Using Adversarial Defense Methods to Improve the Performance of Deep-Neural-Network-Controlled Automatic Driving Systems

Mike Camara

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Automatic Driving Systems

Automatic Driving Systems

Safer roads

Automatic Driving Systems

Safer roads

Freeing up time

Automatic Driving Systems

Safer roads

Freeing up time

Mobility for all

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Automatic Driving Systems

Modular pipeline

End-to-end pipeline

Automatic Driving Systems

Modular pipeline

Expensive sensors

Cameras
LiDAR
Radar
HD Maps
GNSS

End-to-end pipeline

Automatic Driving Systems

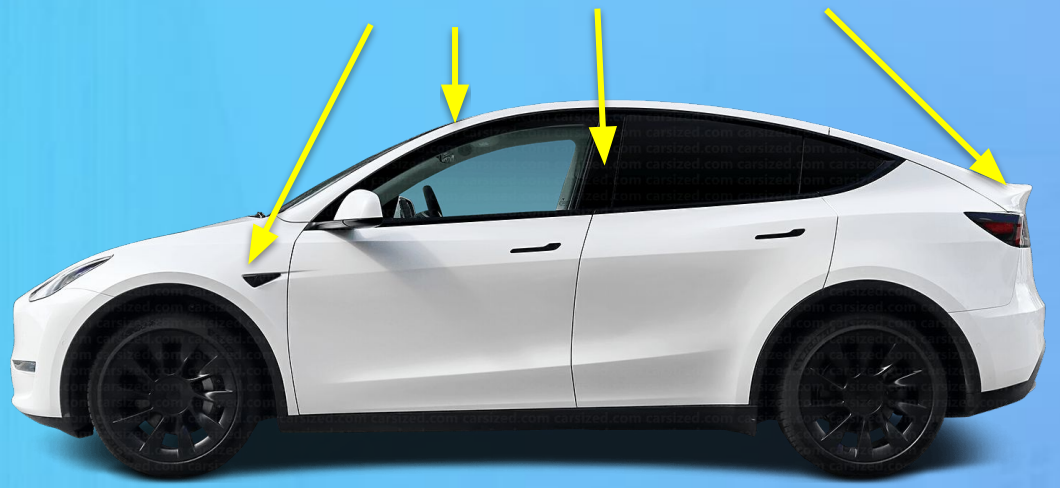
Modular pipeline

Expensive sensors

Cameras
LiDAR
Radar
HD Maps
GNSS

End-to-end pipeline

Computer Vision



Cameras
Data

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Automatic Driving Systems

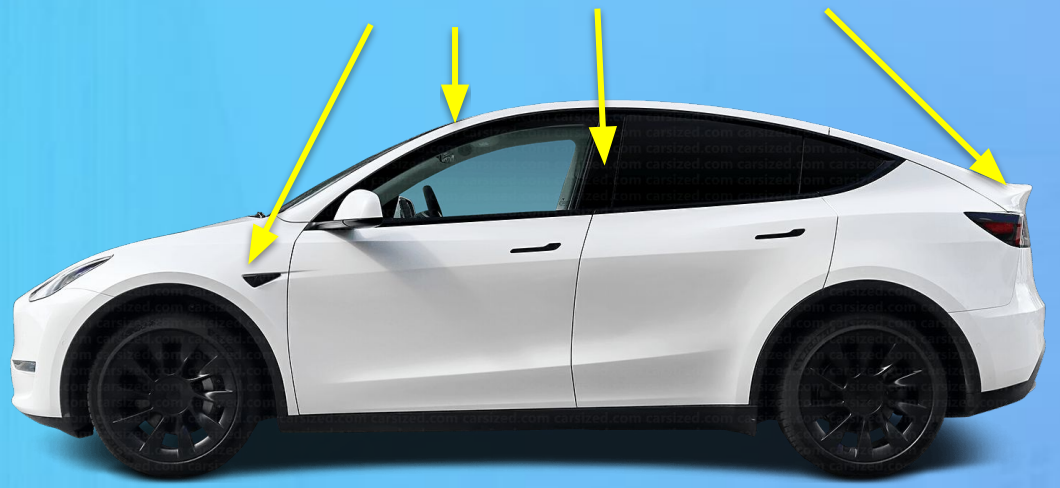
Modular pipeline

Expensive sensors

Cameras
LiDAR
Radar
HD Maps
GNSS

End-to-end pipeline

Computer Vision



Cameras

a lot of **Data**

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

End-to-end Pipeline

Collect

Labelled images

Dataset

End-to-end Pipeline

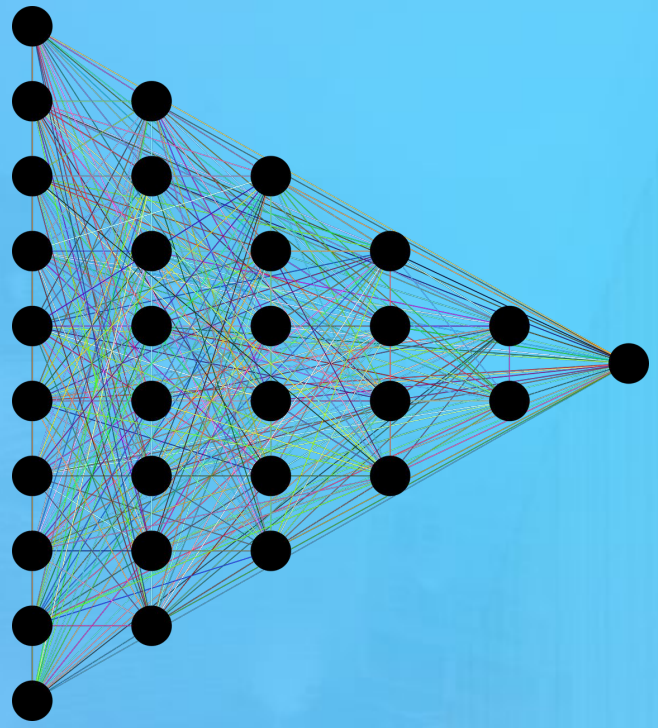
Collect

Labelled images

Dataset

Train

Deep learning model



Convolutional Neural Network

CNN

TAL TECH

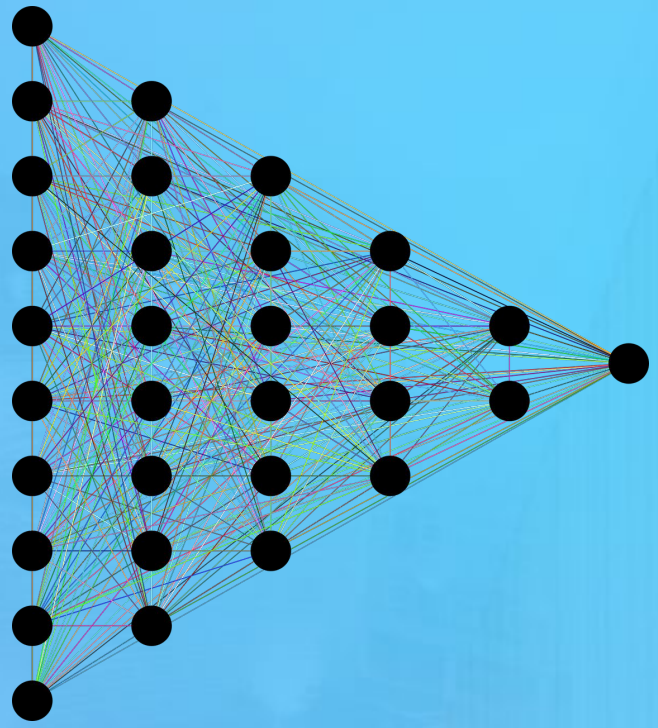
TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

End-to-end Pipeline

Collect
Labelled images

Dataset

Train
Deep learning model



Convolutional Neural Network

CNN

Output
Model

autopilot.h5

TAL TECH

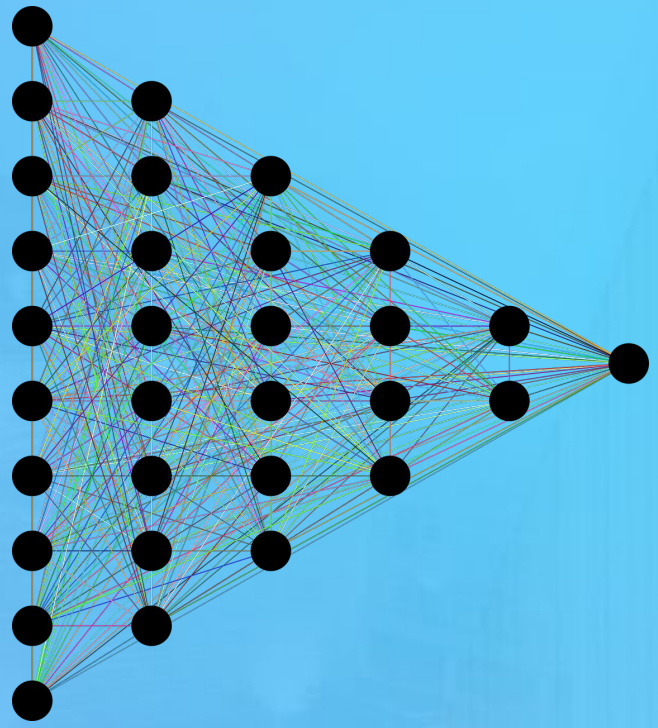
TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

End-to-end Pipeline

Collect
Labelled images

Dataset

Train
Deep learning model



Convolutional Neural Network

CNN

Output
Model

autopilot.h5

Deploy
Model in production

ADS

Steering
Acceleration
Brake

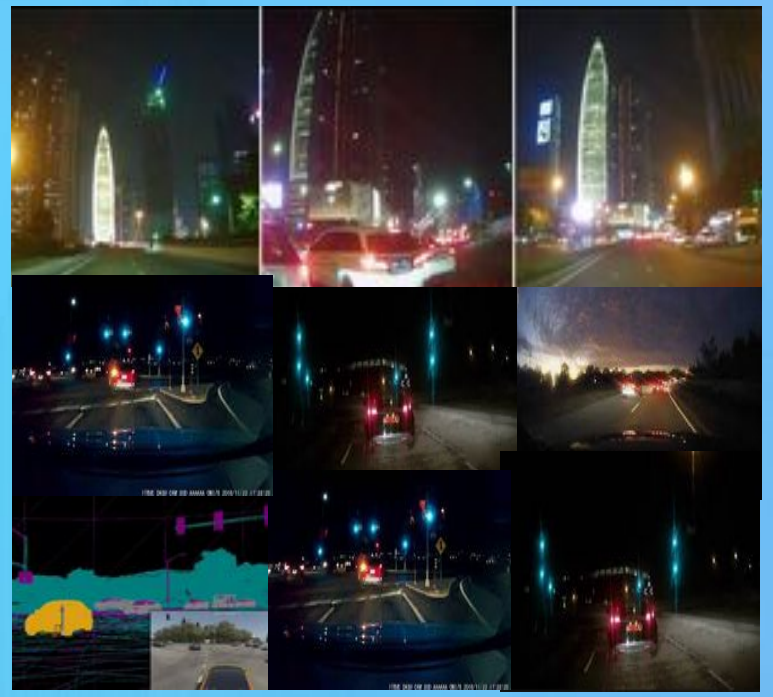
End-to-end Pipeline

Large dataset

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

End-to-end Pipeline



End-to-end Pipeline

Collect the dataset in
different lighting levels

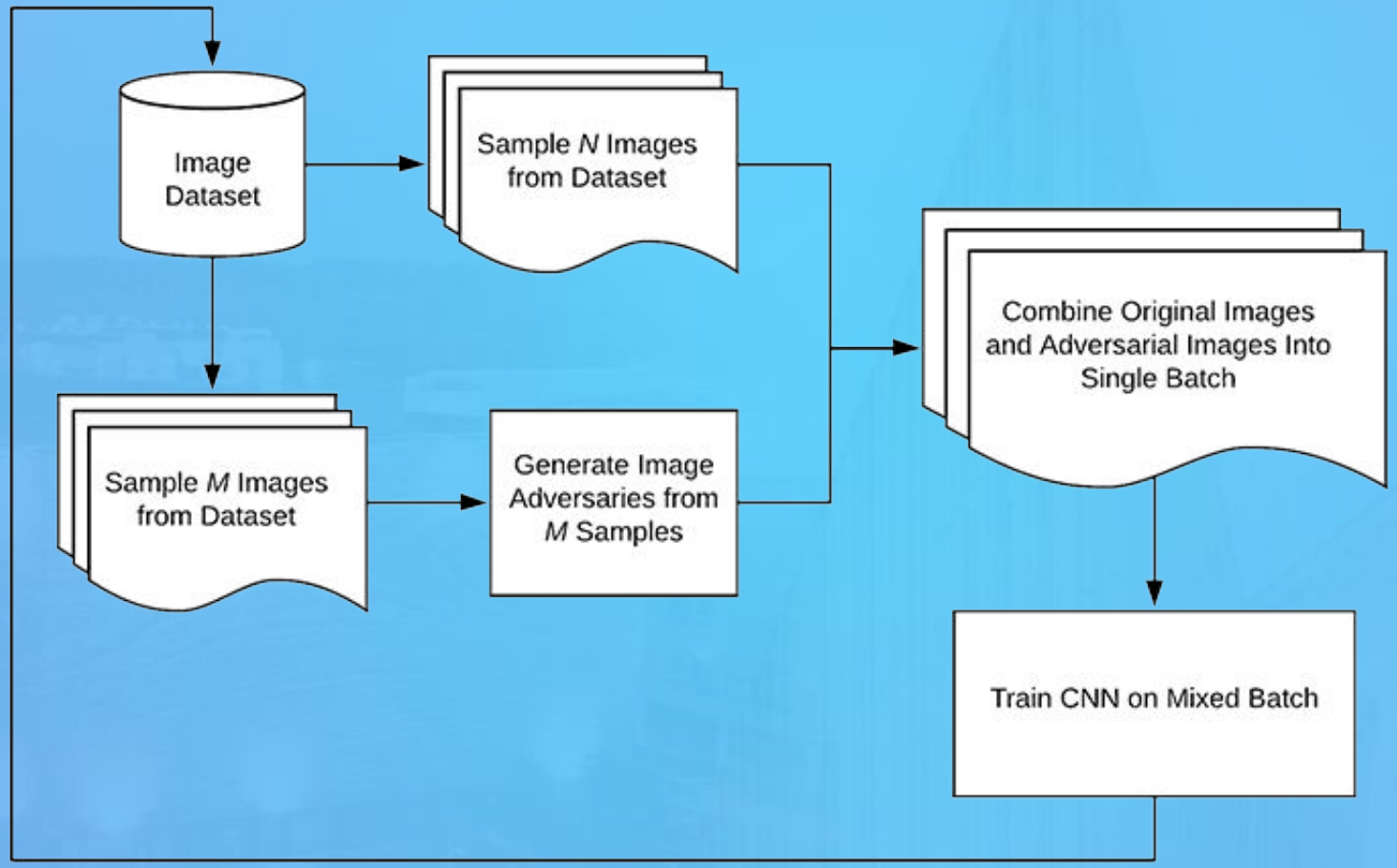
End-to-end Pipeline

Collect the dataset in
different lighting levels

TAL TECH

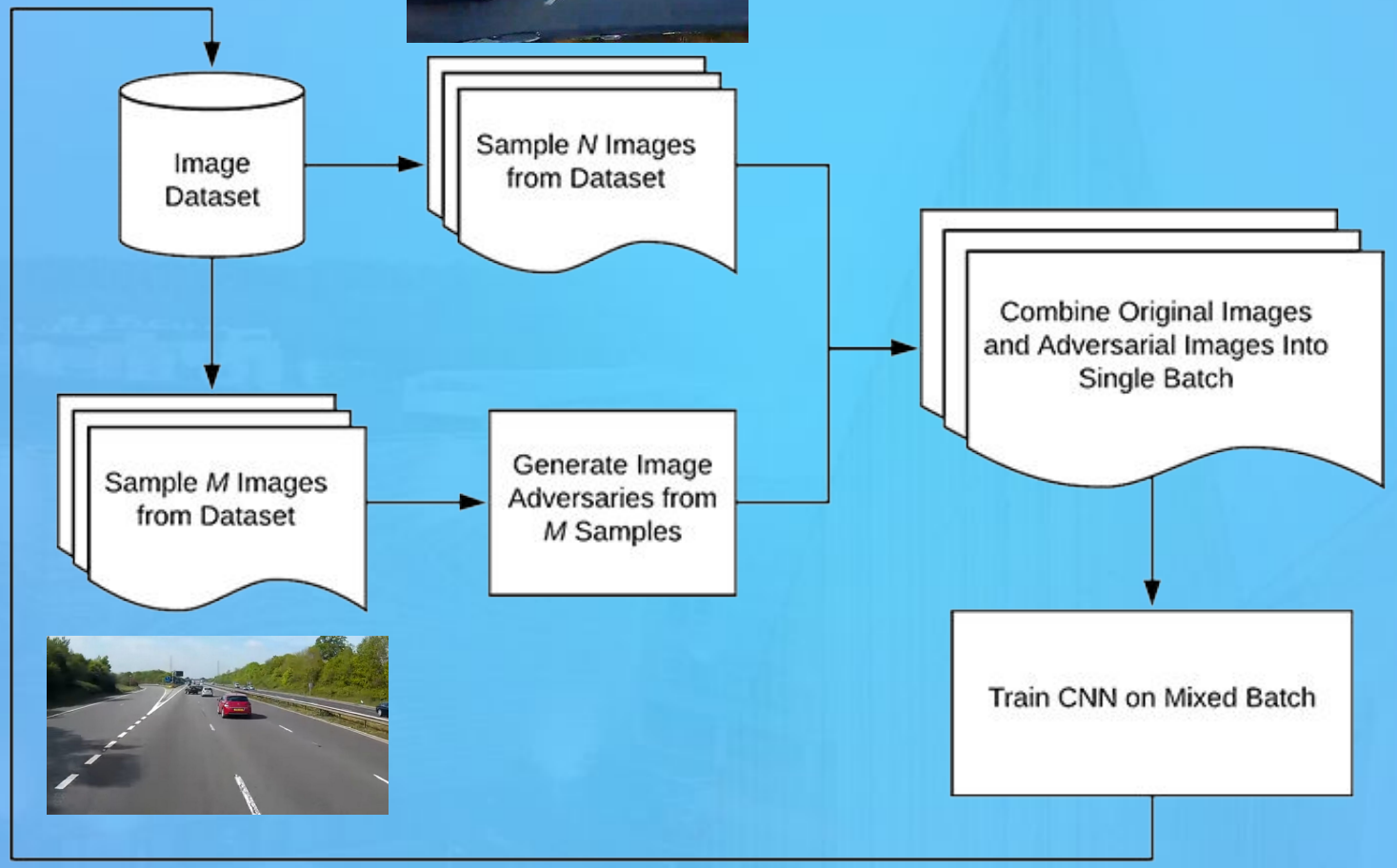
TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

CNN's Adversarial Defense Training Method



(Rosebrock, 2021)

