Università della Svizzera italiana

erc

37th IEEE/ACM International Conference on Automated Software Engineering

Software Institute

ThirdEye: Attention Maps for Safe Autonomous Driving Systems

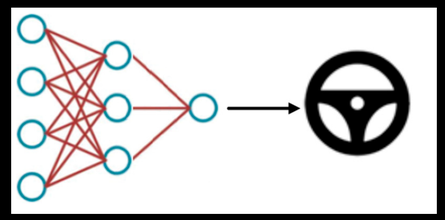



Andrea Stocco, USI *Paulo J. Nunes, UFPE* *Marcelo d'Amorim, UFPE* *Paolo Tonella, USI*

@tsigalko18


Lane keeping & detection
Object recognition
vehicles
pedestrians
/Autonomous
/Sensing
/Communication
/Battery
/Navigation
/Mirrorless
/Ecology
Self-Driving
48
mph

Model Construction
(training)



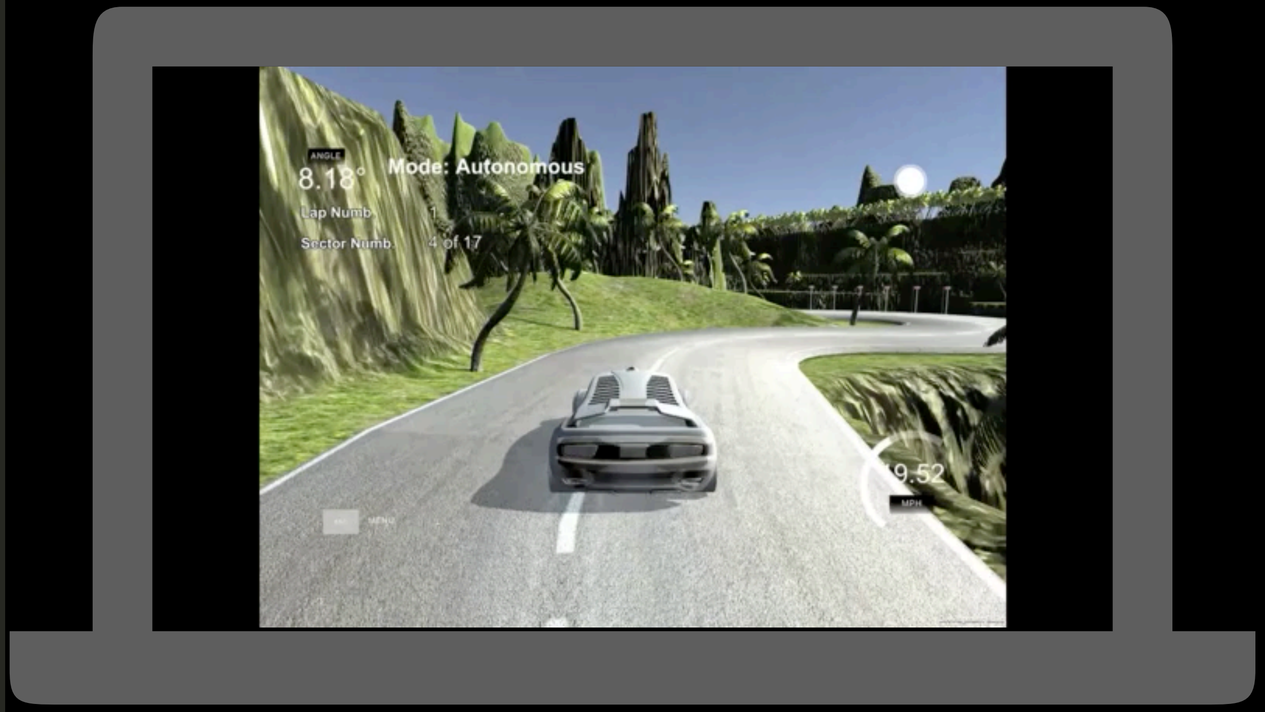
Data Collection
(in-field driving)

In-field testing
(evaluation)

In-house testing
(simulation)

Steering Angle predicted by the DNN

```
-0.16933852434158325
-0.2953818142414093
-0.2953818142414093
-0.2953818142414093
-0.24482464790344238
-0.24482464790344238
-0.24482464790344238
-0.2340604066848755
-0.2340604066848755
-0.2340604066848755
-0.2876757085323334
-0.2876757085323334
-0.2876757085323334
-0.28597092628479004
-0.28597092628479004
-0.28597092628479004
-0.280177503824234
-0.280177503824234
-0.280177503824234
-0.1850987821817398
-0.1850987821817398
-0.1850987821817398
-0.2626234292984009
-0.2626234292984009
-0.2626234292984009
-0.20239685475826263
```



