KI ABSICHERUNG

Safe AI for Automated Driving

Final Event | June 23rd 2022

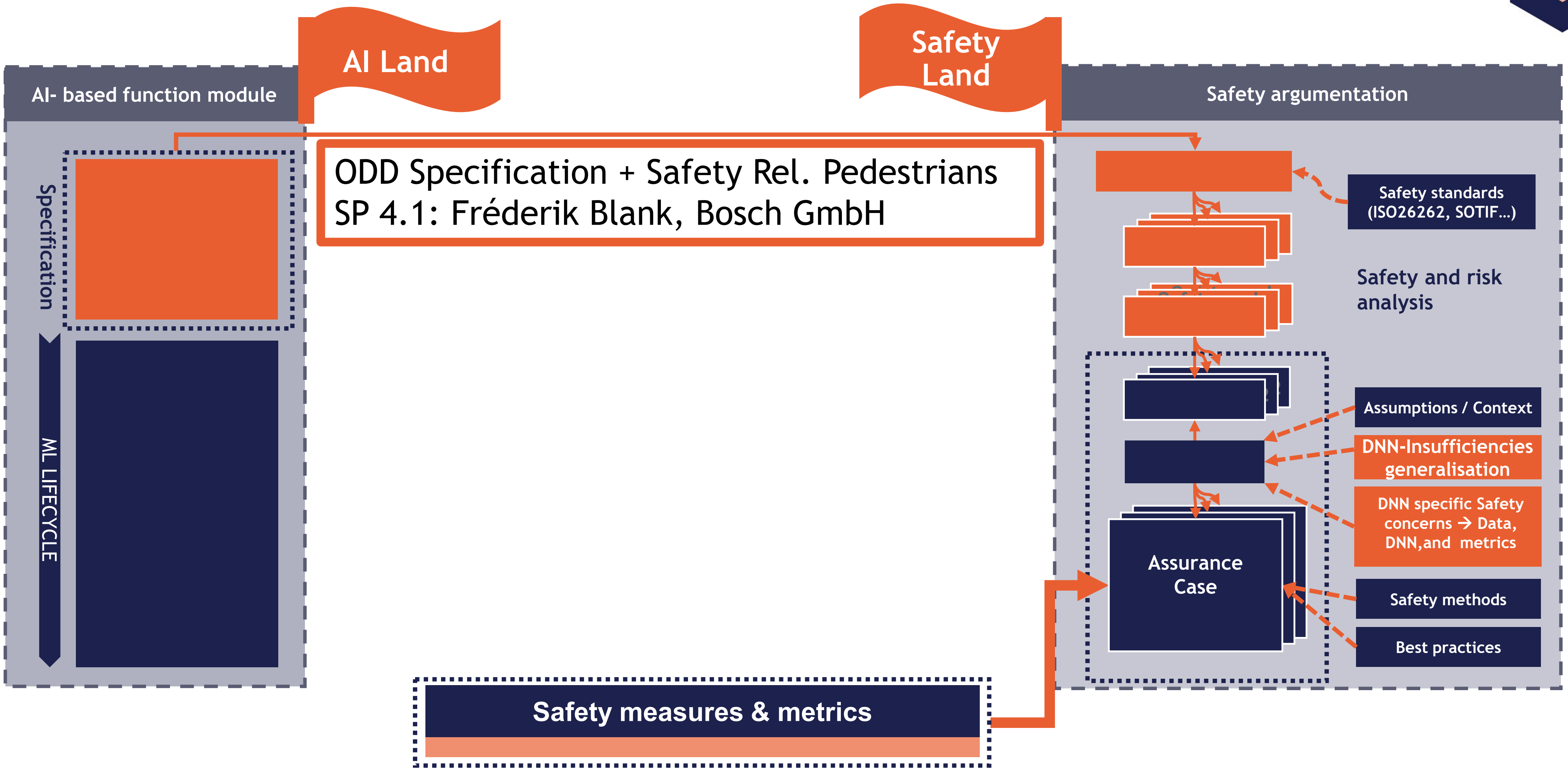
KI Absicherung: Results from the Sub-Projects

PD Dr. Michael Mock (Fraunhofer IAIS), Frédérik Blank (Robert Bosch GmbH),
Dr. Thomas Stauner (BMW AG), Fridolin Bauer (BMW AG), Dr. Fabian Hüger (Volkswagen AG),
Andreas Rohatschek (Robert Bosch GmbH)

Big Picture

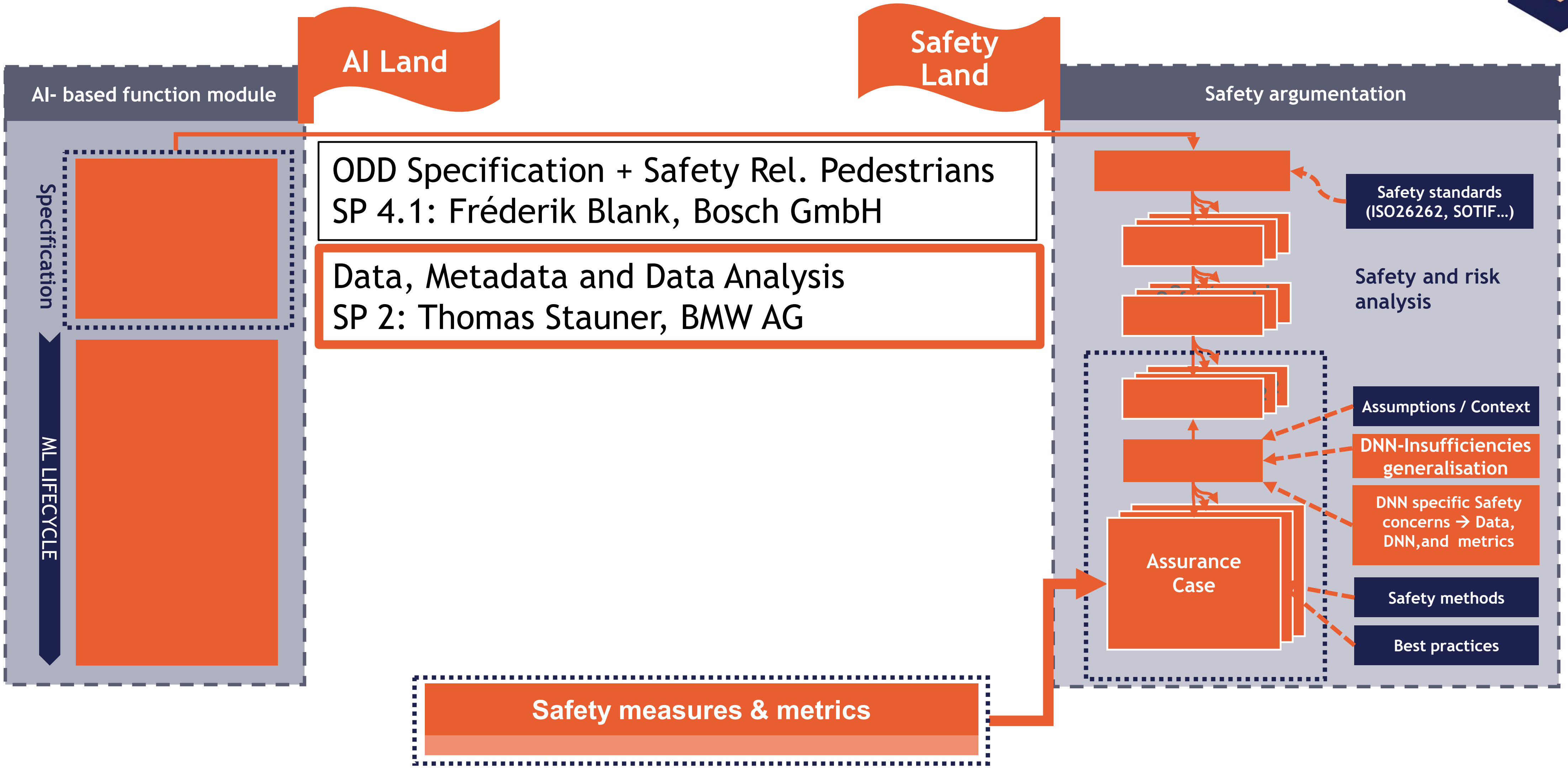
Michael Mock, Fraunhofer IAIS

KI-Absicherung: Results from the Sub-Projects

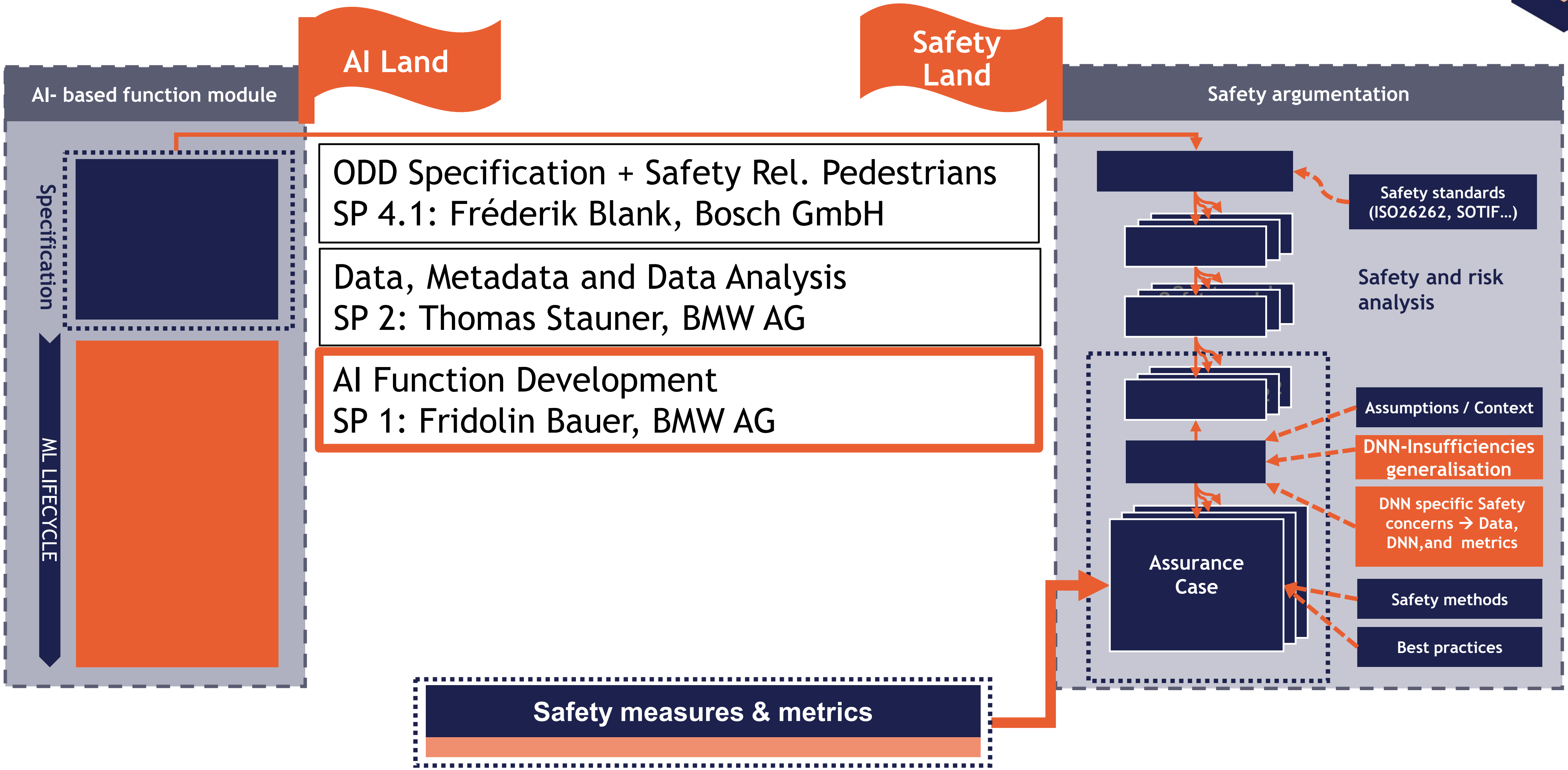


INTERNAL

KI-Absicherung: Results from the Sub-Projects



KI-Absicherung: Results from the Sub-Projects

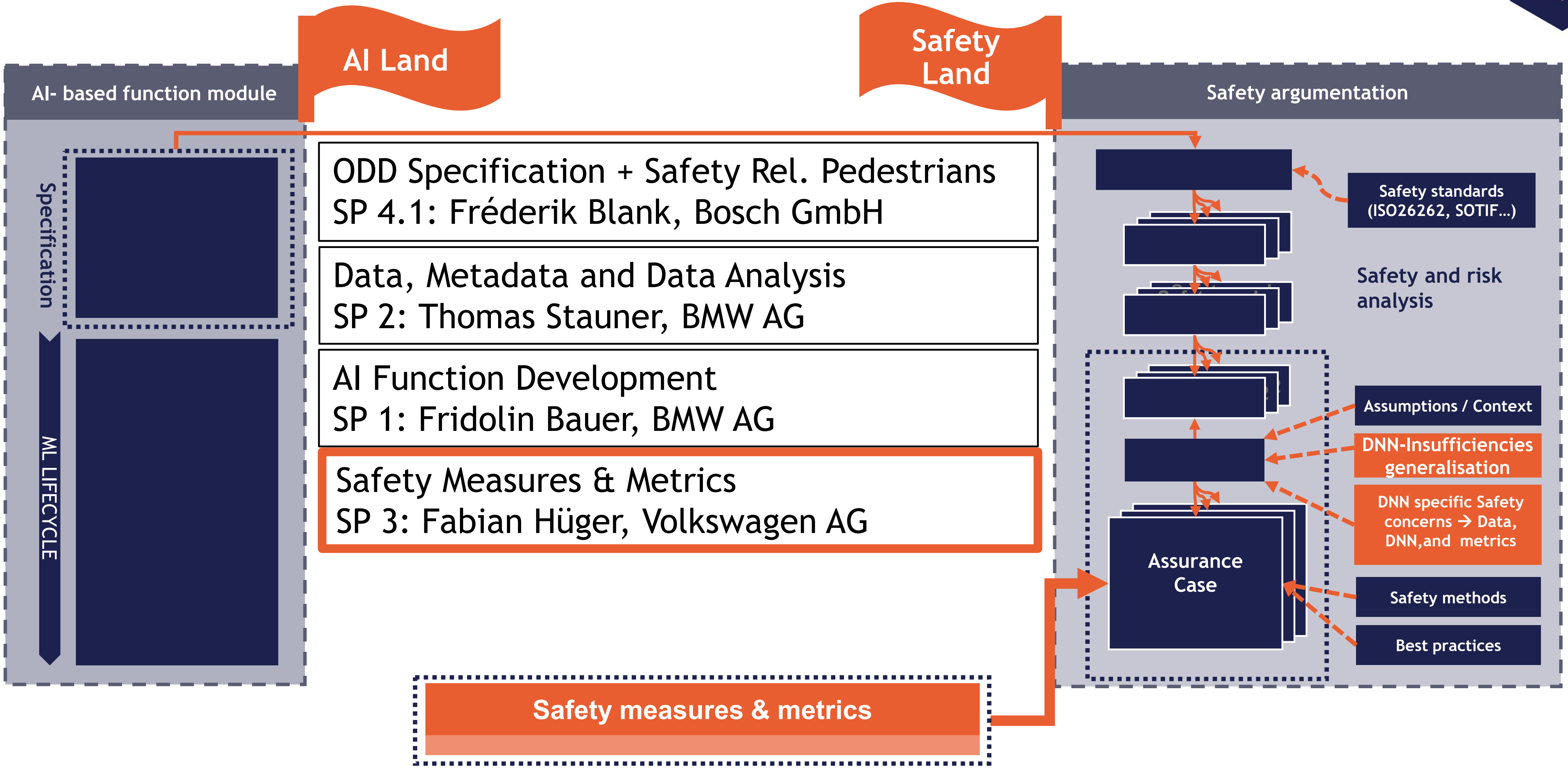


ODD Specification + Safety Rel. Pedestrians
SP 4.1: Fréderik Blank, Bosch GmbH

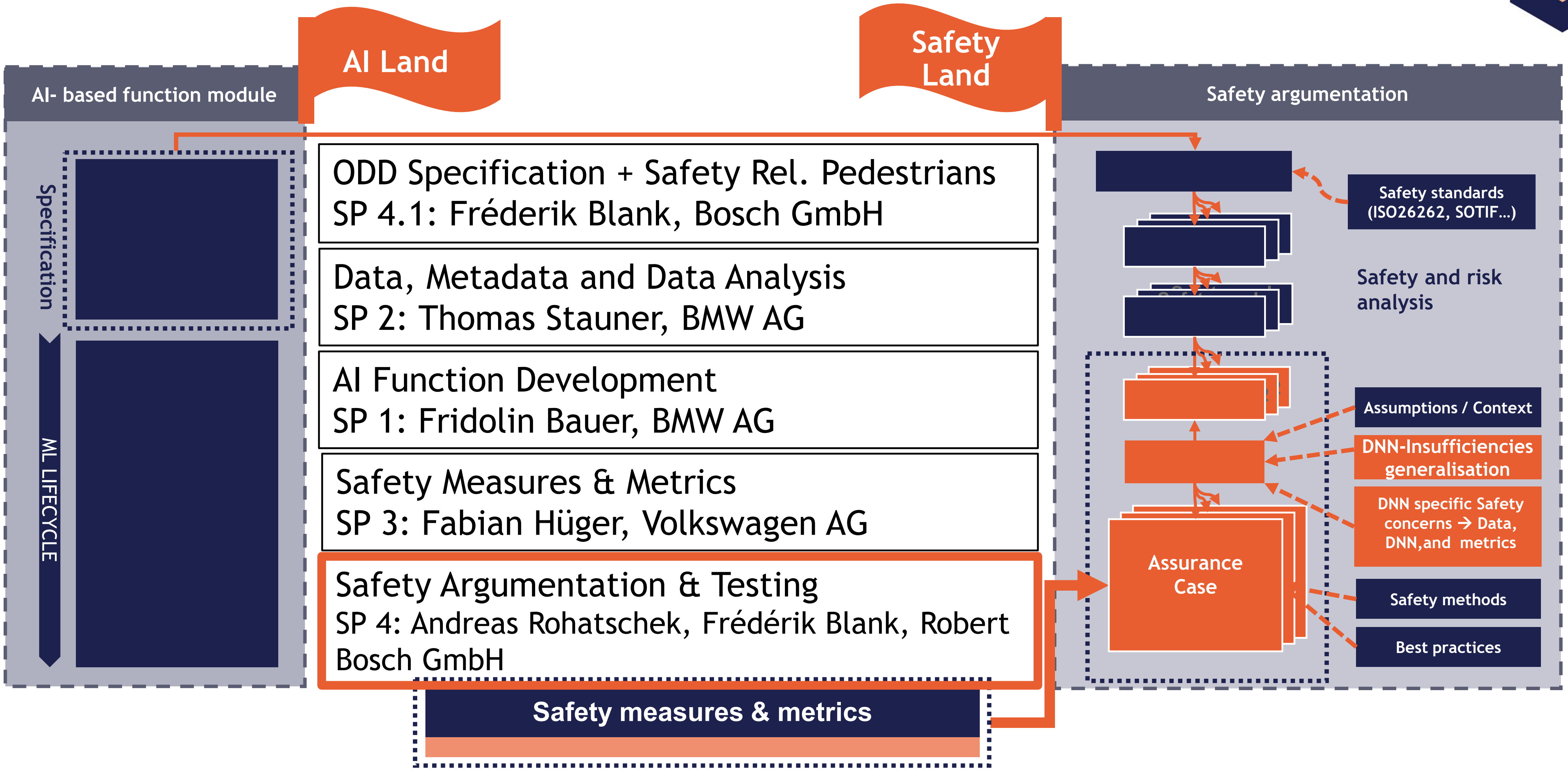
Data, Metadata and Data Analysis
SP 2: Thomas Stauner, BMW AG