CNN's Adversarial Defense Training Method

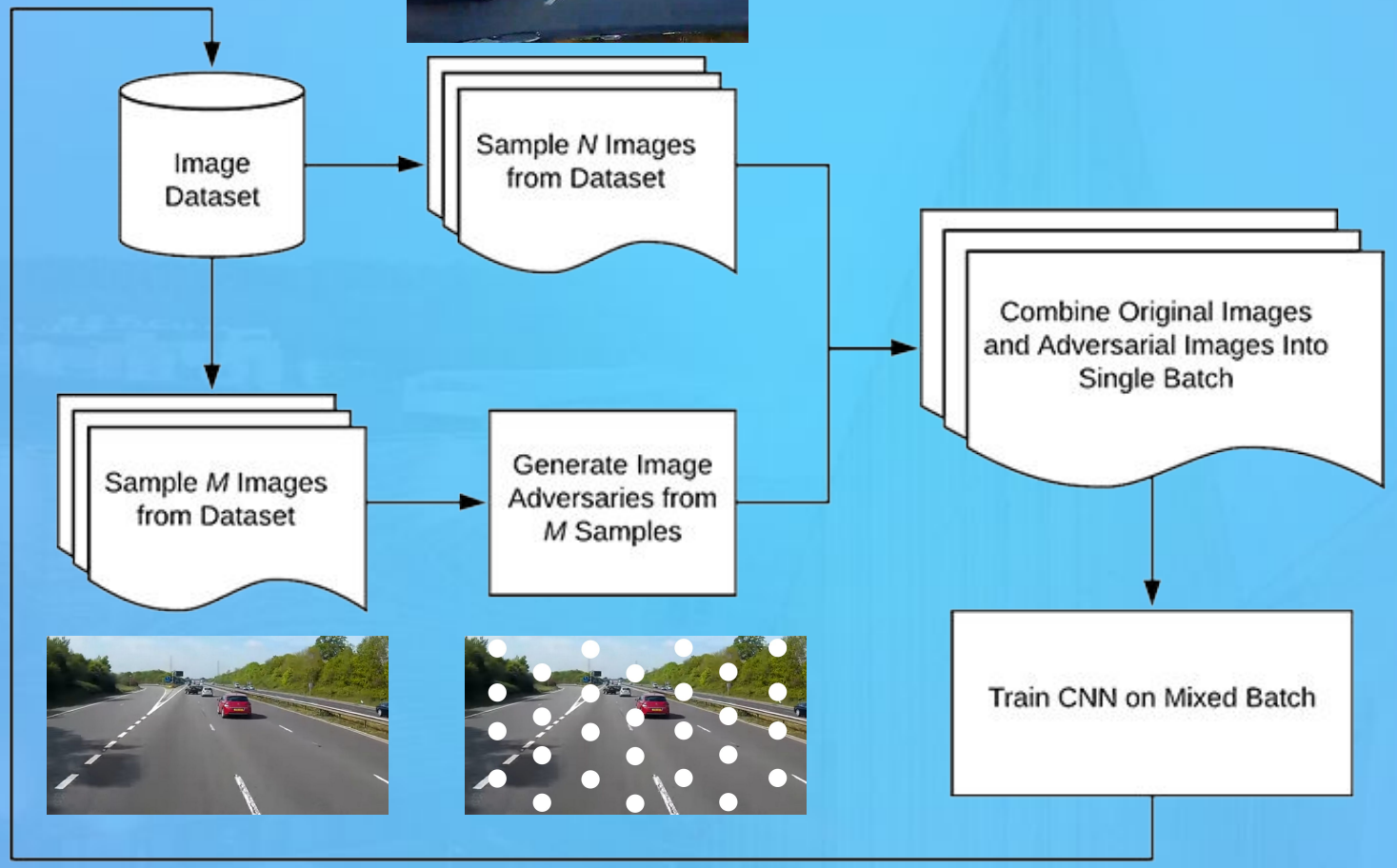


(Rosebrock, 2021)

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

CNN's Adversarial Defense Training Method



(Rosebrock, 2021)

CNN's Adversarial Defense Training Method

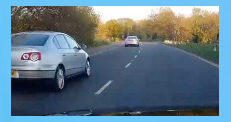
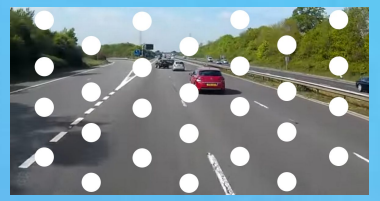


Image Dataset

Sample *N* Images from Dataset

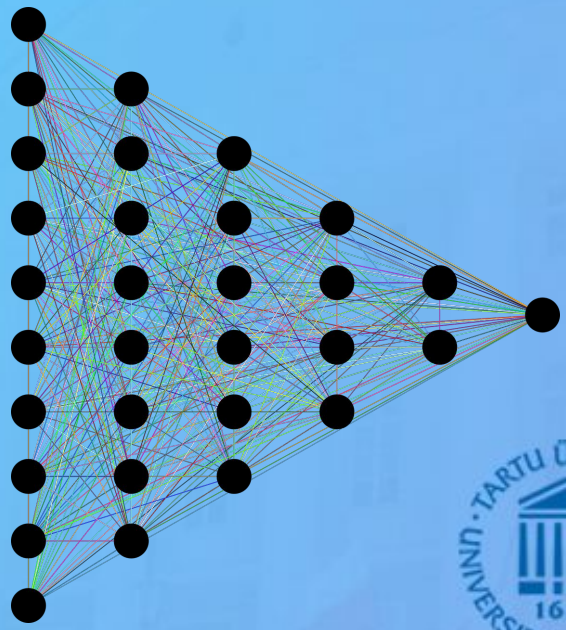
Sample *M* Images from Dataset

Generate Image Adversaries from *M* Samples



Combine Original Images and Adversarial Images Into Single Batch

Train CNN on Mixed Batch

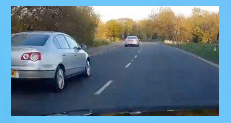


(Rosebrock, 2021)

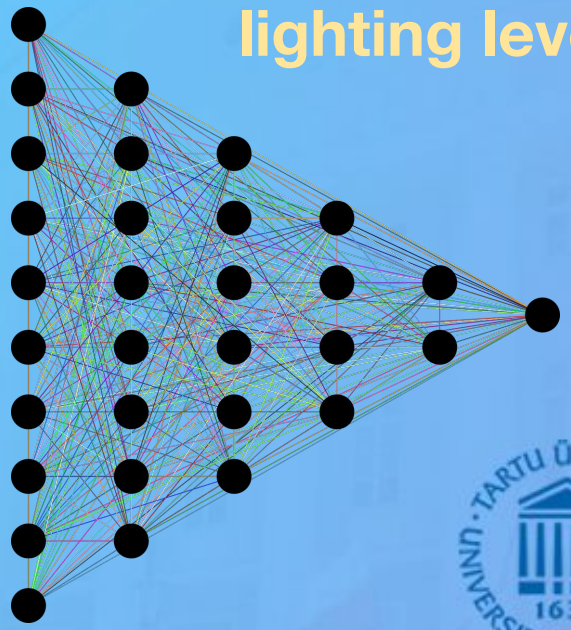
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

CNN's Adversarial Defense Training Method



Model **better at** generalizing to **unseen lighting levels**



(Rosebrock, 2021)

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Goal

To **improve the performance** of CNN models
in unseen **lighting conditions**

TAL TECH

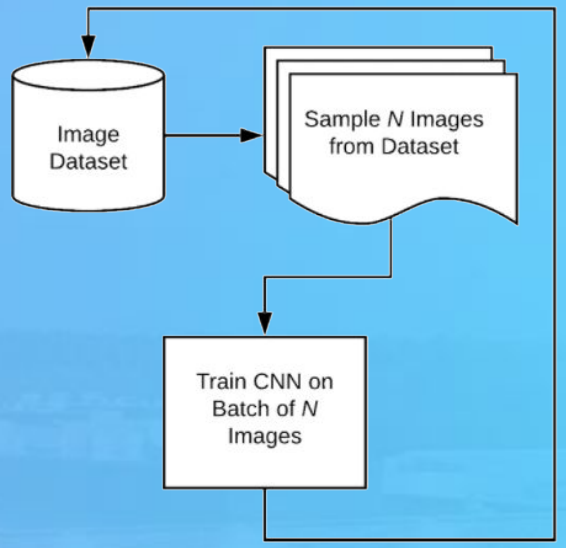
TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Research Question

Can adversarial defense training methods improve the **neural network** generalization skills to unseen **lighting conditions**?

Hypothesis

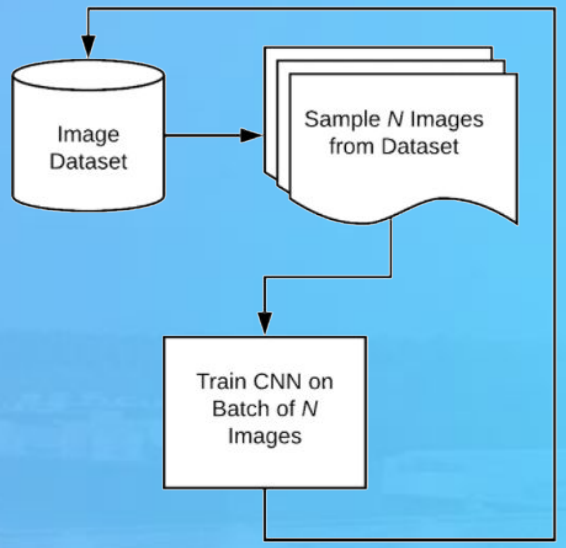
Train model



Conventional

Hypothesis

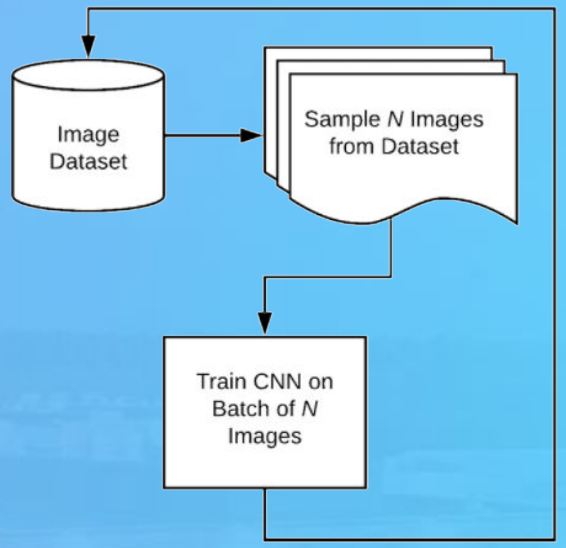
Train model

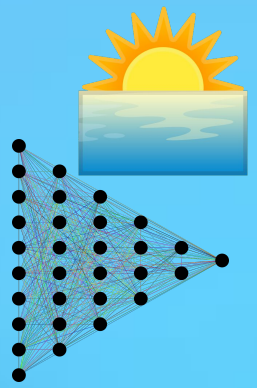


Conventional

Hypothesis

Train model



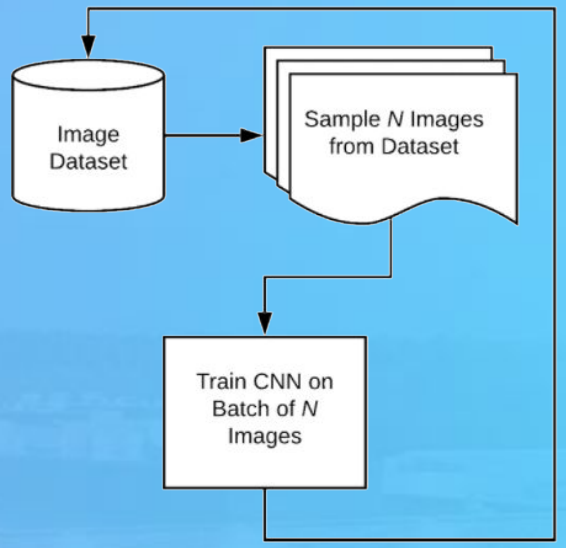


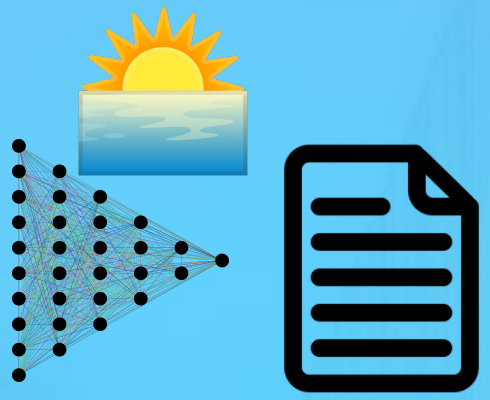
Conventional

Hypothesis

Train model

Output model





Conventional

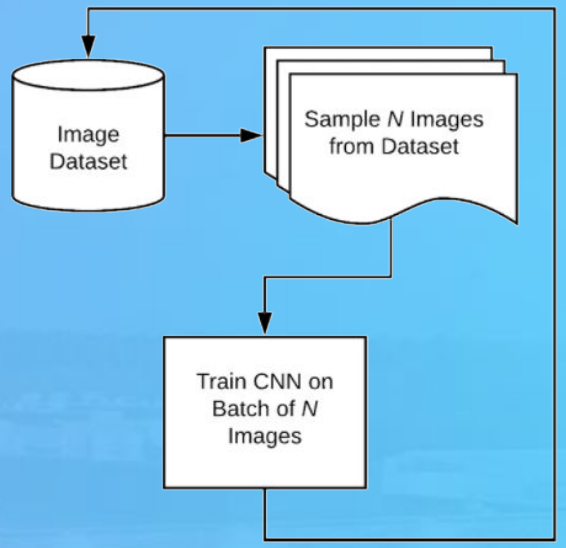
Hypothesis

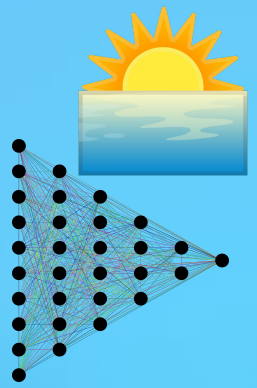
Train model

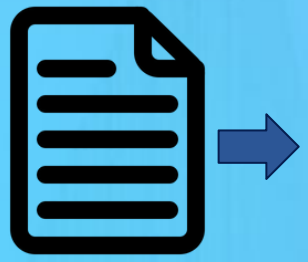
Output model

Deploy model in production

Conventional







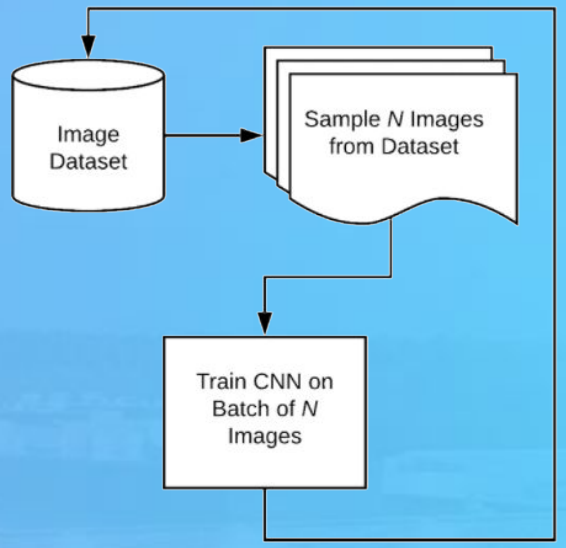
Hypothesis

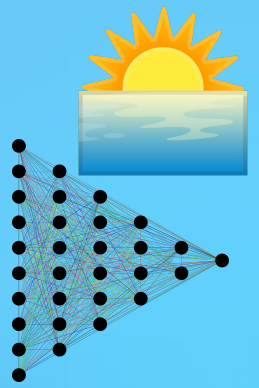
Train model

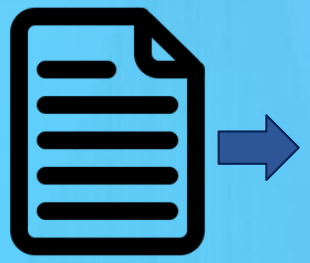
Output model

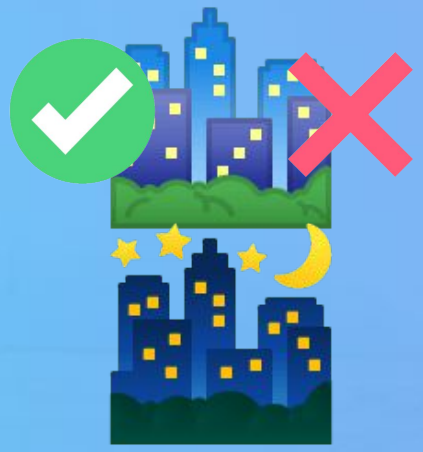
Deploy model in production

Conventional







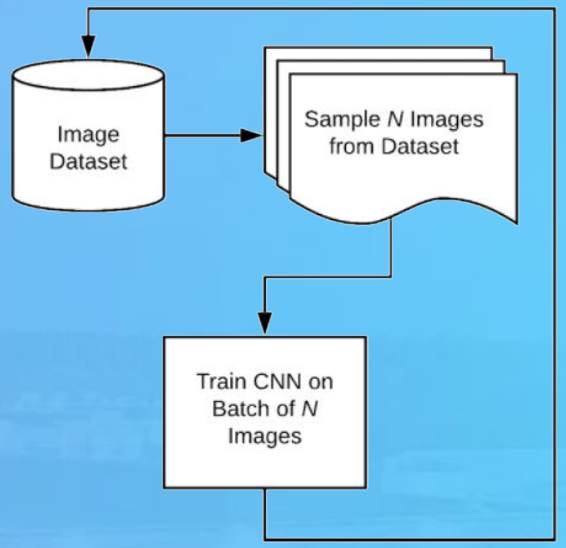


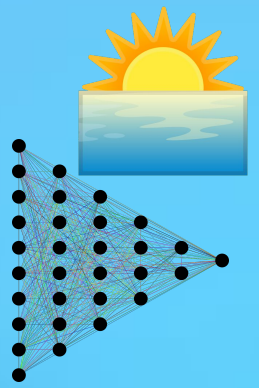
Hypothesis

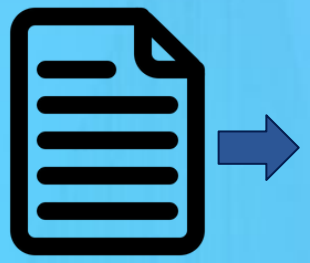
Train model

Output model

Deploy model in production

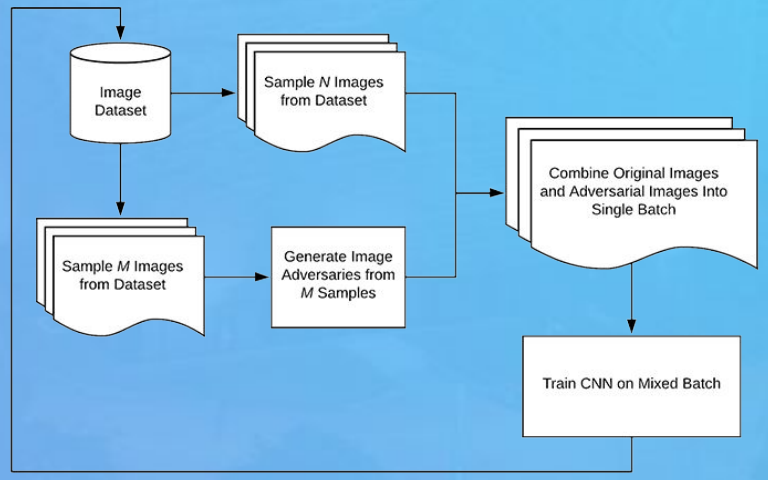






Conventional

Proposed



(Rosebrock, 2021)