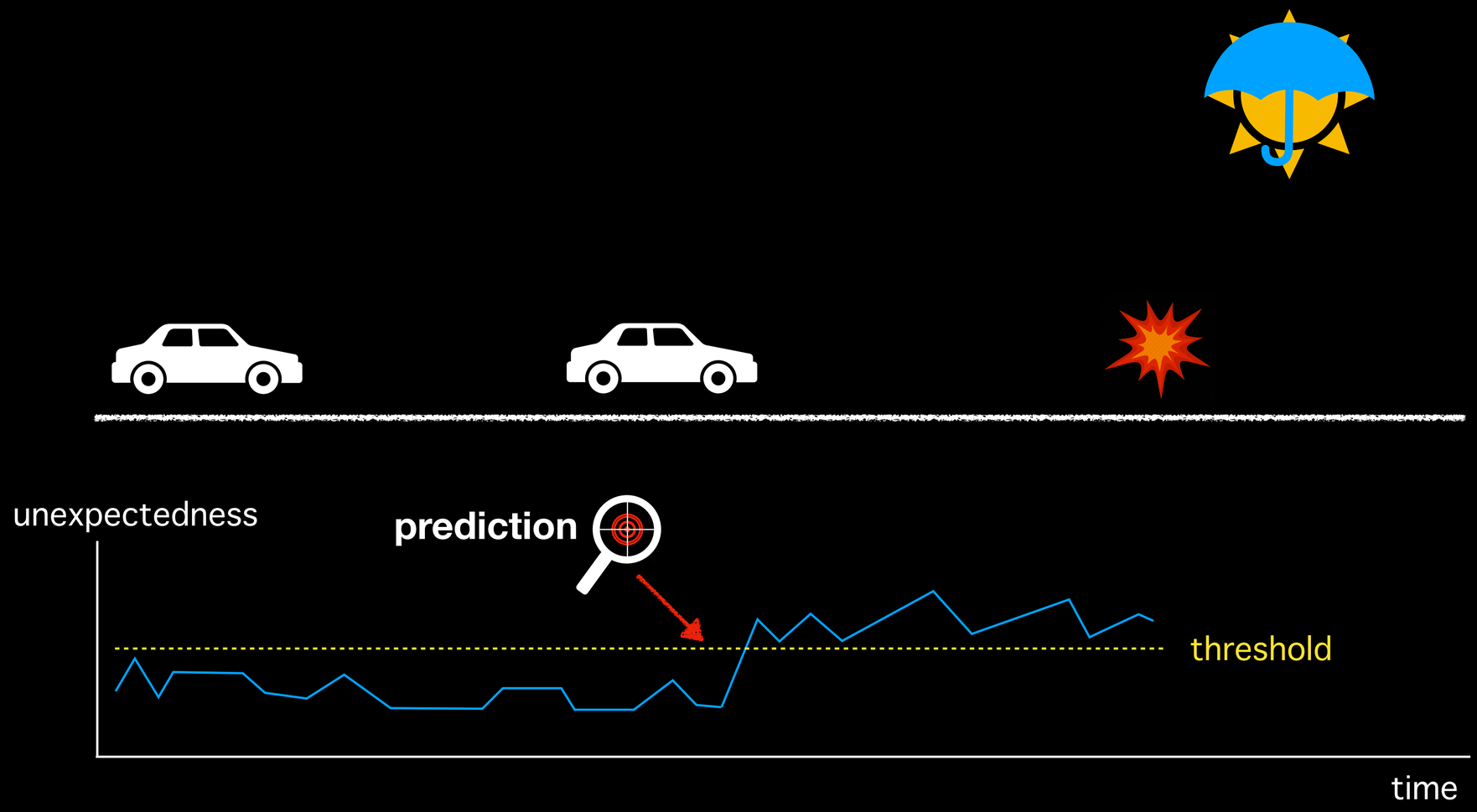
Training set cannot contain
ALL possible driving conditions!

*Can we **predict***
***unexpected** conditions**?

* ***Unexpected condition:***
Unseen, potentially hazardous and misbehaviour-inducing driving condition

