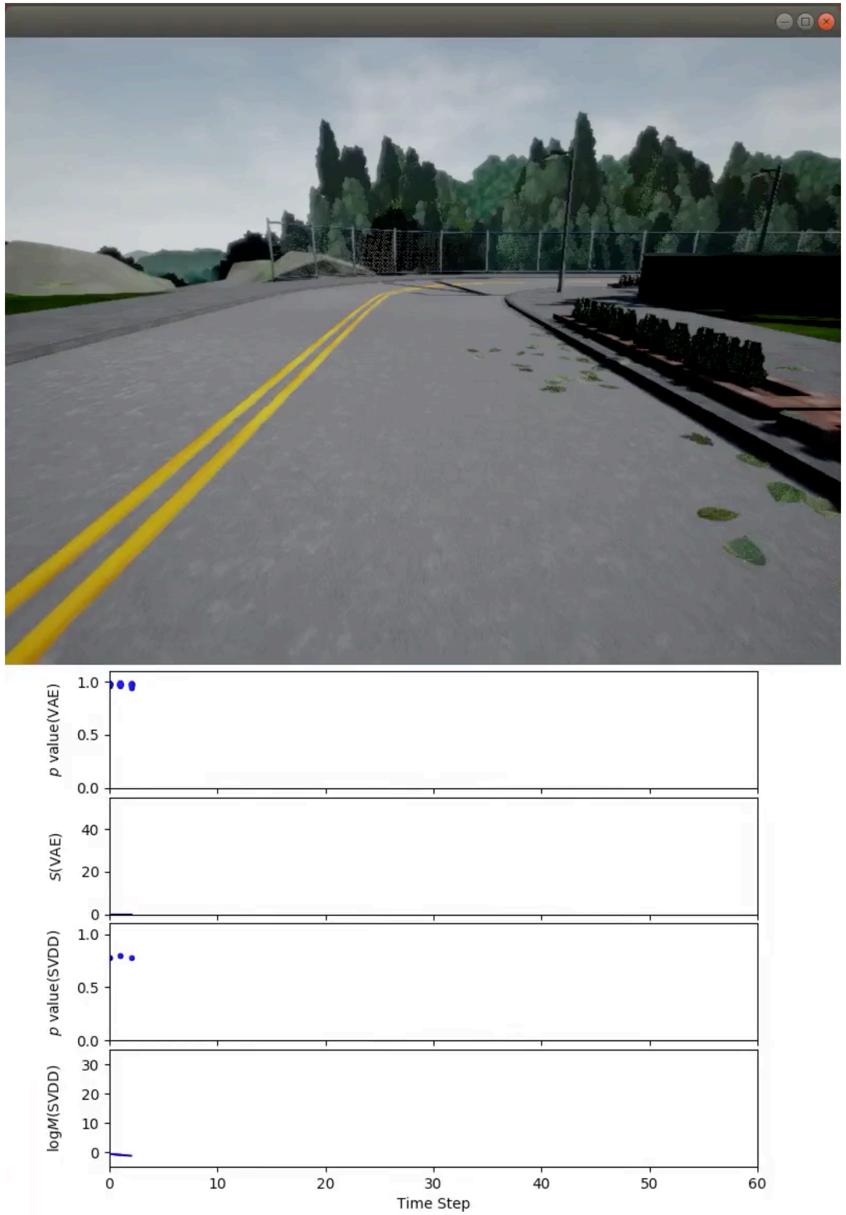
Evaluation
- Detected 105 out of 105 episodes with different positions and rotations of the two black lines which are chosen to cause traffic infraction

