Generative AI for Autonomous Driving Systems Testing

Andrea Stocco

TUM fortiss

Image generated with DALL·E 3

Automated Software Testing

We like to generate tests, monitor them, and make them real!

Dr. Andrea Stocco
Head of AST @ fortiss
Prof. @ TUM
stocco@fortiss.org
andrea.stocco@tum.de

Stefano Carlo Lambertenghi
Generative AI Testing /
Reality Gap Assessment and Mitigation
lambertenghi@fortiss.org

Davide Yi Xian Hu
Generative AI Testing
hu@fortiss.org

Lev Sorokin
Algorithm Optimization /
Cross-Simulation Testing
sorokin@fortiss.org

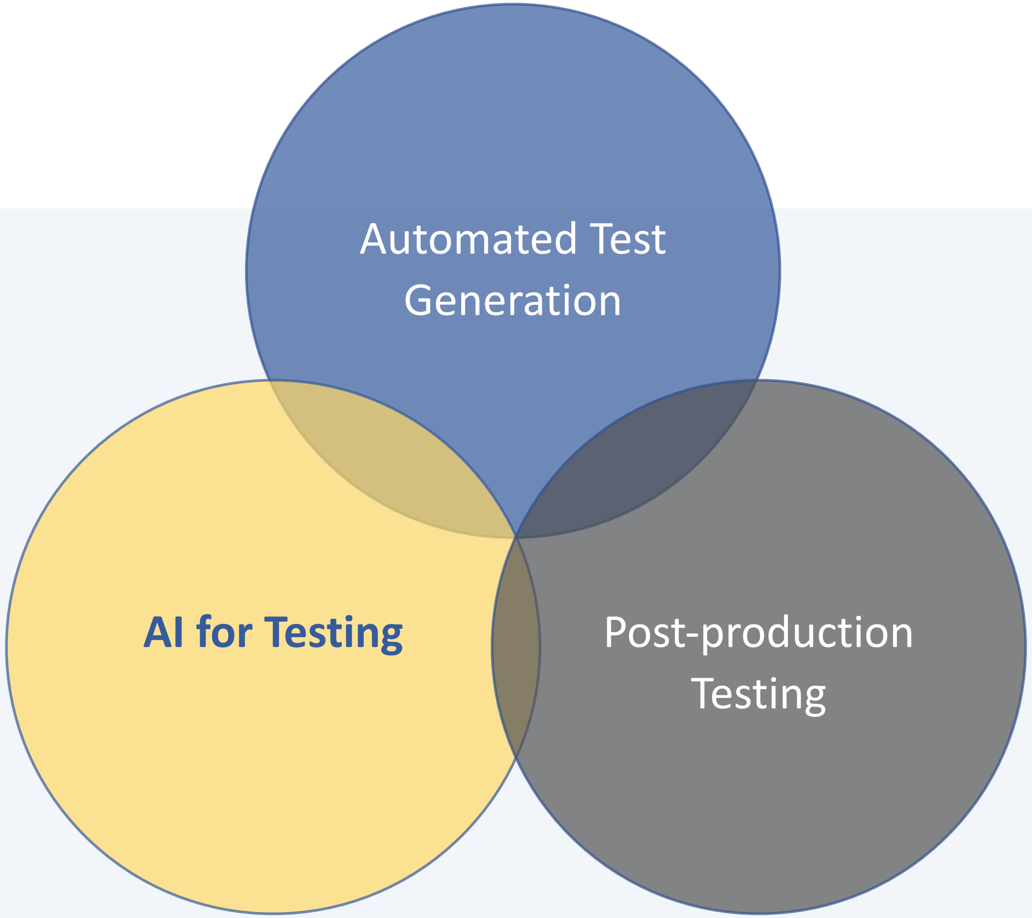
Xingcheng Chen
eXplainable Artificial Intelligence (XAI) /
Post-Production Testing
xchen@fortiss.org

Oliver Weißl
Algorithm Optimization
weissl@fortiss.org

Automated Software Testing

Core Research

- **Automated Test Generation**
  How can we automatically generated complex scenario-based tests efficiently and effectively?
- **AI for Testing**
  How can we leverage GenAI techniques, uncertainty quantification and explainable AI for testing CPS?
- **Post-production Testing**
  How to ensure a high dependability of deep neural network driven-cyber-physical systems (CPS) in production?



- *Transferability between Virtual vs Physical-world Testing*
- *Assessing Quality Metrics Reality Gap Input Mitigation with GenAI*
- *GenAI for Test Domain Augmentation*

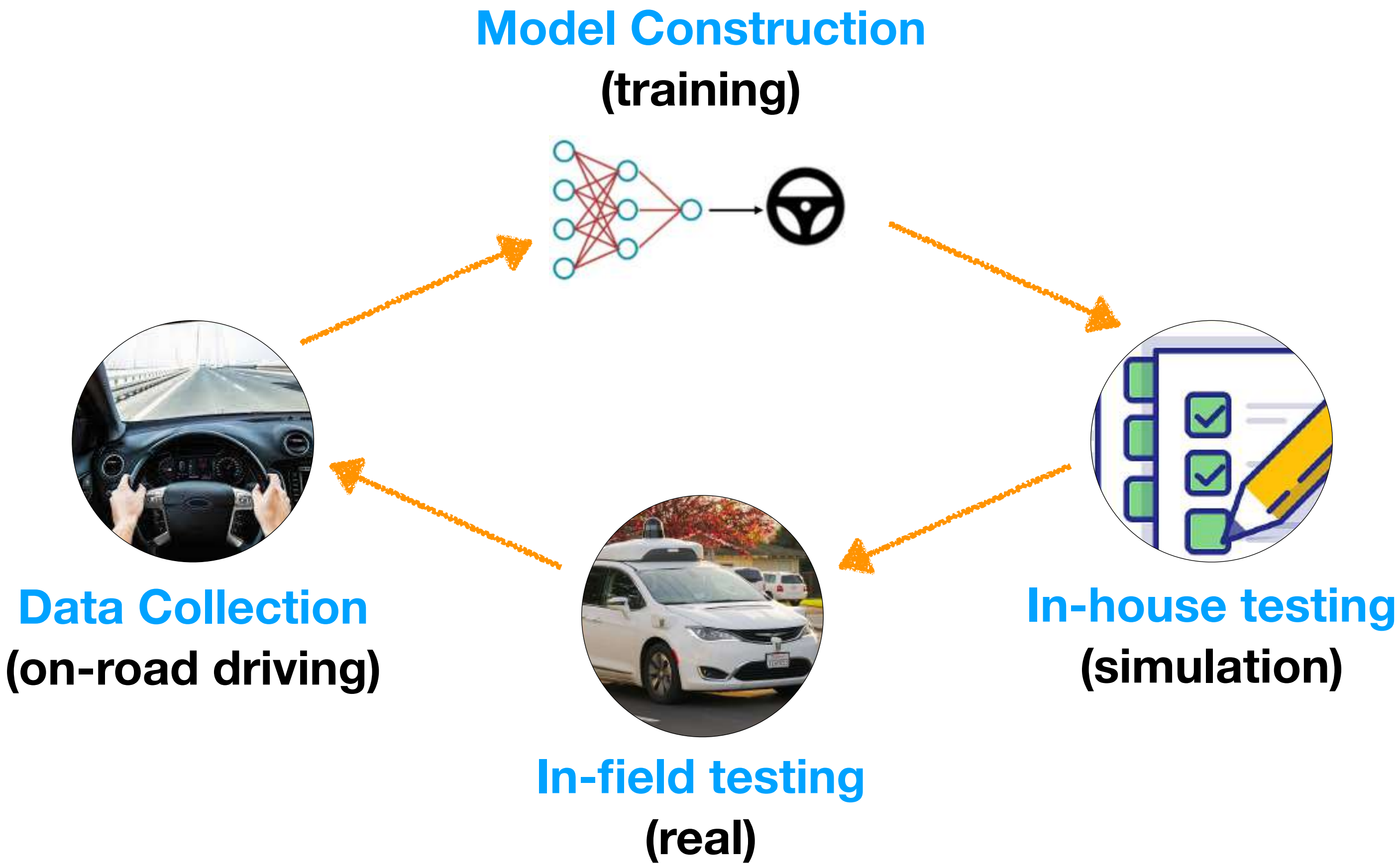
Mind the Gap! A Study on the Transferability of Virtual vs Physical-world Testing of Autonomous Driving Systems

Stocco, Pulfer, Tonella.

In IEEE Transactions on Software Engineering. 2023

Automated Driving System (ADS) testing

How to ensure that an ADS system is ready for deployment?



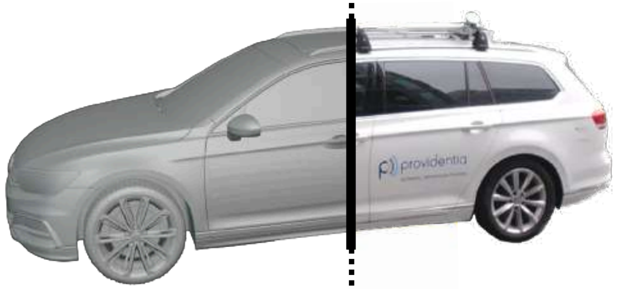
Automated Driving System (ADS) testing

How to ensure that an ADS system is ready for deployment?