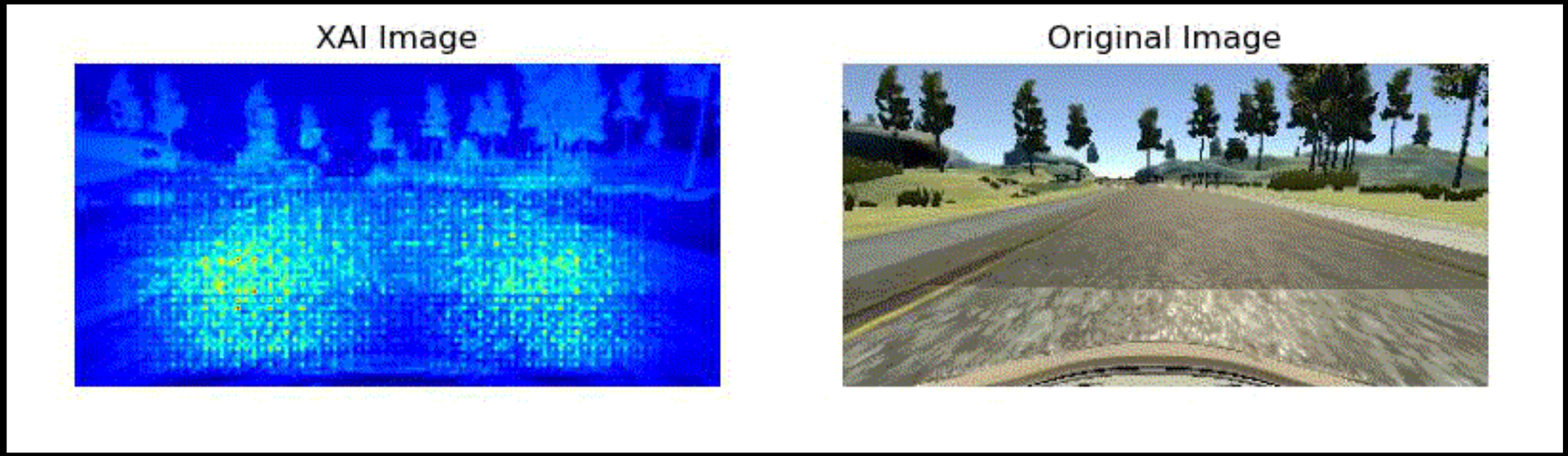
Unexpectedness metric

Anomaly detection



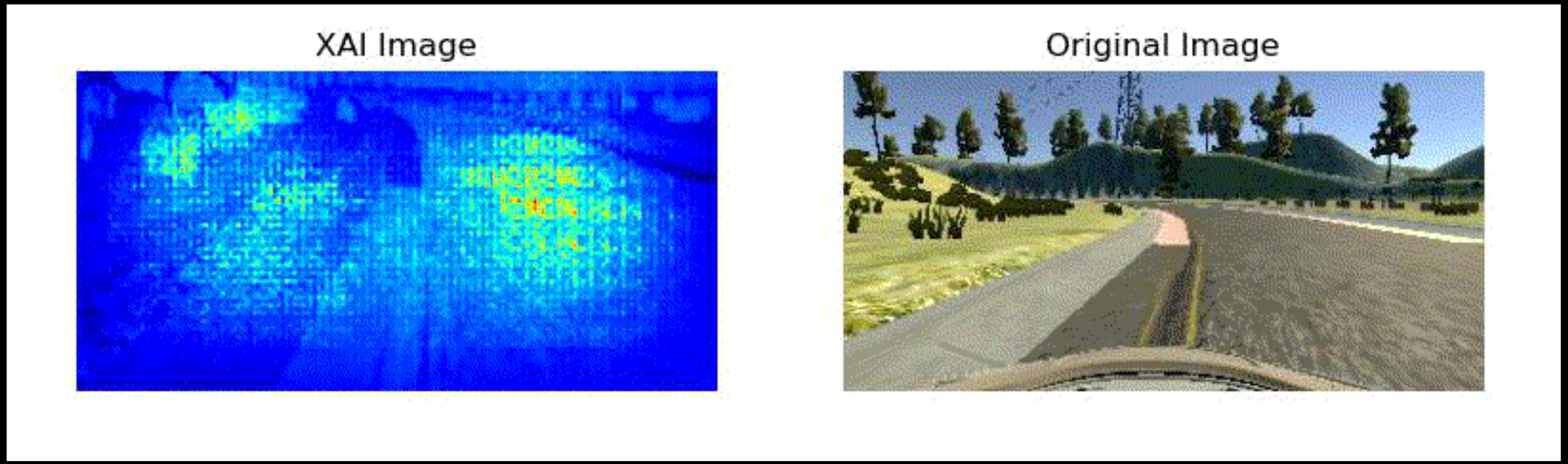
XAI for failure prediction

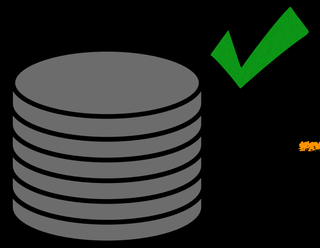
Nominal



XAI for failure prediction

Uncertain/Unexpected





Training set

1. Failure Predictor Training

2. Probability distribution fitting

3. Treshold estimation

Trained Predictor

Training set

1. Failure Predictor Training

1.1 Attention Map Generation

2. Probability distribution fitting

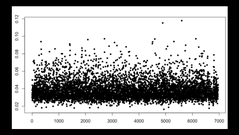
3. Treshold estimation

1. Failure Predictor Training

1.1 Attention Map Generation

2. Probability distribution fitting

2.1 Confidence Score Synthesis



2.2 Gamma distribution Fitting

3. Treshold estimation

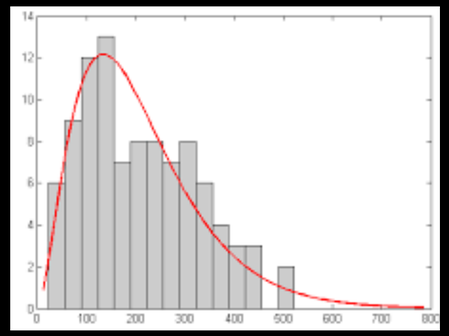
1. Failure Predictor Training

1.1 Attention Map Generation

2. Probability distribution fitting

2.1 Confidence Score Synthesis

2.2 Gamma distribution Fitting



3. Treshold estimation

1. Failure Predictor Training

1.1 Attention Map Generation

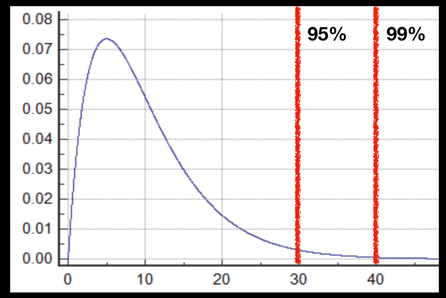
2. Probability distribution fitting

2.1 Confidence Score Synthesis

2.2 Gamma distribution Fitting

3. Treshold estimation

3.1 w/ Maximum Likelihood Estimation



1. Failure Predictor Training

1.1 Attention Map Generation

2. Probability distribution fitting

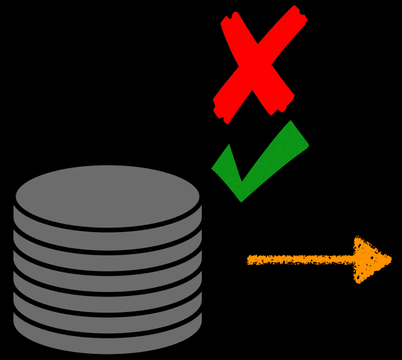
2.1 Confidence Score Synthesis

2.2 Gamma distribution Fitting

3. Treshold estimation

3.1 w/ Maximum Likelihood Estimation

4. Testing



Evaluation set

time series analysis

Confidence Score Synthesis

$$\overline{\boldsymbol{h}} = \frac{1}{WHC} \sum_{i=1,j=1,c=1}^{W,H,C} h_{[i][j][c]}$$