AI Function Development
SP 1: Fridolin Bauer, BMW AG

KI-Absicherung: Results from the Sub-Projects



KI-Absicherung: Results from the Sub-Projects



INTERNAL

ODD Specification + Safety Rel. Pedestrians

SP 4.1: Frédérik Blank, Robert Bosch GmbH

KI ABSICHERUNG
Safe AI for Automated Driving

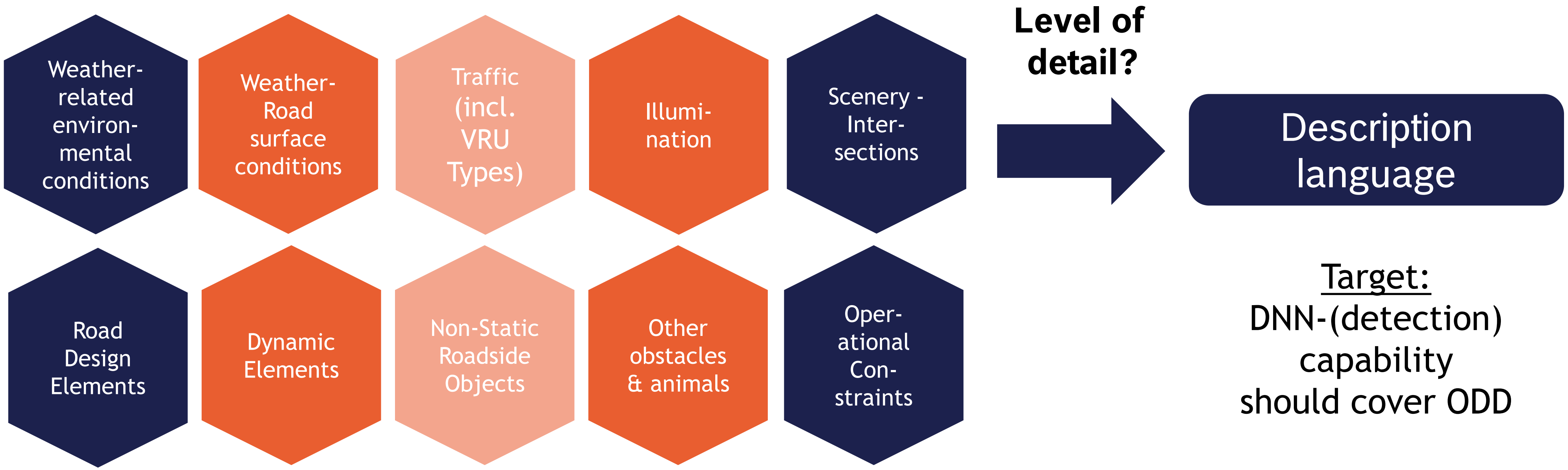
Final Event | June 23rd 2022

Structuring the input Space & Safety-relevant pedestrians (AP4.1/P1)

Frédérik Blank, Robert Bosch GmbH

Structuring the input Space - Operational design domain (ODD)

- An ODD describes / specifies operating conditions under which a given automated driving system or feature is specifically designed to function [...]
  - Taxonomy and Definitions for Terms Related to Driving Automation Systems (examples)



<u>Target:</u>
DNN-(detection) capability should cover ODD

A description language & semantic input space modeling is needed to…

Complexity of language

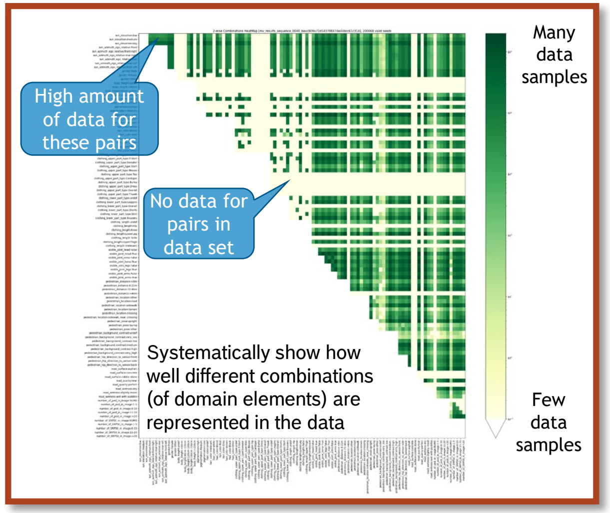
Be able to describe / specify operating conditions (and edges of ODD*) as of PAS 1883:2020 and others

Systematically capture important knowledge and describe the (expected) key input space dimensions and their possible variations (→ Ontology / Semantic domain model)

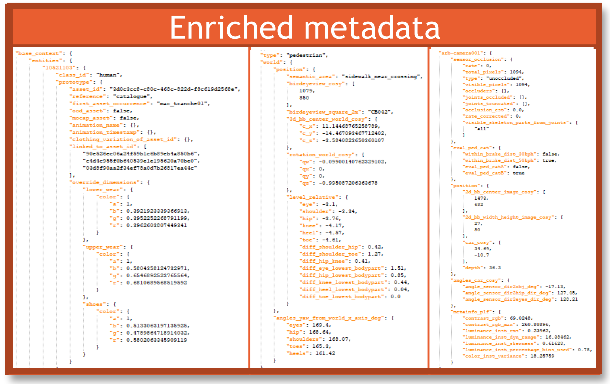
Perform training and assurance data coverage analyses for data driven AI-based systems

Systematically describe training & test data sets including safety-relevant Corner cases / rare critical situations to be considered

For synthetic perception data production & metadata: describe data dimensions that should be variated & incrementally generate new data

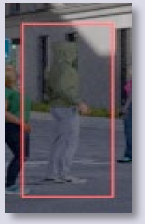


Visualization of an exemplary data coverage analysis

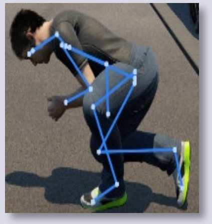


Extract of an enriched metadata JSON for one pedestrian instance

Performance Limiting Factors (PLFs)



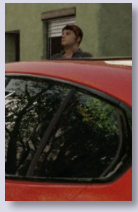
low contrast (e.g. similar color to background)



Un-common poses

uncommon person **clothing,** strong patterns

Light induced image **artefacts** (e.g. reflections)



(strong) **occlusions** by objects, light, …

Distant persons (depth)

High range of light intensities

Low contrast due **weather conditions**

PLFs support to

- identify important data dimensions
- prioritize and constrain useful training & test-space

PLFs in images ideally to be tagged/ labeled (semi)automatically

Definition

A measurable factor, either

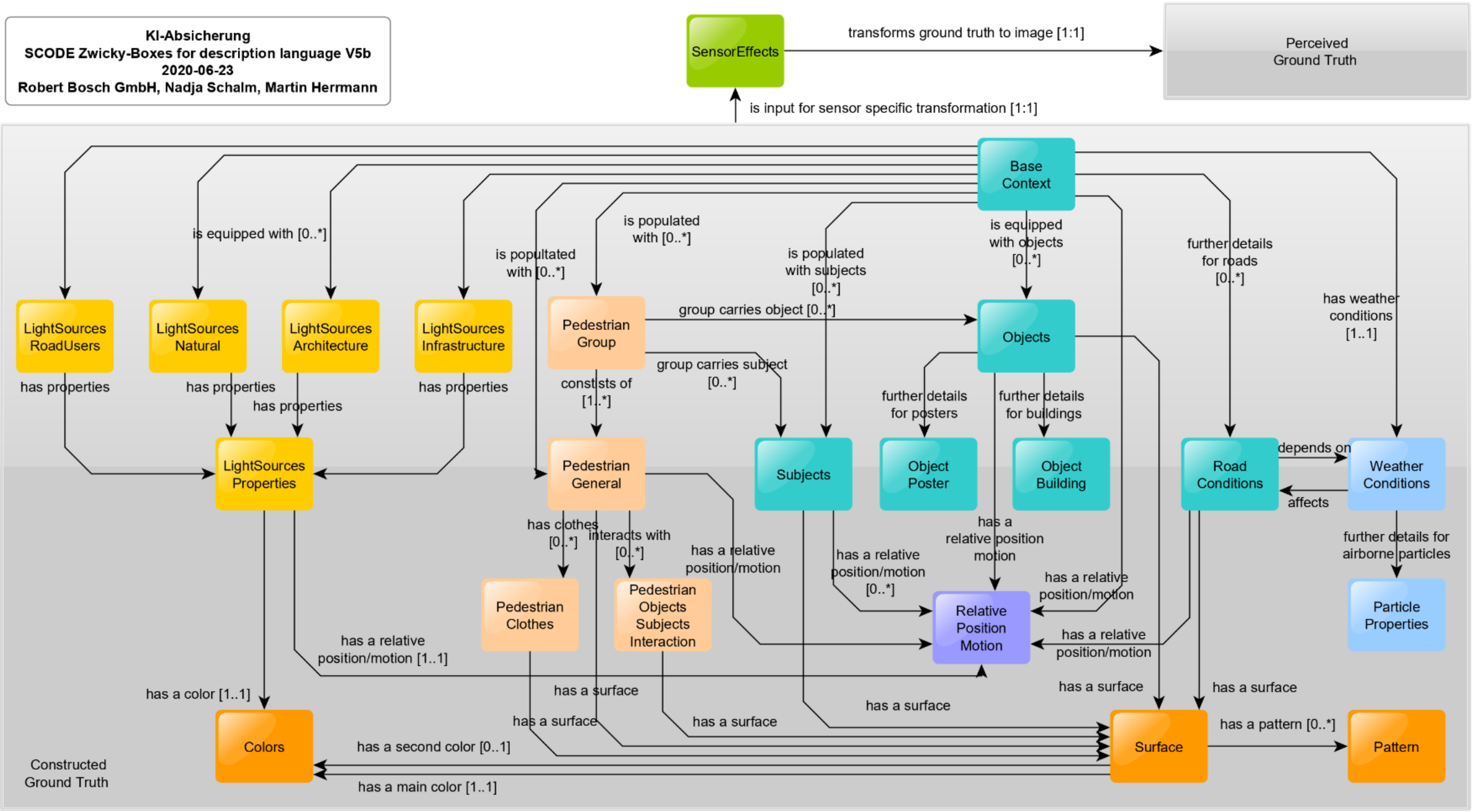
- Direct physical effect or
- Model of effect

that leads to drops in perception performance

Further examples of PLFs:

- uncommon person locations, above or below ground
- uncommon person motion
- groups of persons, occlusion
- person depictions on images and posters
- person reflections in specular surfaces
- …

High Level View of domain model / Ontology



Source: Bosch

Ontology as **semantic description** of input space to describe Operational Design Domain (ODD) & input data

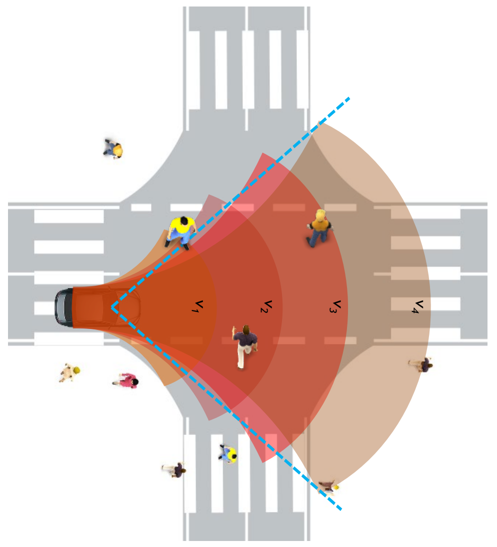
Total

- ~10 subdomains
- ~250 dimensions
- ~1000 variations / alternatives

Approach (iterative)

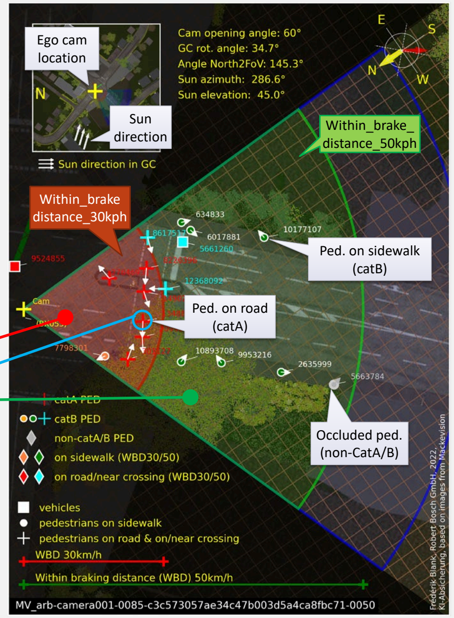
- Review of public data sources / existing standards
- Brainstorming with experts
- Expert interviews
- Iterative refinement
- Needs to be challenged / extended by identified corner cases

Safety relevant pedestrians



- From a safety perspective and risk assessment, not every pedestrian is "equally" at risk.
  - ➢ Include safety relevance of a pedestrian into ML-based metrics
- Description language, ontology & metadata to provide means to:
  - ➢ Describe pedestrians and their possible safety-related characteristics
- Starting point: Definition of *relevance* based on purely positional considerations:
  - Braking distance / distance of person to ego-vehicle
  - Ego-Camera opening angle
  - Semantic area of pedestrian location → Road / Sidewalk / Crossing
- Each pedestrian was annotated with safety-related metadata (eval_CatA / eval_CatB / other)

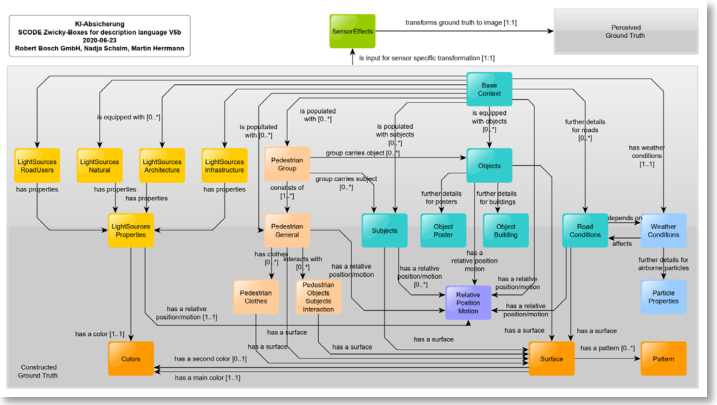

```
"13488544": {
   ...
  "depth": 18.04,
  "angle": -4.92,
  "pos_cc_x": 17.97,
  "pos_cc_y": -1.55,
  "pos_bev_col": 1120,
  "pos_bev_row": 1177,
  "within_brake_dist_30kph": true,
  "within_brake_dist_50kph": true,
  "semantic_area": "crossing",
  "eval_catA": true,
  "eval_catB": true
  }
```

Summary

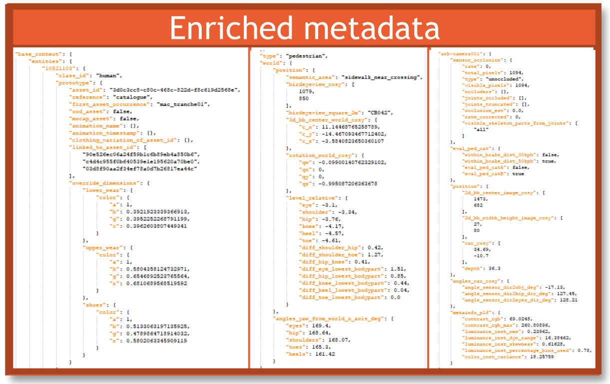
Key enablers in the overall approach for assurance of AI-based functions…

Domain model & Ontology



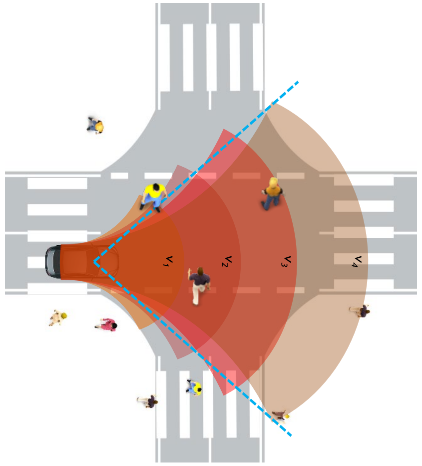
10 sub-somains, 250 dimensions

Enriched metadata



>50 enriched meta-annotations per pedestrian

Safety relevant pedestrians



3 safety categories

KI ABSICHERUNG

Safe AI for Automated Driving

Frédérik Blank, Robert Bosch GmbH

Frederik.Blank@de.bosch.com

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

Supported by:

Federal Ministry for Economic Affairs and Climate Action

on the basis of a decision by the German Bundestag

www.ki-absicherung.vdali.de @KI_Familie KI Familie

2

Data, Metadata and Data Analysis
SP 2: Thomas Stauner, BMW AG

KI ABSICHERUNG
Safe AI for Automated Driving

Final Event | June 23rd 2022

TP2: Generieren von synthetischen Lern- und Testdaten.

Thomas Stauner, BMW AG

Synthetic Data Offers Unique Features for Training and Assurance of ML

- Facilitated GDPR compliance
- Simulation of sensor variants and mounting positions
- Explicit control of coverage and bias
- Influence factor analysis based on targeted variations, esp. for safety analysis
- Design of data of dangerous and/or rare situations
- Provision of rich ground truth and metadata

Same sensor position, different sensor patameters

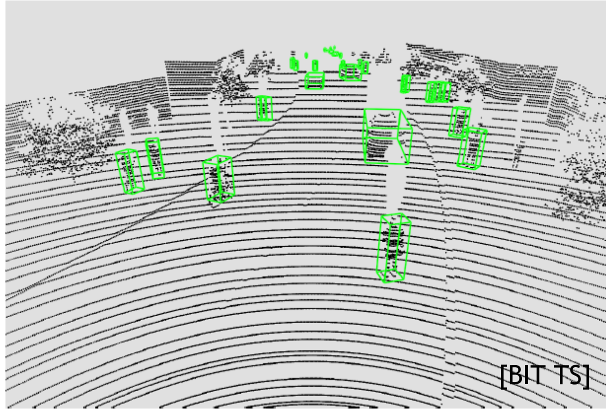
Support for Influence Factor Analysis

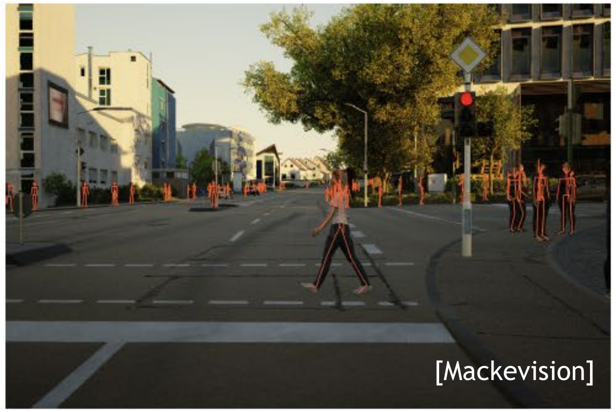
- Same camera position, different sun position

Control of the Data Distribution and Generation of Dangerous Situations

- Varying pedestrian distribution, close pedestrians on the road

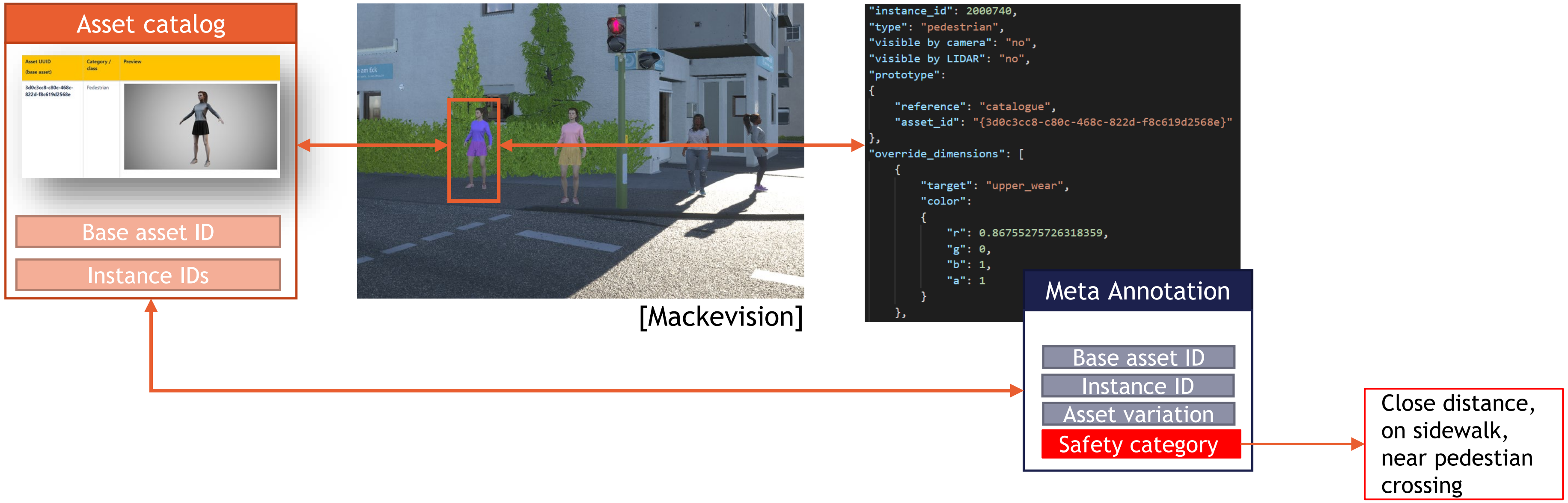
Provision of Rich Ground Truth and Metadata





For safety analysis, diverse meta data for synthetic images can be computed

- For the systematic analysis of weaknesses of an AI function, rich meta information is required
- It allows the engineer to retrieve semantic information w.r.t. an ontology for the situation depicted in a frame. Examples are body size of pedestrians or safety relevance w.r.t. the function under development