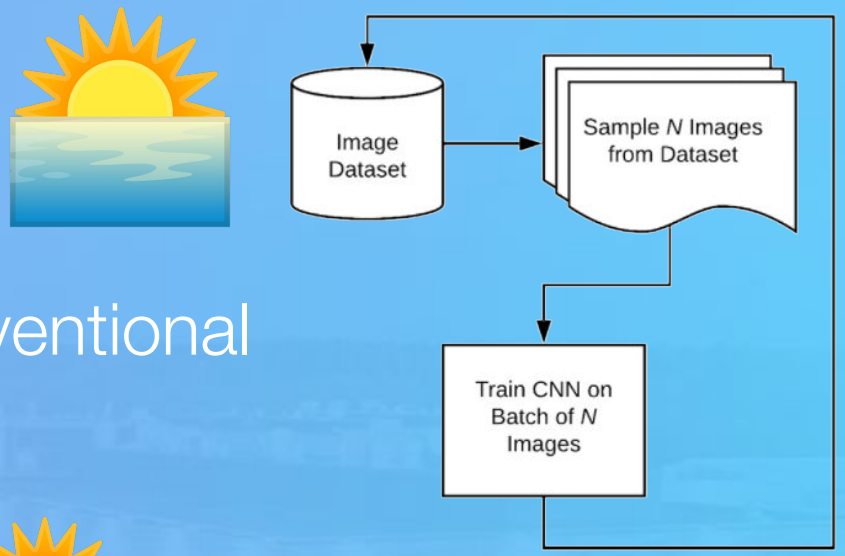
Hypothesis

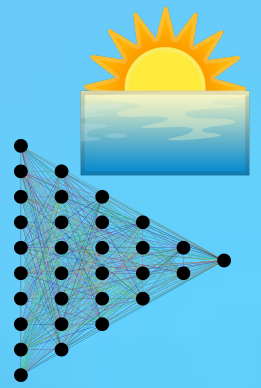
Train model

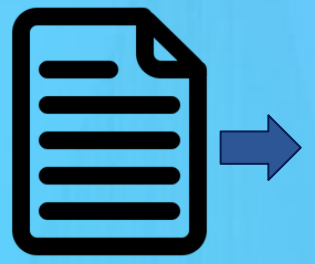
Output model

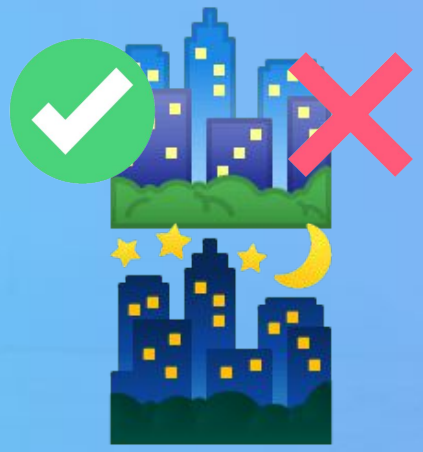
Deploy model in production

Conventional

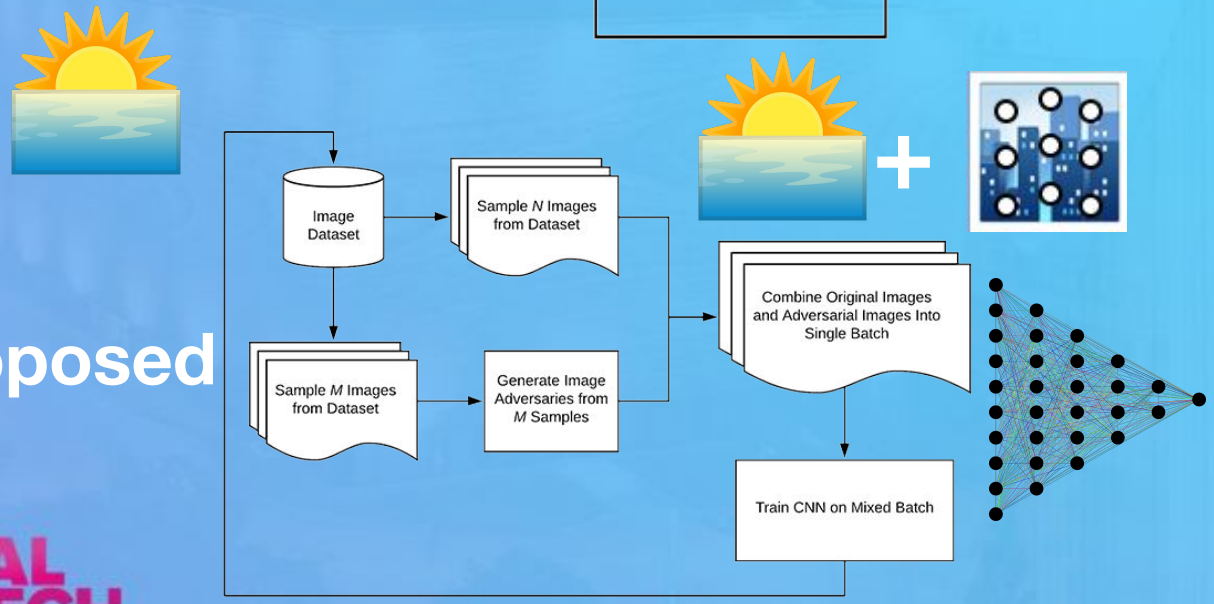








Proposed



(Rosebrock, 2021)

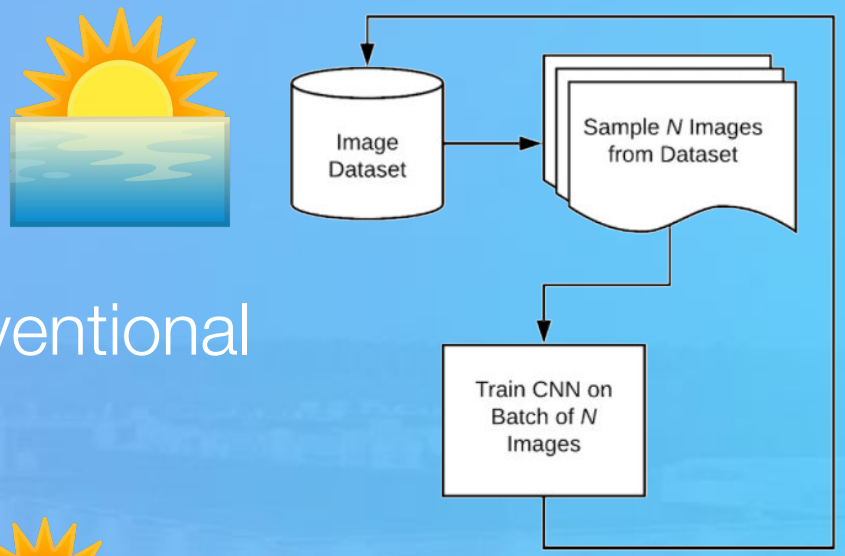
Hypothesis

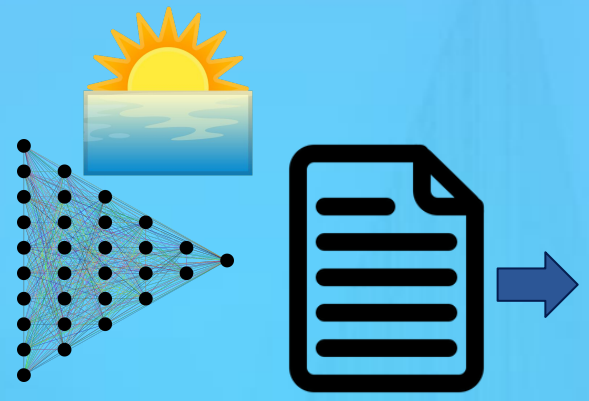
Train model

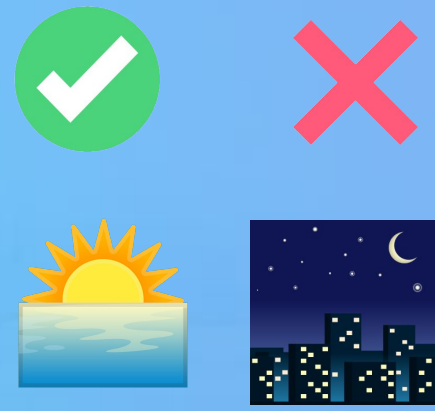
Output model

Deploy model in production

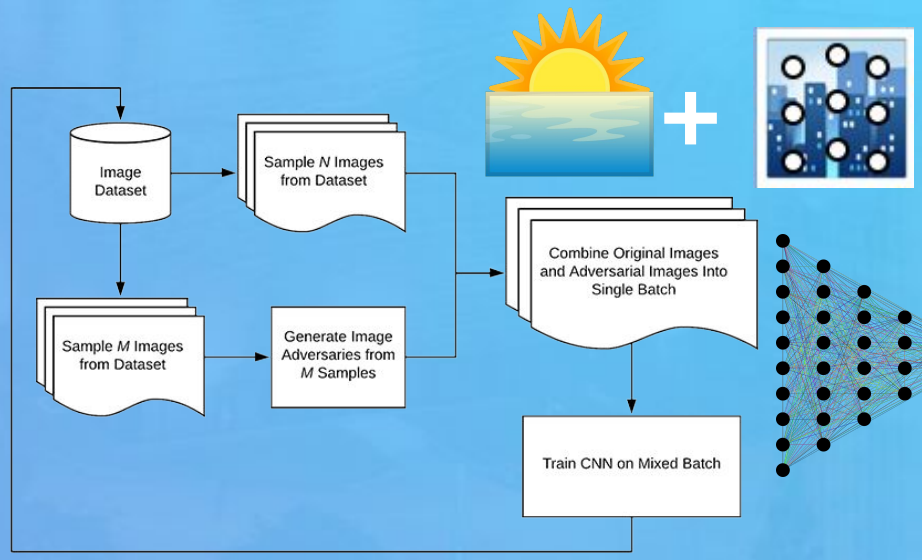
Conventional

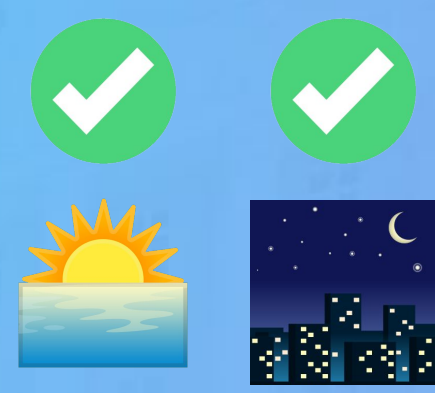






Proposed

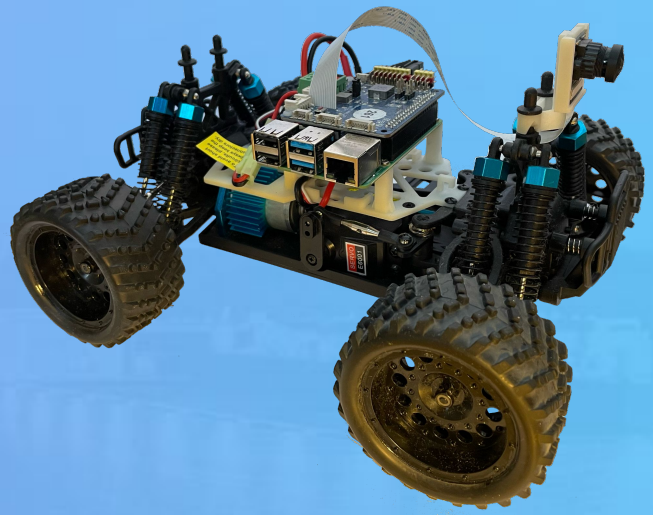




(Rosebrock, 2021)