Simulation testing

⊕ Cheaper
Safer
More versatile

⊖ Approximation of reality



Real-world testing

⊕ Evaluate ADS realistically

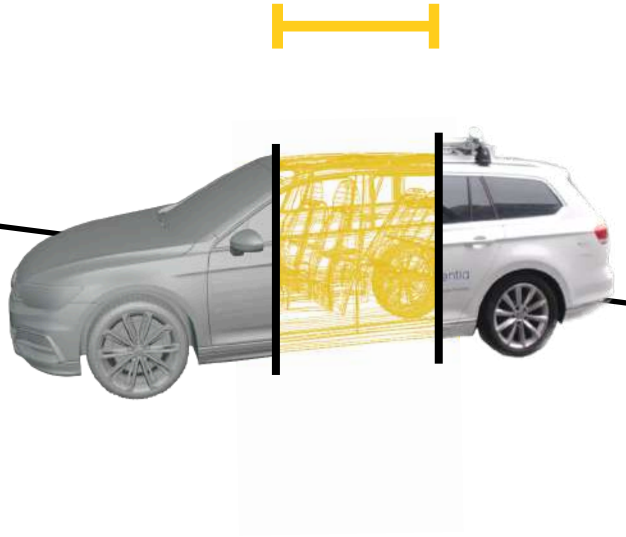
⊖ High cost
High risk
Limited

Reality Gap

Difference between simulated and real vehicle

Simulation testing | **Gap** | **Real-world testing**

Simulated — Behaviour, Scenarios, Sensors



Real — Behaviour, Scenarios, Sensors

ADS **Simulated Behaviour** ≠ ADS **Real Behaviour**

Perception Reality Gap

Difference between simulated and real input images

Simulation

Real-world

Perception Gap

ADS

Simulated Behaviour

≠

ADS

Real Behaviour

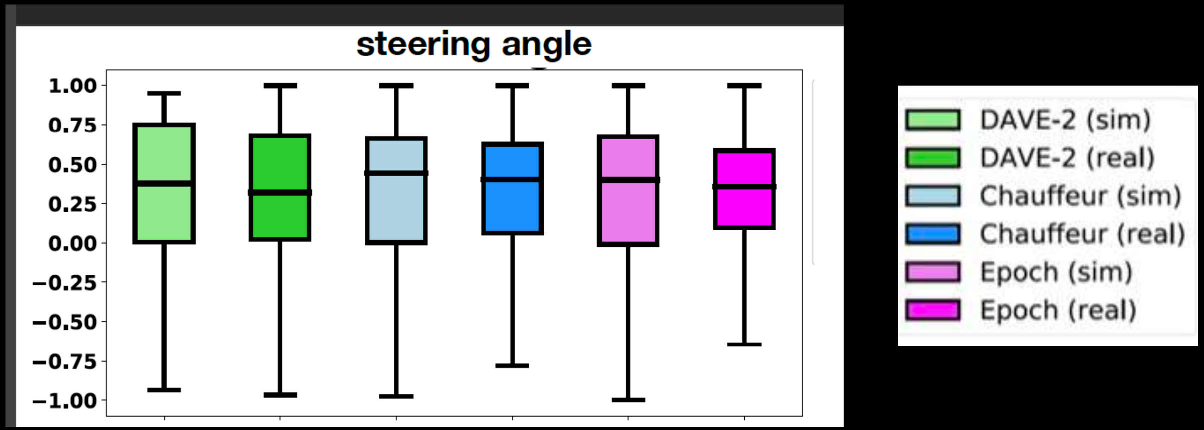
Gaidon, A et al. 2016
Geiger, A et al. 2013

When considering simulated and real-world environments…

- Would the same driving model behave the same?
- Would it fail the same?

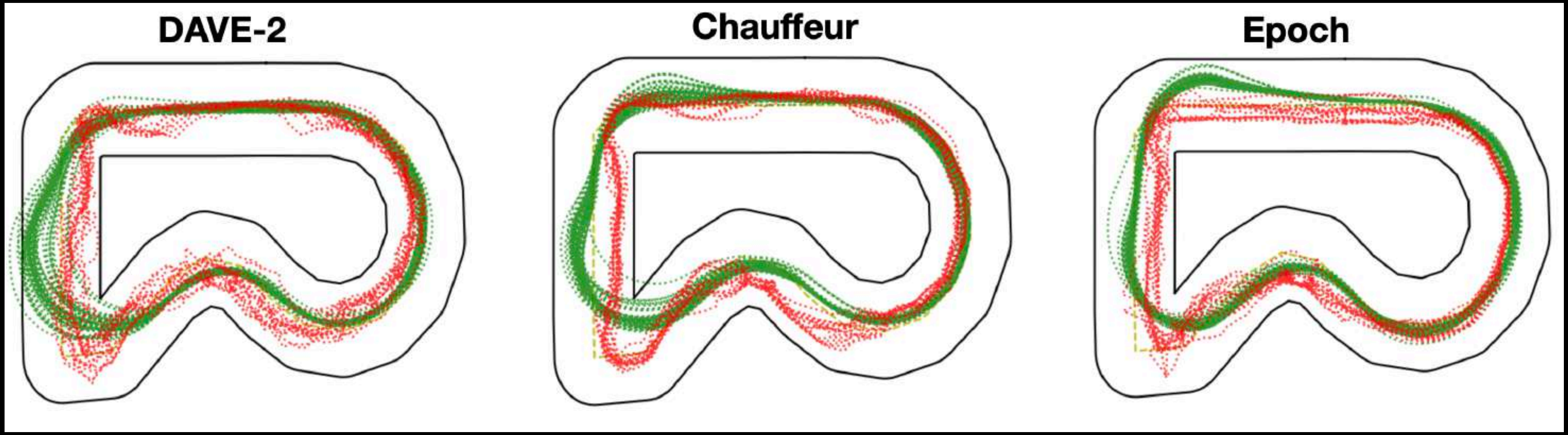
Same lane-keeping model architecture, two worlds

Would the same driving model behave the same?



Steering angle distributions do transfer across simulated and real-world environments

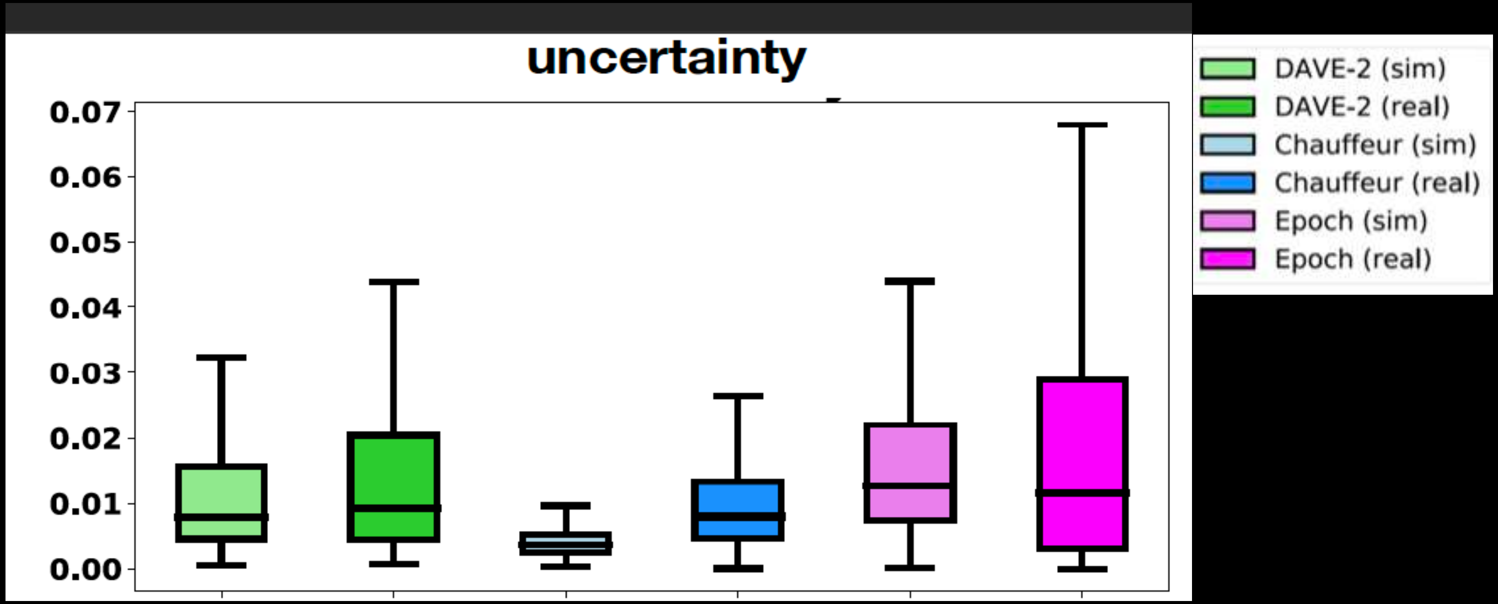
Expected as we aligned the two environments.
It may suggests that component-level testing is an option but…



Virtual (green) and physical (red) trajectories.