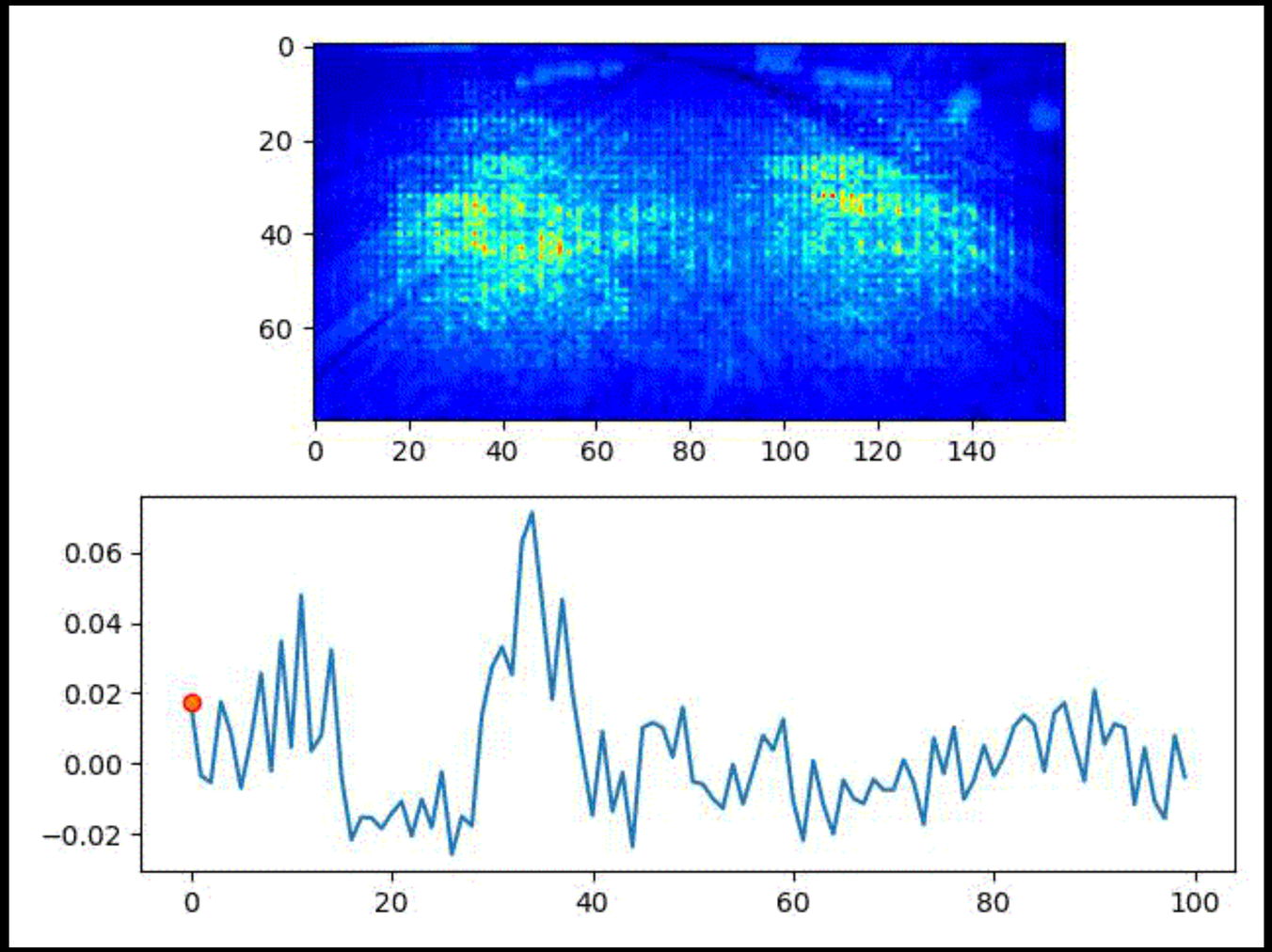
HA = Heatmap
Average Function

$$\overline{\nabla \boldsymbol{h}_t} = \frac{1}{WHC} \sum_{i=1,j=1,c=1}^{W,H,C} h_{t[i][j][c]} - h_{t-1[i][j][c]}$$

HD = Heatmap
Derivative Function

$$\boldsymbol{h_e} = \mathscr{L}(\boldsymbol{h}, dec(enc(\boldsymbol{h})))$$

HRL = Heatmap
Reconstruction Loss

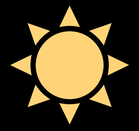

0
20
40
60
0
20
40
60
80
100
120
140
0.06
0.04
0.02
0.00
−0.02
0
20
40
60
80
100

Experimental Study

Procedure, Main Results

Udacity sim

Dave-2 NVIDIA.

Predictor (Thirdeye) 

ThirdEye

3 confidence scores

- Average
- Derivative
- Rec. Loss

Baseline: SelfOracle (ICSE 2020)

Two Experiments

External Unknown Scenarios 

Internal Uncertain Scenarios 

Metrics

Precision

Recall

F-3

Udacity sim

Dave-2 NVIDIA

Predictor (Thirdeye)

ThirdEye

3 confidence scores

- Average
- Derivative
- Rec. Loss

Baseline: SelfOracle (ICSE 2020)

Two Experiments

External Unknown Scenarios

Internal Uncertain Scenarios

Metrics

Precision

Recall

F-3

Udacity sim

Dave-2 NVIDIA

Predictor (Thirdeye)

ThirdEye

3 confidence scores

- Average
- Derivative
- Rec. Loss

Baseline: SelfOracle (ICSE 2020)

Two Experiments

External Unknown Scenarios

Internal Uncertain Scenarios

Metrics

Precision

Recall

F-3

Udacity sim

Dave-2 NVIDIA

Predictor (Thirdeye)

ThirdEye

3 confidence scores

- Average
- Derivative
- Rec. Loss

Baseline: SelfOracle (ICSE 2020)