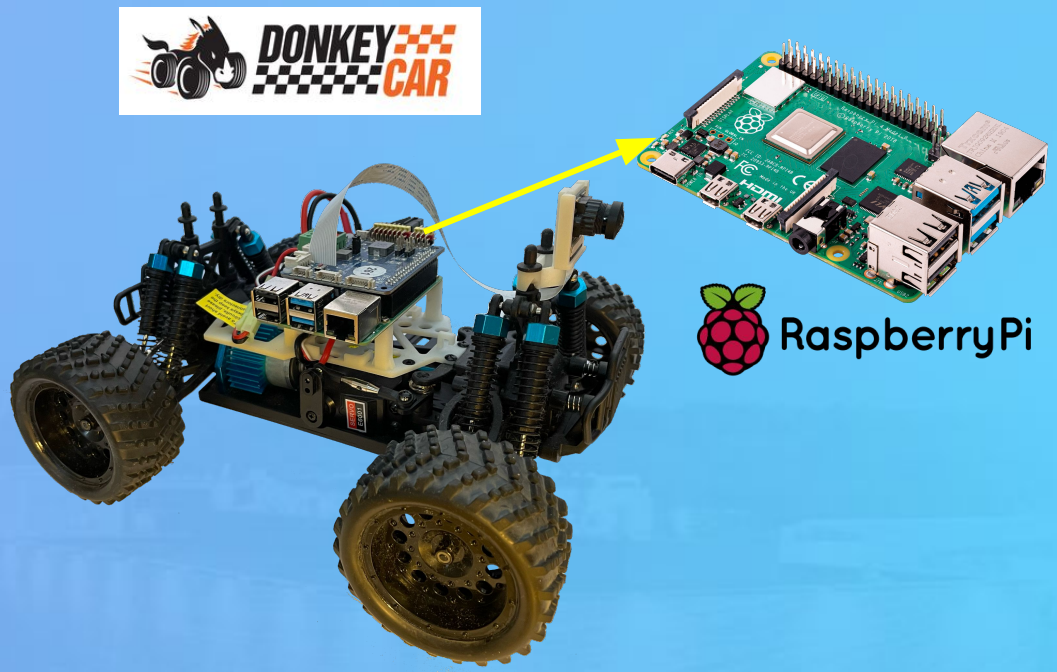
Method

Training CNN models



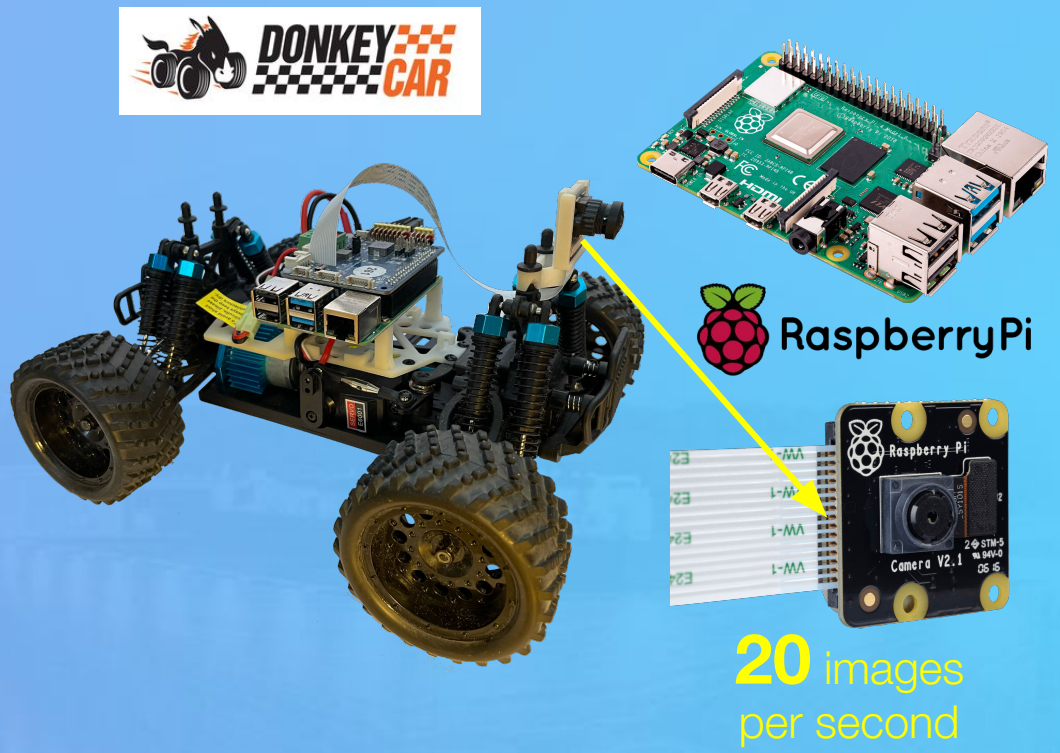
Method

Training CNN models



Method

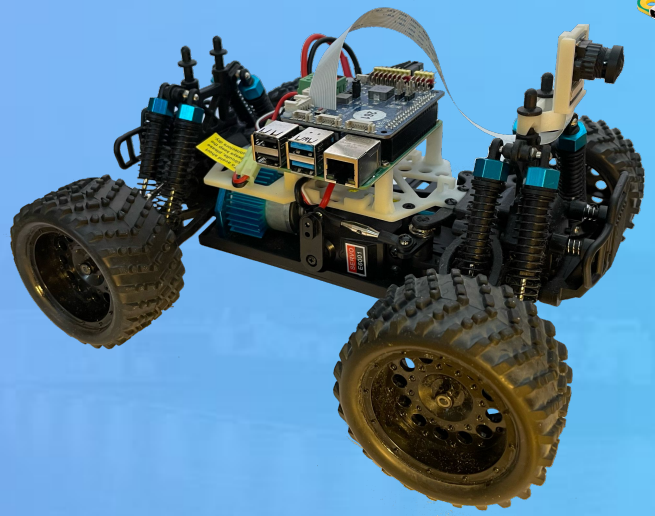
Training CNN models

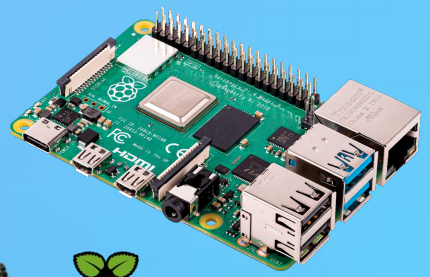


20 images
per second

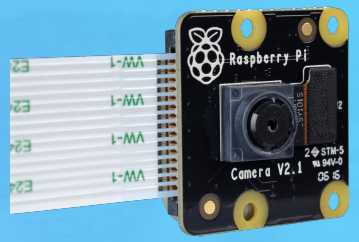
Method

Training CNN models



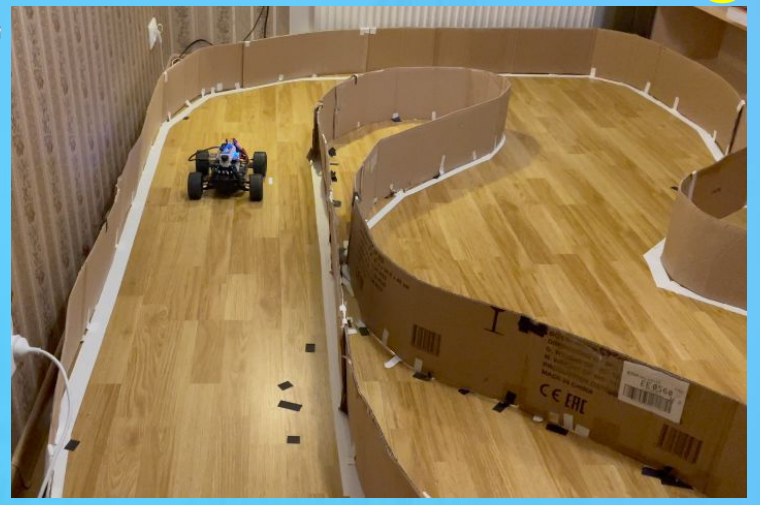


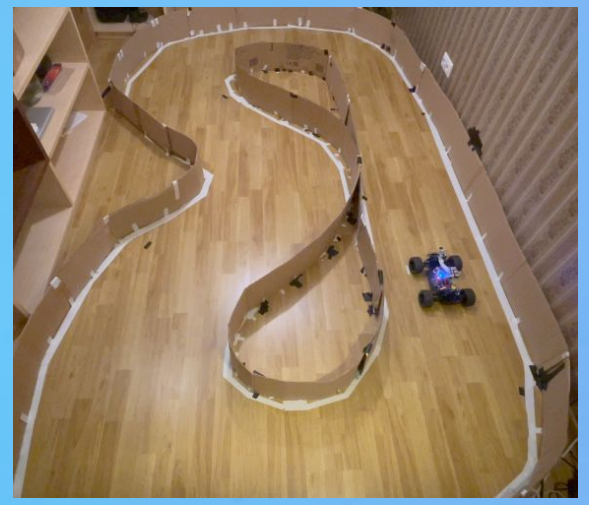
RaspberryPi



20 images
per second

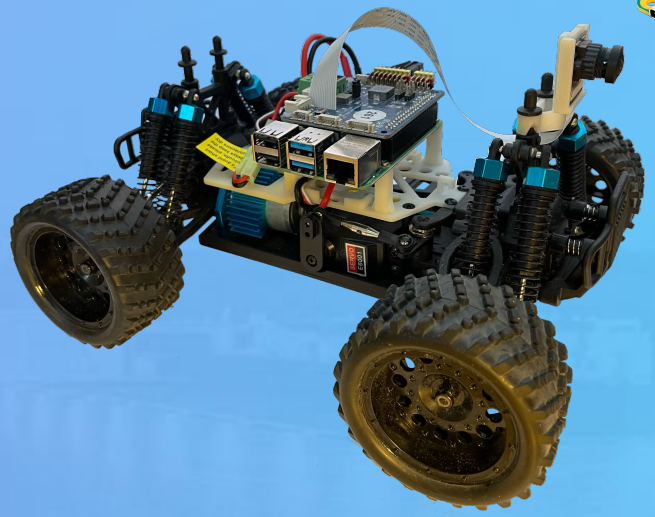
ONLY in lower-lights conditions

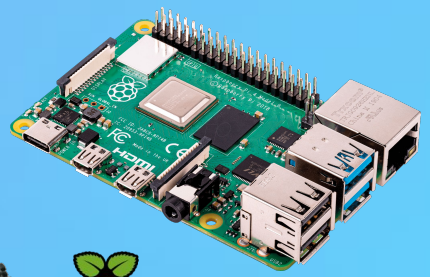




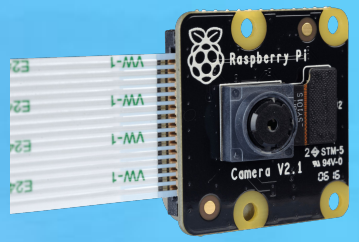
Method

Training CNN models



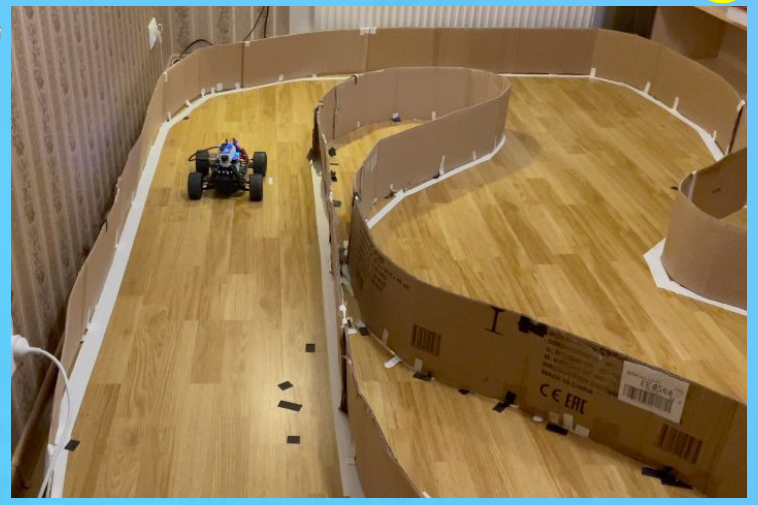


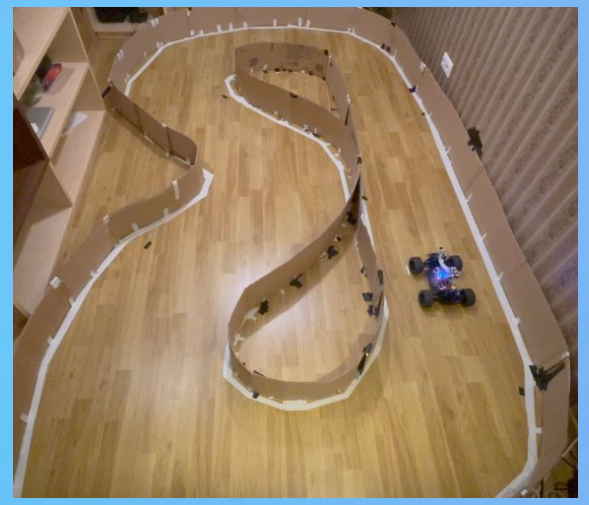
RaspberryPi



20 images per second

ONLY in lower-lights conditions





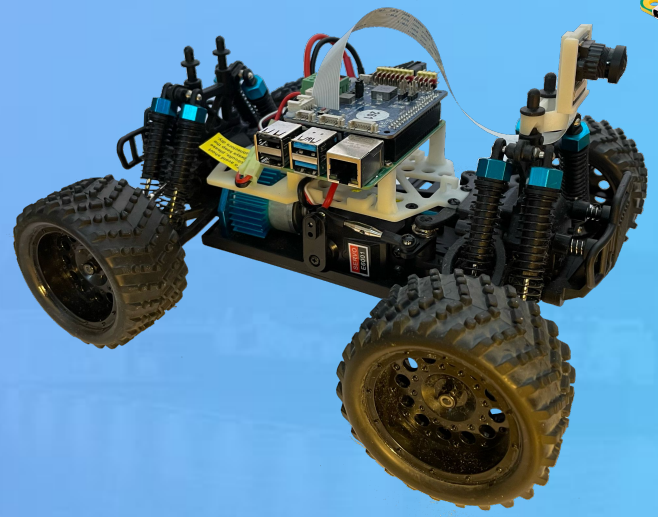
Imitation Learning

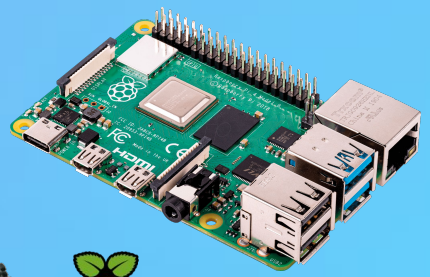
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

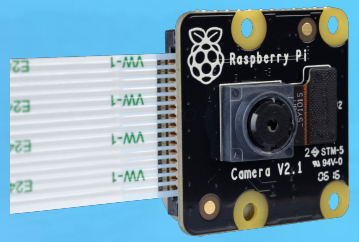
Method

Training CNN models





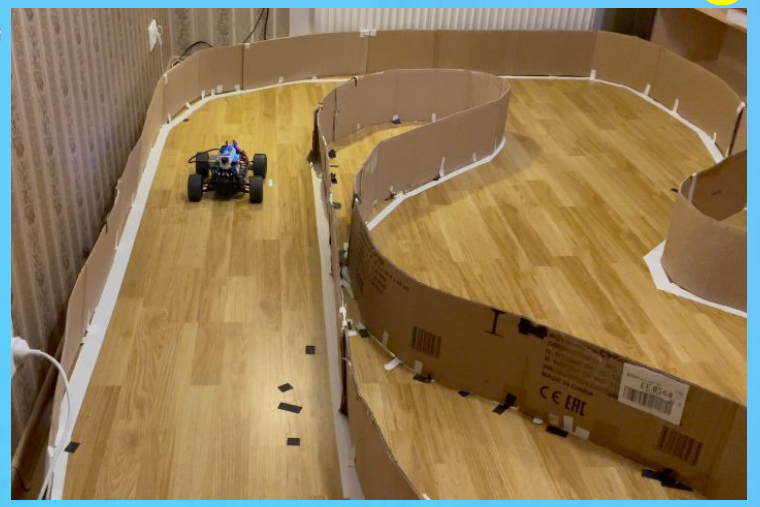
RaspberryPi

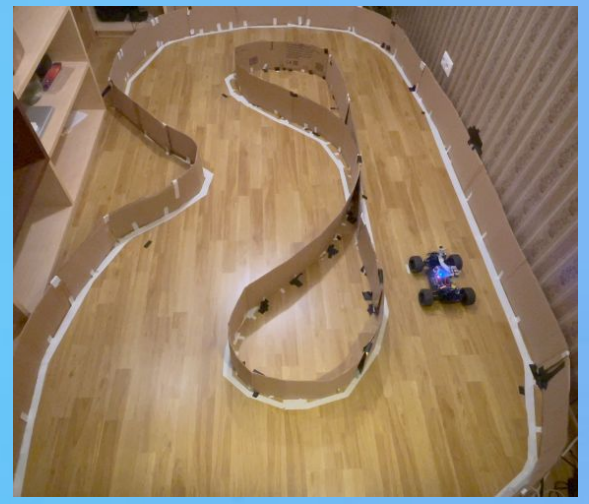


20 images per second

Expert Demonstration

ONLY in lower-lights conditions





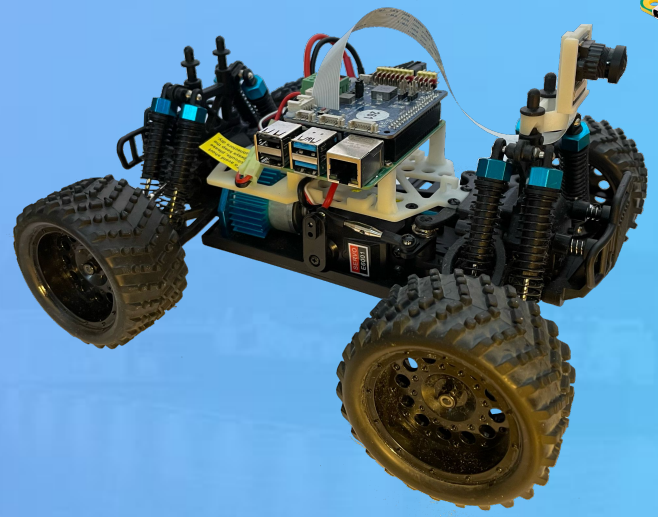
Imitation Learning

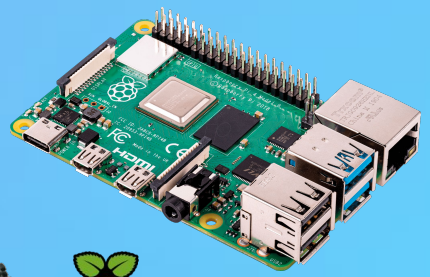
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

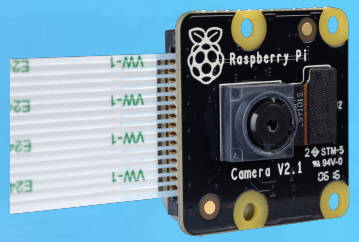
Method

Training CNN models



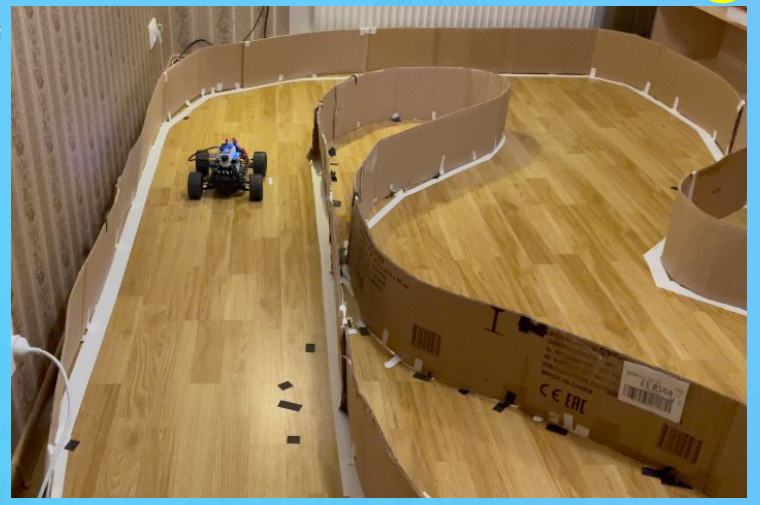


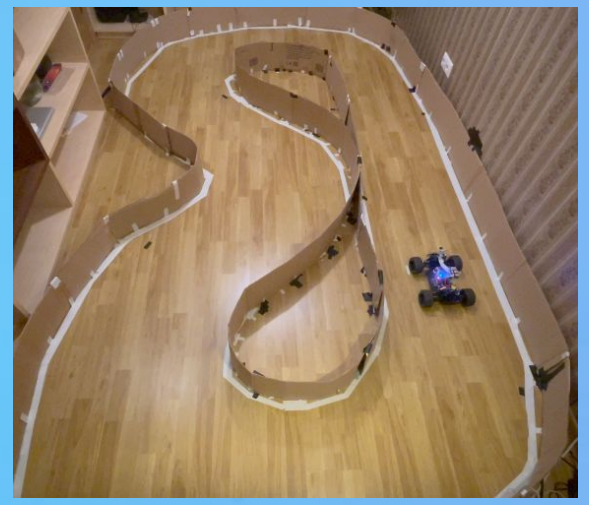
RaspberryPi



20 images per second

ONLY in lower-lights conditions





Expert Demonstration

Dataset Cleaning

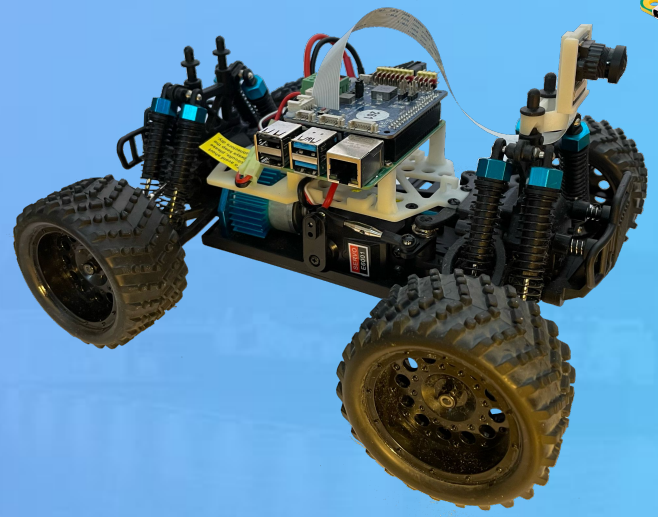
Imitation Learning

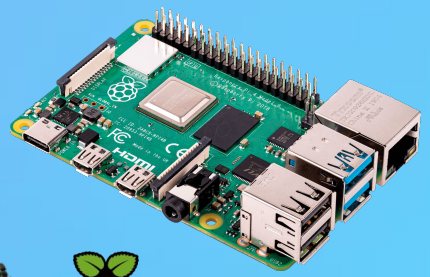
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

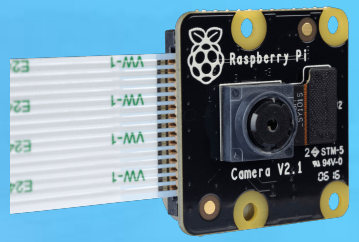
Method

Training CNN models



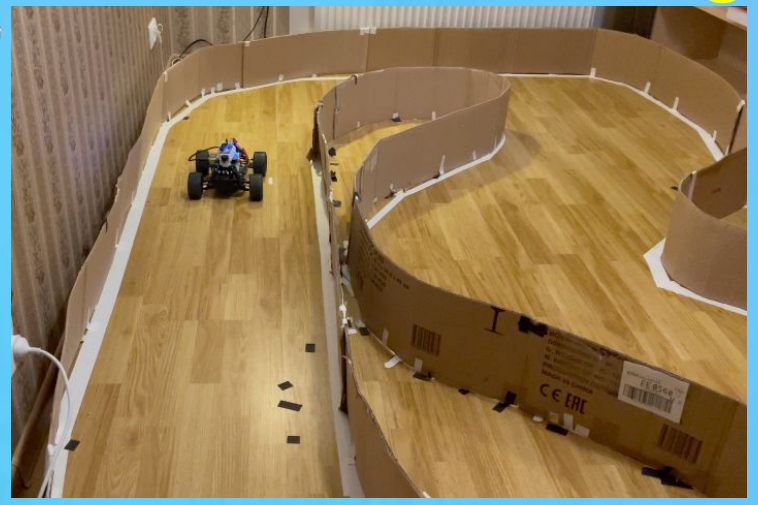


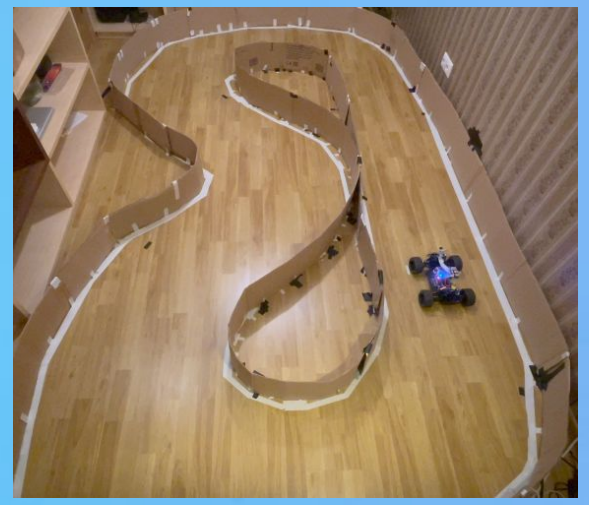
RaspberryPi



20 images per second

ONLY in lower-lights conditions





Expert Demonstration

Dataset Cleaning

CNN Training

Imitation Learning

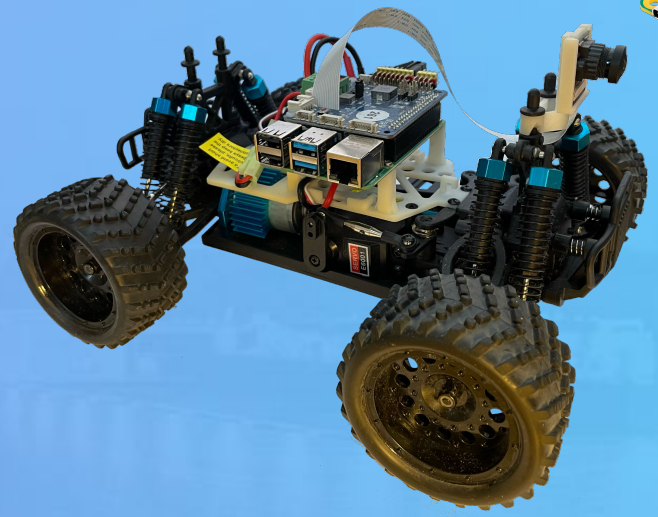
TAL TECH

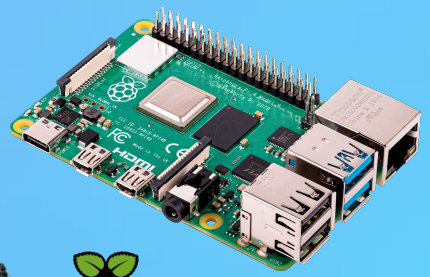
TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Method

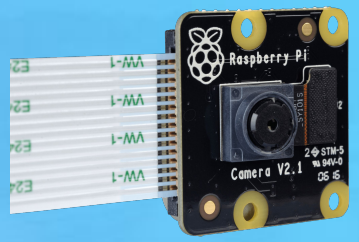
Training CNN models

DONKEY CAR



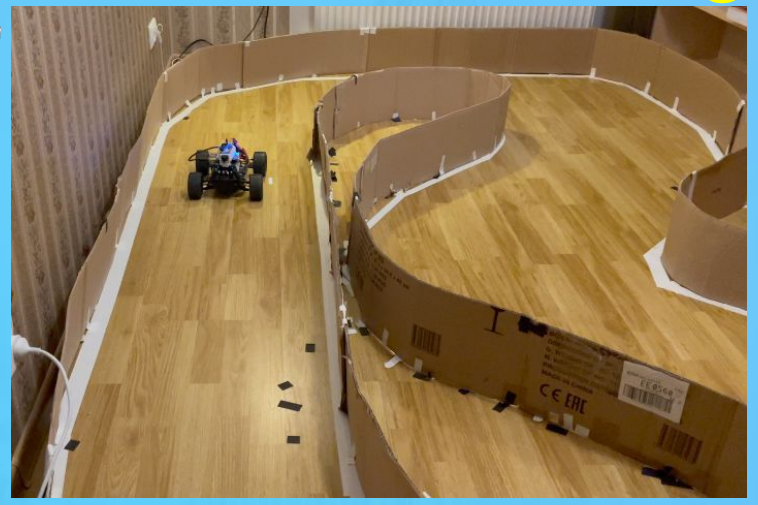


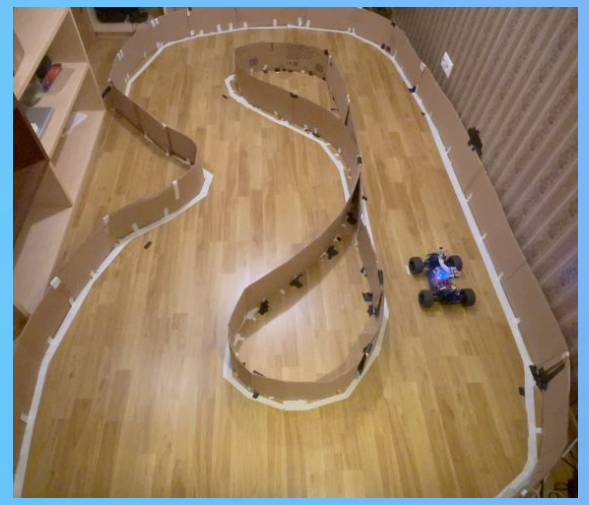
RaspberryPi



20 images per second

ONLY in lower-lights conditions





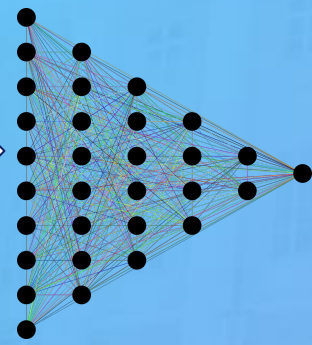
Expert Demonstration

Dataset Cleaning

CNN Training

Model Output

Imitation Learning

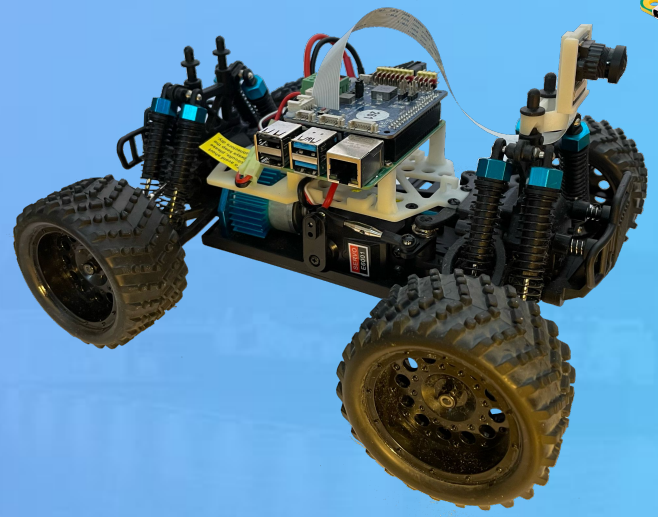


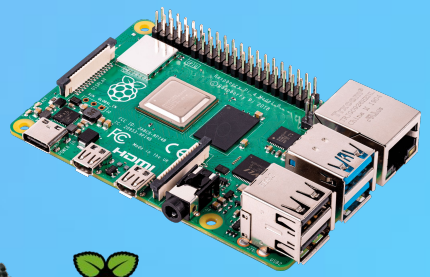
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