SIMULATION RESULTS

No attack

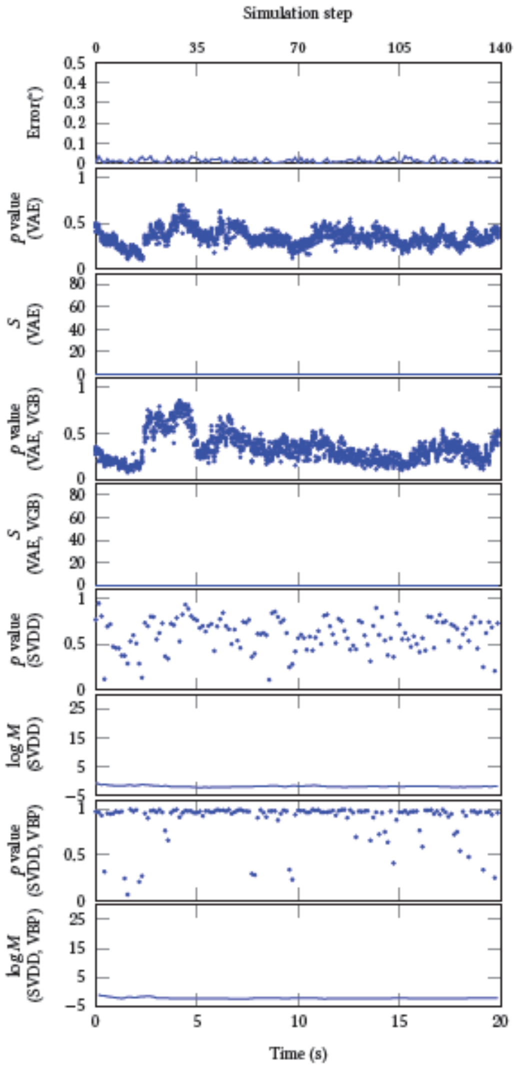


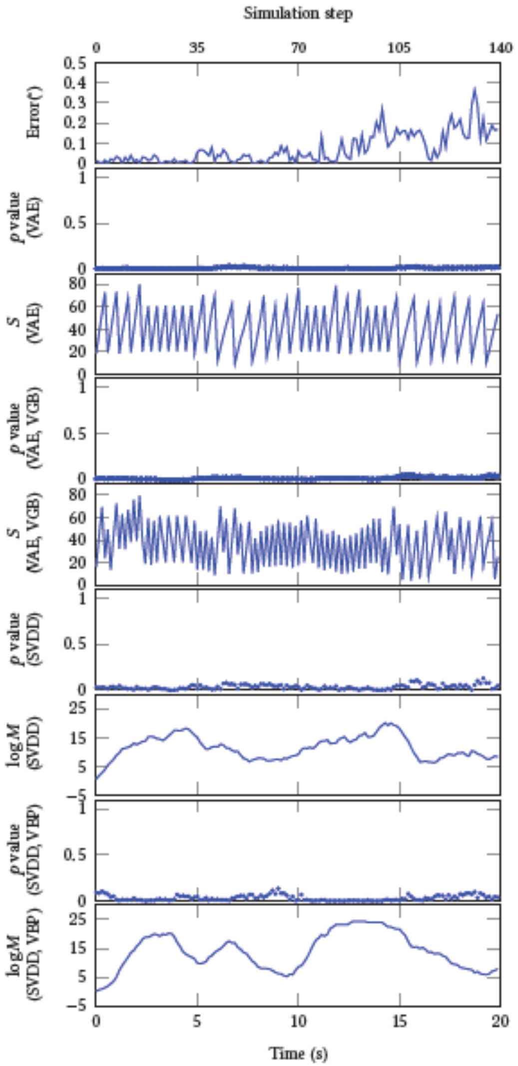
Attack



FORD AUTONOMOUS VEHICLE SEASONAL DATASET

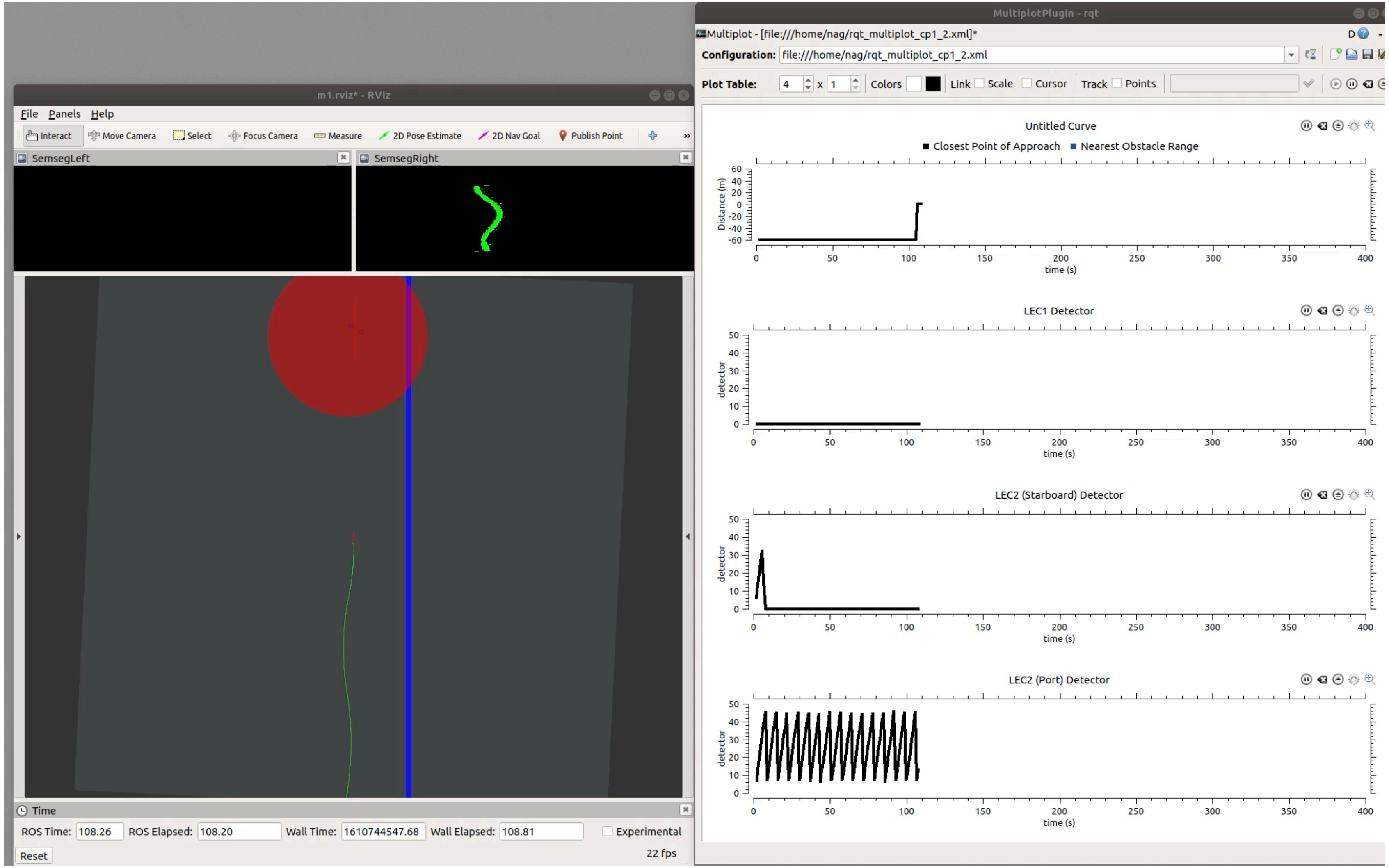
Cloudy weather and freeway driving



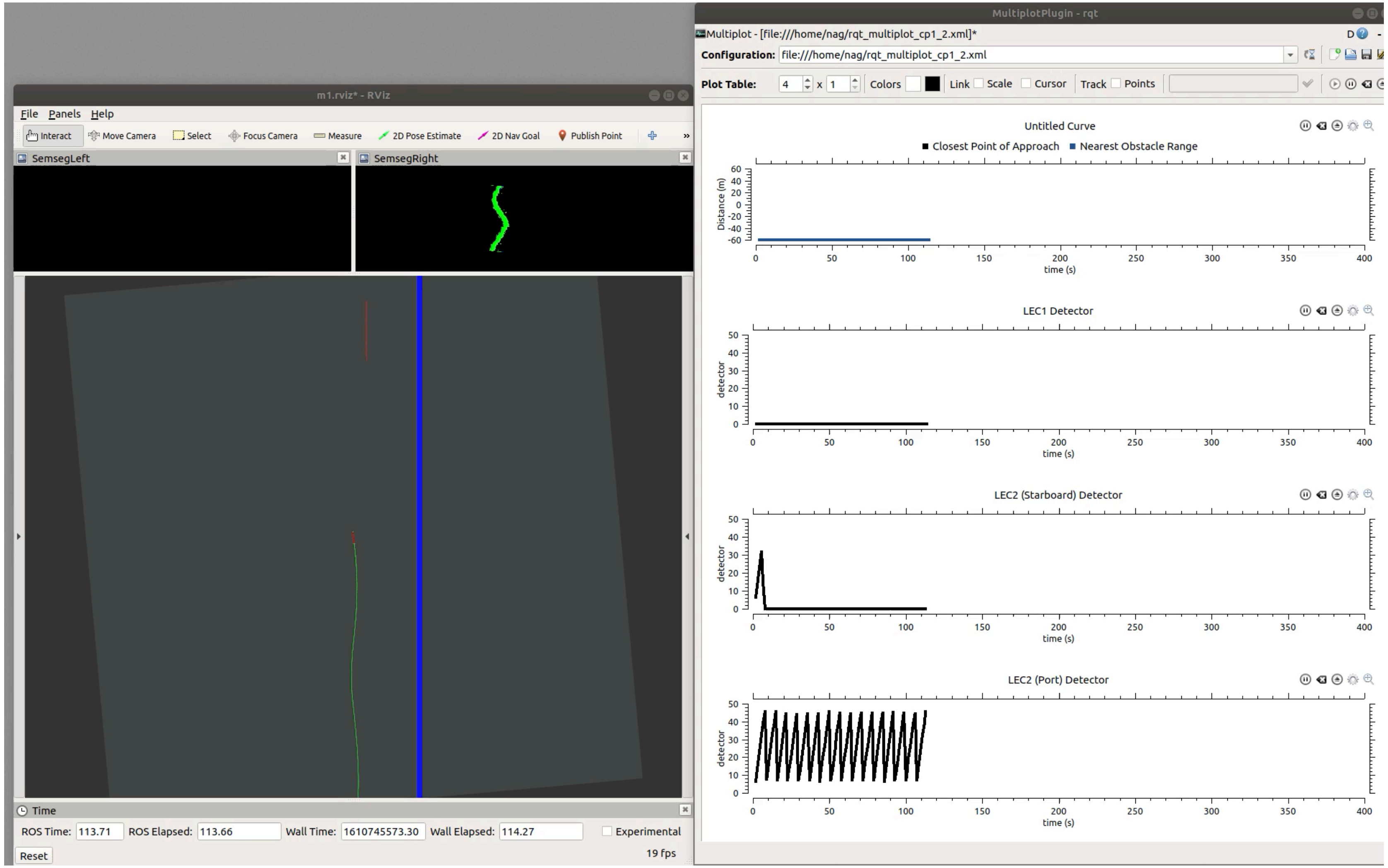


Sunny weather and residential driving

AUV: AVOID OBSTACLE AND COMPLETE PIPELINE INSPECTION



AUV: LOSS OF PIPELINE



HIGHLIGHTS

Learn representations (VAE, AAE, SVDD) that allow effective assurance monitoring based on deep learning and statistical significance testing

Integration into a toolchain for model-based design of cyber-physical systems with learning-enabled components

- Architectural modeling of CPS
- Engineering and integration of LECs
- System software deployment
- Modeling and analysis of assurance cases

Evaluation with open source simulator and open datasets

- Very small number of false positives and detection delay
- Execution time is comparable to the execution time of the original LECs