Two Experiments

External Unknown Scenarios

Internal Uncertain Scenarios

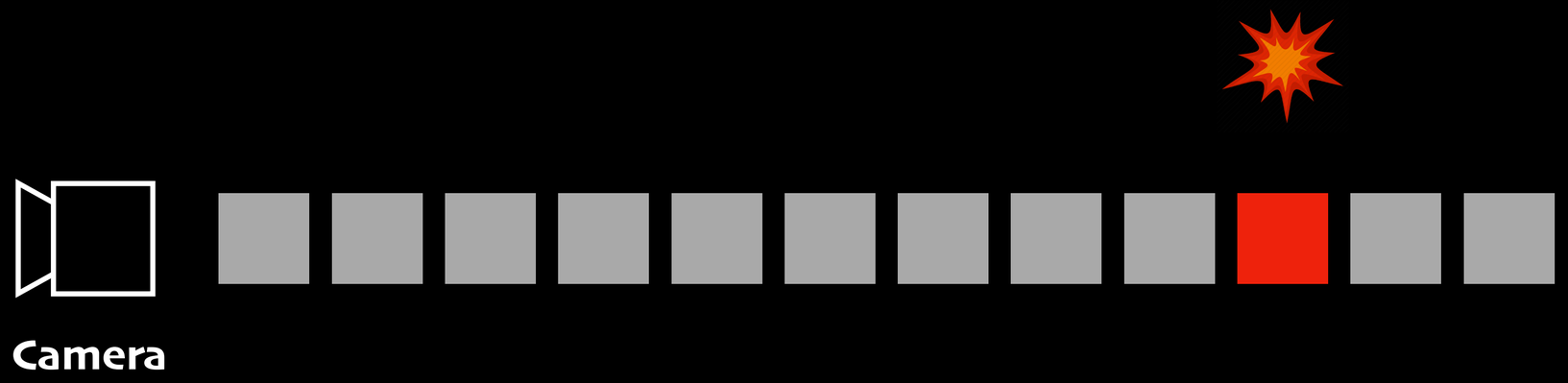
Metrics

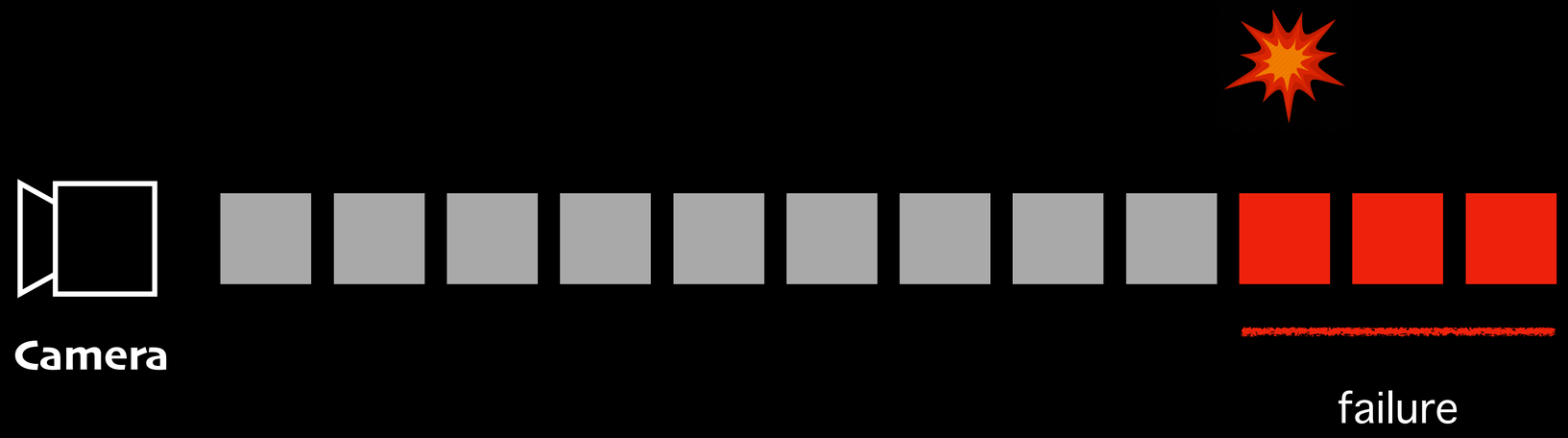
Precision