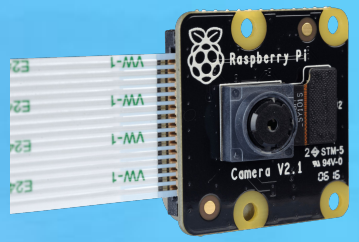
Method

Training CNN models



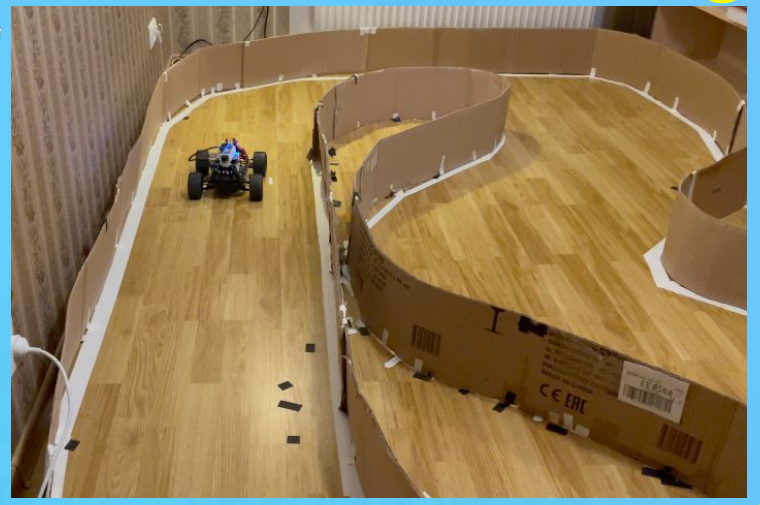


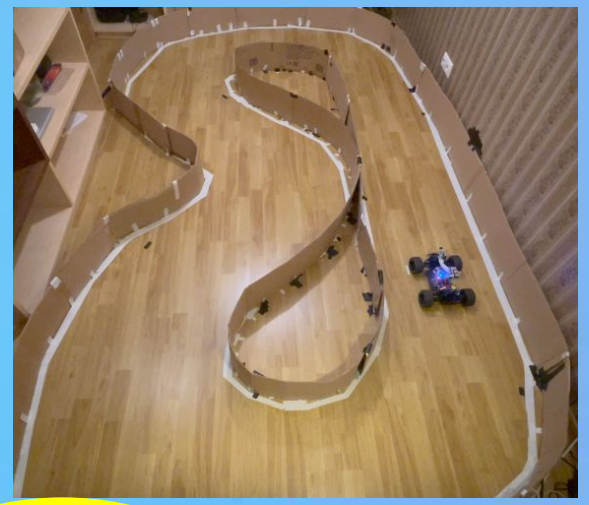
RaspberryPi



20 images per second

ONLY in lower-lights conditions





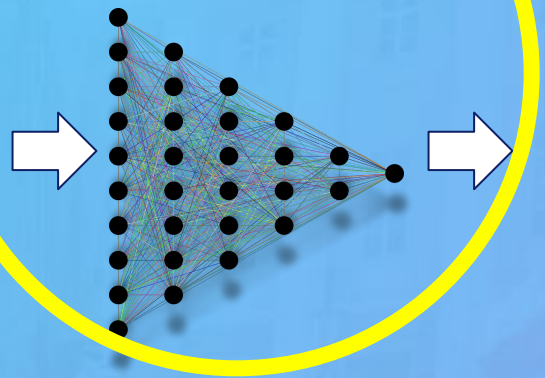
Expert Demonstration

Dataset Cleaning

CNN Training

Model Output

Imitation Learning



TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Method

Training CNN models

Collect dataset
(**only** in lower-lights)

Small
dataset
(~20 laps)

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Method

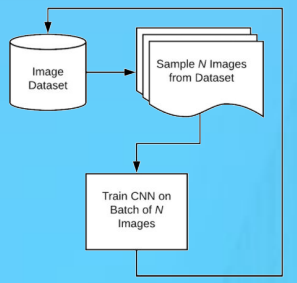
Collect dataset
(**only** in lower-lights)

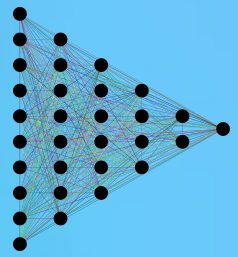
Small
dataset
(~20 laps)

Training CNN models

Method CNN training **CNN** output models

Standard





M-TS

Method

Training CNN models

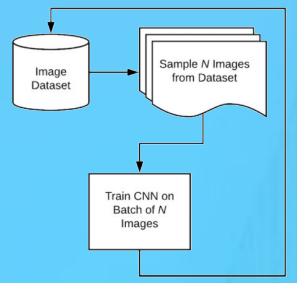
Collect dataset
(**only** in lower-lights)

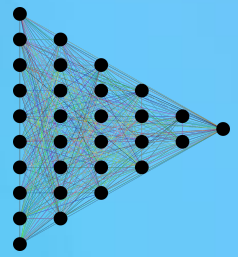
Small

dataset
(~20 laps)

Method CNN training **CNN** output models

Standard

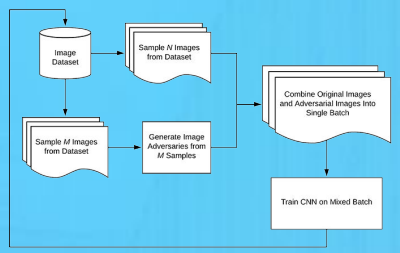


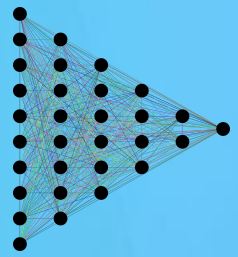


M-TS

Augmented

+





M-TSA

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Method

Training CNN models

Collect dataset
(**only** in lower-lights)

Method CNN training

CNN output models

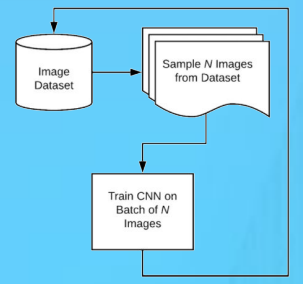
Small

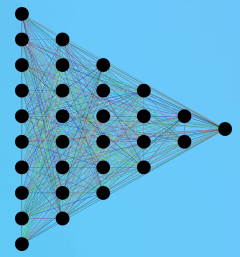
dataset
(~20 laps)

Large

dataset
(~40 laps)

Standard

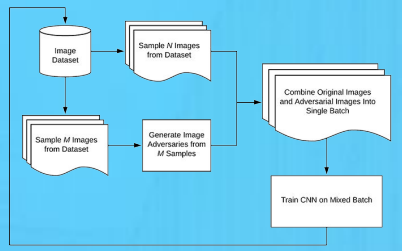


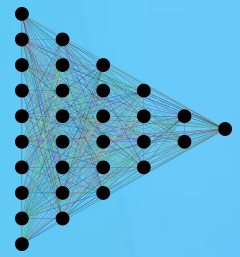


M-TS

Augmented

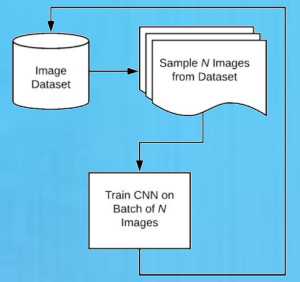
+

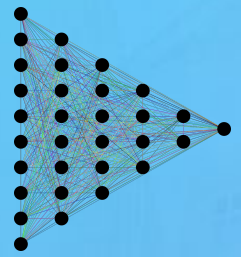




M-TSA

Standard

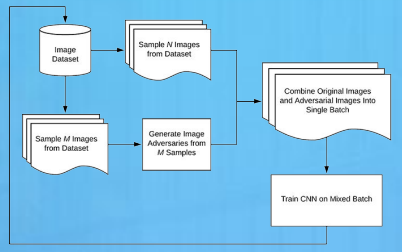


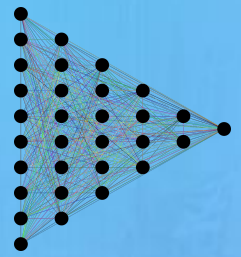


M-TL

Augmented

+





M-TLA

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Method

Evaluating CNN models

Run 2 laps in the following conditions

Lower-lights

L

Higher-lights

H

M-TS

M-TL

M-TSA

M-TLA

Method

Evaluating CNN models

Run 2 laps in the following conditions

Lower-lights

L

Higher-lights

H

M-TS

M-TL

M-TSA

M-TLA

👍

Method

Evaluating CNN models

Run 2 laps in the following conditions

Lower-lights

L

Higher-lights

H

M-TS

M-TL

M-TSA

M-TLA

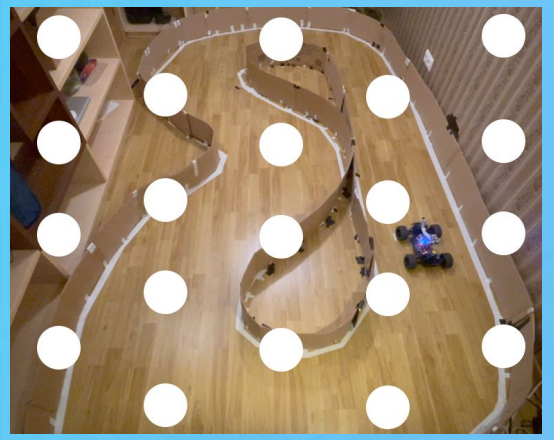
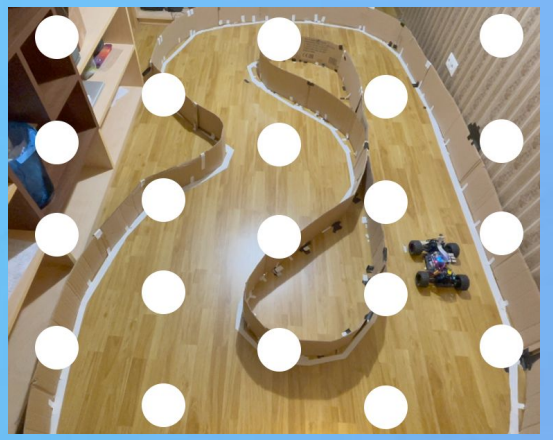
👍

🤞

Method

Evaluating CNN models

Run 2 laps in the following conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	H	LC	HC
M-TS →				
M-TL →				
M-TSA →				
M-TLA →				

👍 🤞

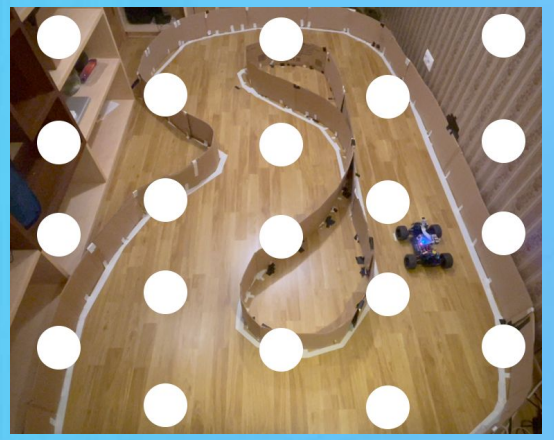
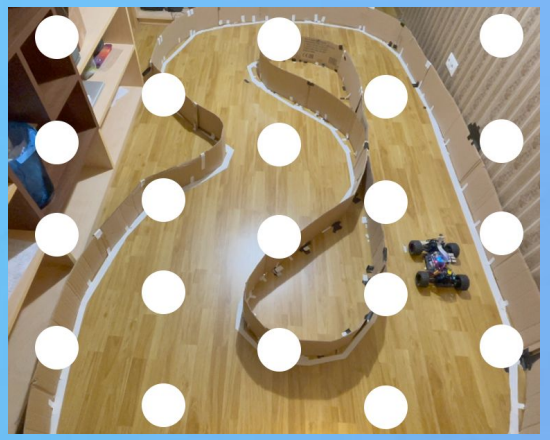
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Method

Evaluating CNN models

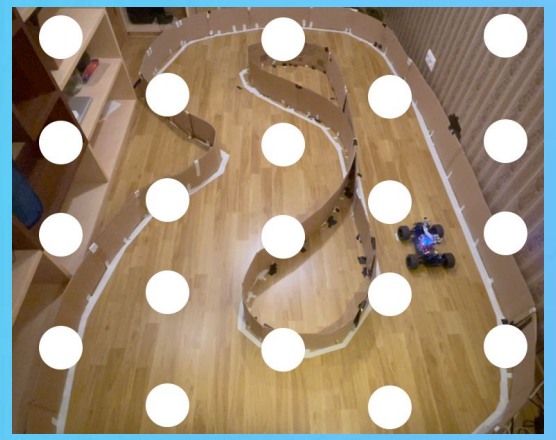
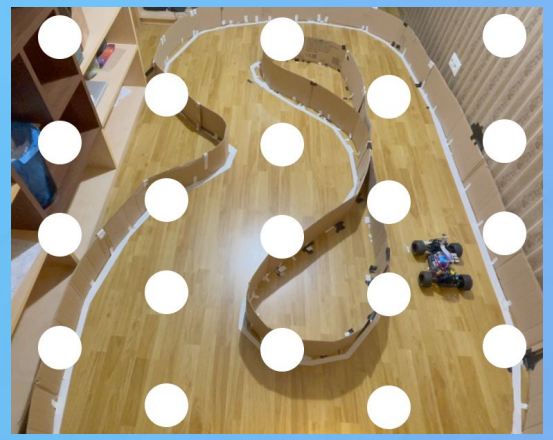
Run 2 laps in the following conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	H	LC	HC
M-TS →				
M-TL →				
M-TSA →				
M-TLA →	👍	🤞	?	?

Method

Evaluating CNN models

Run 2 laps in the following conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	H	LC	HC
M-TS				
M-TL				
M-TSA				
M-TLA				
	👍	🤞	?	?

4 Models **evaluated in** **4** Conditions **=** **16** Evaluations

Method

Evaluating CNN models

Lighting conditions

Lower-lights

L

Higher-lights

H

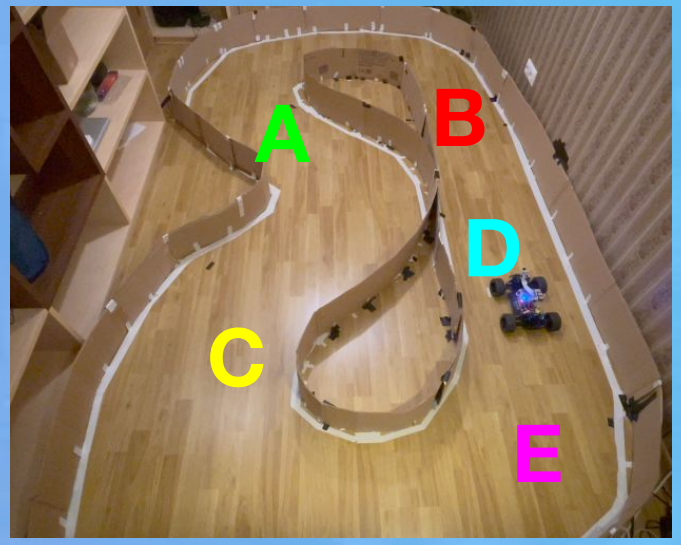
Method

Evaluating CNN models

Lighting conditions

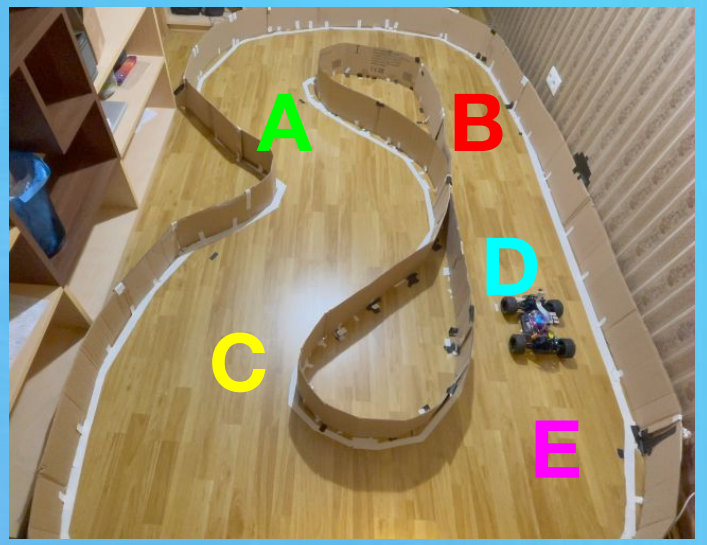
Lower-lights

L



Higher-lights

H



Location	Lower-lights L	Higher-lights H
A	21	110
B	19	86
C	19	75
D	19	75
E	20	60
Average	**19.6 lux**	**81.2 lux**

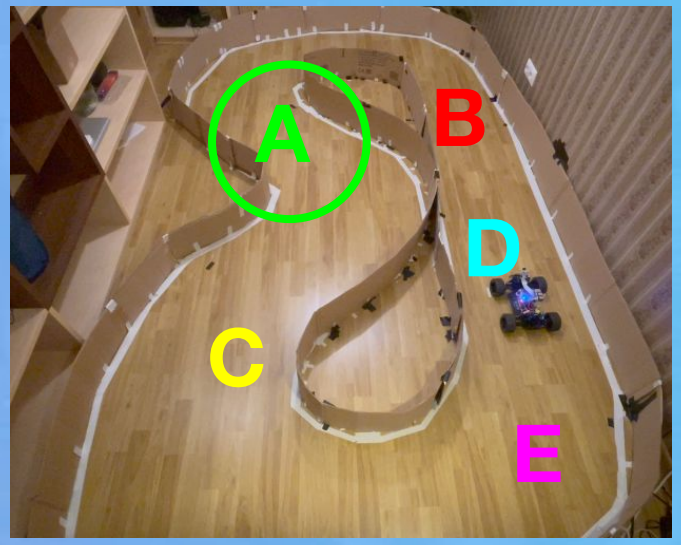
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