Lateral position is different across simulated and real-world environments

Component-level testing is not an option, we need system-level testing

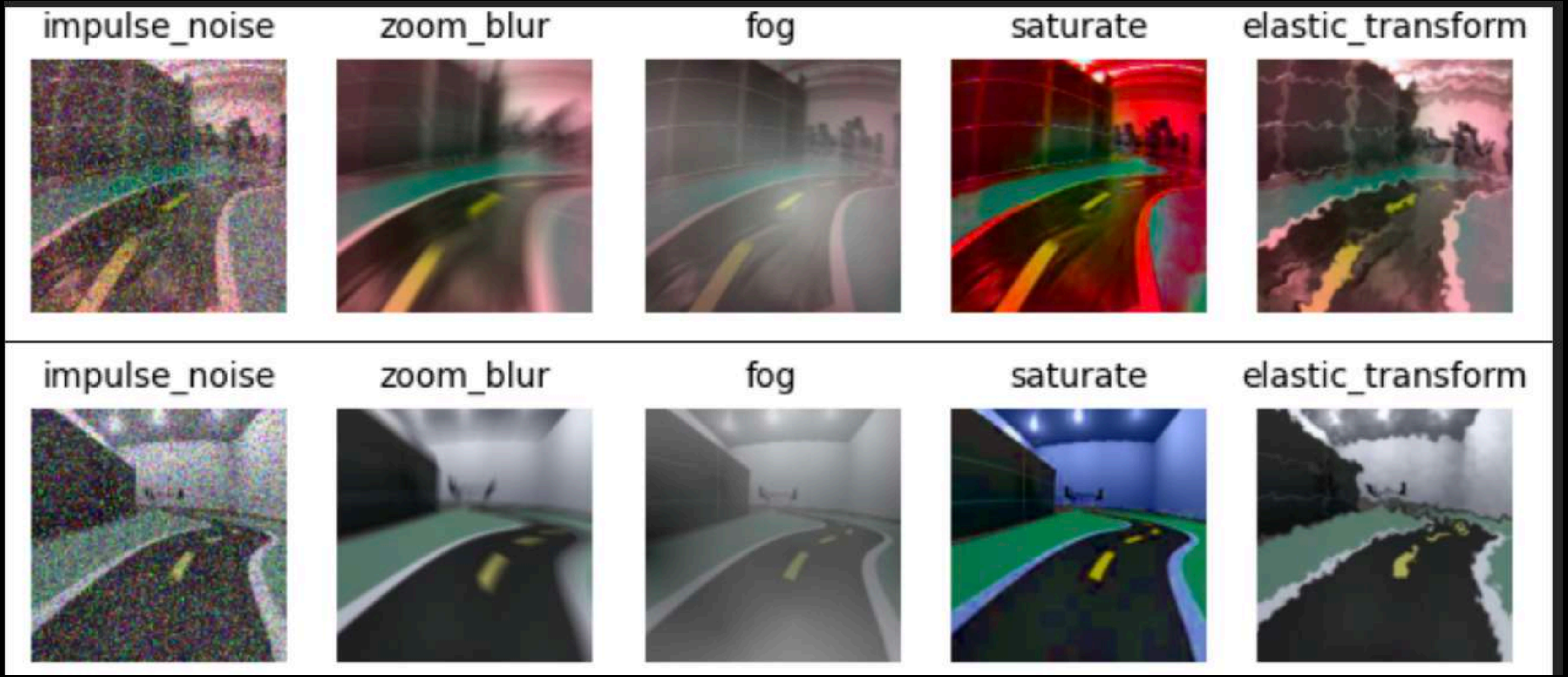


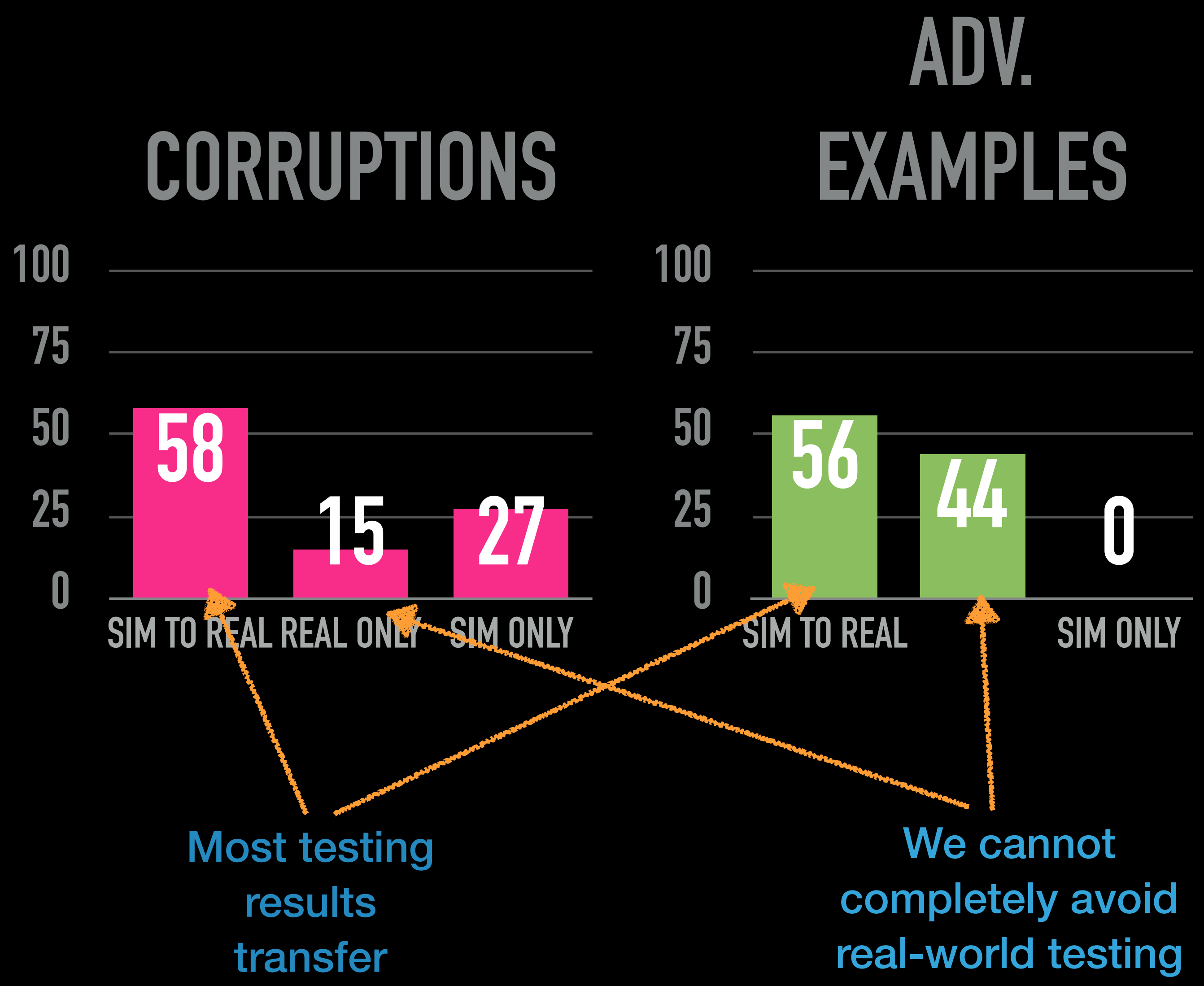
Uncertainty is higher in real-world environments

We need real-world testing (or better simulators)!

Would it fail the same?

We test both simulated and real cars under the same conditions (img corruptions)




CORRUPTIONS
ADV. EXAMPLES
100
75
50
25
0
58
15
27
SIM TO REAL
REAL ONLY
SIM ONLY
100
75
50
25
0
56
44
0
SIM TO REAL
SIM ONLY
Most testing results transfer
We cannot completely avoid real-world testing

Uncertainty useful to prioritize simulations for real-world execution

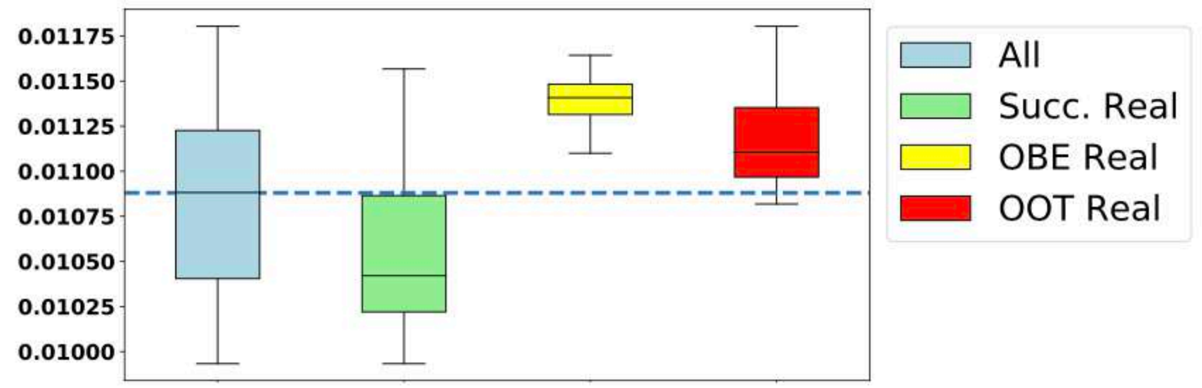


Figure 7: RQ4: test selection for DAVE-2: all tests with uncertainty above the median are not executed in the real-world.