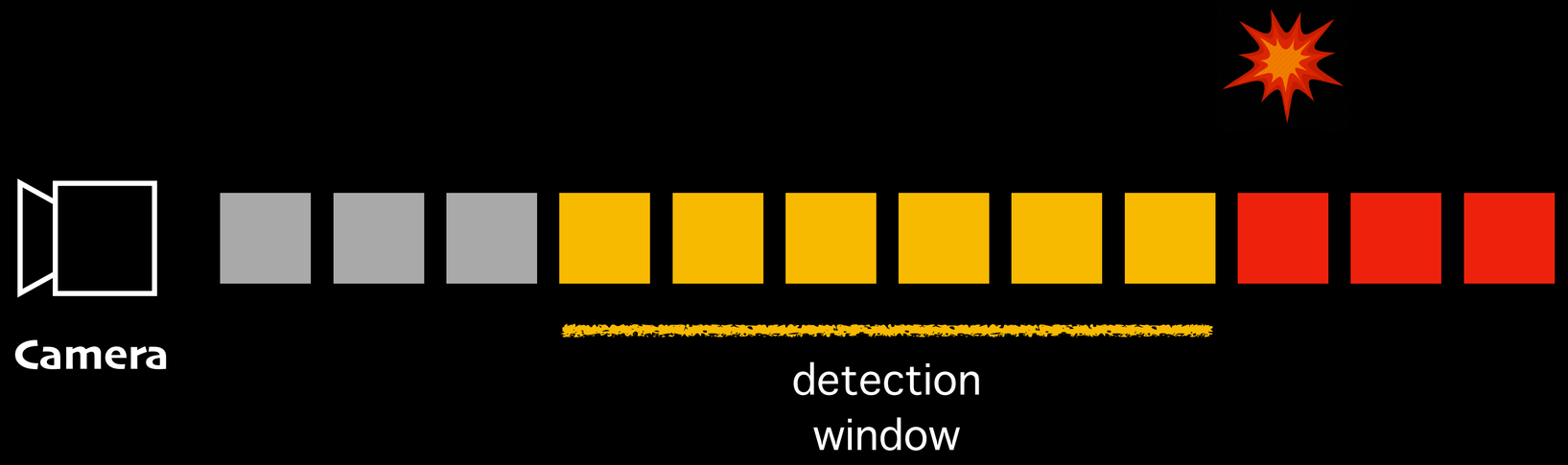
Recall

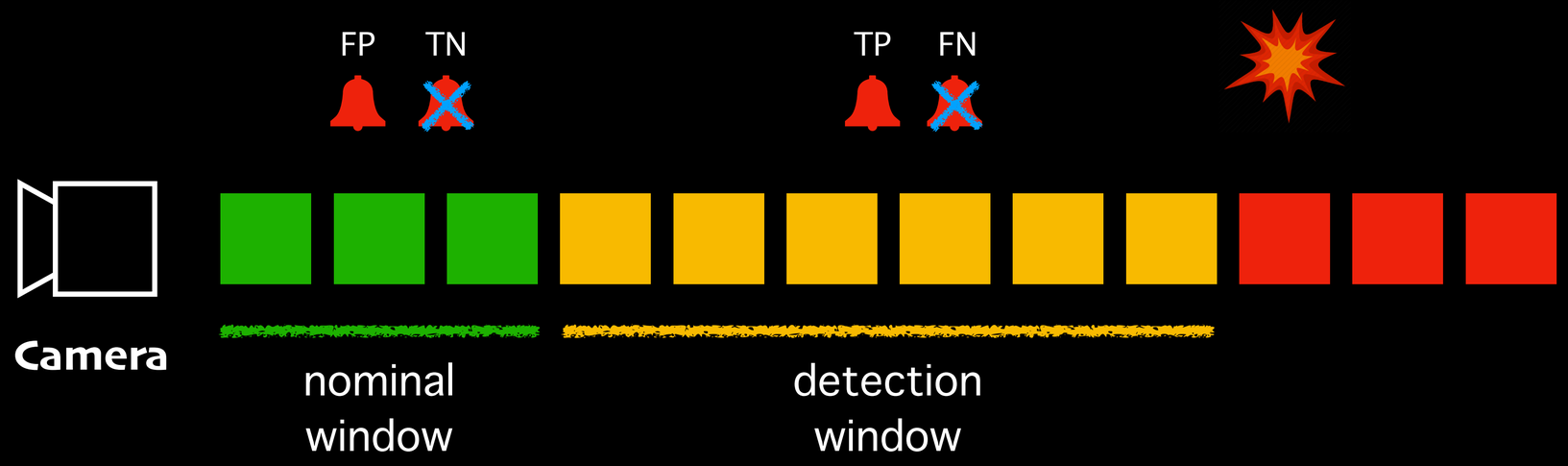
F-3


Camera


Camera


Camera
failure

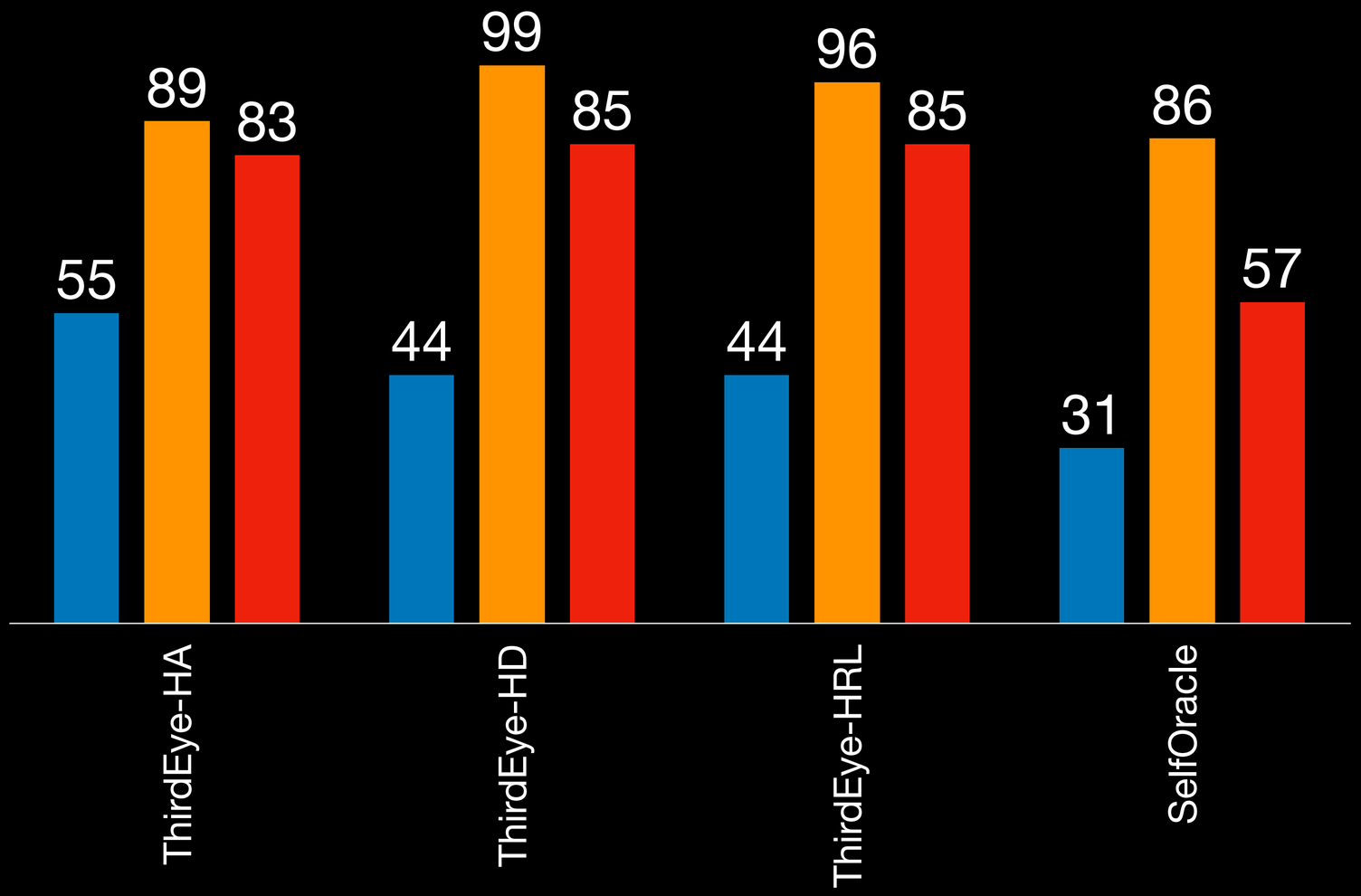

Camera