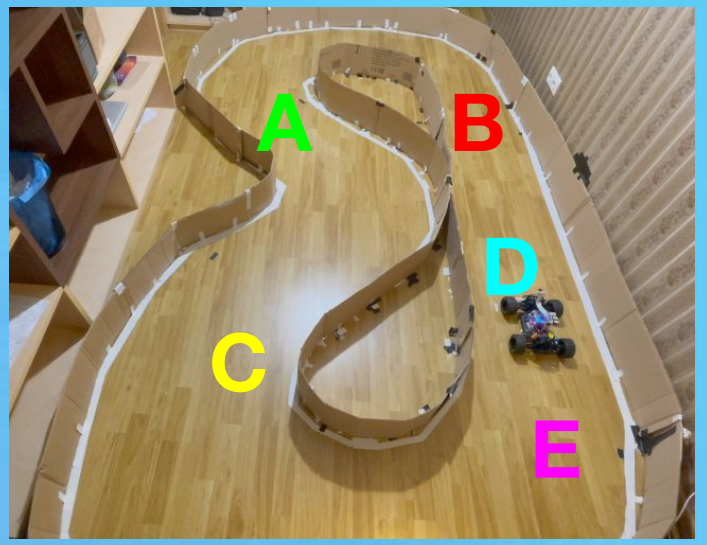
Method

Lighting conditions

Lower-lights
L



Higher-lights
H



Evaluating CNN models

Location	Lower-lights L	Higher-lights H
A	21	110
B	19	86
C	19	75
D	19	75
E	20	60
Average	**19.6 lux**	**81.2 lux**

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

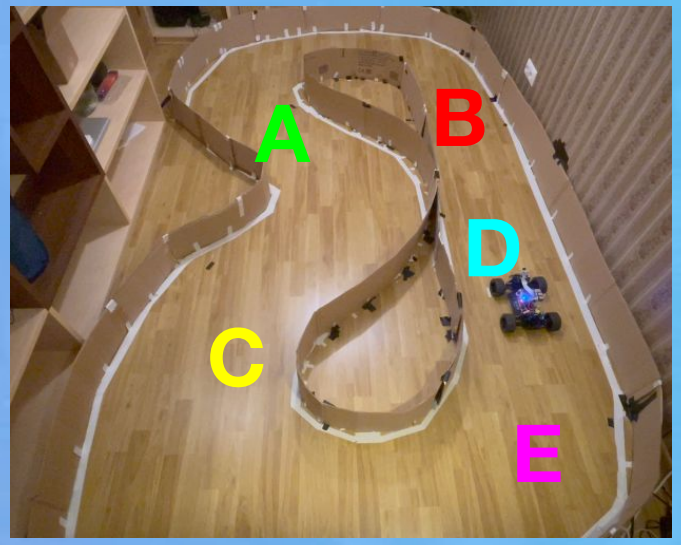
Method

Evaluating CNN models

Lighting conditions

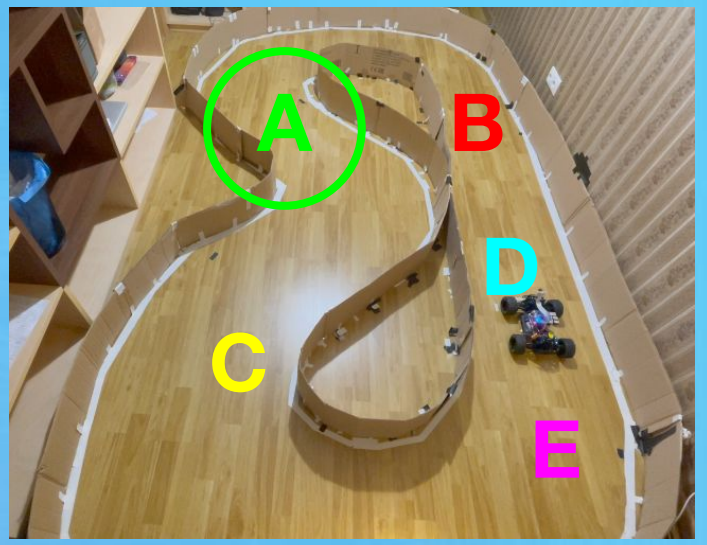
Lower-lights

L



Higher-lights

H



Location	Lower-lights L	Higher-lights H
A	21	110
B	19	86
C	19	75
D	19	75
E	20	60
Average	**19.6 lux**	**81.2 lux**

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

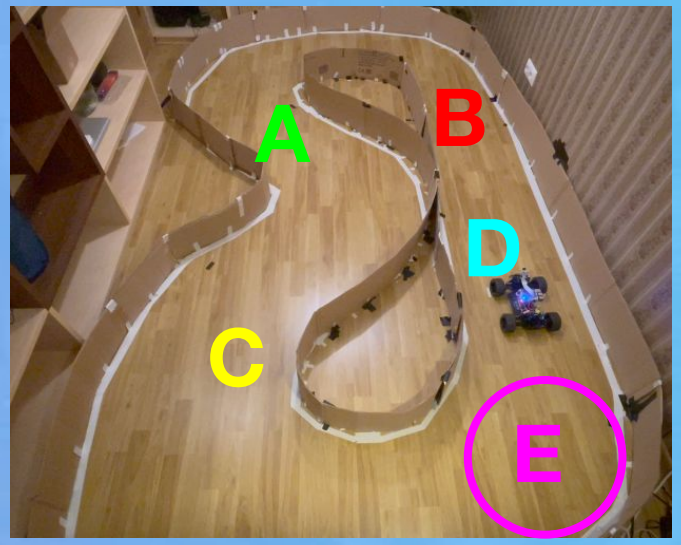
Method

Evaluating CNN models

Lighting conditions

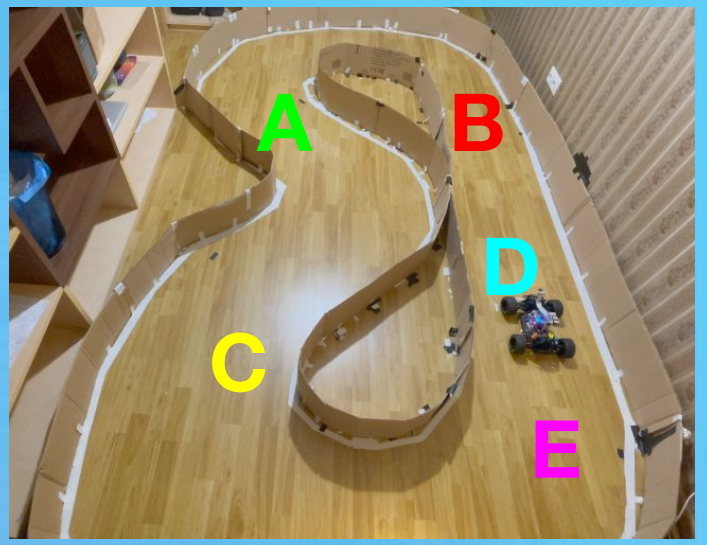
Lower-lights

L



Higher-lights

H

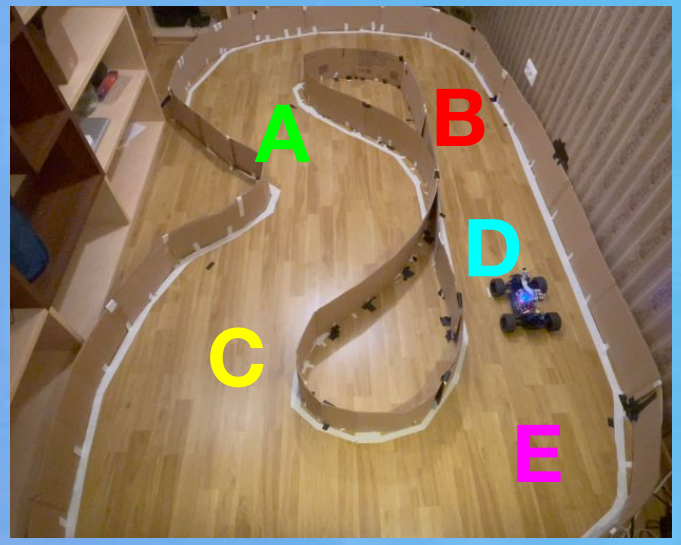


Location	Lower-lights L	Higher-lights H
A	21	110
B	19	86
C	19	75
D	19	75
E	20	60
Average	19.6 lux	81.2 lux

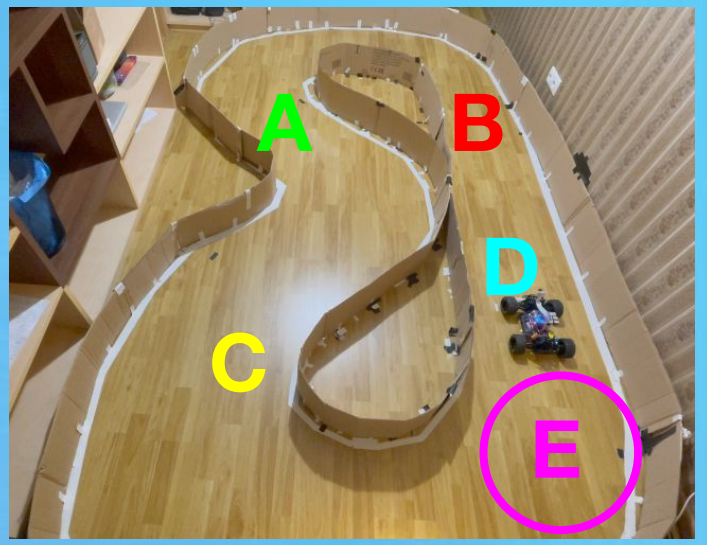
Method

Lighting conditions

Lower-lights
L



Higher-lights
H



Evaluating CNN models

Location	Lower-lights L	Higher-lights H
A	21	110
B	19	86
C	19	75
D	19	75
E	20	60
Average	**19.6 lux**	**81.2 lux**

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Method

Evaluating CNN models

Evaluating conditions (continued)

Independent variable	Values
Laps	**2**
Lights	**low, high**
Vision	**uncorrupted, corrupted**
Dependent variables	Values
Collisions	**[0...n]**

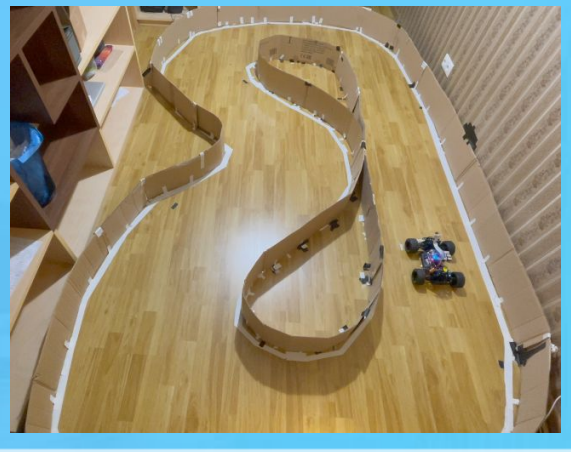
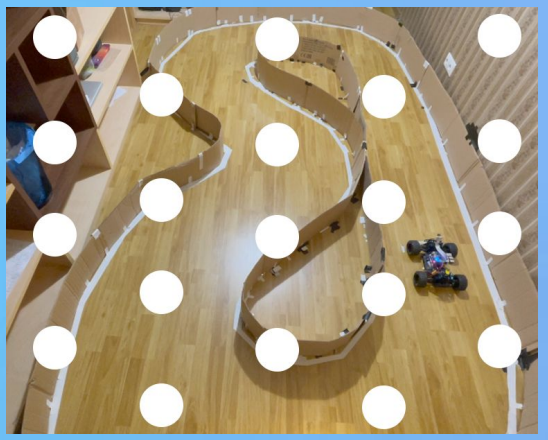
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Results Collisions in **2 laps**

👁️ Seen conditions

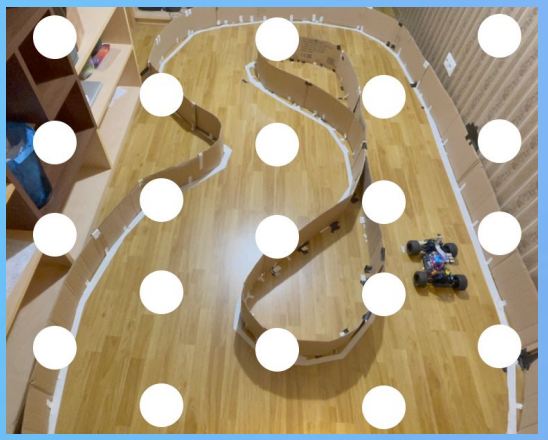
🤔 Never-seen-before conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS				
M-TL				
M-TSA				
M-TLA				

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Results Collisions in **2 laps**

	👁️ Seen conditions		🤔 Never-seen-before conditions	
	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0			
M-TL	✅ 0			
M-TSA				
M-TLA				

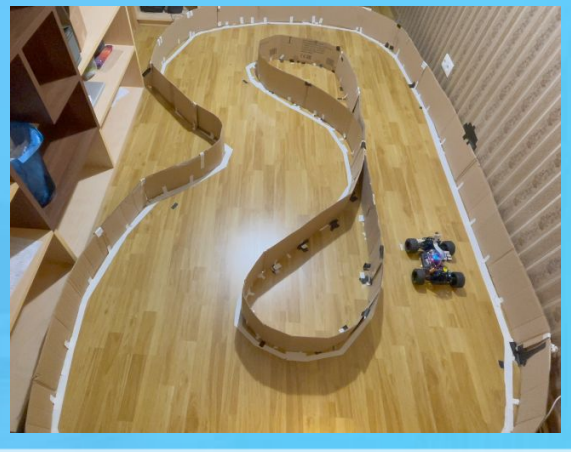
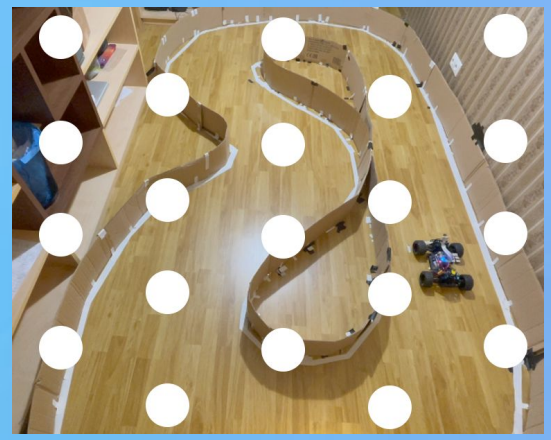
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Results Collisions in **2 laps**

👁️ Seen conditions

🤔 Never-seen-before conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0			
M-TL	✅ 0			
M-TSA	✅ 0			
M-TLA	✅ 0			

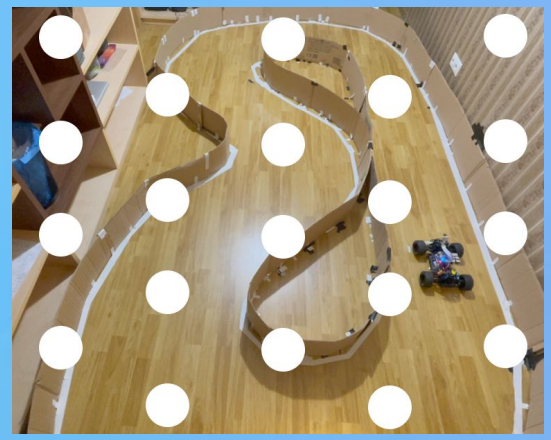
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Results Collisions in **2 laps**

👁️ Seen conditions

🤔 Never-seen-before conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0	❌ 5		
M-TL	✅ 0	❌ 4		
M-TSA	✅ 0			
M-TLA	✅ 0			

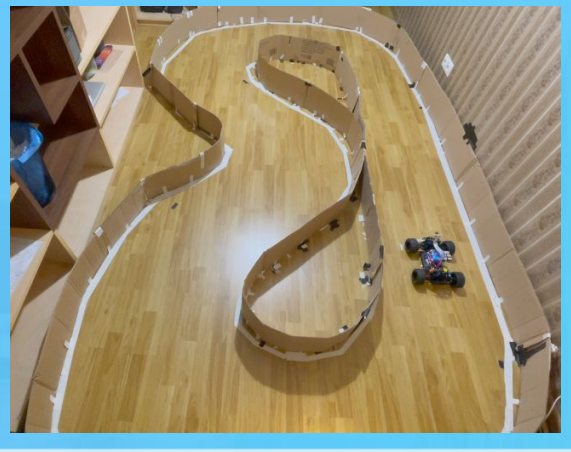
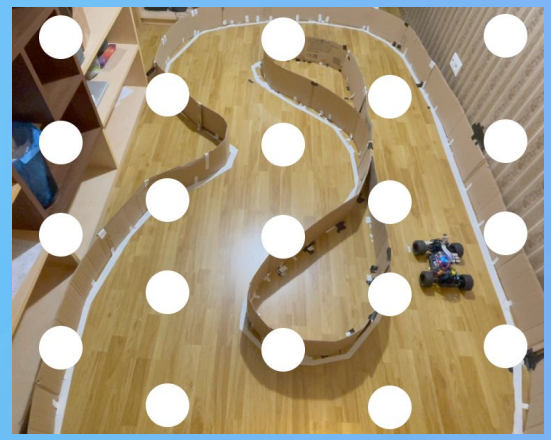
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Results Collisions in **2 laps**

👁️ Seen conditions

🤔 Never-seen-before conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0	❌ 5		
M-TL	✅ 0	❌ 4	🤦	
M-TSA	✅ 0			
M-TLA	✅ 0			

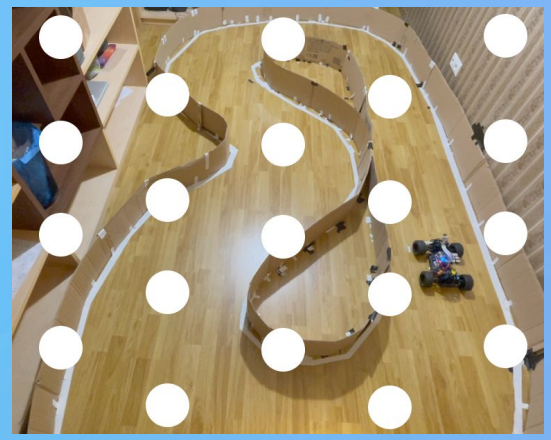
TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Results Collisions in **2 laps**

👁️ Seen conditions

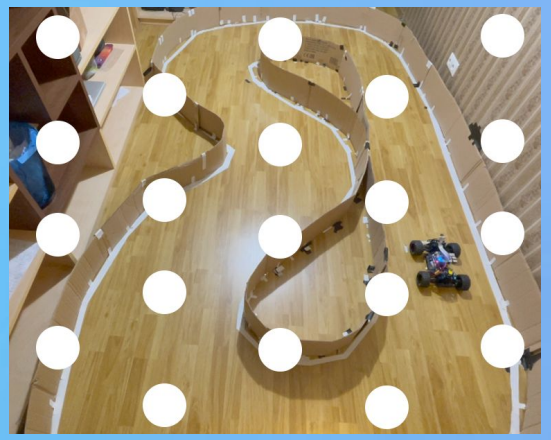
🤔 Never-seen-before conditions

	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0	❌ 5		
M-TL	✅ 0	❌ 4		
M-TSA	✅ 0	✅ 0 👍		
M-TLA	✅ 0	✅ 0		

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

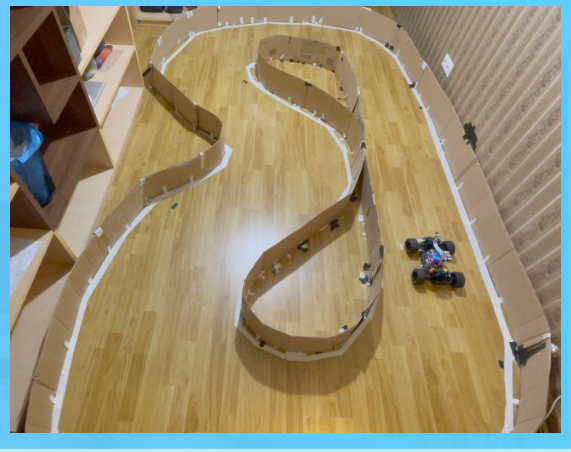
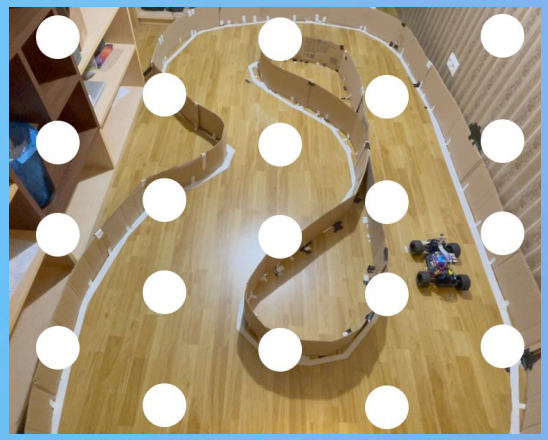
Results Collisions in **2 laps**