Two Toolchains with Distinct Features Have Been Developed: (1) with Physical-Based Rendering

- Target: Accurate simulation of light transport within the virtual scene
- Architecture: Integration of Intel OSPray Studio/BIT TS scene generator with real sensor models from Bosch (Camera) and Valeo (Lidar)
- Special Features
  - Automized scenario generation
  - glTF 3D scene format
  - Realistic Lidar data
  - Procedural, physics-based sun-sky model, support of motion blur
  - Natural motion due to motion capturing on assets

High scene complexity Motion Blur

[Images: Intel/BIT TS]

Two Toolchains with Distinct Features Have Been Developed: (2) with Real-Time Rendering Engine

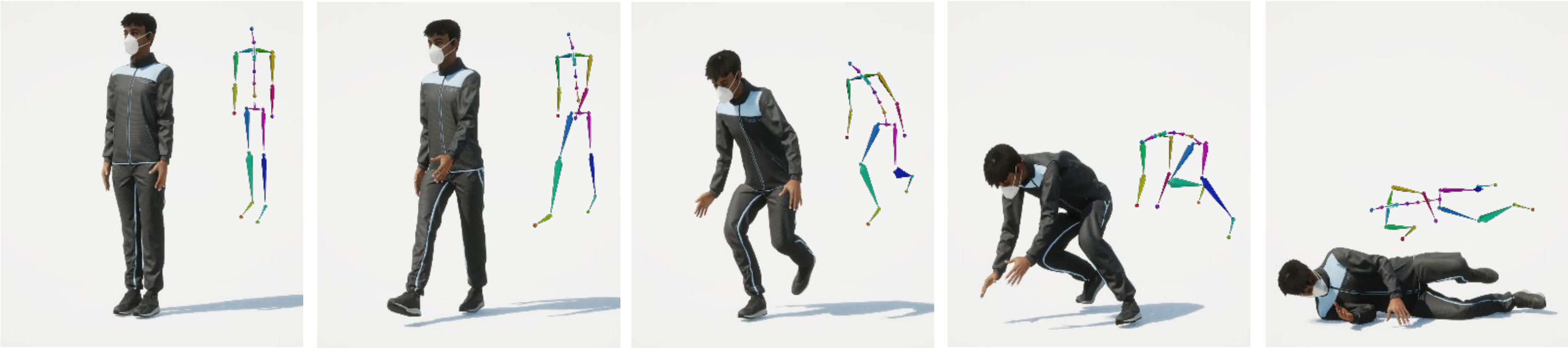
- Target: exploit capabilities and efficiency of State-of-the-art game engine, with high quality lighting, powerful material systems, animation tools, and flexible APIs
- Special features:
  - Automated scene variations, e.g. clothing, parked cars, combination of movements
  - Effects – procedural sun, procedural clouds, wetness, fog, vignetting, lens flare, artificial light
  - Natural motion due to motion capturing
  - Metadata on occlusion
  - Support for automatic scene generation from TP4 format

[Images: Mackevision]

The Toolchains Build on Targeted Asset Generation and Motion Capturing

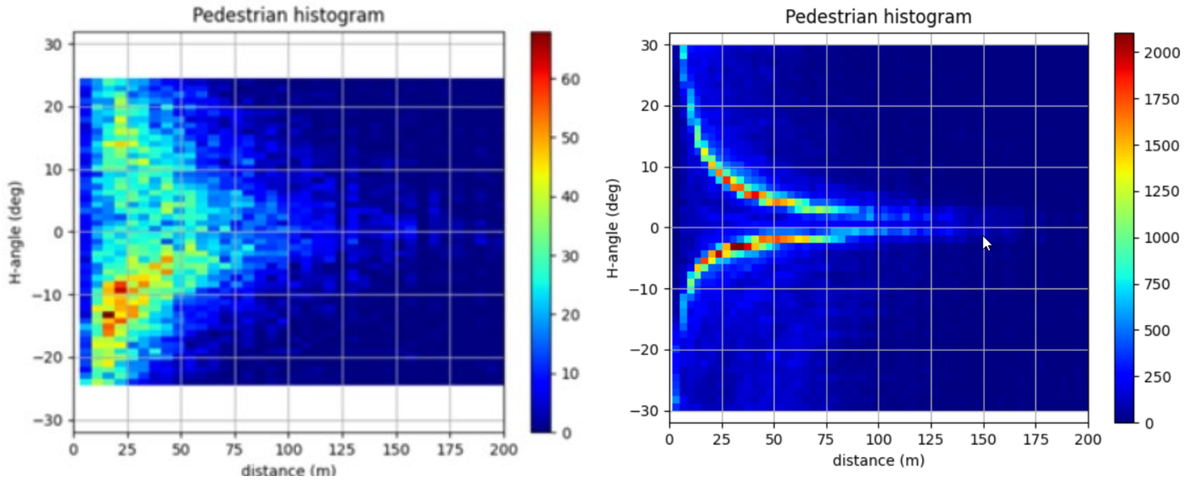
- Pedestrian assets were designed w.r.t. the TP4 ontology targeting on high coverage

- Synthetic data generation with high degree of realism and accuracy motivates measurement of key elements such as pedestrian motion and material characteristics

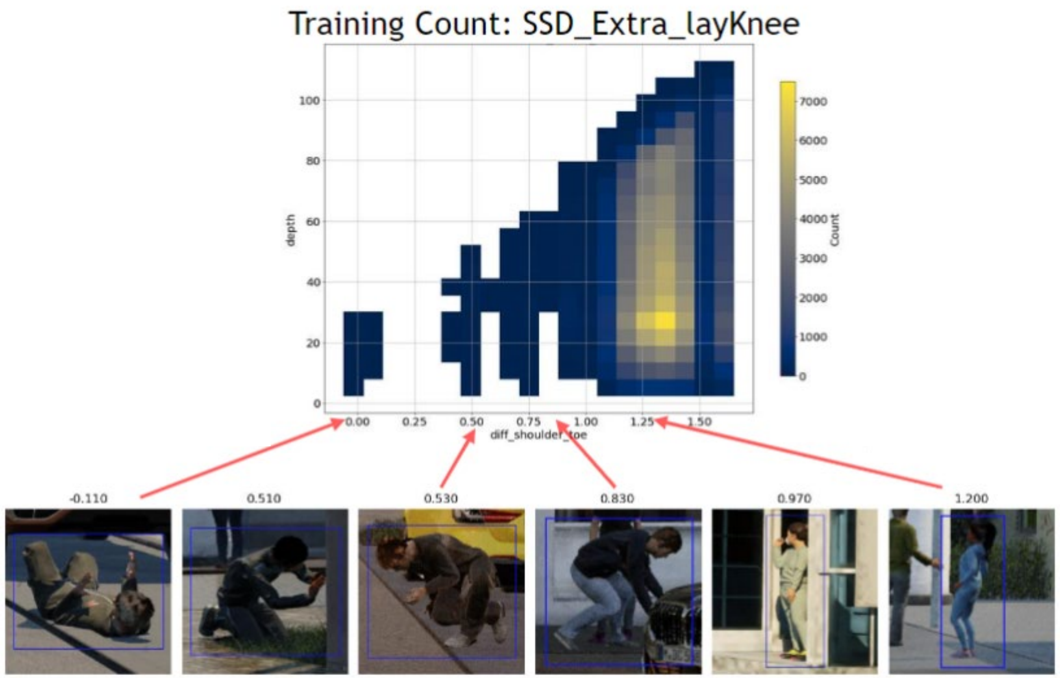


[Images: Mackevision]

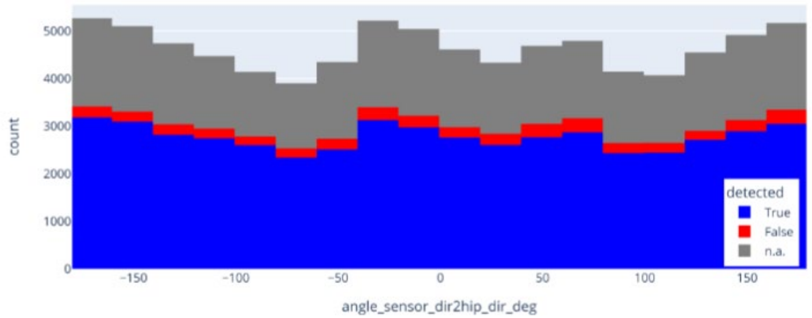
Data Quality Analysis Contributes to Evidence Workstreams on Data Coverage and Performance Limiting Factors, Examples



Evaluation of pedestrian distribution [Intel]



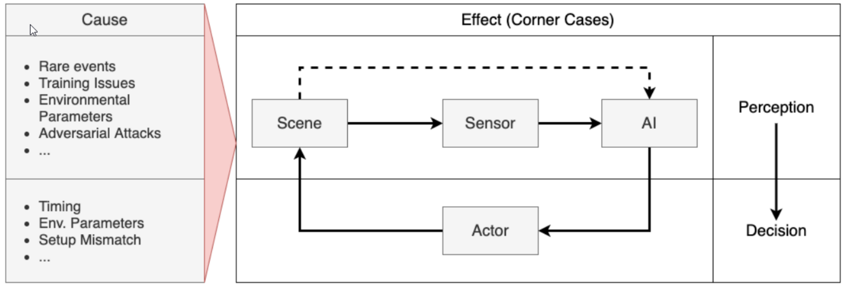
Analysis of pose coverage [Bosch]



Analysis of pedestrian orientation coverage [BMW/Exida]

Summary on Data Generation

- Two toolchains with distinct features developed
- 360.000 frames produced and provided to the project
- Broad contribution to evidence workstreams
- Corner case taxonomy developed and methods for corner case detection explored
- Effects of sensor parameter changes and domain adaptation approaches examined



Base structure for corner case taxonomy [QualityMinds]

Example corner case [QualityMinds/BIT TS]

KI ABSICHERUNG

Safe AI for Automated Driving

Dr. Thomas Stauner, BMW AG

Thomas.Stauner@bmw.de

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

Supported by:

Federal Ministry for Economic Affairs and Climate Action

on the basis of a decision by the German Bundestag

www.ki-absicherung.vdali.de @KI_Familie KI Familie

3

AI Function Development

SP 1: Fridolin Bauer, BMW AG

KI ABSICHERUNG

Safe AI for Automated Driving

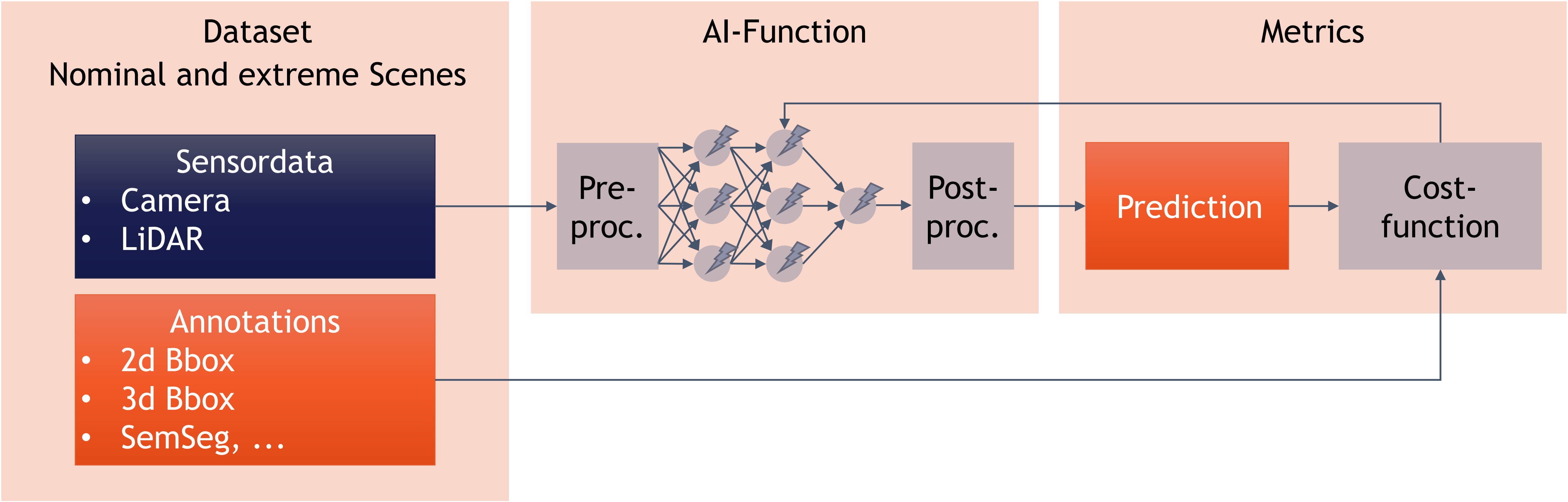
Final Event | June 23rd 2022

TP1: AI-Function

Fridolin Bauer, BMW AG

AI-function Specification

- Specification of synthetic data, AI-Function and metrics
- From DNN-developers perspective

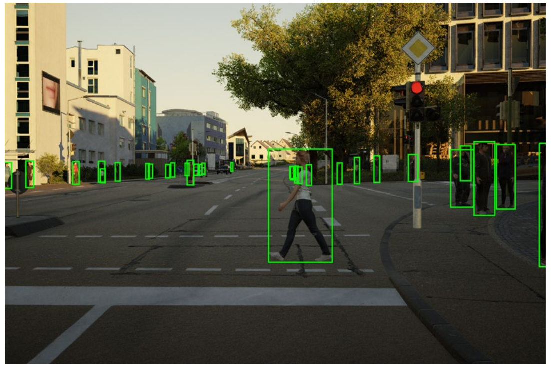


INTERNAL

AI-function Specification

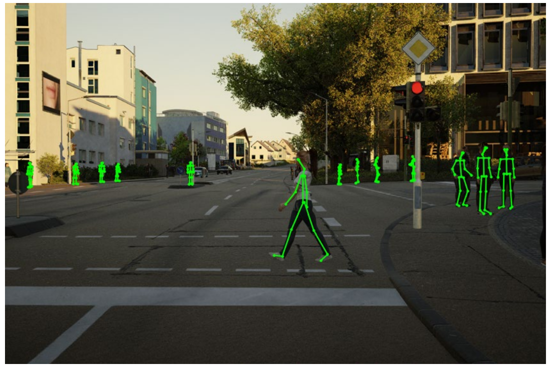
- Examples of data including specified annotation

Mackevision



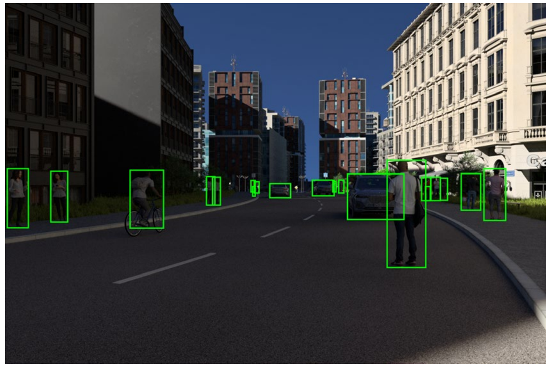
Camera image + 2d Detection

Semantic Segmentation

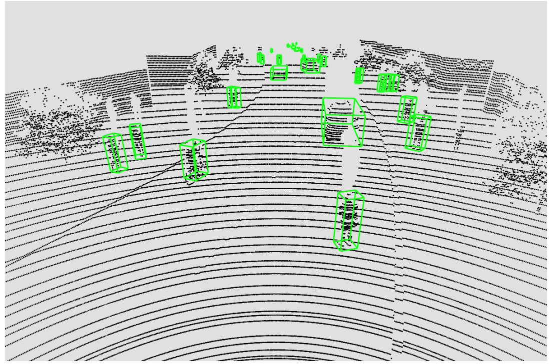


Skeleton- and Pose Data

BIT TS



Camera image + 2d Detection



LiDAR Data + 3d Detection

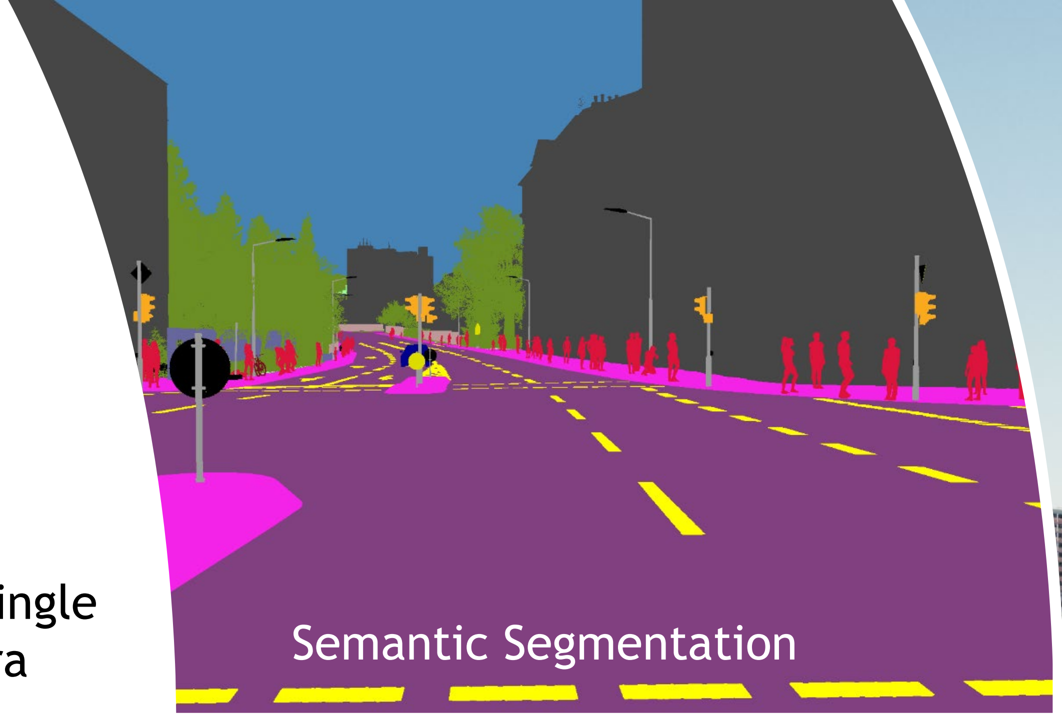
Bodypart Segmentation

Monocular Pedestrian Detection

Task: Detect pedestrians in a single frame from a monocular camera image

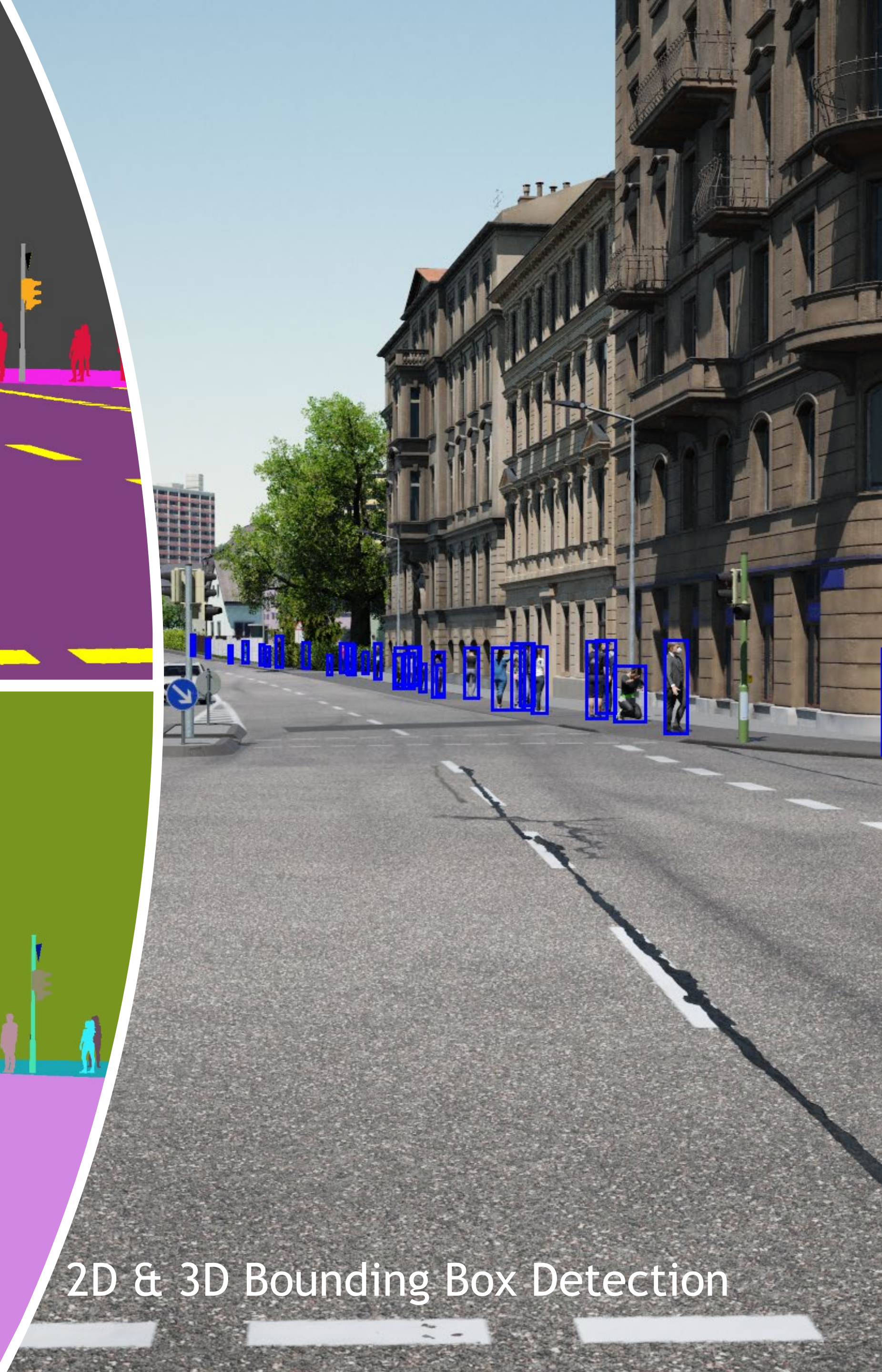
Implemented Algorithms:

