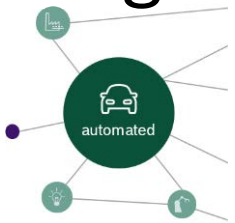
OPEN QUESTIONS IN TESTING OF LEARNED COMPUTER VISION FUNCTIONS FOR AUTOMATED DRIVING,

MATTHIAS WOEHRLE, CHRISTOPH GLADISCH AND CHRISTIAN HEINZEMANN

BOSCH

Testing of learned computer vision function for Automated Driving

Automated Driving



- Level 4/5 automated driving in an urban environment
  - High demands on safety and performance in highly complex scenarios
- Classical software verification methods and coverage criteria not sufficient
  - Tests and coverage need to be defined on the domain, not only on the software structure as today
  - Particularly, if autonomy is supported or (partially) implemented by machine learning

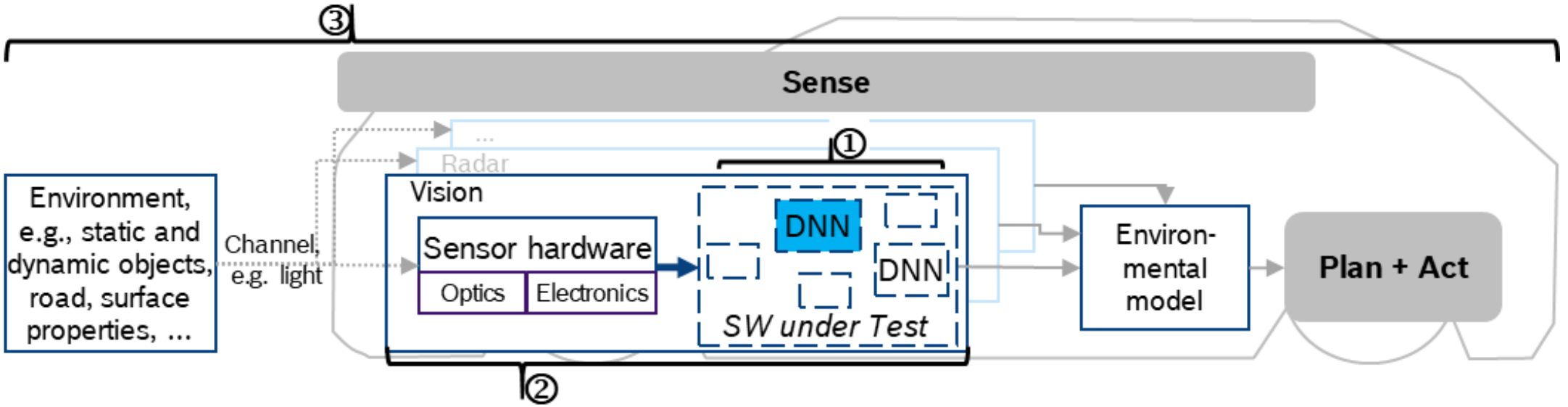
Goal: Make Testing Economically Feasible

∞ *reduce to* → **N**

Coverage of input domain of autonomous system theoretically requires infinitely many test cases due to open context -> **finite test set**

Testing of learned computer vision function for Automated Driving

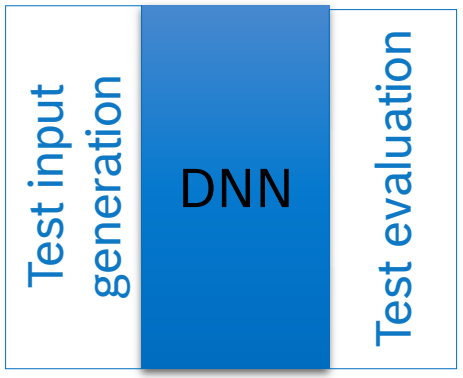
System context: Vision function in an automotive context



How do we create good (relevant and meaningful) test data efficiently for a CV function interpreting images of driving scenes in the physical world? How do we verify relevant properties of the corresponding DNNs?

BOSCH

Testing of learned computer vision function for Automated Driving
Summary of the paper



- Synthesis of work from autonomous driving, software testing, computer vision and machine learning
- Overview of
  1) test generation
  2) test evaluation methods
- 11 Exemplary research questions

Different view point to literature for training & validation in machine learning:

- *In training we focus on comparing average case behavior based on cost metrics - mainly to evaluating competing designs*
- *In verification and testing we are typically concerned with (worst-case) behavior w.r.t. specific properties.*

Testing of learned computer vision function for Automated Driving

Test generation: Leveraging synthetic data

https://github.com/carla-simulator/carla

MIT License

(+) Generate dedicated samples to cover the domain
(+) Labels and meta-data inherently available
(-) Required affordances and fidelity [1]
(-) Residual risk w.r.t simulation and its fidelity [2]