Uncertainty useful to prioritize simulations for real-world execution

Time saving 51% — ca 13 hours

24% useless executions (false positives)

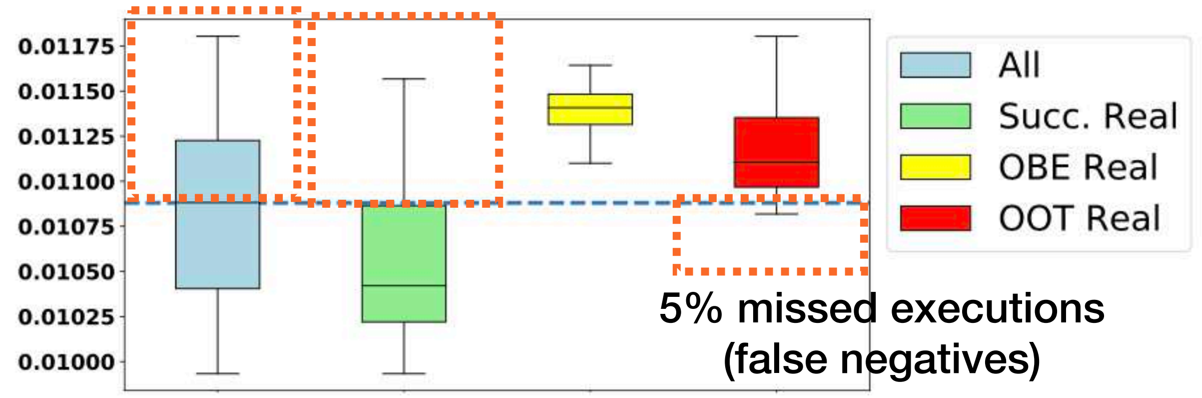


Figure 7: RQ4: test selection for DAVE-2: all tests with uncertainty above the median are not executed in the real-world.

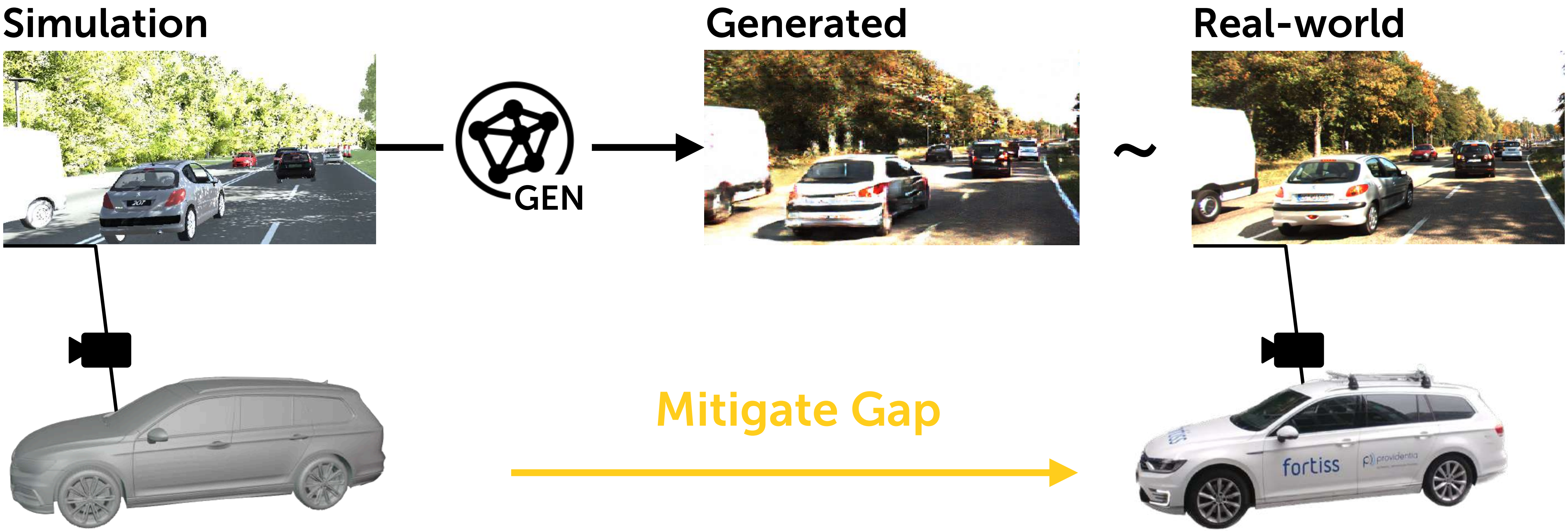
Assessing Quality Metrics for Neural Reality Gap Input Mitigation

Lambertenghi and Stocco.

In Proceedings of 17th IEEE International Conference on Software Testing, Verification and Validation 2024

Generative Image-to-Image Translation

Generative models for perception reality gap mitigation




REWE

Generative Image-to-Image Translation shortcomings

Generated

Real-world

TUM fortiss

Evaluate Image-to-Image Translation models

Measure quality of generated images, considering the target domain

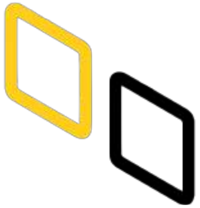
Generated

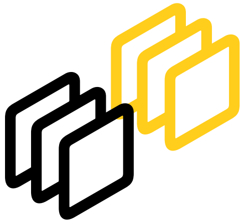
Gap distance?

Real-world

Single-Image Metrics

- ⊕ Precise comparison
- ⊖ Mapping required





Distribution-Level Metrics

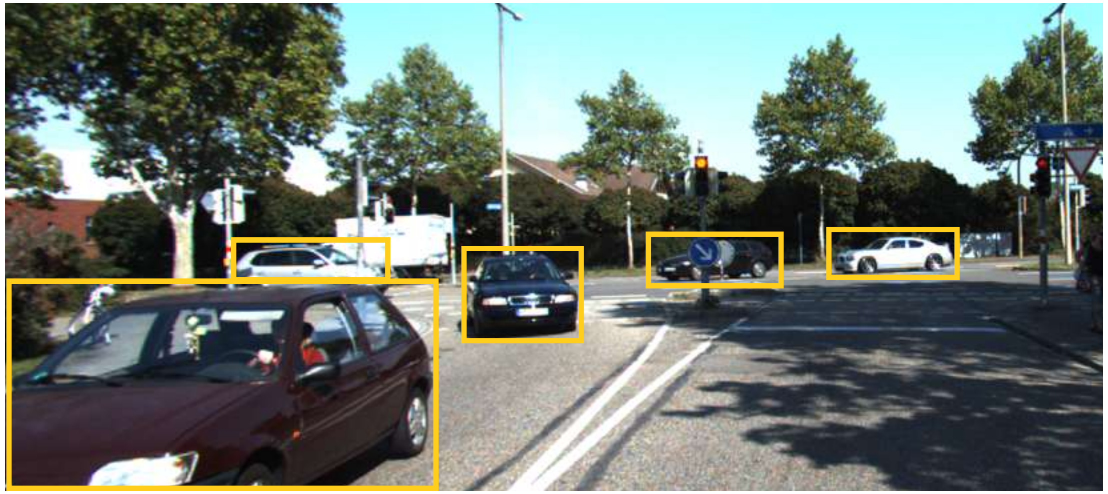
- ⊕ No mapping required
- ⊖ Single value for entire dataset

Methodology

Perception-based ADS Tasks

ADS

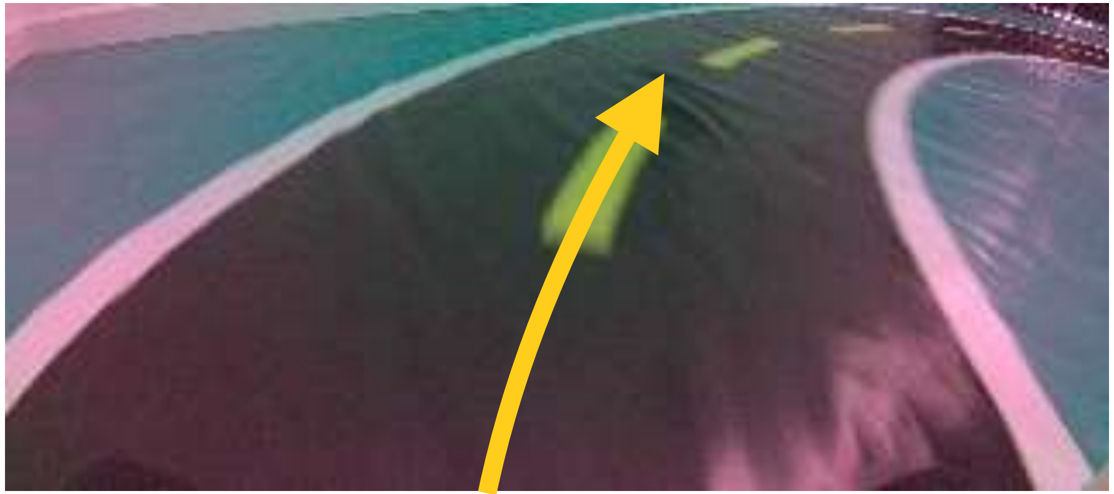
Vehicle detection



YOLOv3

COCO

Lane keeping



Dave-2

Mind
the Gap!