- Single Shot Detector (2D-BB, Opel)
- DeeplabV3+ (Sem-Seg, Intel)
- DeeplabV3 (Sem-Seg, ZF)
- Detectron2 (Instance-Seg, ZF)
- Frustum-PointNets (3D-BB, Valeo)
- Single Shot Detector + pose & posture (2D-BB, HCI)



Semantic Segmentation

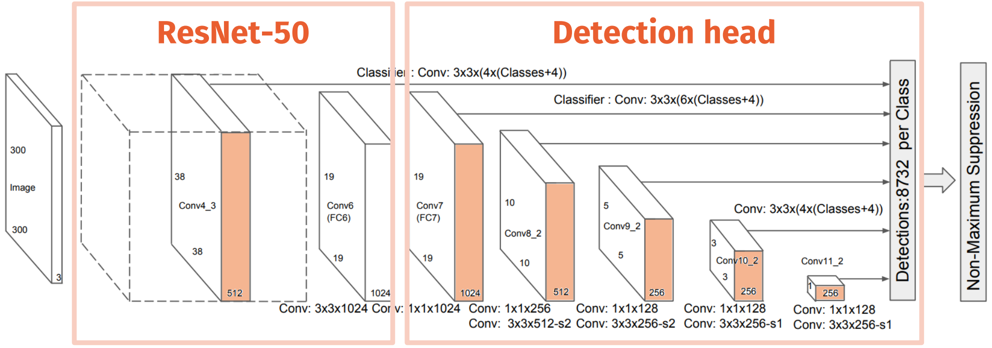
Instance Segmentation

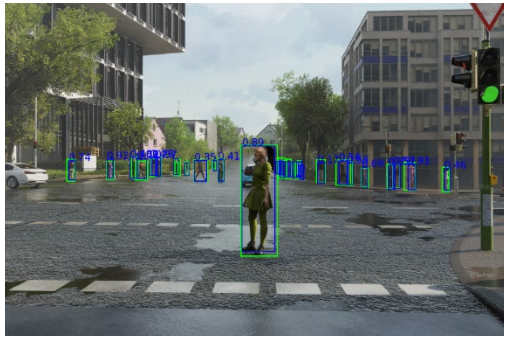


2D & 3D Bounding Box Detection

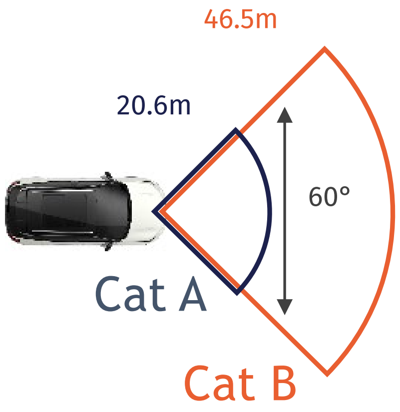
Safety Relevant Pedestrian Evaluation

Single Shot Detector:





Semantic Placement



Breaking Distance to Vehicle

Categorize Detections

Evaluation Filter	Precision =TP/(TP+FP)	Recall =TP/(TP+FN)
Non-difficult (Training)	+	+
Cat B	0 [1]	+
Cat A	– [1]	+ +

[1] FP too high, evaluation filter not applicable

Fusion at different levels

- Task: 3D Pedestrian Detection using LiDAR and Camera Data
- Demonstrated Fusion of Camera and LiDAR Data at different Levels

- Algorithms and Partners in the WP
  - Fusion at Sensor Level (Opel)
  - Fusion at Feature Level (BMW)
  - Fusion at Regression Level (ZF)
  - Single Modality Lidar (TUM)
  - Sequential Fusion (Valeo)
  - Fusion at Temporal Level (DFKI)

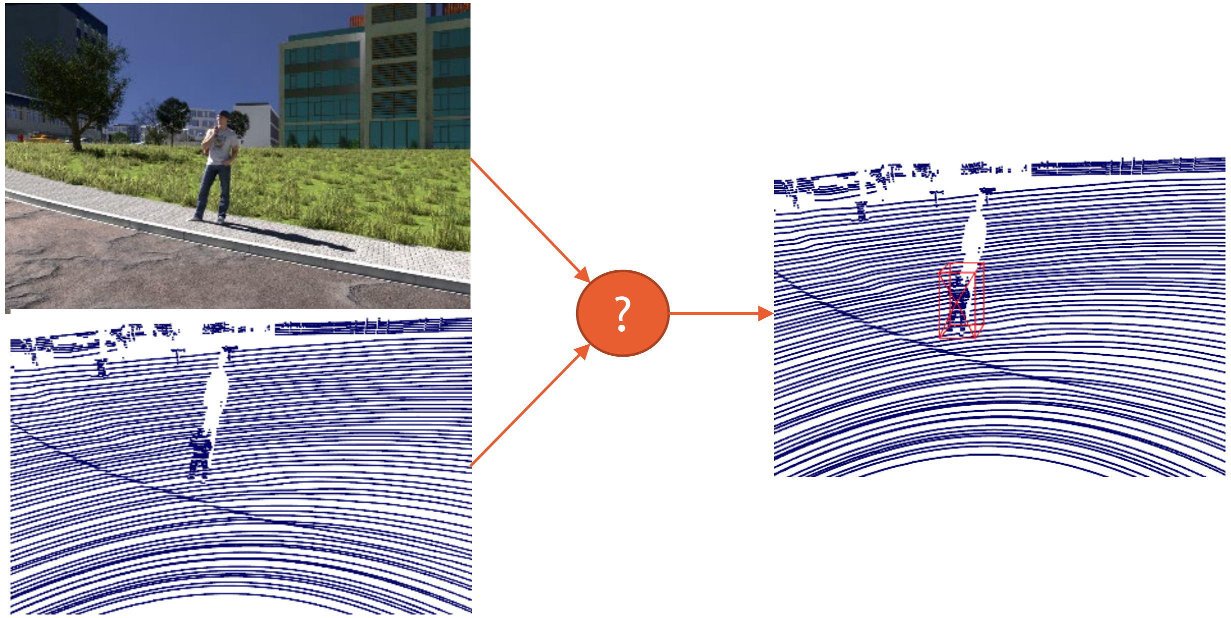
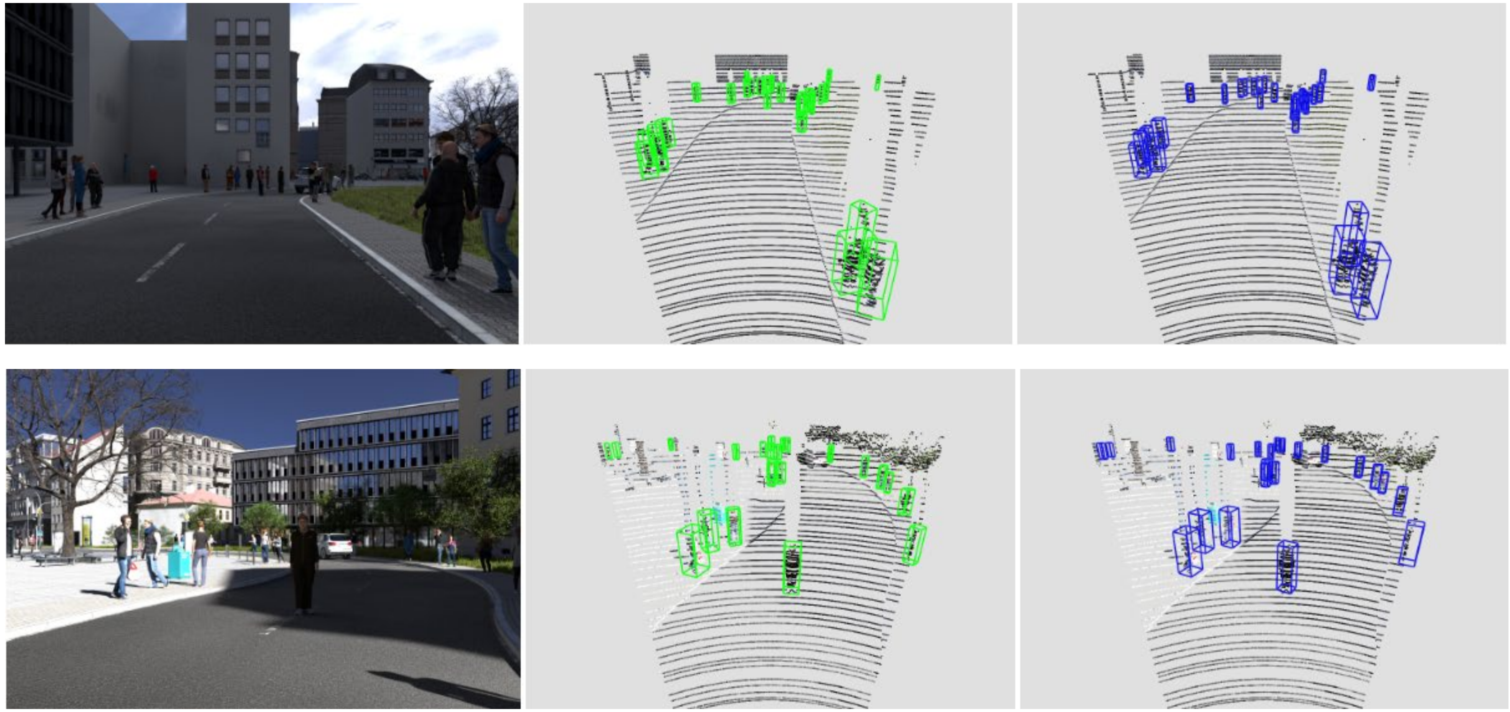


Fig. Task: How to fuse Camera and LiDAR data for 3D Pedestrian detection in LiDAR or Camera space

Fusion at Sensor Level

- Fusion of Camera and LiDAR data at Sensor level
- Fusion: Extended PointPillars by appending LiDAR pointcloud with RGB values from camera



Camera Images | Ground-truth | Predictions

INTERNAL

Fusion at Temporal Level

- Developed LRPD (Long Range Pedestrain Detection) algorithm for mid and long range detection
- Developed a two-step Temporal Fusion algorithm using Particle Filter and Faster-RCNN

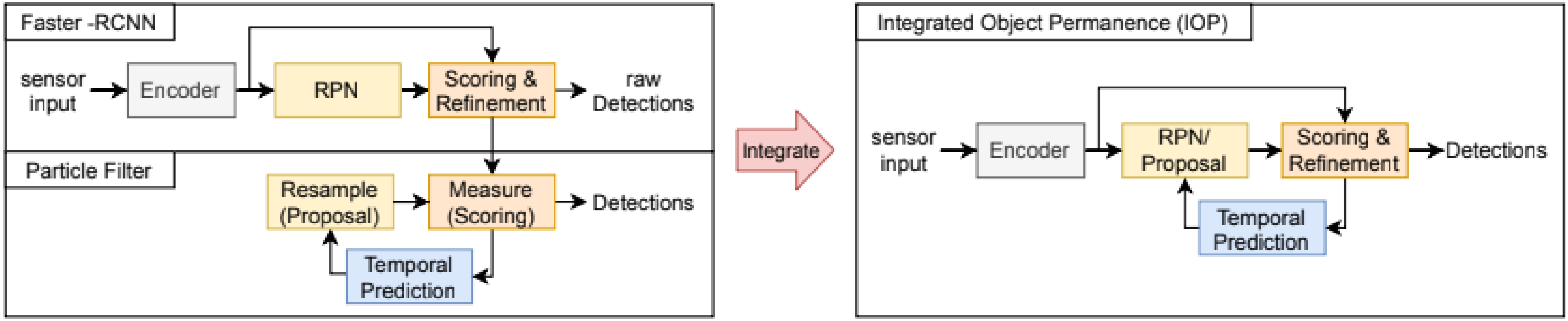


Fig: Combination of two approaches into one interated architecture

Fusion at Temporal Level

- Integrated Object Permanence into Faster-RCNN object detector

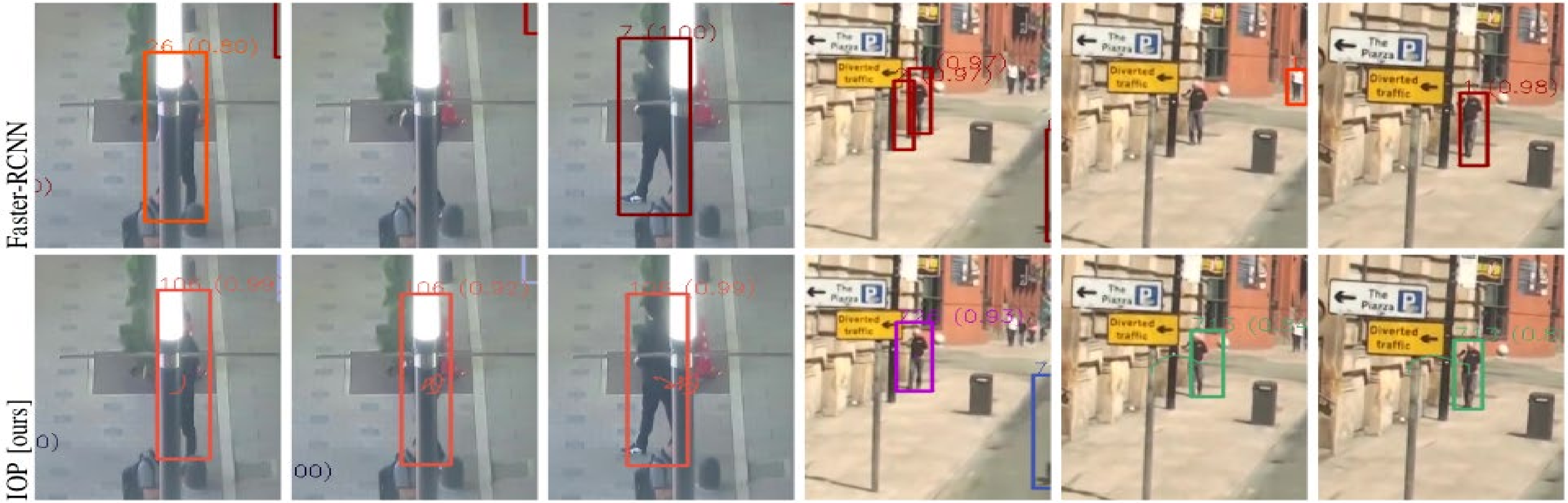


Fig: Comparison: Faster-RCNN and IOP from E1.4.6

Human Pose Estimation

Supervised Human Pose Estimation

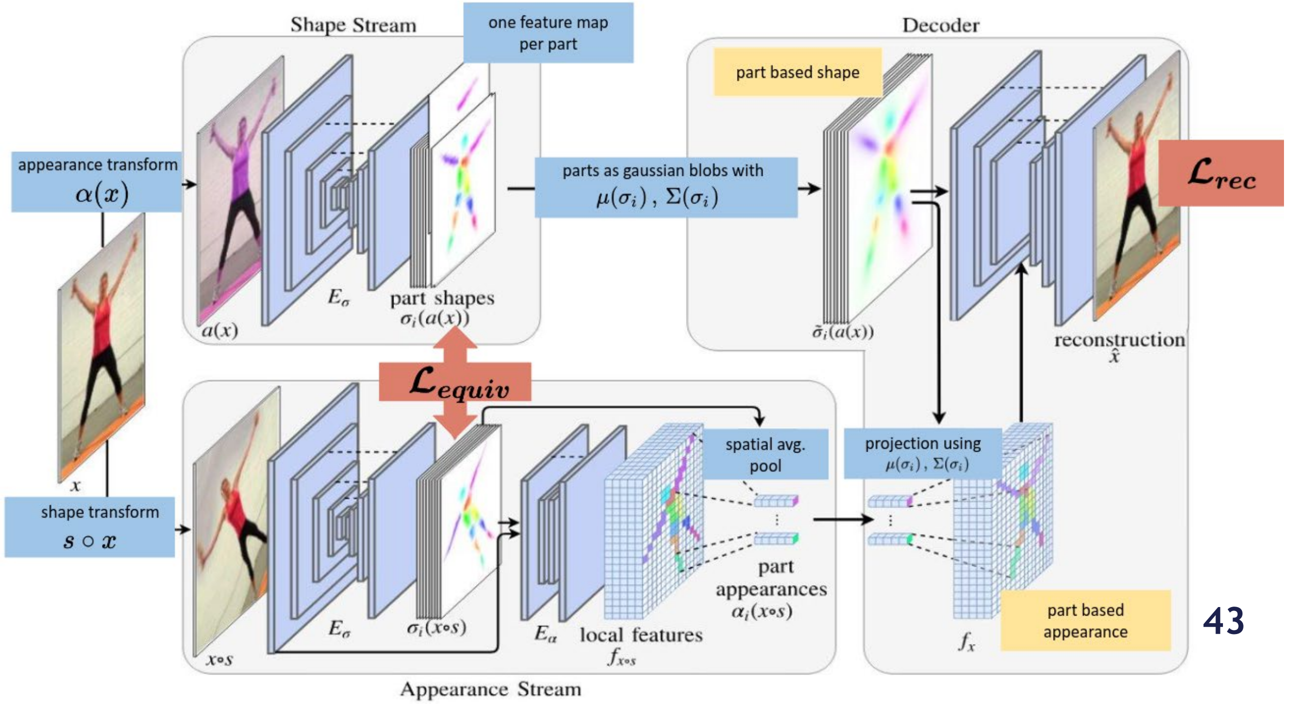
- Far away pedestrians tiny
- Superresolution required
- Hybrid Top-Down/Bottom-Up Approach

Unsupervised Human Pose Estimation

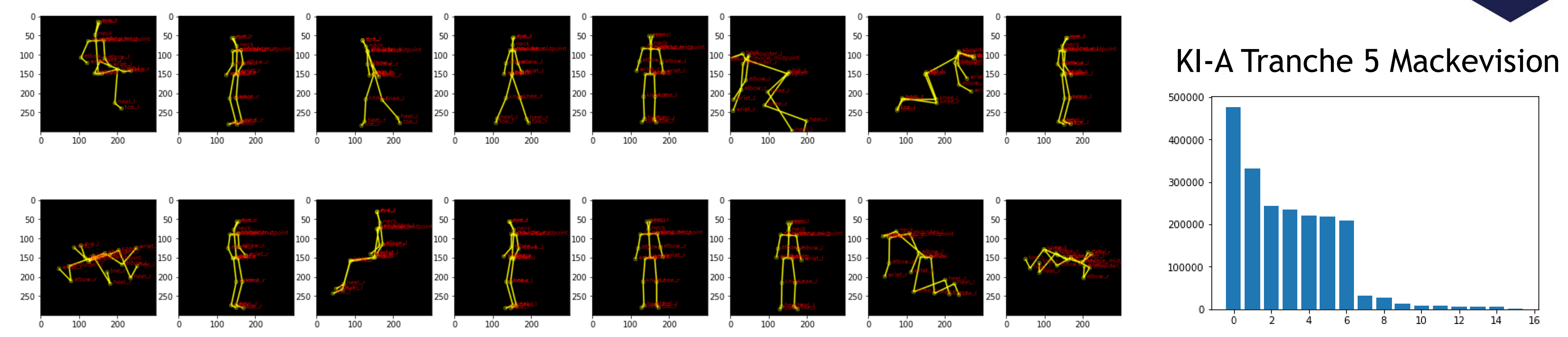
- No Labels
- Geometric Equivariance Loss
- Appearance Invariance Loss

Sensorfusion for Robust Human Pose Estimation

- Depth Ambiguity
- -> "Parallele Highlight Vorträge"