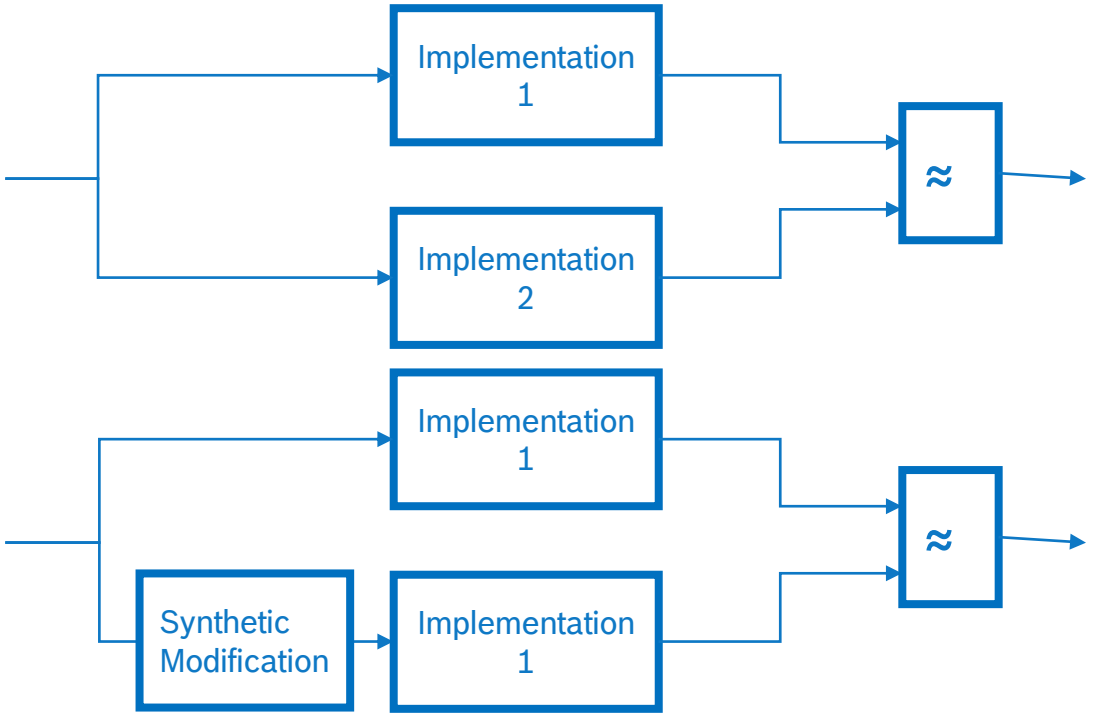
Synthetic \ Real	Desirable	Violation
Test passes	Accepted desirable	Missed violation
Test fails	False alarm	Caught violation

Exemplary Research Question: Which affordances should a simulation provide to support building a good test set with domain coverage?

[1] Hutter, A.: Einsatz von Simulationsmodellen beim Test elektronischer Steuergeräte. In: Sax, E. (ed.) Automatisiertes Testen Eingebetteter Systeme in der Automobilindustrie. Hanser (2008)
[2] Koopman, P.,Wagner, M.: Toward a framework for highly automated vehicle safety validation. Tech. rep., SAE Technical Paper (2018)

Testing of learned computer vision function for Automated Driving

Test evaluation: Invariance



Differential testing

- Requires two implementations
- Only shows inconsistencies
- Example: DeepXplore [3]

Metamorphic testing

- Requires synthetic data modification
- Relies on synthetic data evaluation (above)
- Example: DeepRoad [4]

Exemplary Research Question: How can we use specifications and formal methods to reap a larger benefit from ground truth data to effectively multiply our test set?

[3] Pei K, Cao Y, Yang J, Jana S. DeepXplore: Automated whitebox testing of deep learning systems. Proc. SOSP 2017,1–18.
[4] Zhang M et al.. DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems. ASE 2018,132–142.

BOSCH

THANK YOU

CHRISTOPH.GLADISCH@DE.BOSCH.COM

BOSCH

Testing of learned computer vision function for Automated Driving

Proposed research questions for test input generation

- Sampling around Labeled Test Images:
  - What notions of robustness and corresponding test images should be included in a good test set?
  - Which kind of coverage criterion could be used to argue exhaustiveness of a test set?
  - What data augmentations should be used on images in a good test set?
- Domain and Data Analysis:
  - What would be a basic check list of nuisance factors and other hazards that should be considered for a good test set?
  - How do we integrate knowledge from the analysis of the ODD and OEDR into designing a good test set?
  - How to concertize abstract tests into concrete images?
- Synthetic Data
  - Which affordances should a simulation provide to support building a good test set with domain coverage?
  - How can we leverage synthetic data to economically scale a good test set?

BOSCH

Testing of learned computer vision function for Automated Driving

Proposed research questions for test evaluation

- How do we obtain ground truth for diverse test data in a cost-effective manner?
- What are relevant, task-and domain-specific evaluation metrics?
- How can we use specifications and formal methods to reap a larger benefit from ground truth data to effectively multiply our test set?

BOSCH