detection
window

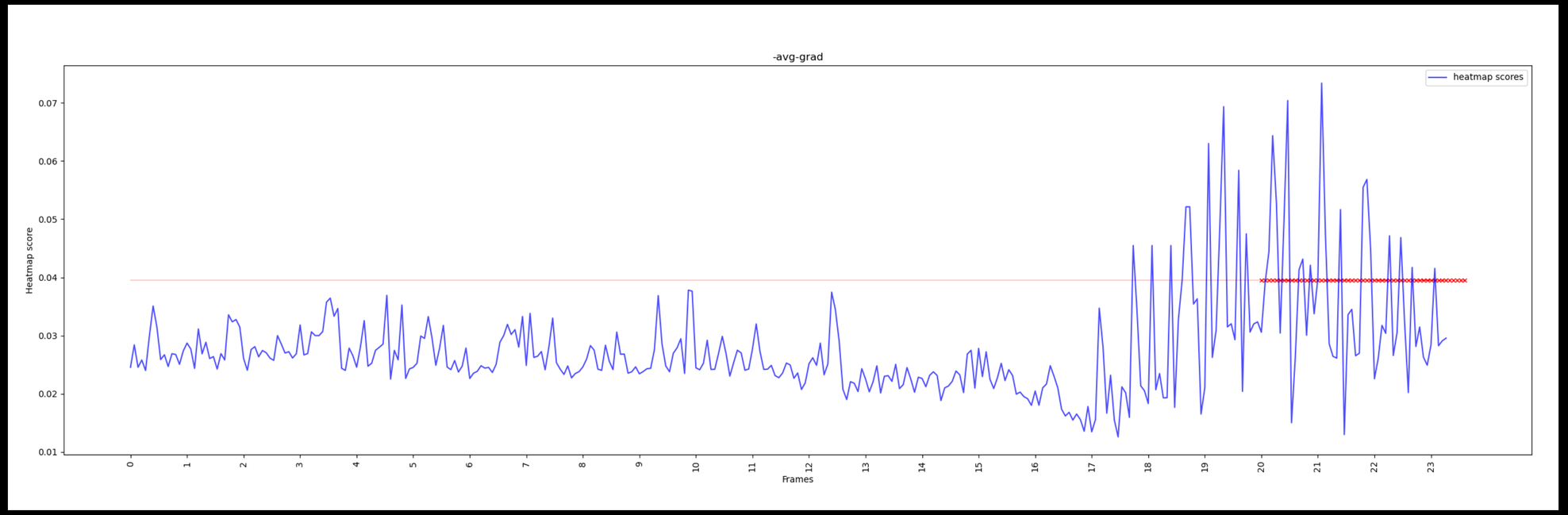

FP
TN
TP
FN
Camera
nominal
window
detection
window

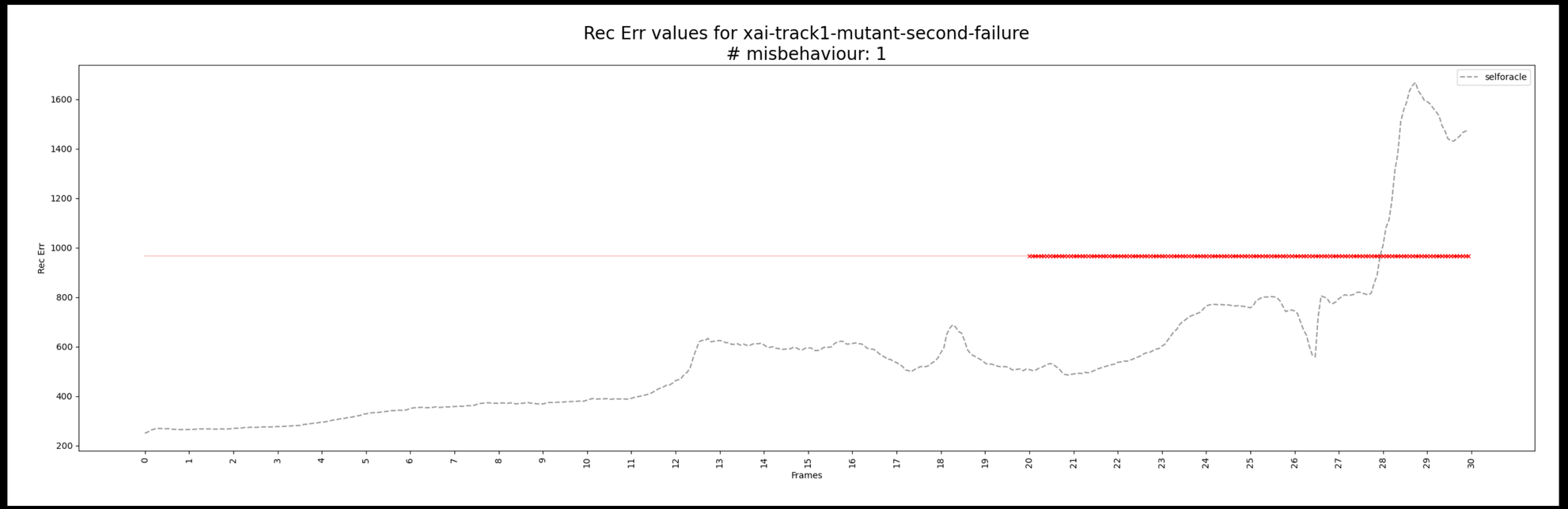
Is our approach effective ?