16 most common human poses per dataset



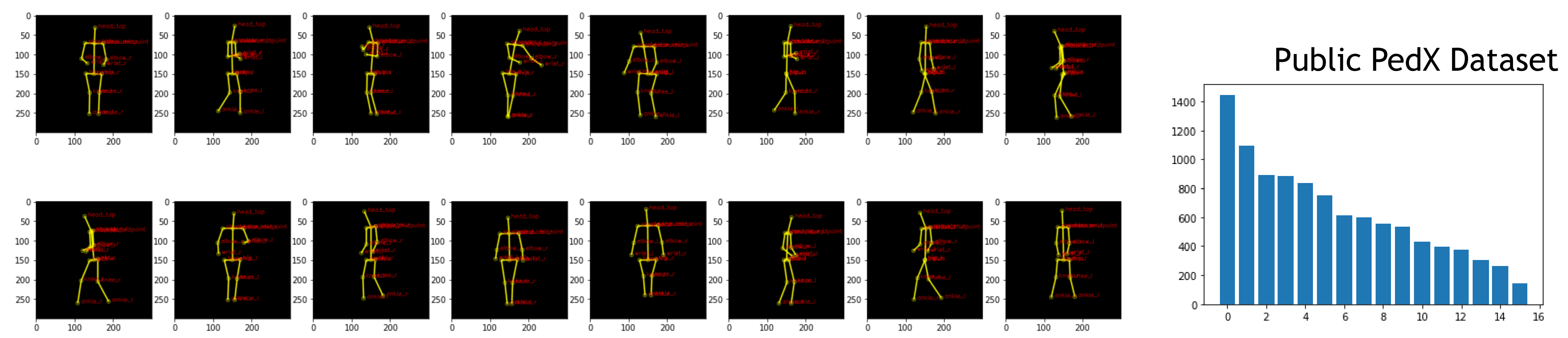


Fig: Distribution of different human poses on different data sets

KI
ABSICHERUNG
Safe AI for Automated Driving

Fridolin Bauer, BMW AG
Fridolin.Bauer@bmwgroup.com

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

www.ki-absicherung.vdali.de @KI_Familie KI Familie

4

Safety Measures & Metrics

SP 3: Fabian Hüger, Volkswagen AG

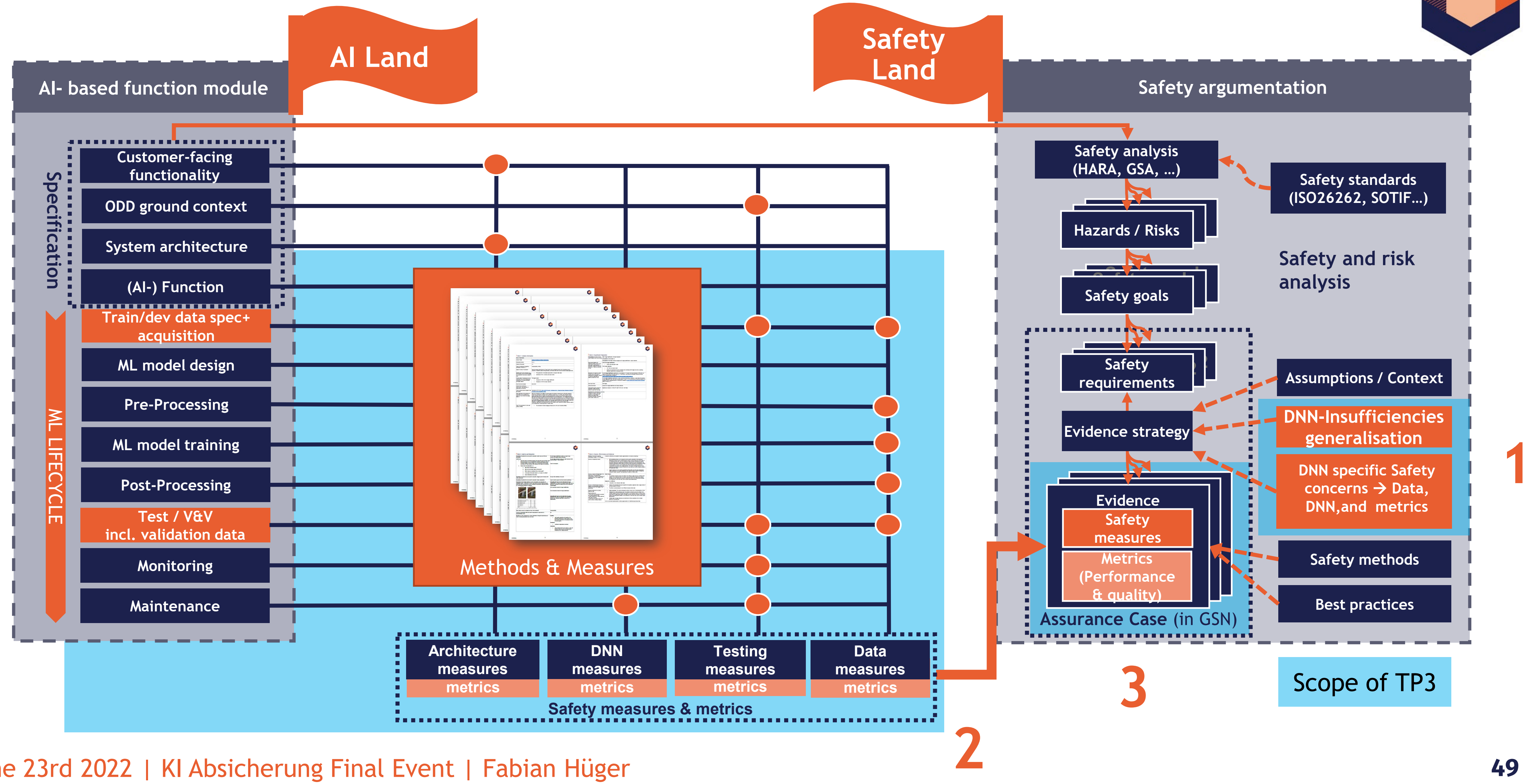
KI ABSICHERUNG
Safe AI for Automated Driving

Final Event | June 23rd 2022

Methods and measures for the verification of the AI function

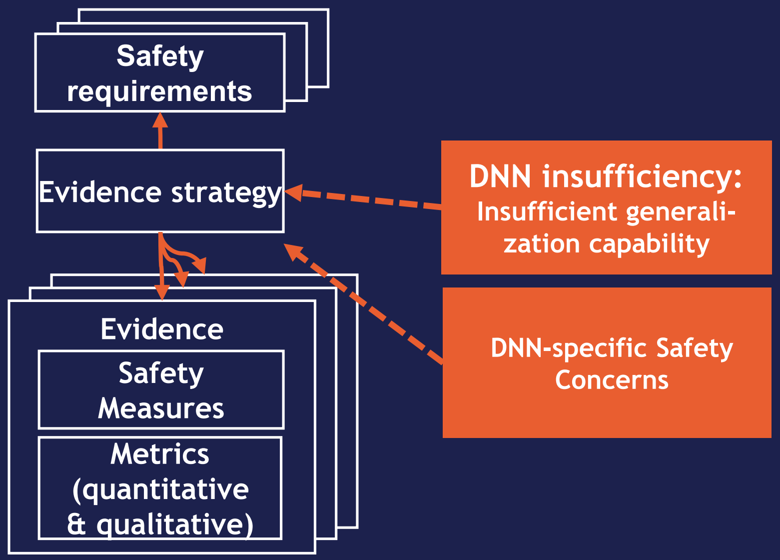
Dr. Fabian Hüger, Volkswagen AG

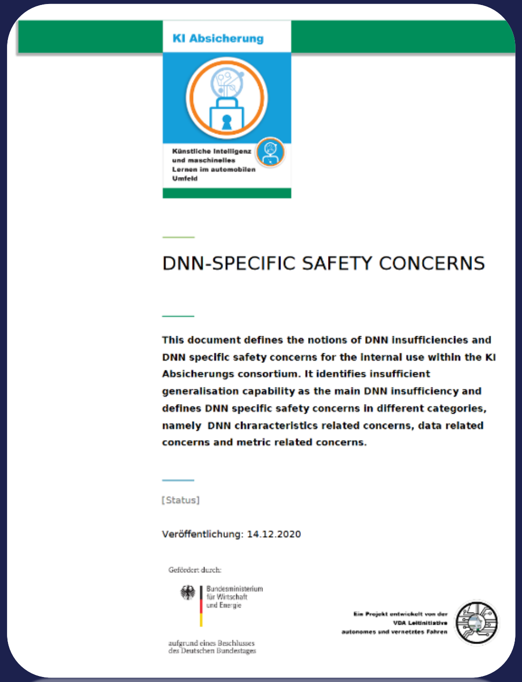
Methods and Measures in context of the KI Absicherung Big Picture



DNN-specific Safety Concerns

1




KI Absicherung

Künstliche Intelligenz und maschinelles Lernen im automobilen Umfeld

DNN-SPECIFIC SAFETY CONCERNS

This document defines the notions of DNN insufficiencies and DNN specific safety concerns for the internal use within the KI Absicherungs consortium. It identifies insufficient generalisation capability as the main DNN insufficiency and defines DNN specific safety concerns in different categories, namely DNN chraracteristics related concerns, data related concerns and metric related concerns.

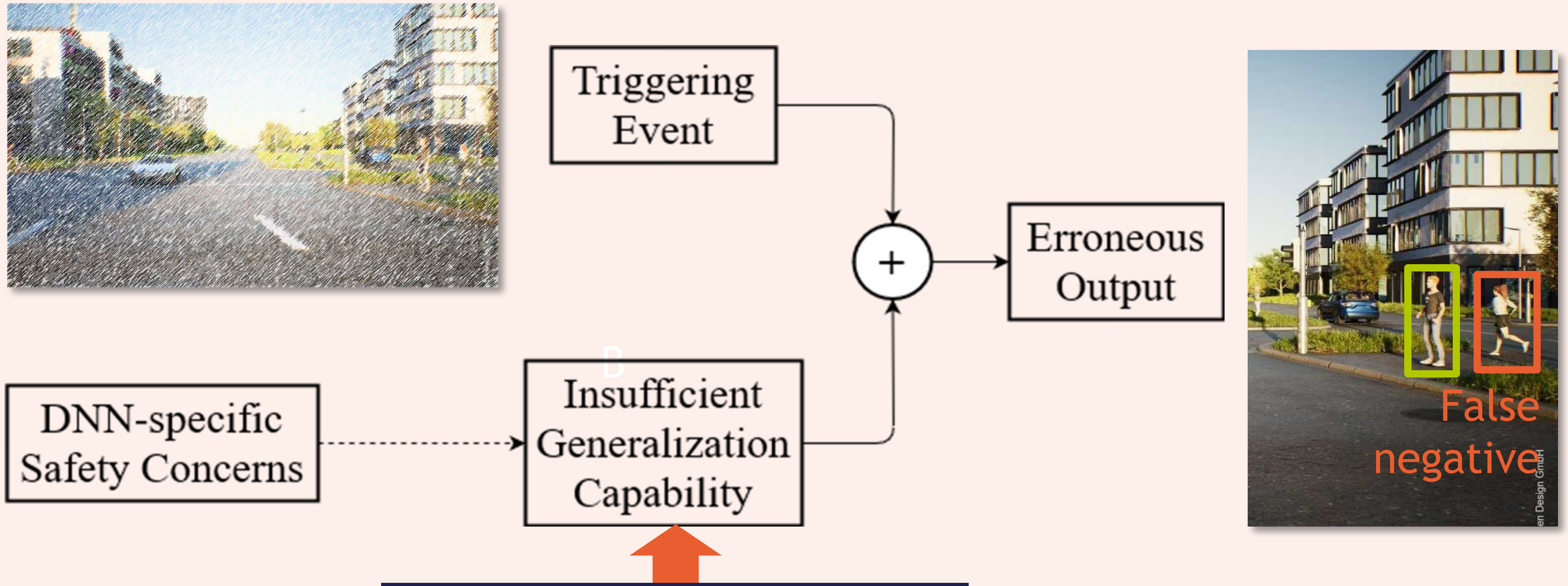
[Status]

Veröffentlichung: 14.12.2020

DNN-specific Safety Concerns (1/2)

We define **DNN-specific Safety Concerns (SCs)** as underlying issues of DNN-based perception which may negatively affect the safety of a system.

e.g., fog, snow, camera issues



Functional Insufficiency according to ISO/PAS 21448

Willers O, Sudholt S, Raafatnia S, Abrecht S (2020) Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks. SAFECOMP 2020 workshop: WAISE 2020

FI-1	**INSUFFICIENT GENERALIZATION CAPABILITY**	Wrong outputs by an AI-based function that was trained on a limited database. Erroneous input to output mapping or wrong approximation.
SC-1.1	**UNRELIABLE CONFIDENCE INFORMATION**	DNNs tend to be overconfident in their predictions under certain conditions or in general outputting unreliable confidence information.
SC-1.2	**BRITTLENESS OF DNNS**	Non-robustness against common perturbations such as noise or certain weather conditions as well as targeted perturbations known as adversarial examples
SC-1.2.1	**LACK OF TEMPORAL STABILITY**	Detection results rapidly changing in time whereas little change occurs in the ground truth
SC-1.3	**INCOMPREHENSIBLE BEHAVIOUR**	Inability to explain exactly how DNNs come to a decision.
SC-1.4	**INSUFFICIENT PLAUSIBILITY**	AI based functions usually lack basic plausibility checks, which are intended to identify detections of the perception function that violate physical laws.
SC-2.1	**DATA DISTRIBUTION IS NOT A GOOD APPROXIMATION OF REAL WORLD**	The distribution of data used in the development should be a valid approximation of the ODD in the real world.
SC-2.2	**INADEQUATE SEPARATION OF TEST AND TRAINING DATA**	Test data might be correlated to training data which might induce overfitting on test data.
SC-2.3	**DEPENDENCE ON LABELLING QUALITY**	Labelling quality can directly affect the resulting model performance. Moreover, due to missing labelling quality, evaluation results might be misleading.
SC-2.3.1	**MISSING LABEL DETAILS OR META-LABELS**	Missing meta-labels or label details possibly leads to improper data selection or insufficient training objectives.
SC-2.4	**SPECIFICATION OF THE ODD**	An incomplete or incorrect ODD specification leads to incomplete data records for training and testing.
SC-2.5	**DISTRIBUTIONAL SHIFT OVER TIME**	A DNN is trained and tested at a certain point in time. Changes will occur naturally and therefore can potentially harm the performance of DNNs.
SC-2.6	**UNKNOWN BEHAVIOUR IN RARE CRITICAL SITUATIONS**	The long tail problem describes the fact that there exists an enormous amount of possibly safety-critical street scenes that have a low occurrence probability.
SC-3.1	**SAFETY-AWARE METRICS**	Some state-of-the-art metrics only evaluate the average performance of DNNs. Safety-aware metrics are required to sophistically evaluate the performance of DNNs.

Based on:

O.Willers, S. Sudholt, S. Raafatnia, S. Abrecht: Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks

T. Sämann, P.Schlicht, F. Hüger: Strategy to Increase the Safety of a DNN-based Perception for HAD Systems

G. Schwalbe, B. Knie, T. Sämann, T. Dobberphul, L. Gauerhof, S., V. Rocco: **Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications**

Functional Insufficiencies

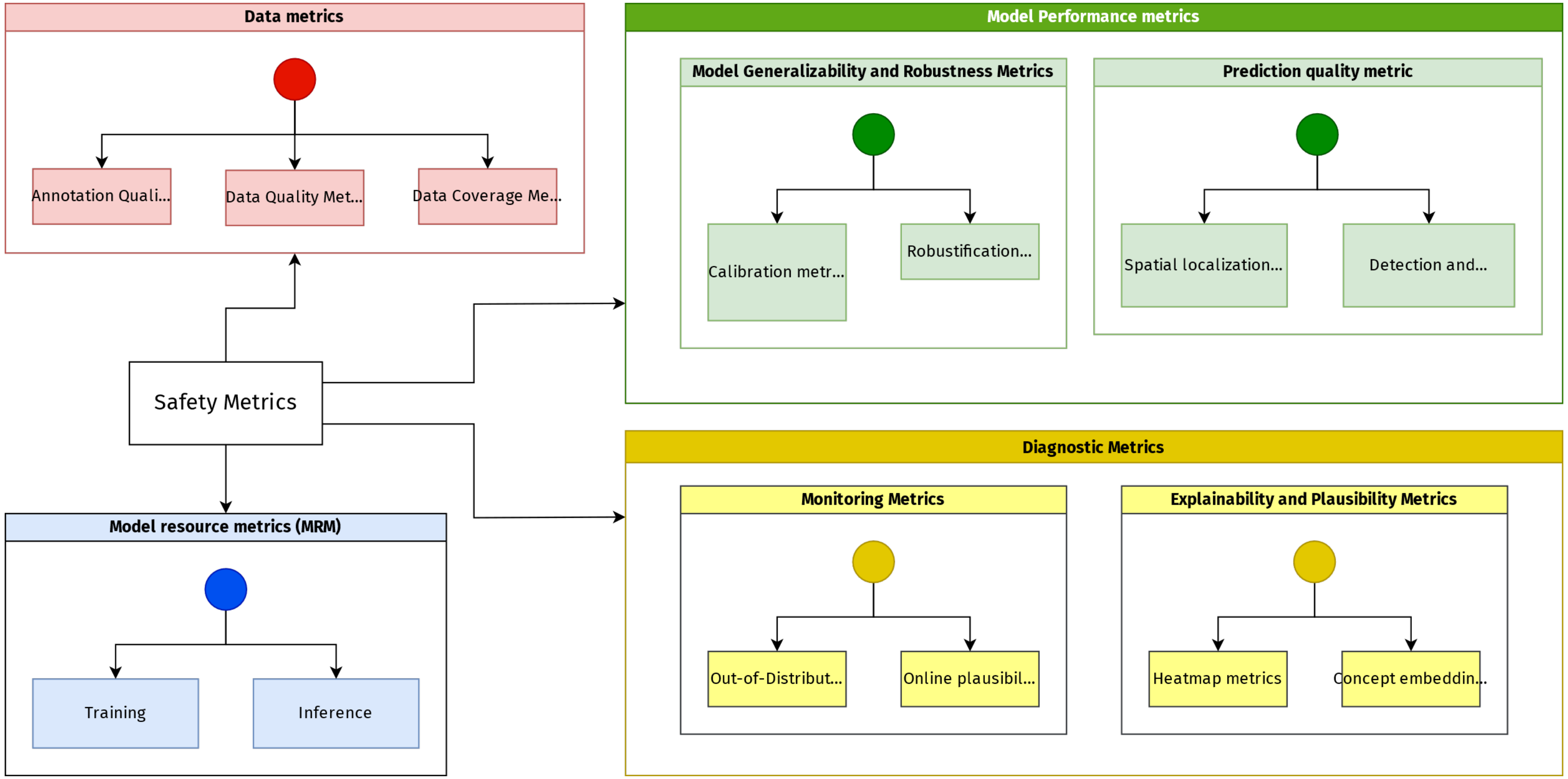
DNN-characteristics-related concerns

Data-related concerns

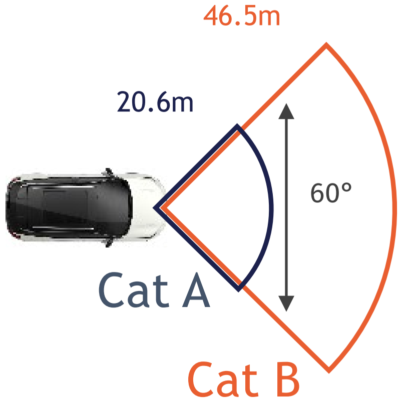
Metric-related concerns

DNN-specific Safety Concerns (2/2)