Redmon, J. et al. 2018
Lin, T. et al. 2014
Bojarski, M. et al. 2016
Stocco, A et al. 2023

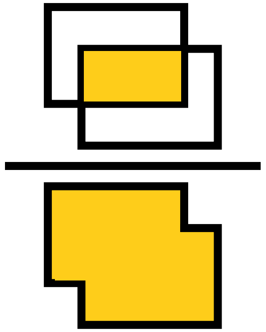
TUM fortiss

Methodology

ADS Evaluation metrics

Prediction Error

IoU



MSE

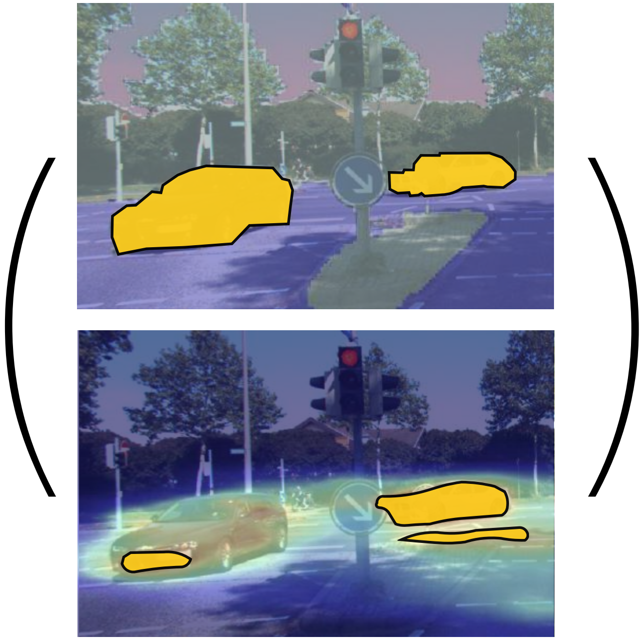
Confidence

P(CAR)

MC
Dropout

Attention Error

MSE



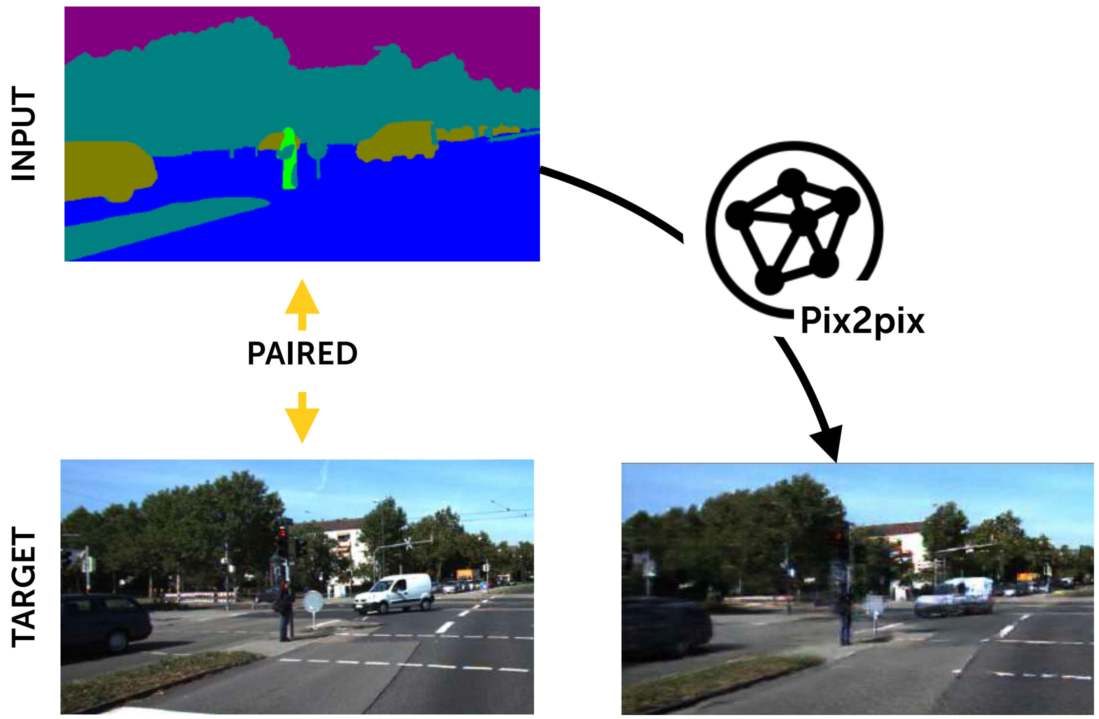
Dean, T. et al. 2013
Gal, Y. et al. 2016

Methodology

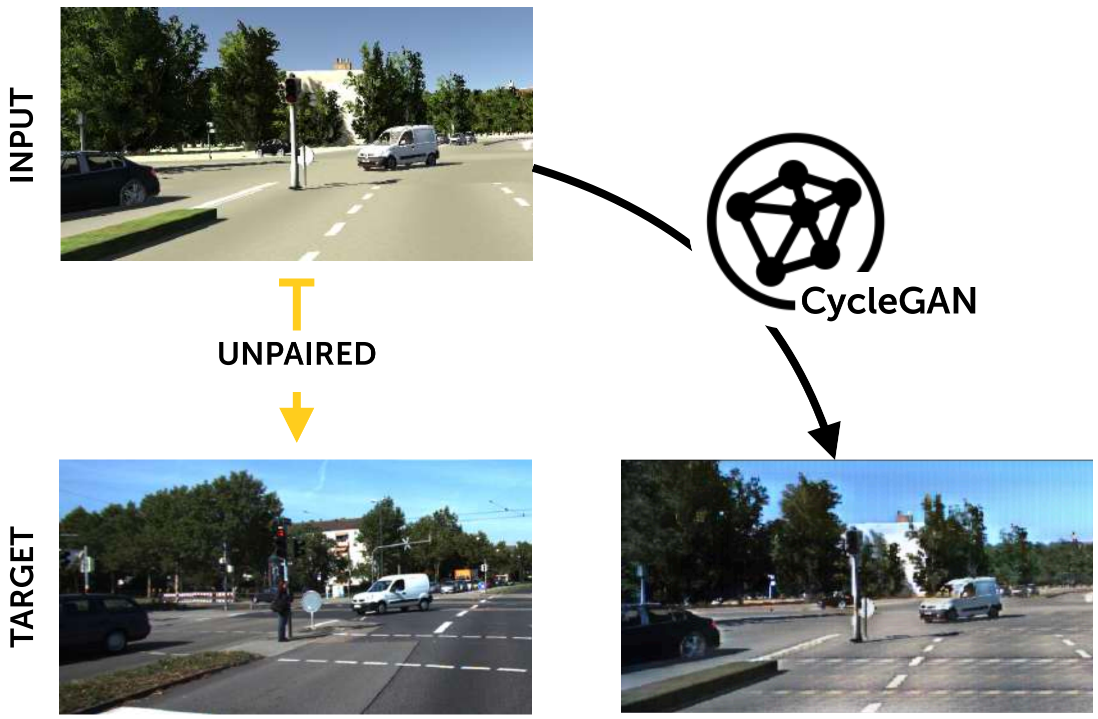
Generative Image-to-Image Translation Models

GEN

Paired training



Unpaired training



Zhu, J.-Y. et al. 2017
Isola, P. et al. 2017

Methodology

Image Quality Metrics

Distribution Level Metrics

IS	Inception-score
FID	Fréchet Inception Distance
KID	Kernel Inception Distance

Single Image Metrics

SSIM	Structural Similarity Index
PSNR	Peak signal-to-noise ratio
MSE	Mean Squared Error
CS	Cosine Similarity
TSI	Texture Similarity Index
WD	Wasserstein Score
KL	KL Divergence
HistI	Histogram Intersection
CPL	Classifier Perceptual Loss
SSS	Semantic Segmentation Score

Borji, A. et al. 2018
Pang, Y. et al. 2022

Empirical evaluation

Image Metrics VS Behaviour Metrics

Correlation

How do existing Image-to-image evaluation metrics correlate with the associated ADS behaviour?