	👁️ Seen conditions		🤔 Never-seen-before conditions	
	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0	❌ 5		
M-TL	✅ 0	❌ 4		
M-TSA	✅ 0	✅ 0	❌ 2	❌ 5
M-TLA	✅ 0	✅ 0	❌ 3	❌ 5

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

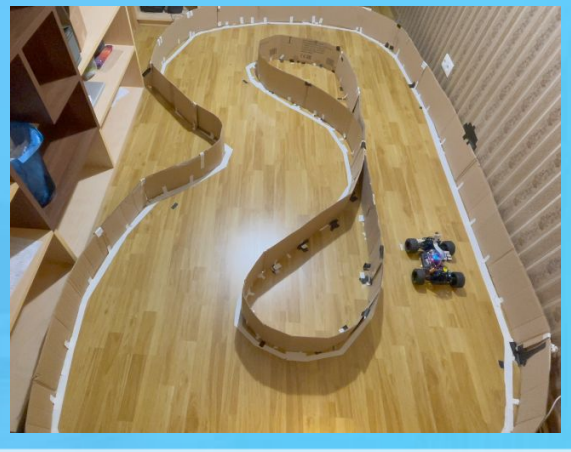
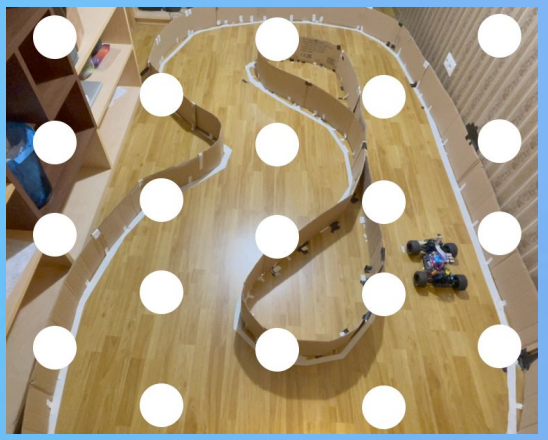
Results Collisions in **2 laps**

	👁️ Seen conditions		🤔 Never-seen-before conditions	
	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0	❌ 5	❌ 12	❌ 7
M-TL	✅ 0	❌ 4	❌ 10	❌ 9
M-TSA	✅ 0	✅ 0	❌ 2	❌ 5
M-TLA	✅ 0	✅ 0	❌ 3	❌ 5

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Results Collisions in **2 laps**

	👁️ Seen conditions		🤔 Never-seen-before conditions	
	Lower-lights	Higher-lights	Lower-lights corrupted	Higher-lights corrupted
	L	**H**	**LC**	**HC**
				
M-TS	✅ 0	❌ 5	❌ 12	❌ 7
M-TL	✅ 0	❌ 4	❌ 10	❌ 9
M-TSA	✅ 0	✅ 0	❌ 2	❌ 5
M-TLA	✅ 0	✅ 0	❌ 3	❌ 5

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Answer to the **Research Question**

"Can adversarial defense training methods improve the neural network generalization skills to unseen lighting conditions?"

Answer to the **Research Question**

"Can adversarial defense training methods improve the neural network generalization skills to unseen lighting conditions?"

Yes, it can.

Answer to the **Research Question**

"Can adversarial defense training methods improve the neural network generalization skills to unseen lighting conditions?"

Yes, it can.

Standard method

- ✔ Seen lower-lights
- ✘ Never seen before higher-lights

Augmented training method

- ✔ Seen lower-lights
- ✔ Never seen before higher-lights

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Answer to the **Research Question**

"Can adversarial defense training methods improve the neural network generalization skills to unseen lighting conditions?"

Yes, it can.

Standard method

- ✔ Seen lower-lights
- ✘ Never seen before higher-lights
- ✘ Vulnerable to adversarial attack

Augmented training method

- ✔ Seen lower-lights
- ✔ Never seen before higher-lights
- 🛡 Reduced collisions under attack

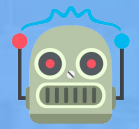
TAL TECH

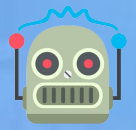
TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Future Work


Measure differences with training dataset and evaluation dataset


Image augmentations


More diverse selection of attack techniques


Use datasets constructed only from corrupted images


Larger volume of samples

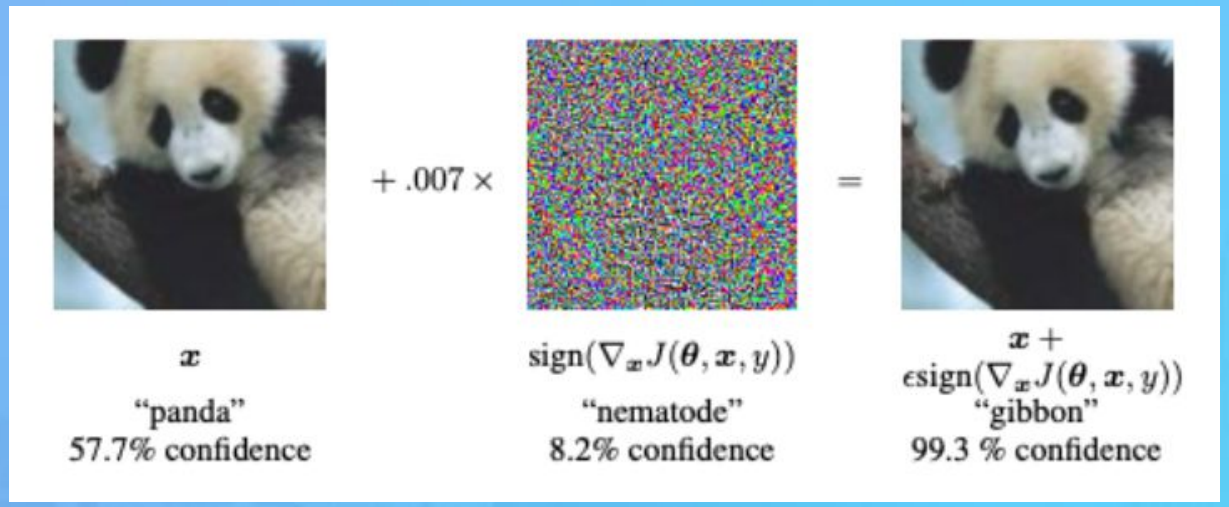
Using Adversarial Defense Methods to Improve the Performance of Deep-Neural-Network-Controlled Automatic Driving Systems

Mike Camara

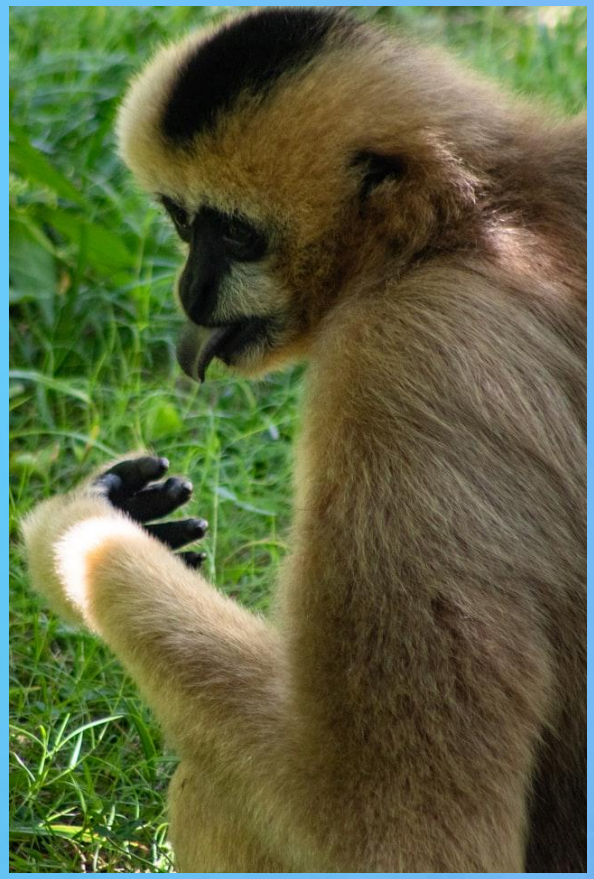
Questions?

Backup
Slides

Machine Learning Adversarial Attacks



Explaining and Harnessing Adversarial Examples
(Goodfellow, Shlens, & Szegedy, 2015)



📸 A Gibbon by @jcotten

White Box

Black Box

Evasion modify input to affect model (FGSM)

Poisoning modify training data to add vulnerabilities

Inference learn confidential private data

Extraction theft of a proprietary model

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Adversarial Robustness Toolbox - ART (IBM Think, 2021)

Attacks

Evasion
Fast Gradient Sign Method (FGSM) – white-box
DeepFool – white-box
Virtual Adversarial Method – white-box
Carlini & Wagner L2 and Linf – white-box
Basic Iterative Method – white-box
Jacobian Saliency Map – white-box
Universal Perturbation – white-box
Decision Tree Attack – white-box
Projected Gradient Descent (PGD) – white-box
NewtonFool – white-box
Elastic Net – white-box
Spatial Transformation – black-box
Query-efficient Black-box – black-box
Zeroth-Order Optimization (ZOO) – black-box
Adversarial Patch – white-box
Boundary Attack – black-box
High Confidence Low Uncertainty (HCLU) – white-box
HopSkipJump – black-box
AutoAttack
AutoPGD – white-box
Brendel&Bethge – white-box
Feature Adversaries – white-box
Shadow Attack – white-box
Simple Black-box Attack (SimBA) – black-box
Wasserstein Attack – white-box
Geometric Decision-based Attack (GeoDA) – black-box

Poisoning
Poisoning Attack on SVM
Adversarial Embedding
Bullseye Polytope
Clean-label Backdoor
Feature Collision

Extraction
Functionally Equivalent Extraction
Copycat CNN
Knockoff Nets

Inference
Membership inference
Attribute inference
Model inversion
Database reconstruction

Defences:

Adversarial Training
Adversarial Training
Adversarial Training Madry-PGD
Fast is Better than Free

Preprocessing
Thermometer encoding
Total variance minimization
PixelDefend
Gaussian data augmentation
Feature squeezing
Spatial smoothing
JPEG compression
Label smoothing

Postprocessing
Reverse Sigmoid
Random Noise
Class Labels
High Confidence
Rounding

Certification
Randomized Smoothing

Metrics
CLEVER
Loss sensitivity
Empirical robustness

Verification
Clique Method Robustness Verification

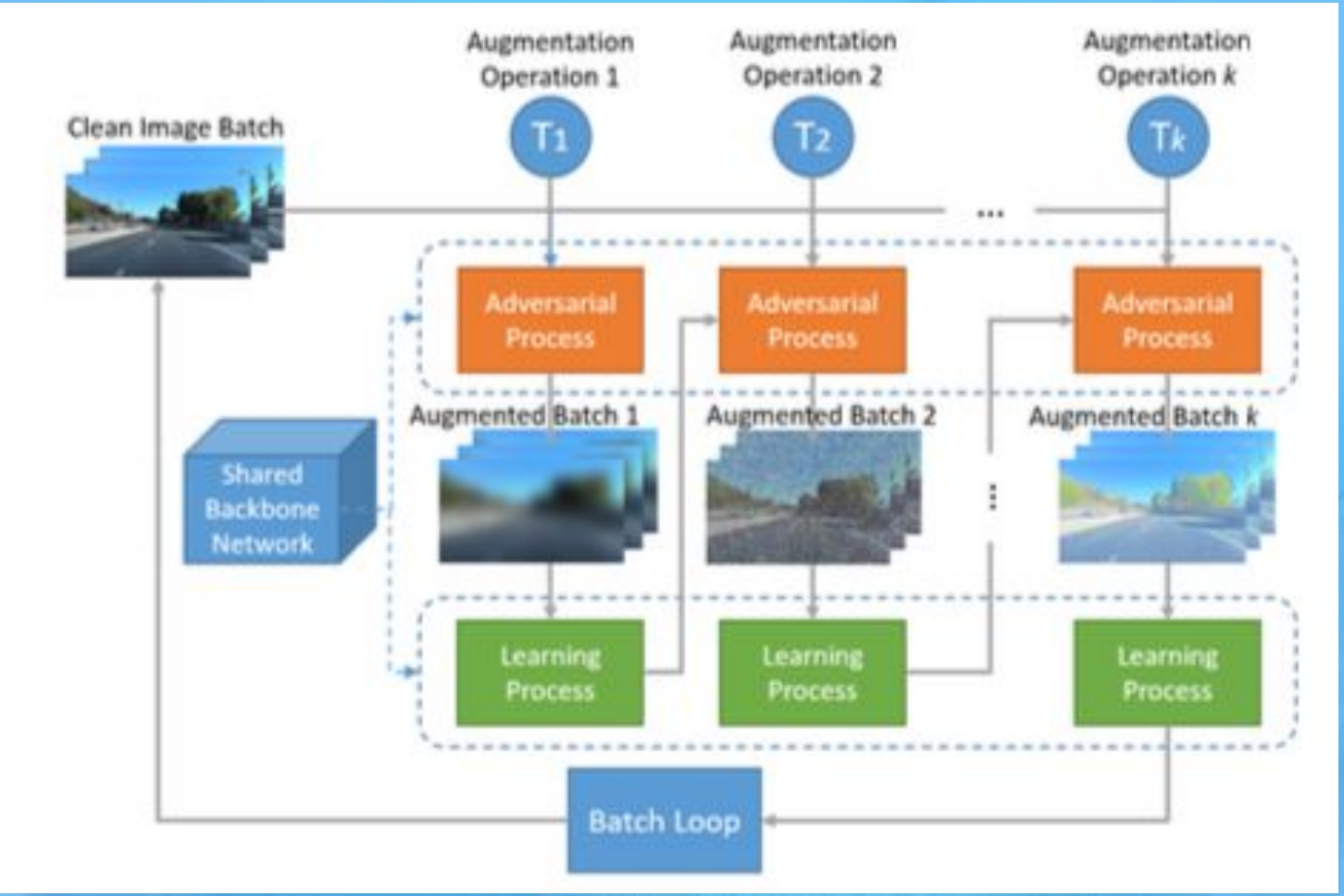
Attack and Fault Injection in Self-Driving Agents on the CARLA Simulator

(Piazzesi, Hong, & Ceccarelli, 2021)

Measured impact of a range of adversarial attacks on the performance of Automatic Driving CNN's deployed in CARLA Simulator environment react to several different adversarial attacks by analyzing the rate of collisions in various circuits in a specific set of predefined conditions.

Adversarial Differentiable Data Augmentation for Autonomous Systems

(M. Shu, Y. Shen, M. C. Lin, and T. Goldstein, 2021)

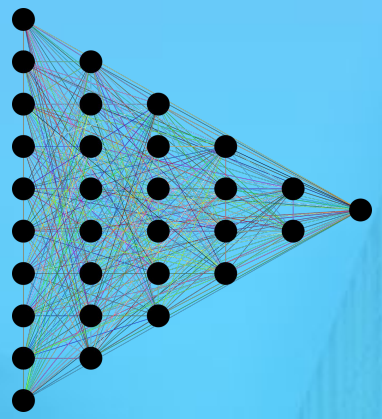


"Adversarially training against image corruptions it results having CNN's with higher levels of robustness and accuracy for a range of settings as compared to baseline and state-of-the-art augmentation methods."

Mixing Normal Images and Adversarial Images when Training CNNs

(A. Rosebrock, 2021)

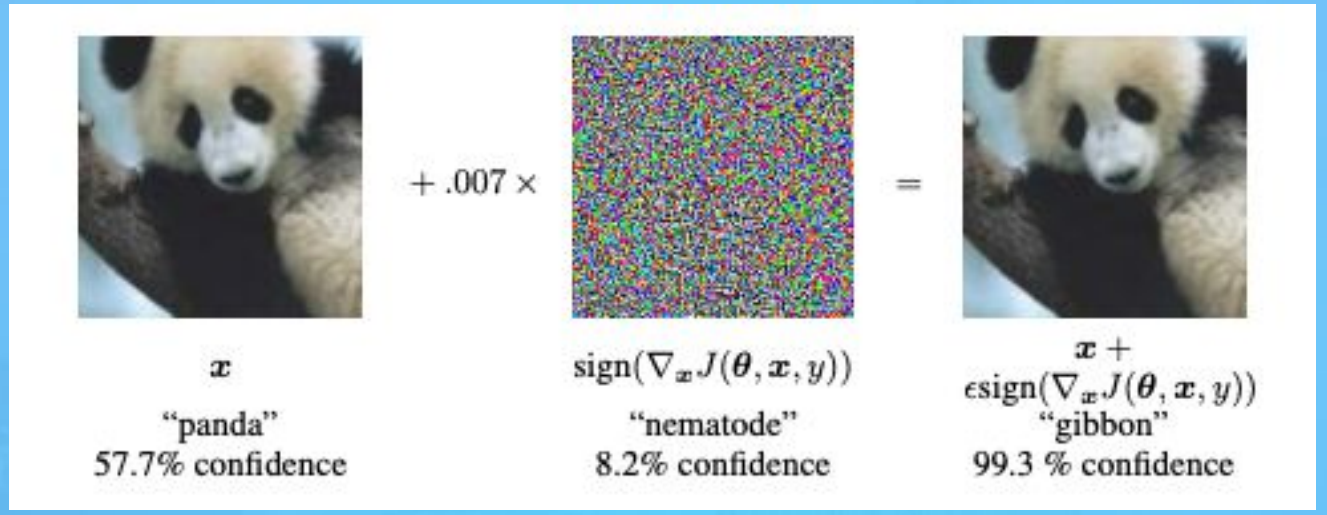
```
normal testing images:
loss: 0.0477, acc: 0.9891

generating adversarial examples with FGSM...

adversarial testing images:
loss: 14.0658, acc: 0.0188
```



```
normal testing images *after* fine-tuning:
loss: 0.0315, acc: 0.9906

adversarial images *after* fine-tuning:
loss: 0.1190, acc: 0.9641
```

"Mixed batch training will make your model become more robust and generalize better" - Rosebrock

Fast Gradient Sign Method (FGSM)



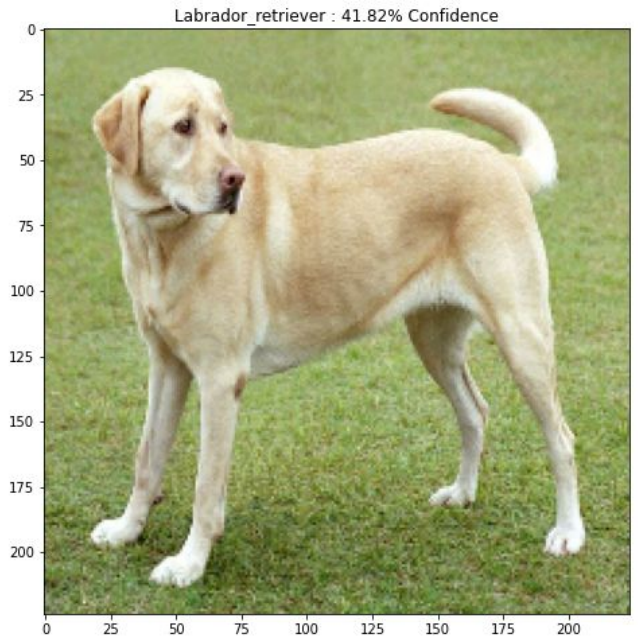
(Goodfellow, Shlens, & Szegedy, 2015)