in predicting misbehaviours

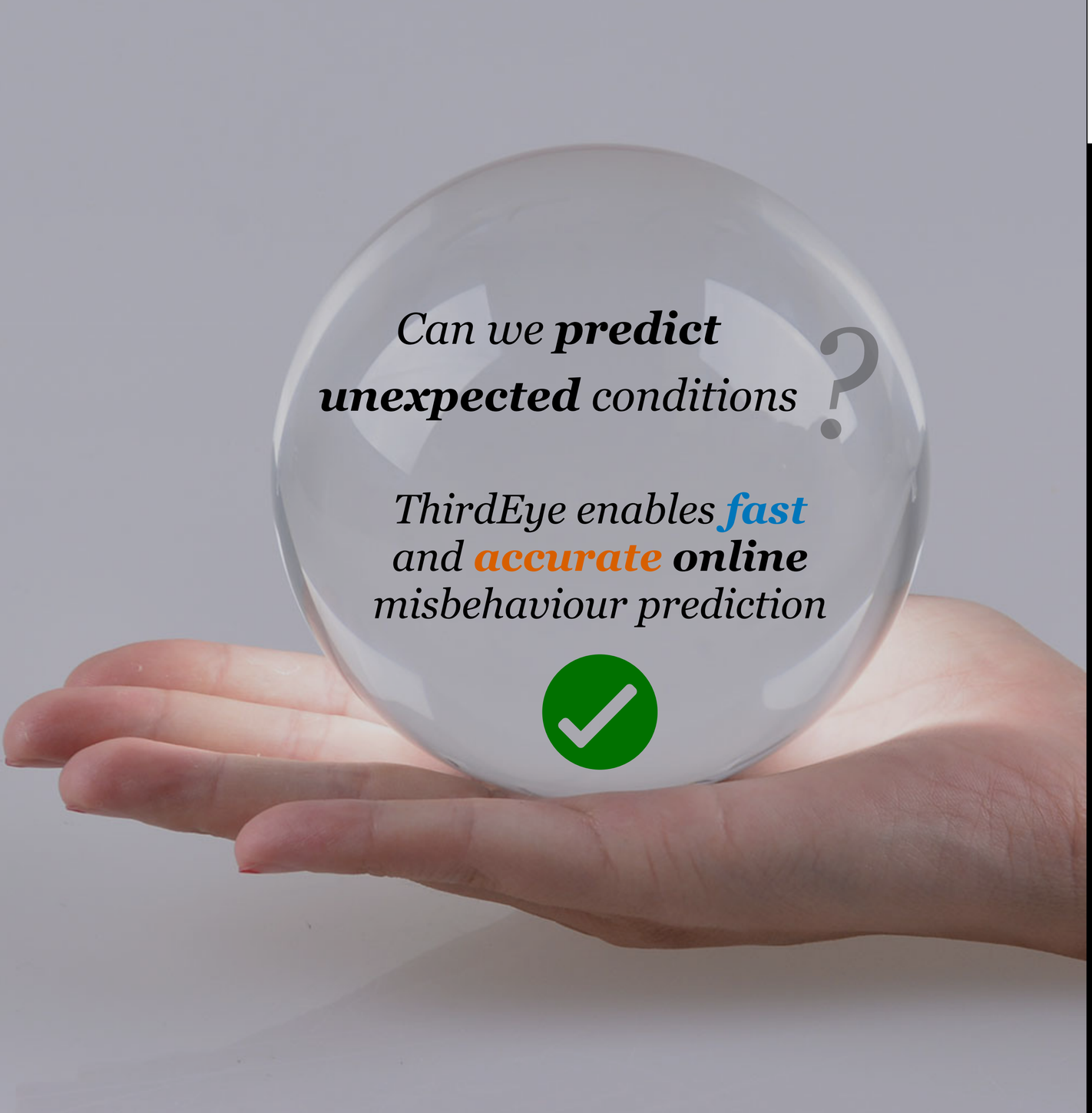
Is our approach effective in predicting misbehaviours ?




-avg-grad
heatmap scores
Heatmap score
Frames


Rec Err values for xai-track1-mutant-second-failure
misbehaviour: 1
selforacle
Rec Err
Frames


*Can we **predict** **unexpected** conditions*
?
*ThirdEye enables **fast** and **accurate** **online** misbehaviour prediction*



Software Institute

37th IEEE/ACM International Conference on Automated Software Engineering

ThirdEye: Attention Maps for Safe Autonomous Driving Systems

Code

https://github.com/tsigalko18/ase22

Simulator

https://github.com/tsigalko18/self-driving-car-sim/tree/USI_v1.0.0