RQ2 (Correlation)

Distribution Level Metrics

	Inception-score (IS)		Fréchet Inception Distance (FID)		Kernel Inception Distance (KID)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.41	0.14	0.24	0.64	0.54	0.54
Confidence	0.37	0.72	0.21	0.86	0.64	0.74
Attention Error	0.41	0.65	0.32	0.78	0.86	0.60

Pearson’s correlation coefficient (0,1)

RQ2 (Correlation)

Distribution Level Metrics

	Inception-score (IS)		Fréchet Inception Distance (FID)		Kernel Inception Distance (KID)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.41	0.14	0.24	0.64	0.54	0.54
Confidence	0.37	0.72	0.21	0.86	0.64	0.74
Attention Error	0.41	0.65	0.32	0.78	0.86	0.60

1 IS and FID are inconsistent across tasks

Pearson's correlation coefficient (0,1)

RQ2 (Correlation)

Distribution Level Metrics

	Inception-score (IS)		Fréchet Inception Distance (FID)		Kernel Inception Distance (KID)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.41	0.14	0.24	0.64	0.54	0.54
Confidence	0.37	0.72	0.21	0.86	0.64	0.74
Attention Error	0.41	0.65	0.32	0.78	0.86	0.60

1 IS and FID are inconsistent across tasks

2 KID is consistent across tasks

Pearson's correlation coefficient (0,1)

RQ2 (Correlation)

Distribution Level Metrics

	Inception-score (IS)		Fréchet Inception Distance (FID)		Kernel Inception Distance (KID)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.41	0.14	0.24	**0.64**	**0.54**	0.54
Confidence	0.37	0.72	0.21	**0.86**	**0.64**	0.74
Attention Error	0.41	0.65	0.32	**0.78**	**0.86**	0.60

Pearson’s correlation coefficient (0,1)

1. IS and FID are inconsistent across tasks
2. KID is consistent across tasks
3. FID is best performer for Lane Keeping, KID for Vehicle detection

RQ2 (Correlation)

Single Image Metrics (2 BEST PERFORMERS)

	Classifier Perceptual Loss (CPL)		Semantic Segmentation Score (SSS)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.29	0.30	0.23	0.25
Confidence	0.23	0.30	0.21	0.30
Attention Error	✕	0.16	✕	0.26

ρ Pearson's correlation coefficient (0,1) [Best of 6 models]

✕ At least 1 of 6 datasets has wrong correlation direction

RQ2 (Correlation)

Single Image Metrics (2 BEST PERFORMERS)

	Classifier Perceptual Loss (CPL)		Semantic Segmentation Score (SSS)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.29	**0.30**	0.23	0.25
Confidence	0.23	**0.30**	0.21	0.30
Attention Error	✗	0.16	✗	**0.26**

1 All metrics have weak or negligible correlation

n Pearson's correlation coefficient (0,1) [Best of 6 models]

✗ At least 1 of 6 datasets has wrong correlation direction

RQ2 (Correlation)

Single Image Metrics (2 BEST PERFORMERS)

	Classifier Perceptual Loss (CPL)		Semantic Segmentation Score (SSS)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.29	0.30	0.23	0.25
Confidence	0.23	0.30	0.21	0.30
Attention Error	✕	0.16	✕	0.26

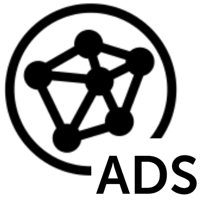
1. All metrics have weak or negligible correlation
2. Multiple metrics have the wrong correlation direction

n Pearson's correlation coefficient (0,1) [Best of 6 models]

✕ At least 1 of 6 datasets has wrong correlation direction

Empirical evaluation

Image Metrics +



Context

Fine-tuning

Does fine-tuning of I2I perception-based metrics improve the sim2real mitigation measurement?

RQ3 (Fine-tuning)

Generated | **Real-world**

SEG **Semantic segmentation model**

Targeted Semantic Segmentation **TSS** =

One Class **OC-TSS** =

RQ3 (Fine-tuning)

	Semantic Segmentation Score (SSS)		Targeted Semantic Segmentation (TSS)		One Class - Targeted Semantic Segmentation (OC-TSS)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.23	0.25	0.26	0.20	**0.57**	**0.42**
	Pix2pix	CycleGAN	BOTH		BOTH	

Pearson's correlation coefficient (0,1) [Best of 6 models]

1. Fine-tuned metrics correlate on both domains
2. TSS improves correlation on Vehicle detection
3. OC-TSS is the only metric with moderate correlation

TUM fortiss

Takeaways

GEN

REAL

Relative Behaviour Metrics

ADS

Image Metrics

GEN

1. Image-to-image GenAI tools effectively tackle domain adaptation in ADS
2. Current GenAI metrics don't align well with the software behavior that relies on their output
3. We need more domain-informed, semantic-aware metrics

Efficient Domain Augmentation for Autonomous Driving Testing Using Diffusion Models

Baresi, Hu, Stocco, Tonella.
https://arxiv.org/abs/2409.13661

ADS requires extensive coverage of the ODD

From regulations to implementation

Existing Standards and Regulations

- ISO/PAS 21448 Safety of the Intended Function (SOTIF)
- UN Regulation No 157 (2021/389)
- ISO 34505 “Scenery Elements (Section 9)” and “Environmental Conditions (Section 10)”

Operational Design Domain (ODD)

- roadway types
- geographic area
- speed range
- environmental conditions (weather as well as day/night time)

Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI

Simulators

- Scalable Testing Environments
- Cost-Effective Data Generation
- Enhanced Control and Repeatability

...enhanced with Generative AI

- Domain-to-Domain transformations (e.g., CycleGAN)
- Text-to-Image transfomations (e.g., Stable Diffusion)
- Edit-Instruction transformations (e.g., InstructPix2Pix)
- Control-conditioned transformations (e.g., Controlnet)