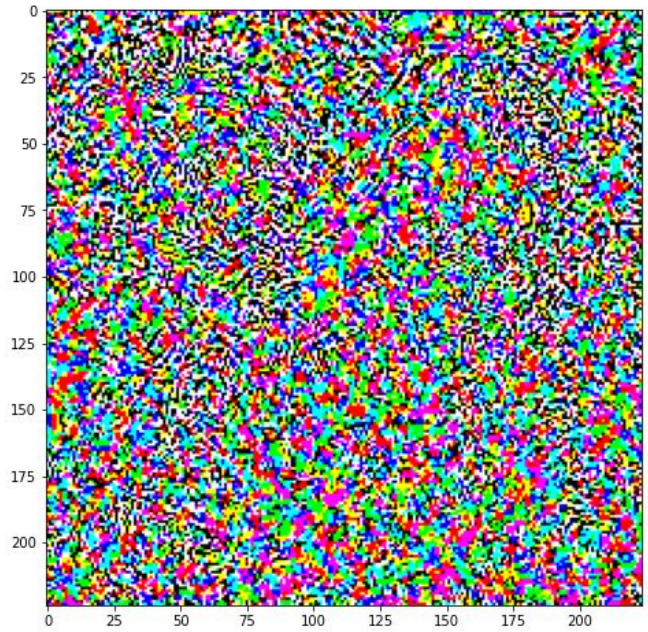
ε
0.8

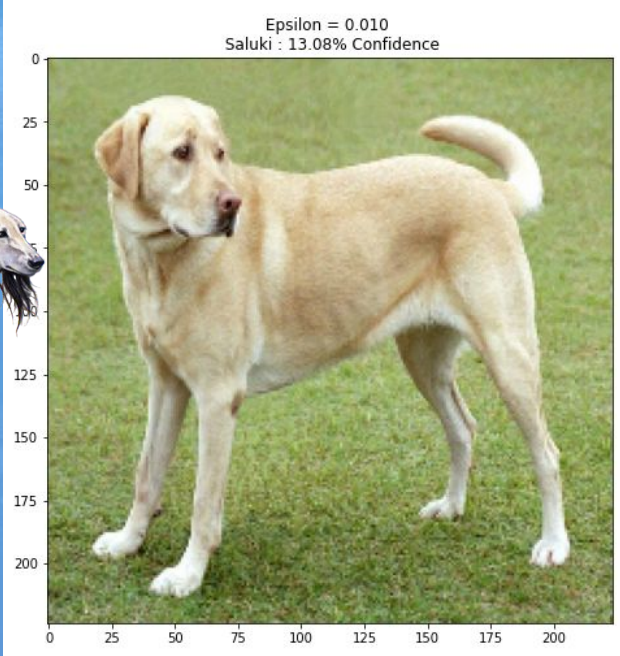
Normal

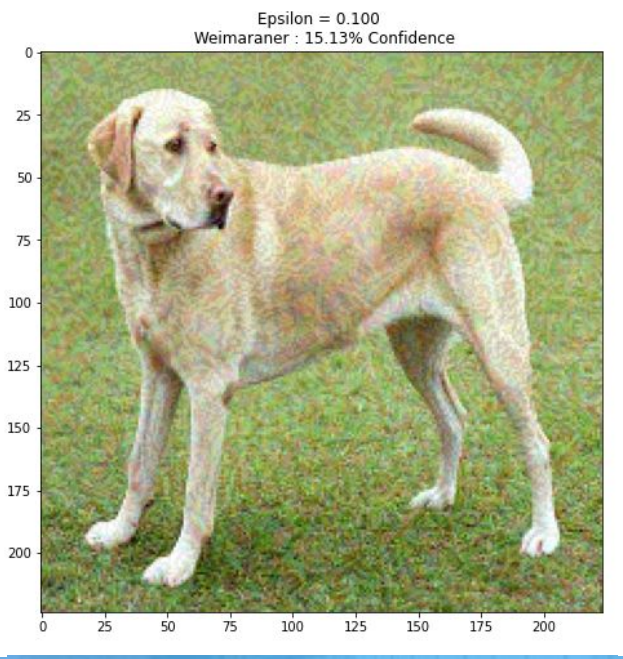
Corrupted

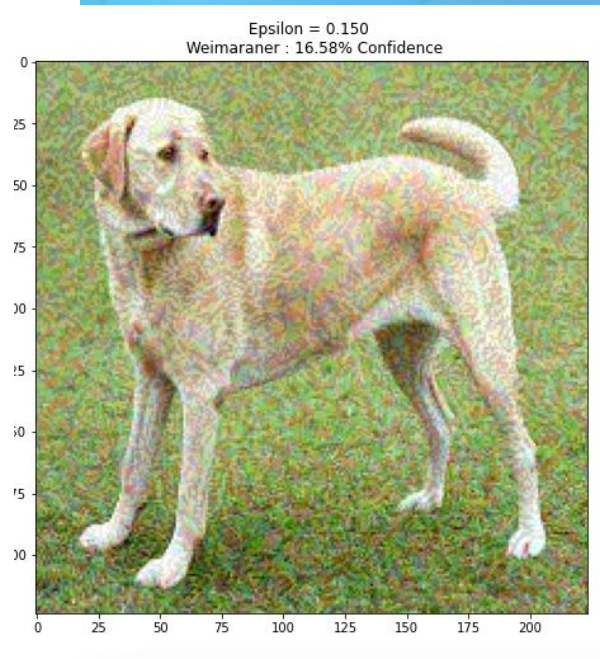
Adversarial Example Using FGSM (TensorFlow, 2022)



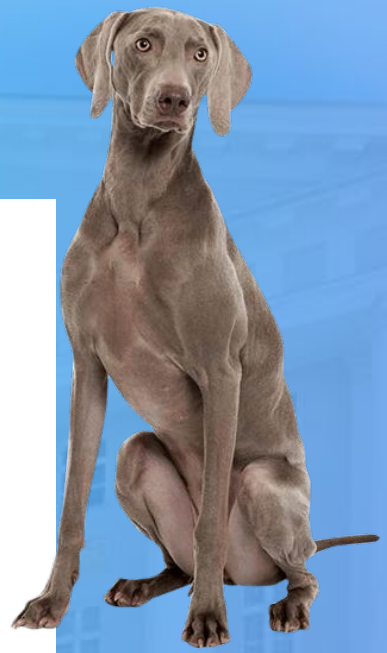






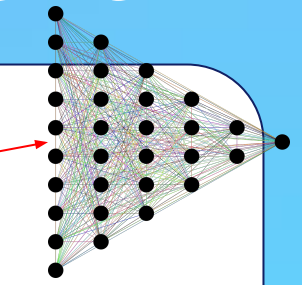


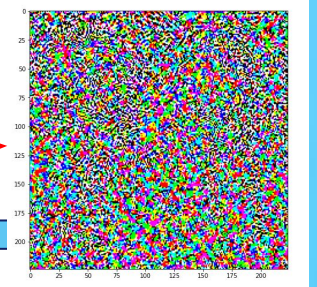
MobileNetV2 model and the ImageNet class names



Adversarial Machine Learning

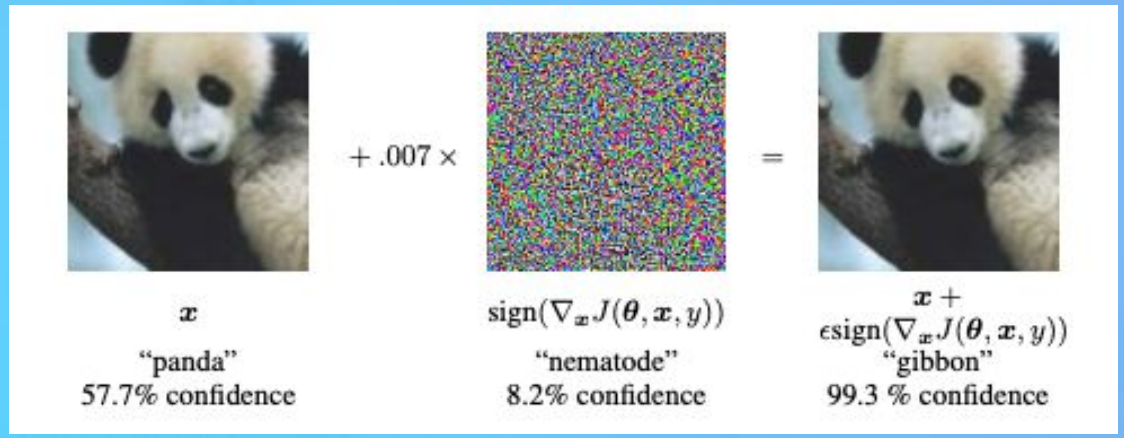
```python
def adversarial_pattern(self, model, image, label):
    image = tensorflow.cast(image, tensorflow.float32)
    with tensorflow.GradientTape() as tape:
        tape.watch(image)
        prediction = model(image)
        loss = tensorflow.keras.losses.MSE(label, prediction)

    gradient = tape.gradient(loss, image)
    signed_grad = tensorflow.sign(gradient)

    return signed_grad
```





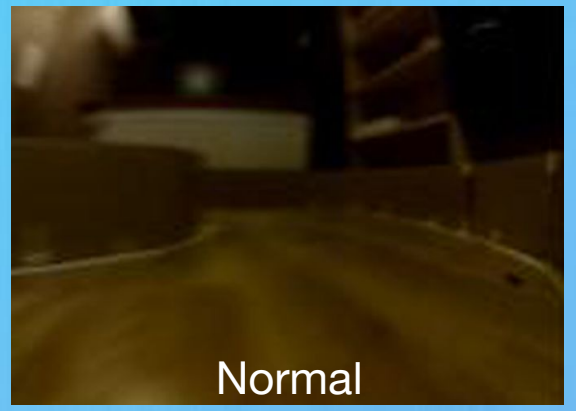
```python
new_data_folder = Tub(os.path.join(base_path, 'corrupted_images/'),
                      inputs=['cam/image_array', 'user/angle', 'user/throttle'],
                      types=['image_array', 'float', 'float'])

for key, record in enumerate(tub1):
    t_record = TubRecord(config=cfg,
                         base_path=tub1.base_path,
                         underlying=record)
    img_arr = t_record.image(cached=False)
    record['user/angle'] = record['user/angle']
    record['user/throttle'] = record['user/throttle']

    image = img_arr.reshape((1,) + img_arr.shape)
    label_to_pass = model.predict(image)
    perturbation = self.adversarial_pattern(model, image, label_to_pass).numpy()
    perturb = ((perturbation[0]*0.5 + 0.5)*255)-50
    adv_img = np.clip(img_arr + (perturb*0.8), 0, 255)
    adv_img = adv_img.astype(int)
    record['cam/image_array'] = adv_img
    new_data_folder.write_record(record)
```

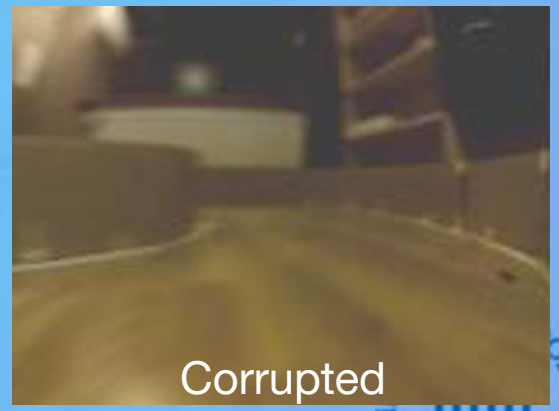
FGSM



(Goodfellow, Shlens, & Szegedy, 2015)

Batch of adversarial images generation



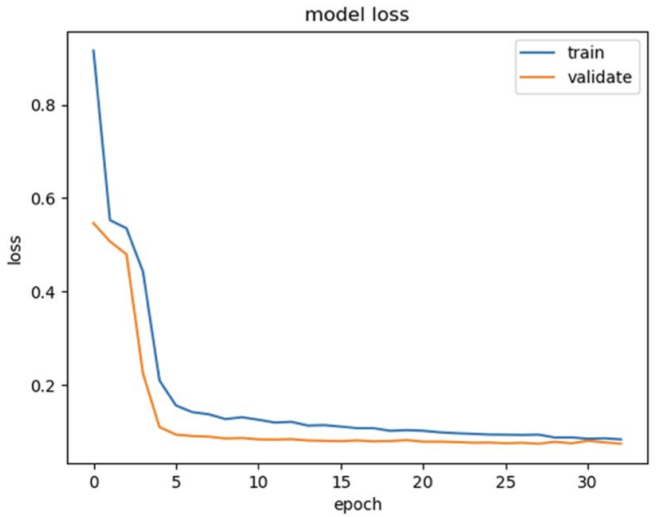


Data Cleaning

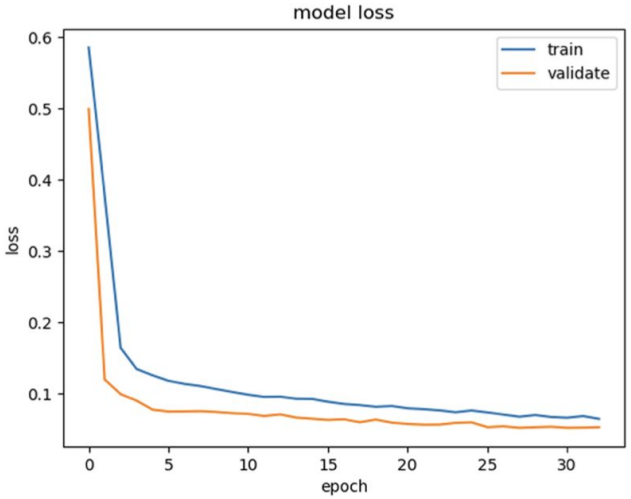
Dataset	Total images	Removed images
Small	5593	842
Large	10418	1491

Training

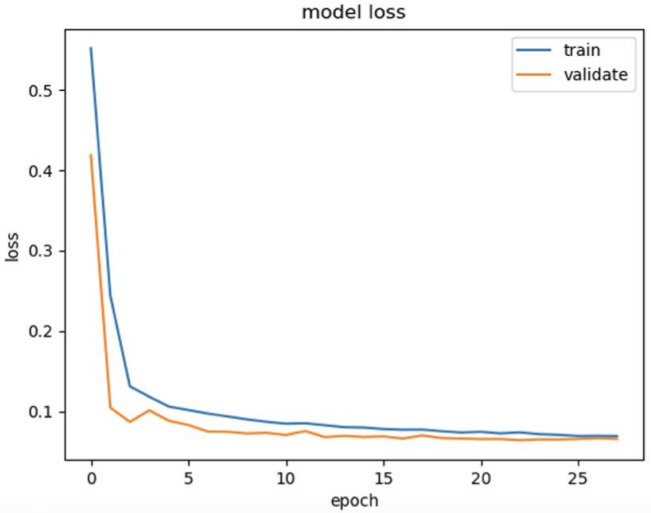
a) M-TS



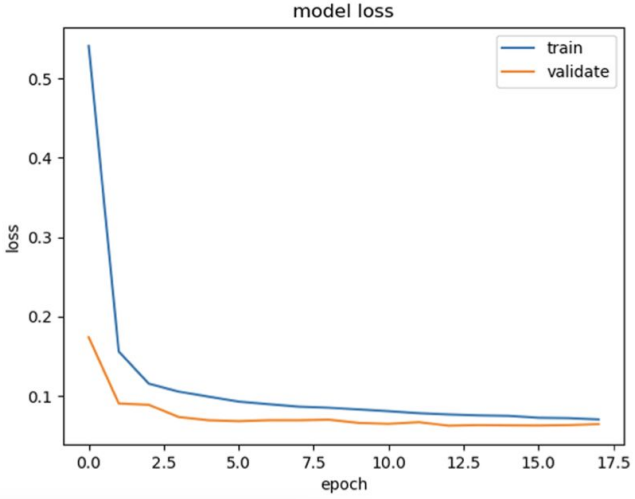
b) M-TSA



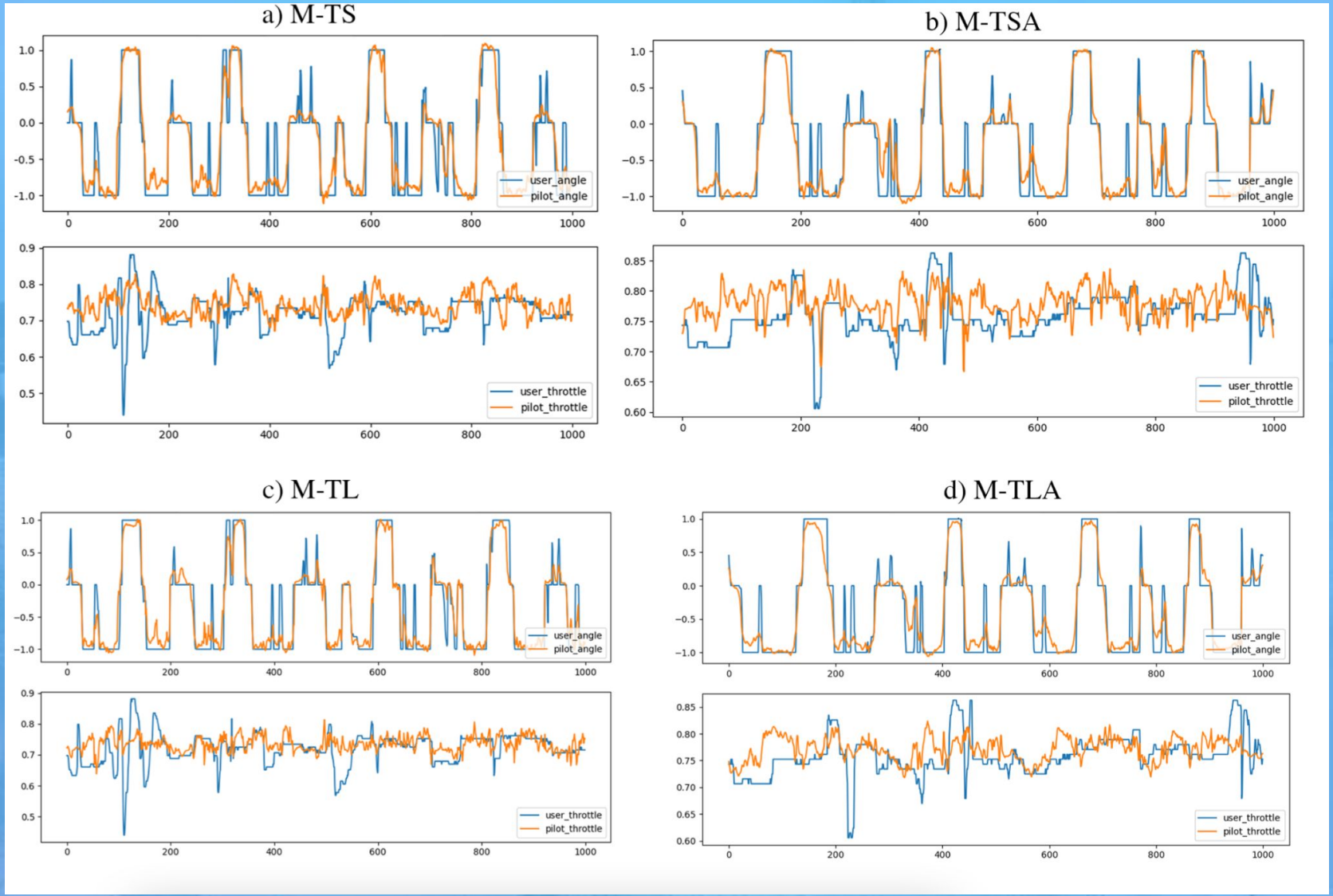
c) M-TL



d) M-TLA



Performance



Evaluation

CNN Score - Quantifying the quality of predictions		
CNN Model	**Prediction**	**RMSD**
M-TS	Steering angle Throttle	0.24630 0.05860
M-TSA	Steering angle Throttle	0.22733 0.05193
M-TL	Steering angle Throttle	0.23190 0.04488
M-TLA	Steering angle Throttle	0.23527 0.04012

Society of Automotive Engineers (SAE) - Automation Levels

0	1	2	3	4	5
No Automation	Driver Assistance	Partial Automation	Conditional Automation	High Automation	Full Automation

End-to-end Pipeline

- ✅ Traffic-Aware Cruise Control
- ✅ Navigate on Autopilot
- ✅ Autosteer
- ✅ Traffic Light and Stop Sign Control
- ✅ Autopark
- ✅ Smart Summon

End-to-end Pipeline

TESLA

WAYVE

✅ End-to-end driving a vehicle in traffic on public UK roads

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

End-to-end Pipeline

✅ **ALVINN**

The first neural network powered automatic driving system.

(Pomerleau, 1988)

End-to-end Pipeline

TESLA

WAYVE

Carnegie Mellon University

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Deep Learning Approach

to Automatic Driving Systems (ADS)

TAL TECH

TARTU ÜLIKOOL · UNIVERSITAS TARTUENSIS · 1632

Thank you


University of Tartu
TalTech
Autonomous
Driving Lab
Delta Centre
Friends