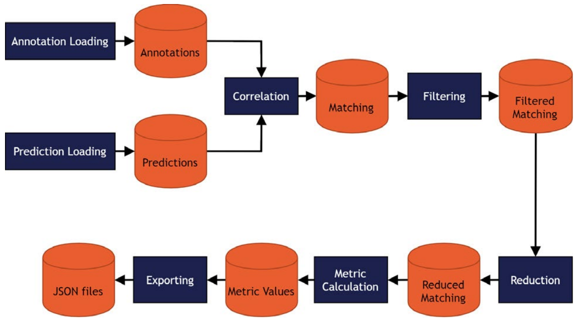
Safety Metrics



Metric Taxonomy & Catalogue



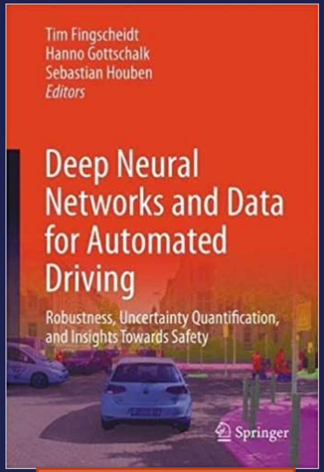
Safety Relevant Pedestrian



Metric Tool

2

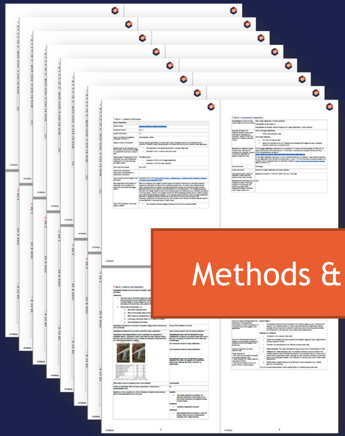
Methods and Measures



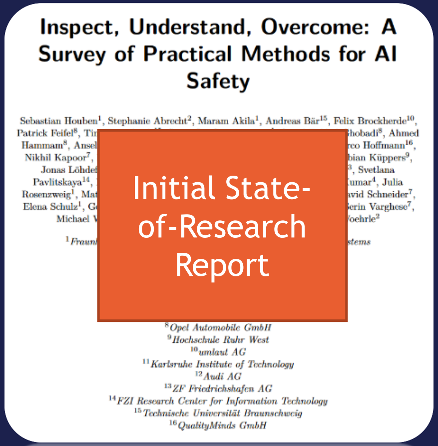
Springer Book

Architecture Measures	DNN Measures	Testing Measures	Data Measures
Metrics	Metrics	Metrics	Metrics

Safety Measures & Metrics

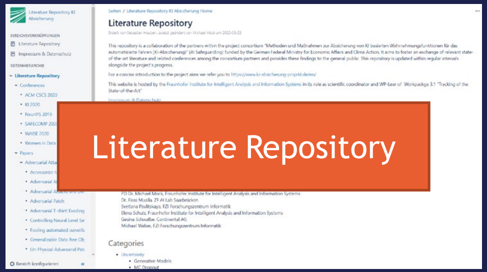


Methods & Measures



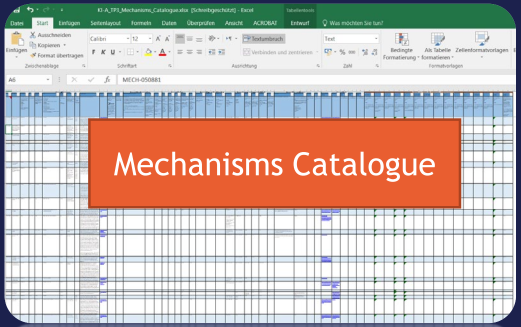
Initial State-of-Research Report

Survey available at
www.ki-absicherung-projekt.de/



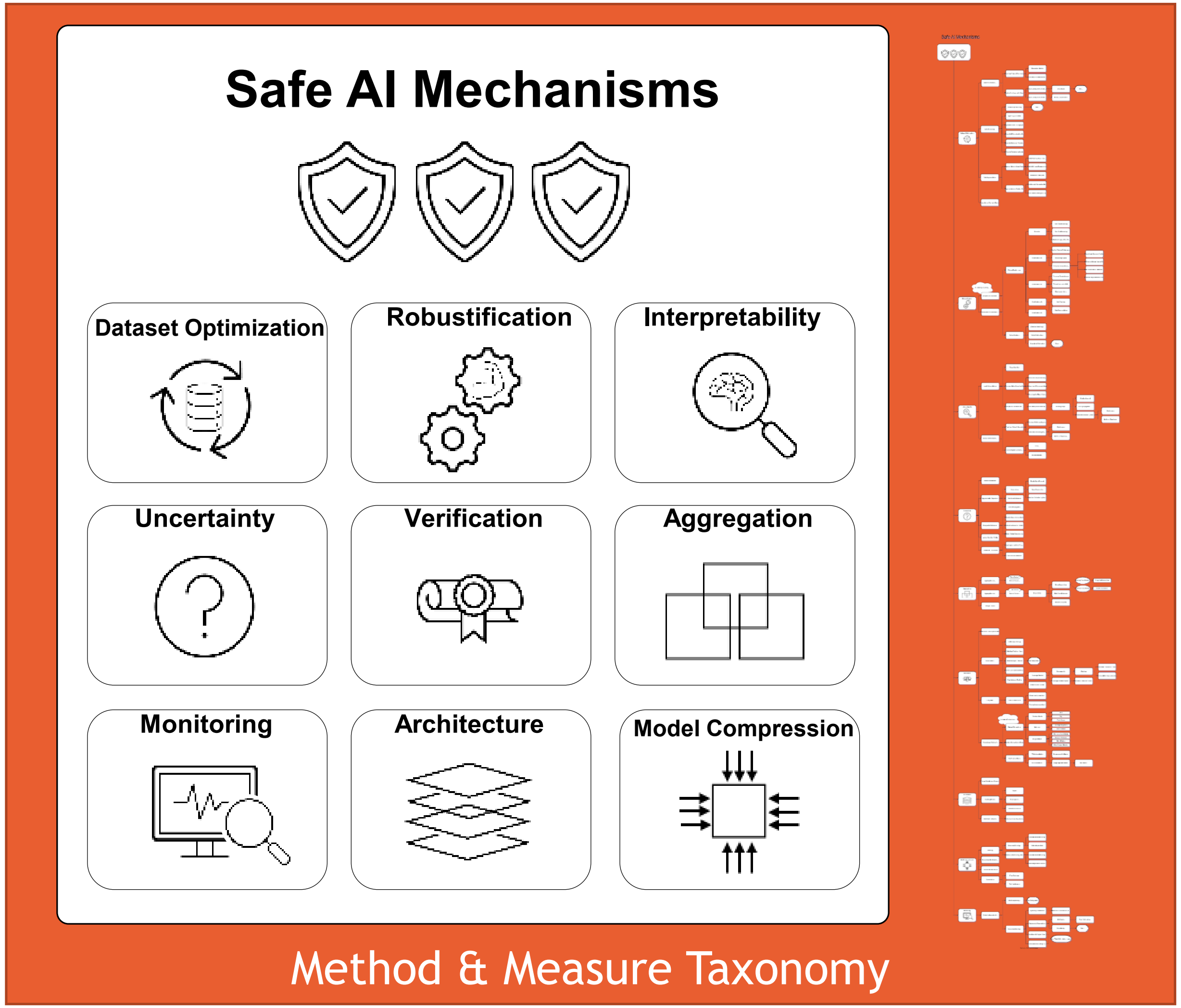
Literature Repository

Literature Repository available at:
tinyurl.com/e3y4pmxs

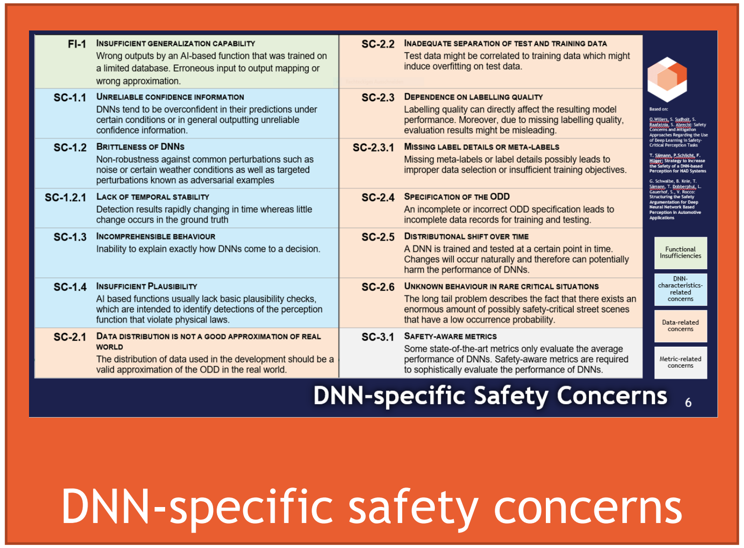


Mechanisms Catalogue

Safe AI Mechanisms addressing the DNN-specific Safety Concerns



addressing



FI-1	**Insufficient generalization capability** Wrong outputs by an AI-based function that was trained on a limited database. Erroneous input to output mapping or wrong approximation.	SC-2.2	**Inadequate separation of test and training data** Test data might be correlated to training data which might induce overfitting on test data.
SC-1.1	**Unreliable confidence information** DNNs tend to be overconfident in their predictions under certain conditions or in general outputting unreliable confidence information.	SC-2.3	**Dependence on labelling quality** Labelling quality can directly affect the resulting model performance. Moreover, due to missing labelling quality, evaluation results might be misleading.
SC-1.2	**Brittleness of DNNs** Non-robustness against common perturbations such as noise or certain weather conditions as well as targeted perturbations known as adversarial examples	SC-2.3.1	**Missing label details or meta-labels** Missing meta-labels or label details possibly leads to improper data selection or insufficient training objectives.
SC-1.2.1	**Lack of temporal stability** Detection results rapidly changing in time whereas little change occurs in the ground truth	SC-2.4	**Specification of the ODD** An incomplete or incorrect ODD specification leads to incomplete data records for training and testing.
SC-1.3	**Incomprehensible behaviour** Inability to explain exactly how DNNs come to a decision.	SC-2.5	**Distributional shift over time** A DNN is trained and tested at a certain point in time. Changes will occur naturally and therefore can potentially harm the performance of DNNs.
SC-1.4	**Insufficient Plausibility** AI based functions usually lack basic plausibility checks, which are intended to identify detections of the perception function that violate physical laws.	SC-2.6	**Unknown behaviour in rare critical situations** The long tail problem describes the fact that there exists an enormous amount of possibly safety-critical street scenes that have a low occurrence probability.
SC-2.1	**Data distribution is not a good approximation of real world** The distribution of data used in the development should be a valid approximation of the ODD in the real world.	SC-3.1	**Safety-aware metrics** Some state-of-the-art metrics only evaluate the average performance of DNNs. Safety-aware metrics are required to sophistically evaluate the performance of DNNs.

Based on:

O. Willers, S. Sudholt, S. Raafatnia, S. Abrecht: Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks

T. Sämann, P. Schlicht, F. Hüger: Strategy to Increase the Safety of a DNN-based Perception for HAD Systems

G. Schwalbe, B. Knie, T. Sämann, T. Dobberphul, L. Gauerhof, S., V. Rocco: Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications

Functional Insufficiencies
DNN-characteristics-related concerns
Data-related concerns
Metric-related concerns

DNN-specific Safety Concerns 6

DNN-specific safety concerns

Mechanism Descriptions

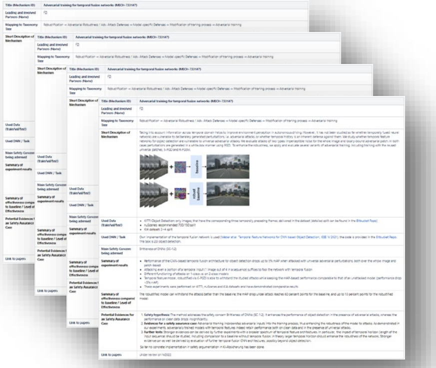
1 Block 1: General Information

2 Block 2: Experiment Preparation

3 Block 3: Metrics and Evaluation

4 Block 4: Results, Effectiveness and Evidences

+ 1-Page-Summaries in public project report (appendix)



Mechanism Catalog

Section 1: General Info			Section 2: Safety Assurance Case	Section 9: Mechanism Rating by Developer						
Mechanism Name	Cluster	Short description	Evidences for the Safety Assurance Case	Main Safety Concern being adressed	Estimated Time to Series Production	Level of effectiveness	Performance Degradation compared to baseline	Changes to DNN architecture	Additional computational overhead at inference time	Additional computational overhead at training time
Confidence Calibration for Object Detection	Uncertainty	The Detection Expected Calibration Error (D-ECE) measures the deviation between average confidence and observed accuracy by means of the object's position/scale. Additionally, there are several methods to post-process the confidence estimates of a network in order to obtain a better match (calibration) of the confidence and the observed accuracy. We propose an extension of common methods to perform a calibration that also takes the position/scale of an object into account.	This mechanism shows miscalibration of DNNs and helps to recalibrate DNNs in a post-hoc step. This is useful to elaborate calibration and thus statistical evidence of DNNs output prediction scores.	Unreliable confidence information (SC-1.1)	1-2 years (some improvements needed)	High	0: equal performance	No changes	Very low	Medium
Aggregation based dependency analysis of neural networks with Visual Analytics	Explainability	The overall goal of the mechanism is to address the problem of DNN insufficient generalisation capability by understanding semantic concepts of the data. Insufficiencies in DNN predictions on the one hand might stem from independent weaknesses (due to stochastic training), but on the other hand might stem from systematic weaknesses like learned shortcuts or flaws in the data. Finding such correlated insufficiencies and identifying and distinguishing outliers from systematic weaknesses leads to gaining insights into the decision of networks. This can be achieved by understanding the semantic concepts of the data. As an automated analysis of semantics is difficult, we are making use of the human tacit and expert knowledge to examine the semantic features visually. We propose to support and guide the human expert within the analyzation process by methods of Visual Analytics to enable a stringent safety argumentation that can be built upon human understandable arguments.	The mechanism most likely contributes to the interpretability of DNNs. The interactive visual analysis makes it possible to conduct a semantic analysis of the DNN predictions w.r.t. meta data and therefore gain insights into the decisions of networks. The iterative analyzation process can lead to a feedback loop between data generation and meta data generation, DNN development and training and metric/mechanism development. All in all, a stringent safety argumentation could be build upon arguments that are understandable by humans The evidence therefore would be something like "no systematic weaknesses found after evaluation by X safety experts". This scenario was depicted in the mini-GSN developed during the evidence workshop of this mechanism.	Incomprehensible behavior (SC-1.3)	1-2 years (some improvements needed)	Medium effect	N/A: cannot compare VA Tool to baseline model	No changes	N/A: cannot compare VA Tool to baseline model	N/A: cannot compare VA Tool to baseline model
Robustness Testing Framework	Robustness Testing	A black box model can be tested on its robustness to a variety of data augmentations and transferred adversarial attacks via this method. This includes: Augmentations like colour jitter, noise, croping, resizing, transferred black box adversarial attacks, pixel blurring, pixel masking, class-specific augmentations etc. Evaluating different networks, both provided by TP1 and open source implementations, on the robustness against adversarial attacks and different data augmentation techniques. Visualization of attacks and responses of the network. Modular, easily extendable software architecture. Mature experiment parameter configuration setup using hydra (https://hydra.cc/). This mechanism does not support training of the model, but does supports its evaluation.	This method addresses the safety concern "Brittleness of DNNs" (SC-1.2). It provides a platform to test the performance of DNNs against corruptions and check their robustness as compared to clean unperturbed data setting. The performance drop between unperturbed and perturbed dataset is slightly less in the robustified VW model as compared to baseline Opel model which does not include any kind of robustification method. Thus this evaluation framework identifies the level of brittleness in DNNs. Further evidences can be derived by identifying the scenes in dataset where the perturbations are negatively affecting the performance.	Brittleness of DNNs (SC-1.2)	< 1 year (slight improvements needed)	High	0: equal performance	No changes	Very low	Low

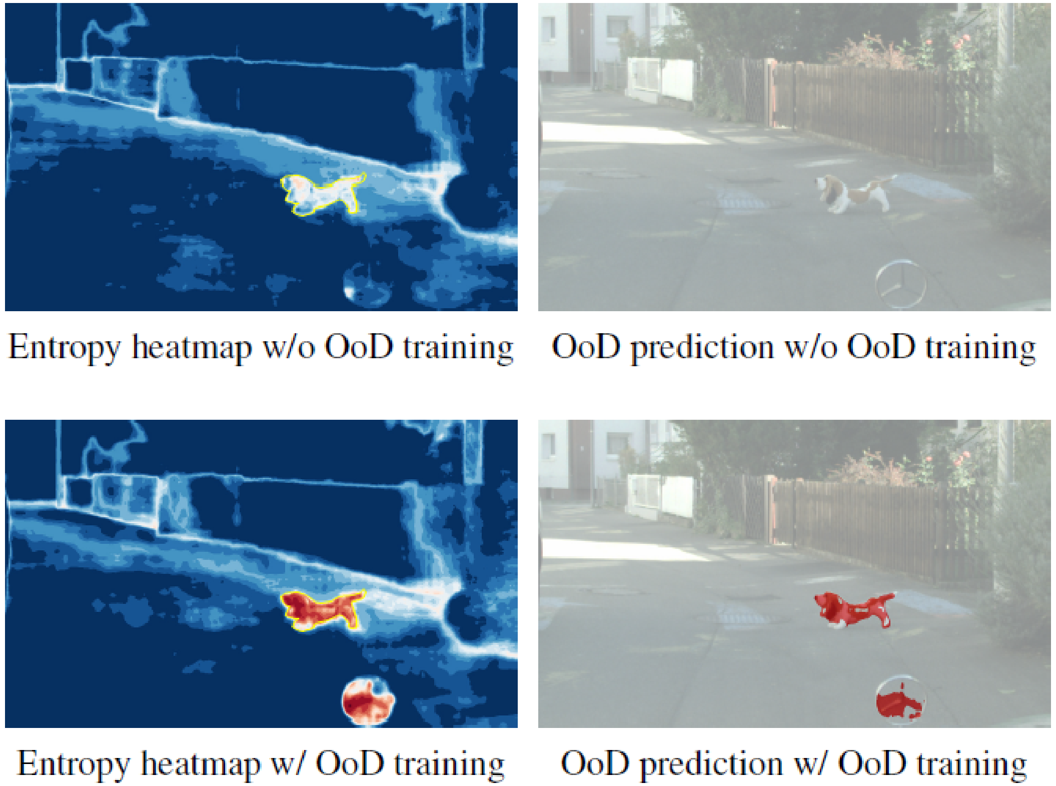
TP3 Safe AI Mechanism Catalog (excerpt, shortened, mechanisms chosen randomly)

Addressed Safety Concerns: Unreliable confidence information

Entropy Maximization and Meta Classification for Out-of-Distribution Detection in Semantic Segmentation

Enforce segmentation networks to output high prediction uncertainty on **Out-of-Distribution inputs** by means of a modified loss function [BUW]

Figure 2: Comparison of softmax entropy heatmap and OoD prediction mask with our OoD training (*top row*) and without (*bottom row*). The yellow lines in the entropy heatmaps mark the annotation of the OoD object. The OoD object prediction is obtained by simply thresholding on the entropy heatmap (in this example at $t = 0.7$ yielding the red pixels in the OoD prediction masks).