Solution: Diffusion Models

Usage for Test Set Augmentation in simulation platforms

Augmentation: **Lightning Strikes**

Input Image

Instruction-edited

Inpainting

Inpainting with Refining

Augmentation: **Autumn Season**

Input Image

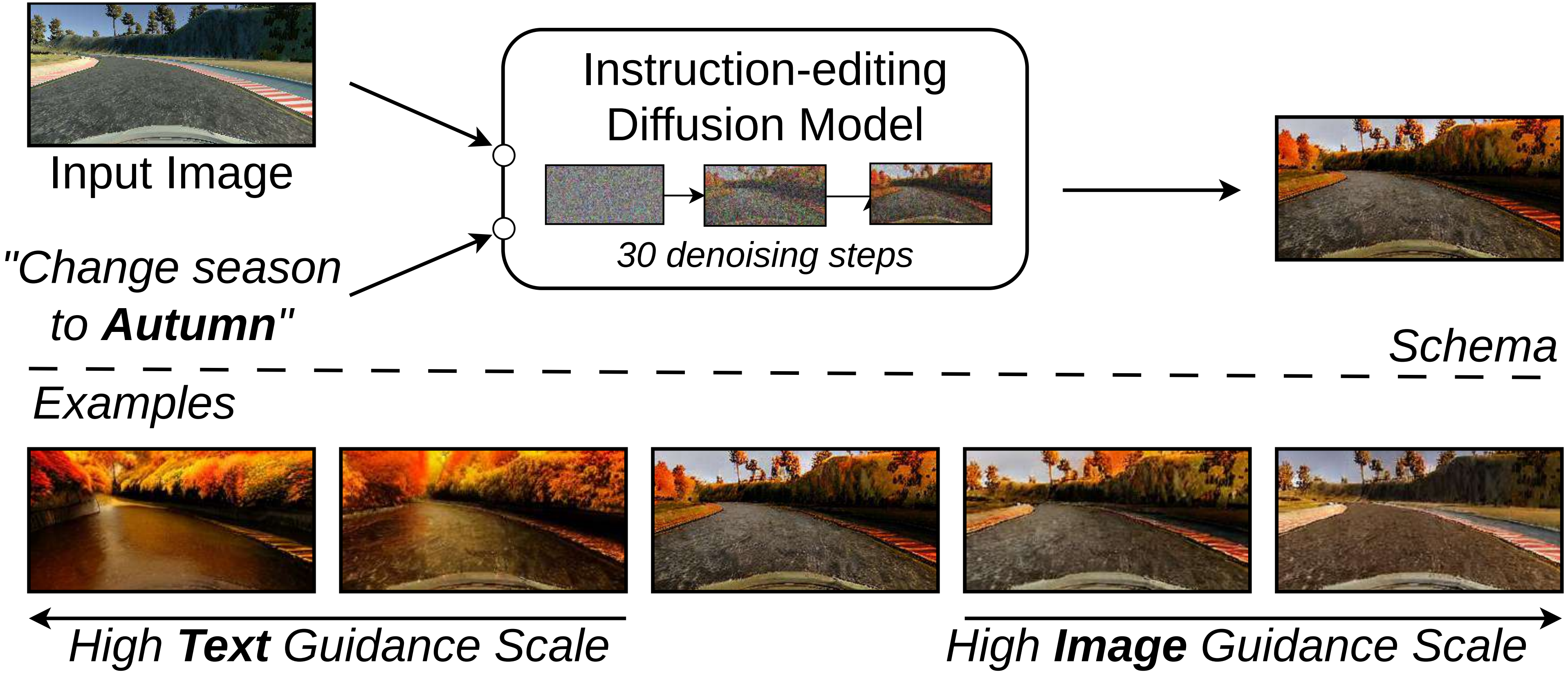
Instruction-edited

Inpainting

Inpainting with Refining

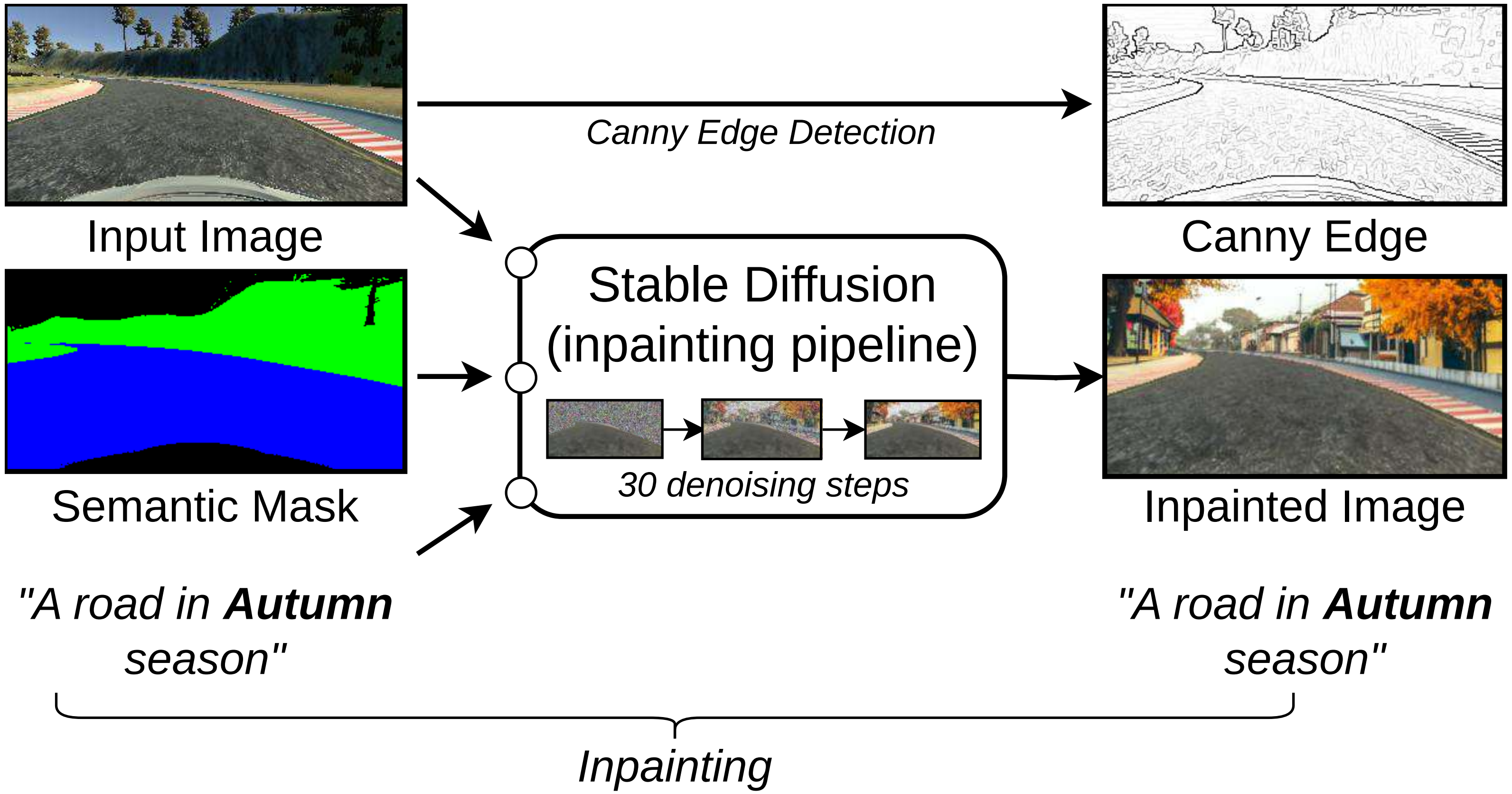
Instruction-editing

Prompt: Textual



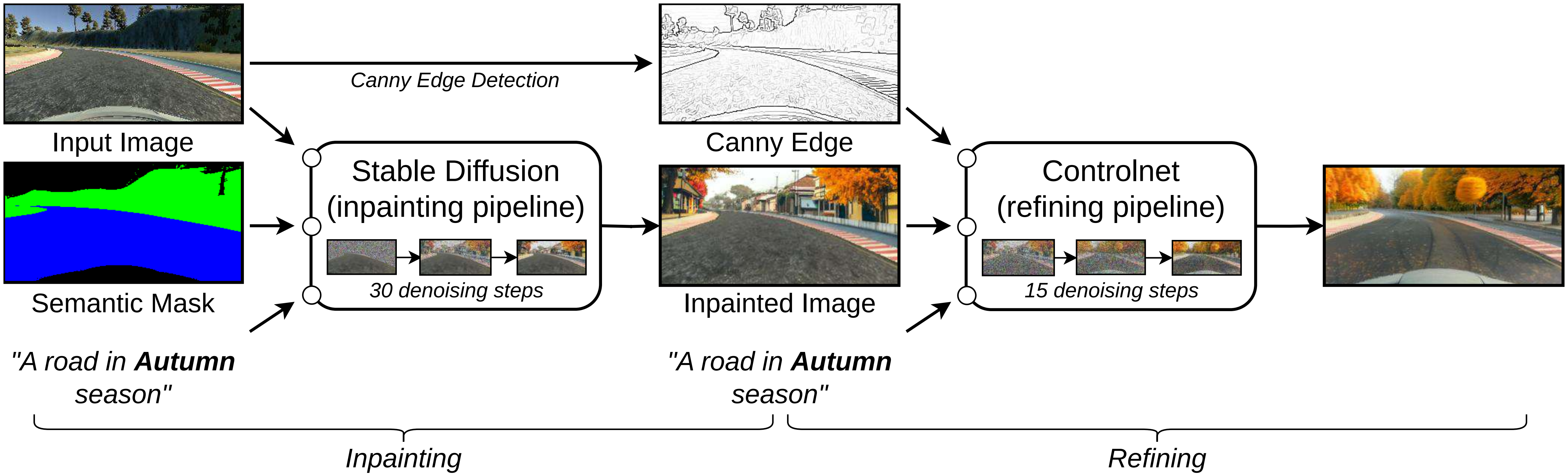
Inpainting

Prompt: Textual + Mask



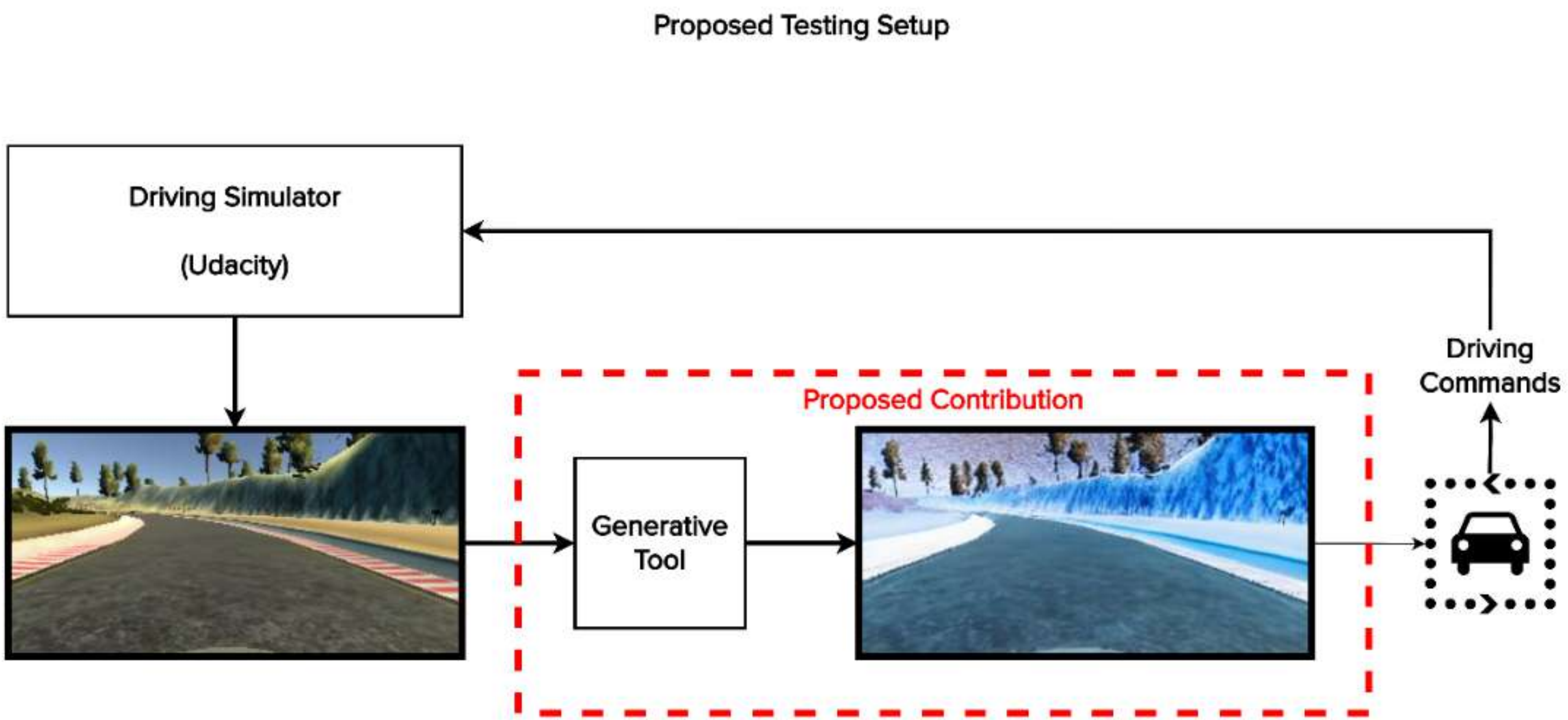
Inpainting with Refinement

Prompt: Textual + Mask



Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI



Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI (naïve integration)

InstructPix2Pix
(Diversity, No Temporal Consistency)

Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI (knowledge distillation)

Our Proposition based on Knowledge Distillation
(Diversity and Temporal consistency)