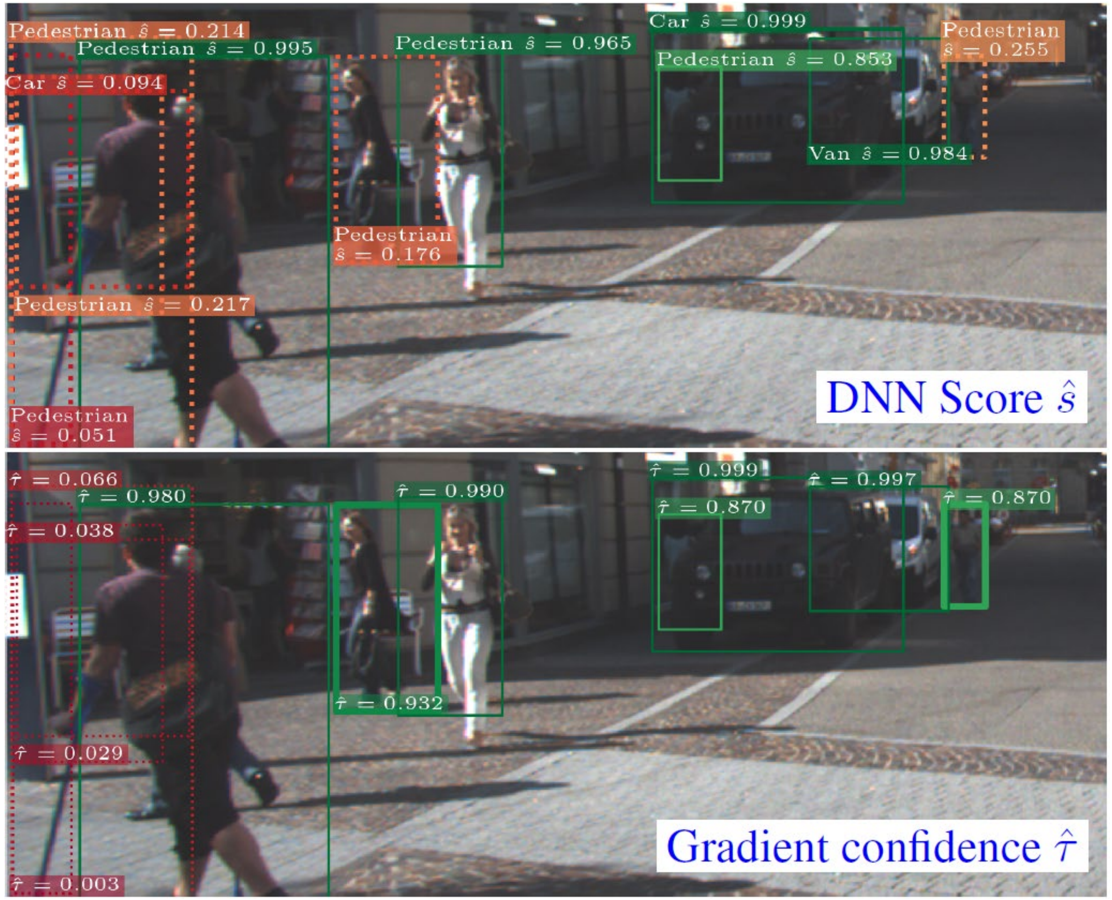


Chan et al., Entropy Maximization and Meta Classification for Out-Of-Distribution Detection in Semantic Segmentation, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5128-5137

Chan, R., Uhlemeyer, S., Rottmann, M., Gottschalk, H. (2022). Detecting and Learning the Unknown in Semantic Segmentation. In: Fingscheidt, T., Gottschalk, H., Houben, S. (eds) Deep Neural Networks and Data for Automated Driving. Springer, Cham. https://doi.org/10.1007/978-3-031-01233-4_10

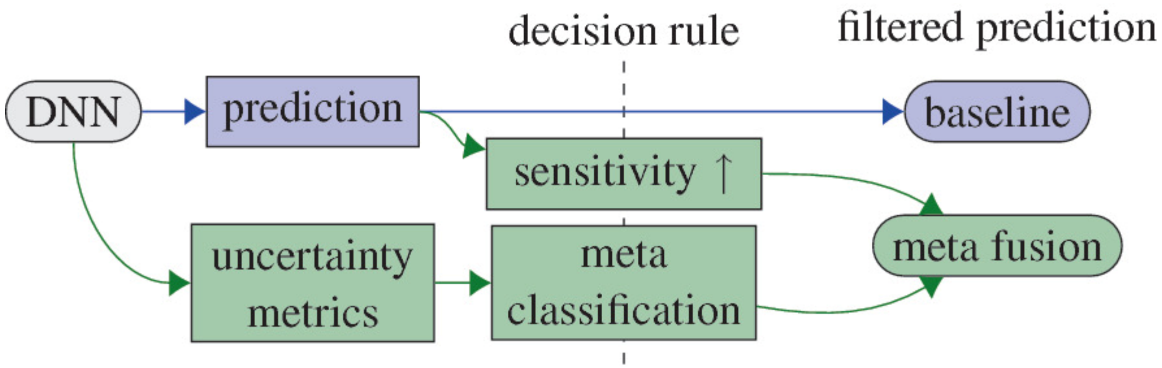
Object Detection Uncertainty based on Gradient Information

Addressed Safety Concerns: Unreliable confidence information



T. Riedlinger et al., Gradient-Based Quantification of Epistemic Uncertainty for Deep Object Detectors,arXiv preprint arXiv:2107.04517v1, 2021

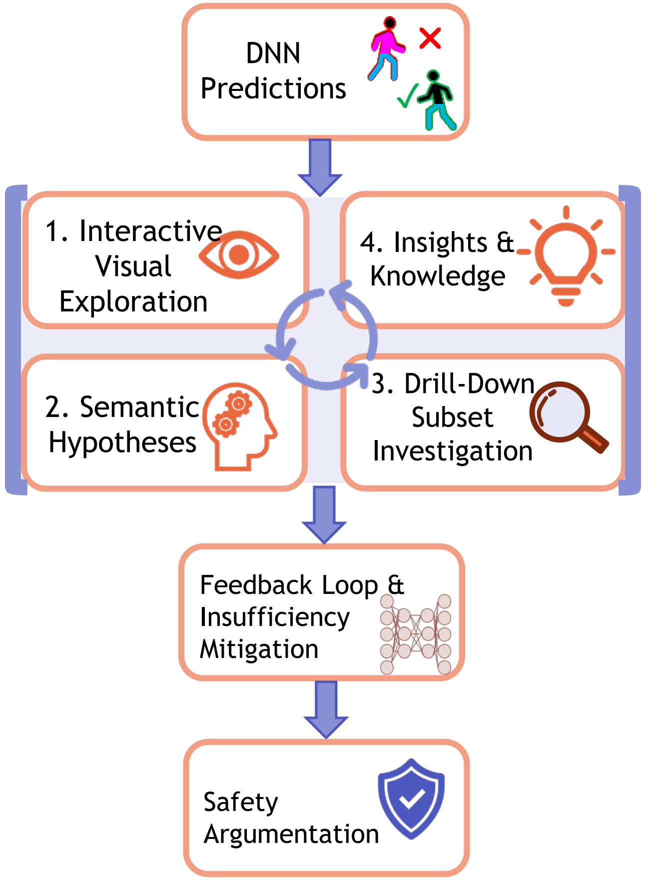
Novel online uncertainty mechanism using gradient information [BUW]

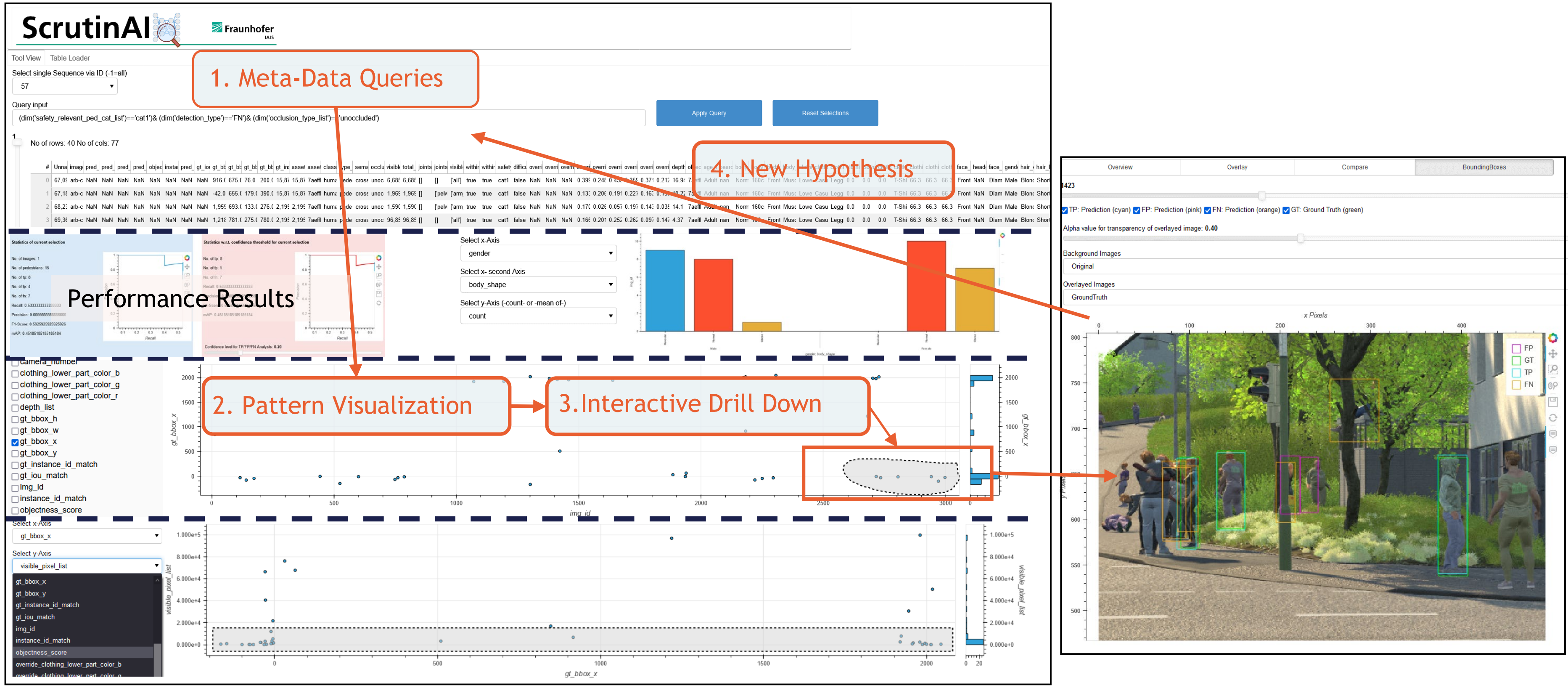


Riedlinger, T., Schubert, M., Kahl, K., Rottmann, M. (2022). Uncertainty Quantification for Object Detection: Output- and Gradient-Based Approaches. In: Fingscheidt, T., Gottschalk, H., Houben, S. (eds) Deep Neural Networks and Data for Automated Driving. Springer, Cham. https://doi.org/10.1007/978-3-031-01233-4_9

Semantic Analysis of DNN Predictions with Visual Analytics and Visual Analytics Tool "ScrutinAI" [IAIS]

Addressed Safety Concerns: Incomprehensible Behavior



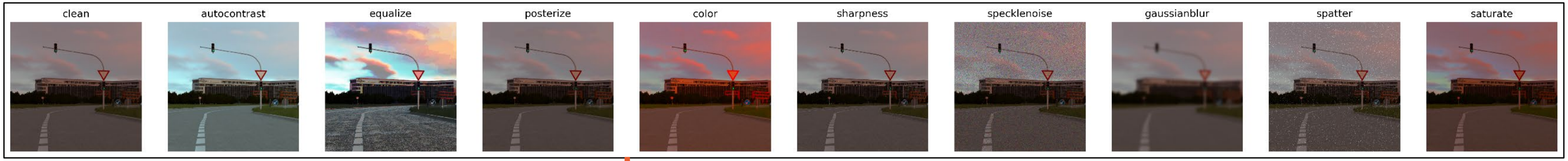


Haedecke, Mock, Akila: „ScrutinAI: A Visual Analytics Approach for the Semantic Analysis of Deep Neural Network Predictions", EuroVis Workshop on Visual Analytics (2022)

[Fraunhofer IAIS]

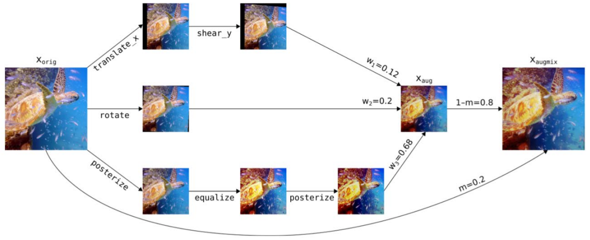
Augmentation Training (AugMix) [Volkswagen]

Addressed Safety Concerns: Brittleness of DNNs



Combined using AugMix

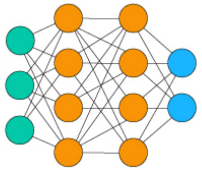
+ Improved robustness
+ Improved generalization
+ Data efficient augmentation strategy



AUGMIX: A SIMPLE DATA PROCESSING METHOD TO IMPROVE ROBUSTNESS AND UNCERTAINTY

Dan Hendrycks[*] DeepMind hendrycks@berkeley.edu
Norman Mu[*] Google normanmu@google.com
Ekin D. Cubuk Google cubuk@google.com
Barret Zoph Google barretzoph@google.com
Justin Gilmer Google gilmer@google.com
Balaji Lakshminarayanan[†] DeepMind balajiln@google.com

Training



DeepLabv3
ResNet 101
(KIA model by Intel)

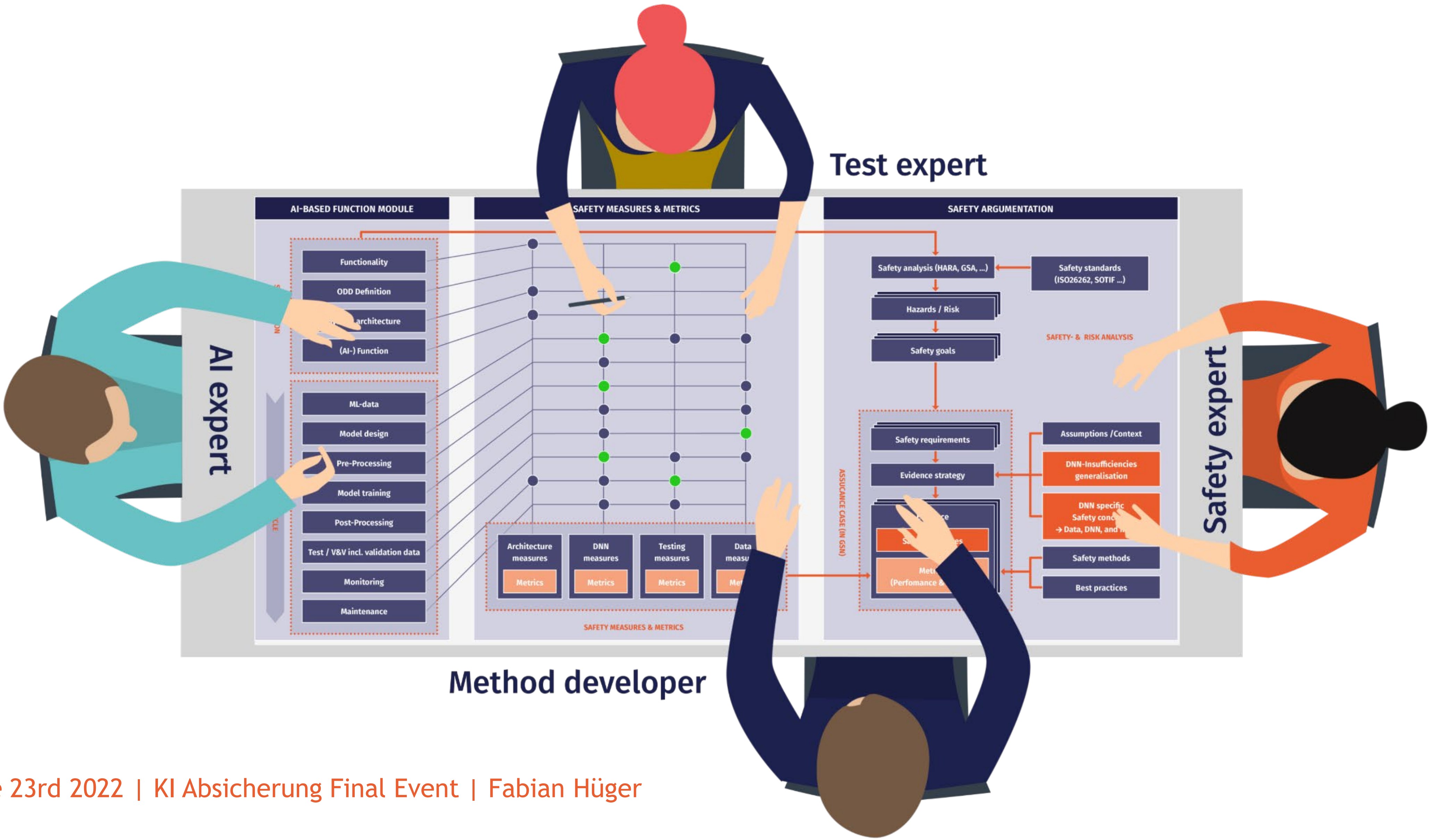
Evaluation on *unseen* „real-world“ corruptions

Based on: Hendrycks et al., AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, D., https://arxiv.org/abs/1912.02781

3

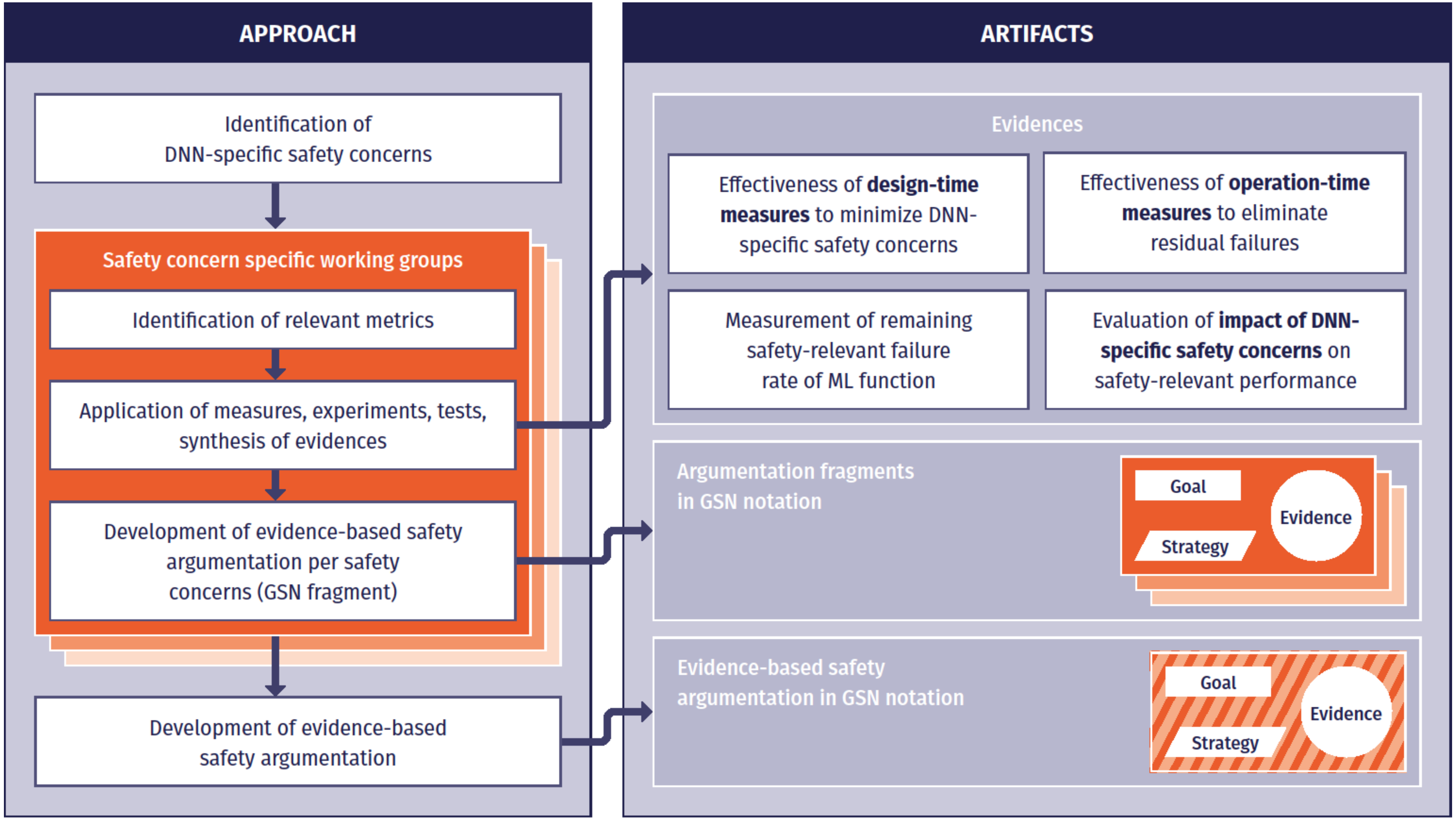
Injecting Mechanisms into the Safety Argumentation: Evidence Workstreams

Evidence Workstreams



INTERNAL

Creation of an evidence-based safety argumentation



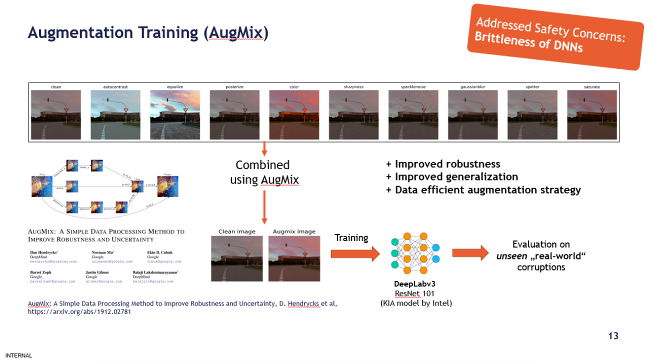
Evidence types:
Burton, S., Hellert, C., Hüger, F., Mock, M., Rohatschek, A. (2022). Safety Assurance of Machine Learning for Perception Functions. In: Fingscheidt, T., Gottschalk, H., Houben, S. (eds) Deep Neural Networks and Data for Automated Driving. Springer, Cham. https://doi.org/10.1007/978-3-031-01233-4_12

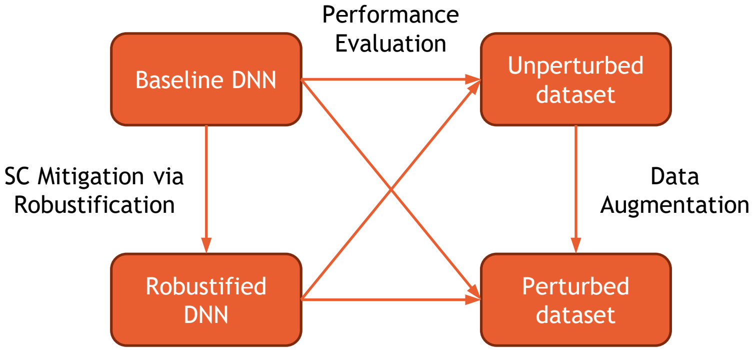
Figure:
F. Blank, F. Hüger, M. Mock, T. Stauner: Methodik zur Absicherung von KI im Fahrzeug, ATZ extra, Springer Verlag, 23.06.2022