Evening
Sunset
Dawn
Morning
Twilight
Night
Sunrise
Dusk

Egypt
Brazil
Australia
Argentina
Canada
Germany
Evening
Sunset
Dawn
England
Morning
Twilight
Night
Sunrise
Dusk


Egypt
Brazil
Australia
Argentina
Canada
Germany
Evening
Sunset
Dawn
England
Beijing
Morning
Twilight
Berlin
Night
Sunrise
Dusk
Dubai


Egypt
Brazil
Australia
Argentina
Canada
Germany
Evening
Sunset
Dawn
England
Beijing
Morning
Twilight
Autumn
Berlin
Night
Sunrise
Dusk
Dust
Dubai
Forest
Blizzard
Snow
Foggy

Contributions

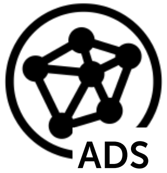
Empirical evaluation

RQ1 Semantic Validity

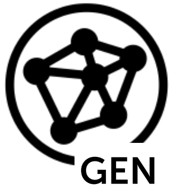
RQ2 Effectiveness

RQ3 Efficiency

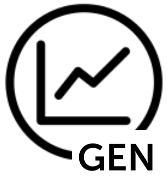
Methodology

4 

Perception-based ADS tasks

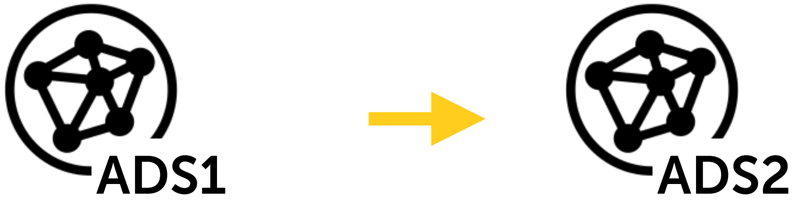
3 

Diffusion models architectures

52 

Operational design domains

Empirical evaluation



Validity

Do diffusion models generate augmented images that are semantically valid ODDs?

How effective is the semantic validator at detecting invalid augmentations?

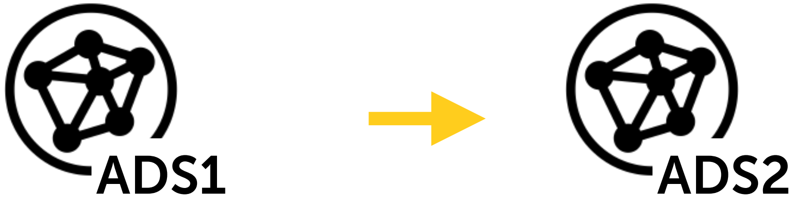
Human Study - Semantic Preservation OC-TSS

- 👉 **Human Study**
  - 🔎 33 participants
  - 🔎 (about 3150 answers)
  - 🔎 66%+1 Agreement

- 👉 Instruction-Edited:
  TP: 18, FN: 2, TN: 16, FP: 0
- 👉 Inpainting:
  TP: 19, FN: 10, TN: 0, FP: 0
- 👉 Refining:
  TP: 10, FN: 4, TN: 4, FP: 3

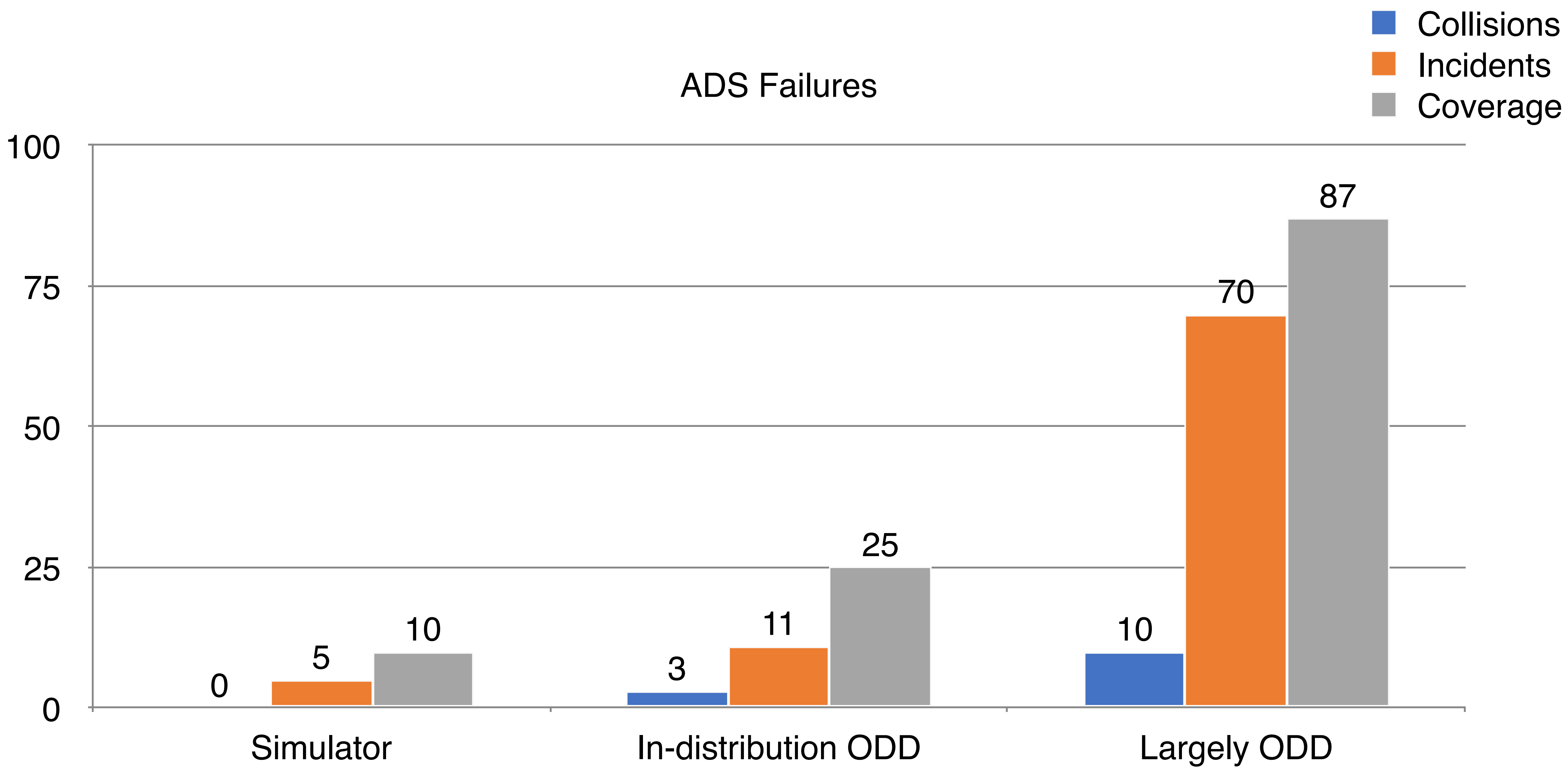
Training Set			
TARGET / OUTPUT	Class0	Class1	SUM
Class0	47 54.65%	3 3.49%	50 94.00% 6.00%
Class1	16 18.60%	20 23.26%	36 55.56% 44.44%
SUM	63 74.60% 25.40%	23 86.96% 13.04%	67 / 86 77.91% 22.09%

Empirical evaluation

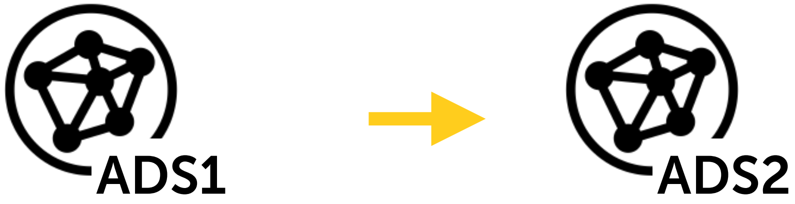


Effectiveness

How effective are augmented images in exposing faulty system-level misbehaviors of ADS?


ADS Failures
Collisions
Incidents
Coverage
100
75
50
25
0
0
5
10
3
11
25
10
70
87
Simulator
In-distribution ODD
Largely ODD

Empirical evaluation



Efficiency

What is the overhead introduced by diffusion model techniques in simulation-based testing?

Does the knowledge-distilled model speed up computation?

Performance Overhead (Inference)

👉 Normal Simulator with **ADS**:
  🔎 **100.24** ± 22.24 milliseconds

👉 AugmentedSim with **Instruction**:
  🔎 **1118.47** ± 114.89 milliseconds (+11X)

👉 AugmentedSim with **Inpainting**:
  🔎 **1370.61** ± 105.95 milliseconds (+13X)

👉 AugmentedSim with **Inpainting with Refinement**:
  🔎 **2029.57** ± 115.03 milliseconds (+20X)

👉 Our Approach (**Knowledge Distillation**):
  🔎 **120.30** ± 0.7 milliseconds (+0.02X)

Takeaways

ORIG ODD

NEW ODD

ADS Behaviour Metrics

GEN Diffusion Models

1. Diffusion models effectively tackle domain generation for ADS testing
2. They complement simulator testing, uncovering failures in areas previously considered error-free
3. Knowledge distillation is key to achieving high simulation efficiency

TUM fortiss

Thank you very much!

Your contact
stocco@fortiss.org