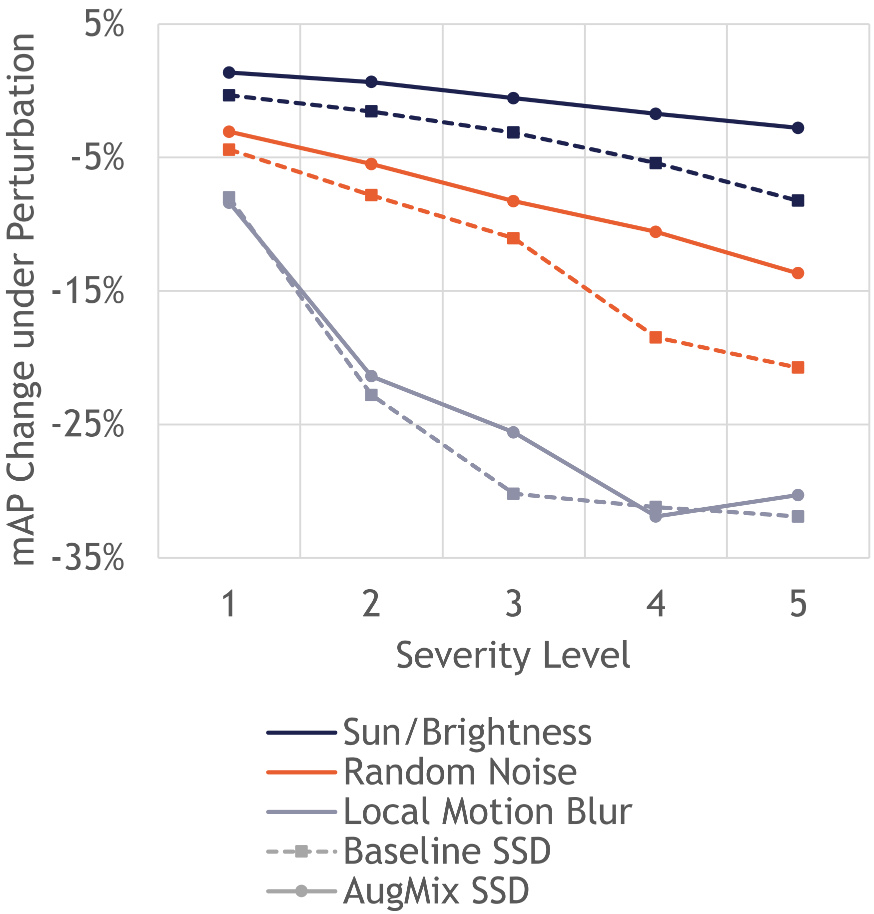
Example Mechanism Evidences
AugMix

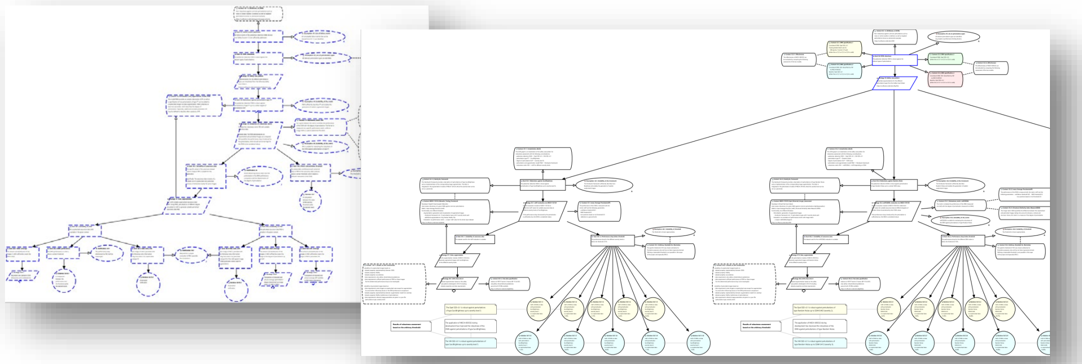
Addressed Safety Concerns:
Brittleness of DNNs

Exemplary requirement: *“The DNN shall be robust against all types of foreseeable noise.”*



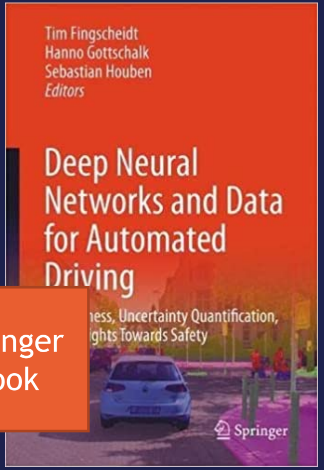






Summary

4



Springer Book

KI Absicherung

Künstliche Intelligenz und maschinelles Lernen im automobilen Umfeld

DNN-SPECIFIC SAFETY CONCERNS

This document defines the notions of DNN insufficiencies and DNN specific safety concerns for the internal use within the KI Absicherungs consortium. It identifies insufficient generalisation capability as the main DNN insufficiency and defines DNN specific safety concerns in different categories, namely DNN characteristics related concerns, data related concerns and metric related concerns.

[Status]

Veröffentlichung: 14.12.2020

Methods & Measures

Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety

Initial State-of-Research Report

Survey available at
www.ki-absicherung-projekt.de/

Literature Repository

Literature Repository

Literature Repository available at:
tinyurl.com/e3y4pmxs

Mechanisms Catalogue

KI ABSICHERUNG

Safe AI for Automated Driving

Dr. Fabian Hüger, Volkswagen AG

fabian.hueger@cariad.technology

KI Absicherung ist ein Projekt der KI Familie und wurde aus der VDA Leitinitiative autonomes und vernetztes Fahren heraus entwickelt.

KI FAMILIE

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

www.ki-absicherung.vdali.de @KI_Familie KI Familie

5.1

Safety Argumentation

SP 4: Andreas Rohatschek, Robert Bosch GmbH

KI ABSICHERUNG

Safe AI for Automated Driving

Final Event | June 23rd 2022

TP4 Part I Safety

Andreas Rohatschek, Robert Bosch GmbH

Our Goal:
Create the Safety Pillar for the bridge between AI Land and Safety Land

"The AI based perception is safe BECAUSE…"
…compelling and convincing arguments and evidence follow

Couple of selected mechanisms

Evidence Workstream 1

Mechanism A

Evidence Workstream 2

Mechanism B

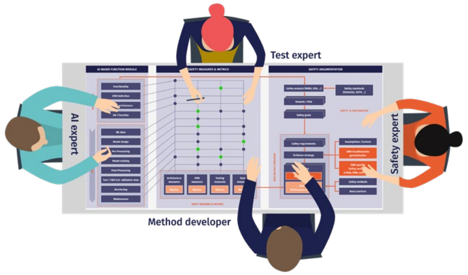
⋮

Evidence Workstream n

Mechanism X

Our Approach: Evidence Workstreams :

Empowering experts from safety engineering and ML to produce measures and evidences



Assurance Case (ISO 15026 – Part 1 Vocabulary):
Reasoned, auditable artefact created that supports the contention that its top-level claim (or set of claims), is satisfied, including systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s)

Our Approach

The path to an evidence-based Safety Argumentation

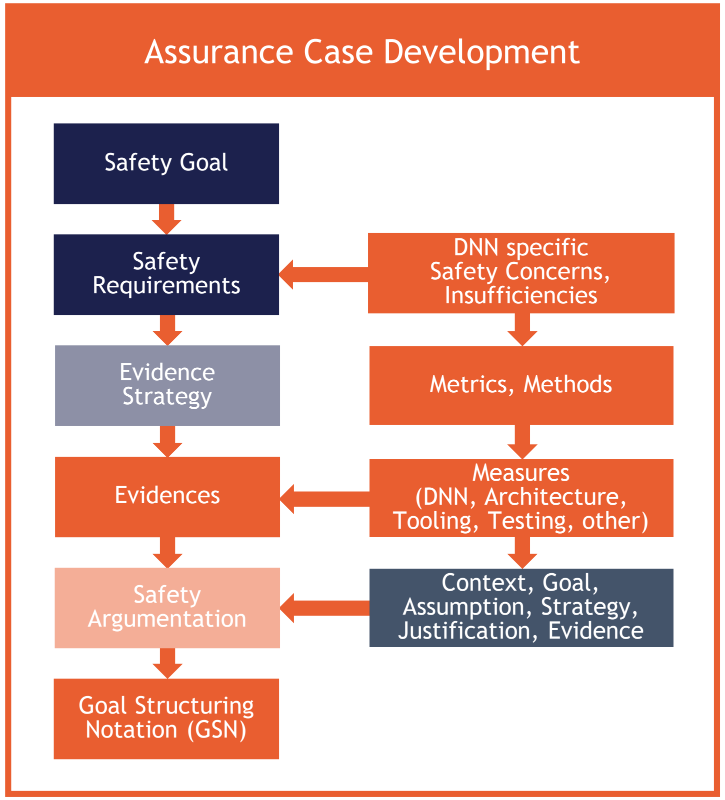
1. Identify potential causes of insufficiencies in the function (in KIA: “DNN-specific safety concerns”)
2. Introduce metrics or some form of judgment to argue that insufficiency was mitigated
3. Develop methods to mitigate the insufficiencies
4. Argue that the residual risk associated with the causes has been reduced to a tolerable level
5. Create the evidence based safety argumentation in a Goal Structuring Notation

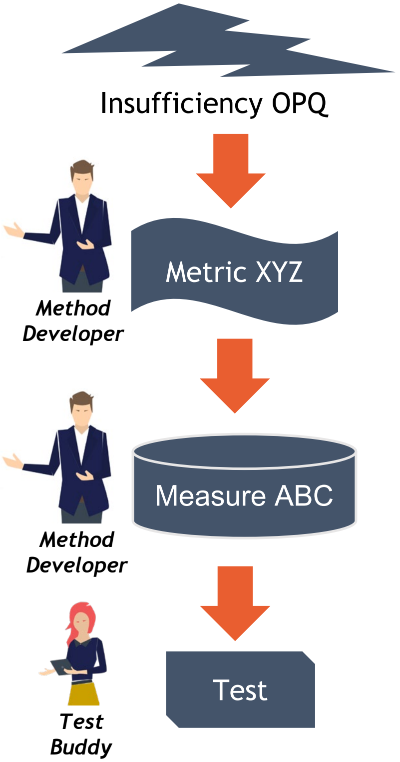
Goal Structuring Notation (GSN)

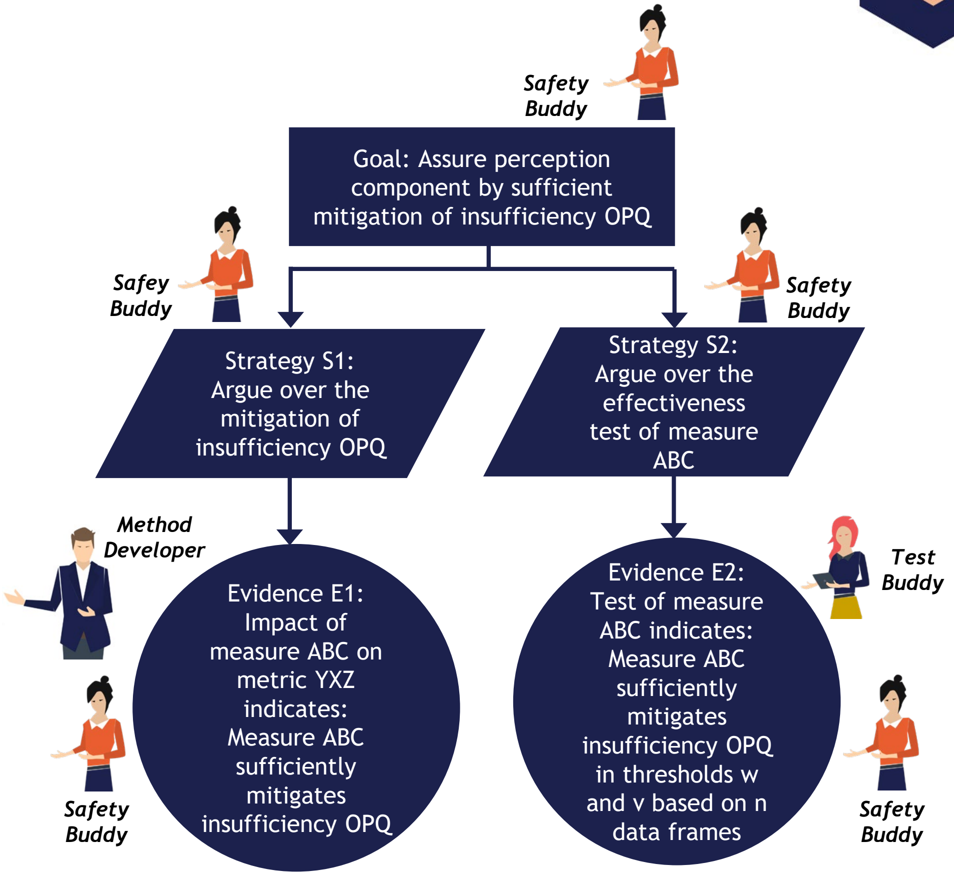
Source: Goal structuring notation, community standard version 3

What are the causes of insufficiencies and what sources of evidence can be used to make this argument?

How to create Evidences from Methods and Tests







Interaction of Method Developer, Safety Buddy and Test Buddy leads to evidences for the safety argumentation

Building Blocks for the Safety Argumentation

Functional

- Safety Goals, Safety Requirements
- DNN related Safety Concerns

Non Functional

- Argumentation Structure, Safety Patterns
- Top-Down (deductive), Bottom-Up (inductive)

Enabler

- GSN Standard 3.0
- Bayesian Belief Networks
- Systems Theoretic Process Analysis

Cluster: Knowledge / competencies

Deliverable: Interfaces and information with the aspects:

- Evidence-relevant information from TP1-3
- Supervision of Safety Buddy work

Knowledge → GSN

Cluster: Design of parts of the safety argumentation

Deliverable: „GSN-Branches“ with the following aspects:

- Synthesis of GSN elements to GSN branches
- Consideration of structure and docking points

GSN GSN GSN

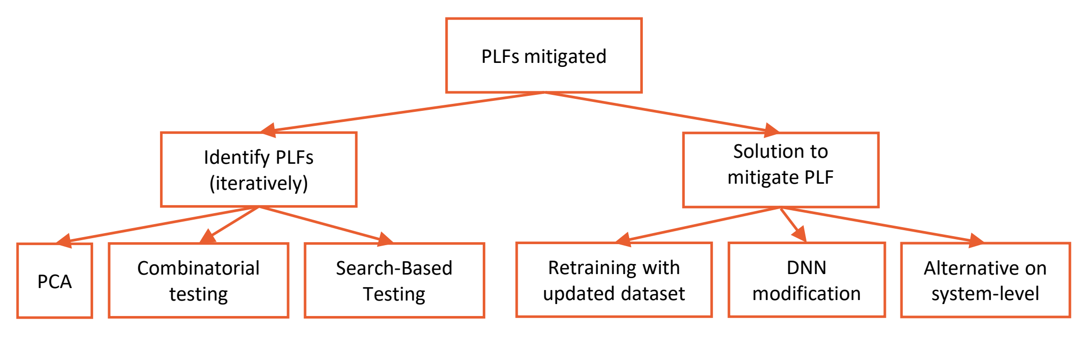
Cluster: Design of the entire safety argumentation

Deliverable: "Overall GSN" with the following aspects:

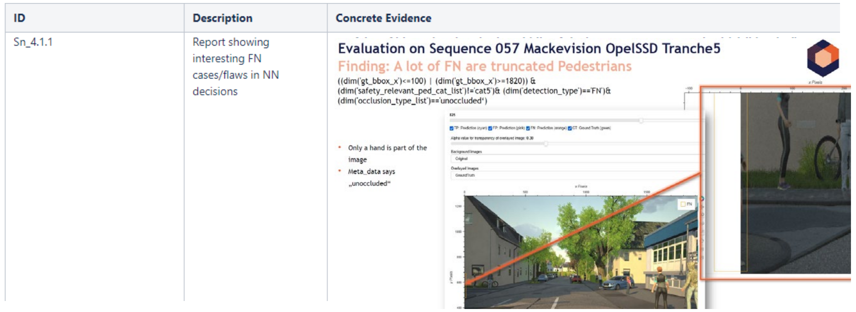
- General structure of the overall GSN
- Docking points for integration of the "GSN branches"

GSN GSN GSN GSN

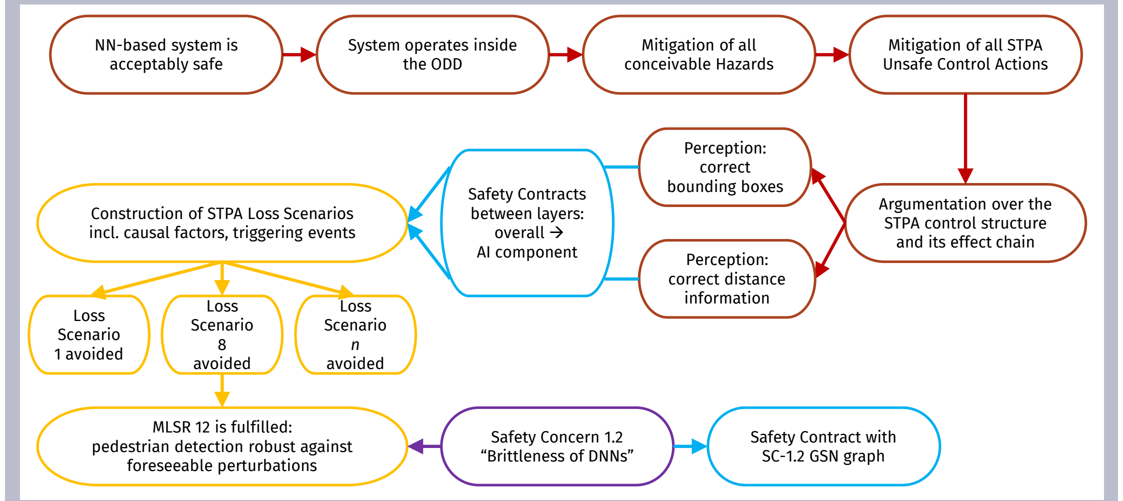
Our Results (Extract)



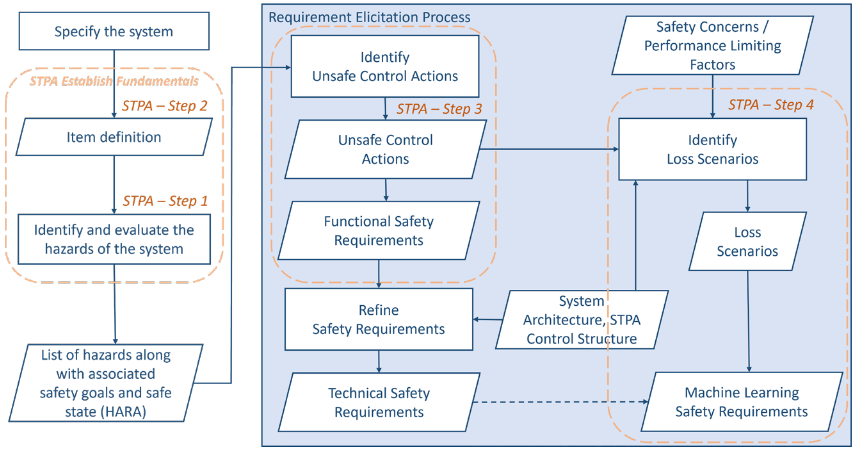
Schematic overview of the GSN safety argumentation for PLF mitigation



Match concrete evidence and solutions in GSN



Overview of the safety argumentation



STPA based approach for the elicitation of ML Safety Requirements

PLF: Performance Limiting Factor; STPA: Systems Theoretic Process Analysis

INTERNAL

Our Achievements

R	We established an evidence-based safety argumentation
E	We learned how to structure the safety argumentation
S	We used Goal Structuring Notation (GSN) to visualize the safety argumentation
U	We investigated several possibilities to create evidences
L	We identified gaps in our argumentation and closed them or take them for future work
T	We integrated argumentations related to DNN-specific safety concerns
S	We considered the combination of qualitative and quantitative evidences

Our deliverable: “Overall Goal Structuring Notation” (structure and argumentation branches)

KI ABSICHERUNG

Safe AI for Automated Driving

Andreas Rohatschek, Robert Bosch GmbH
andreas-juergen.rohatschek@de.bosch.com

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

www.ki-absicherung.vdali.de @KI_Familie KI Familie

5.2

Testing

SP 4: Frédérik Blank, Robert Bosch GmbH

KI ABSICHERUNG

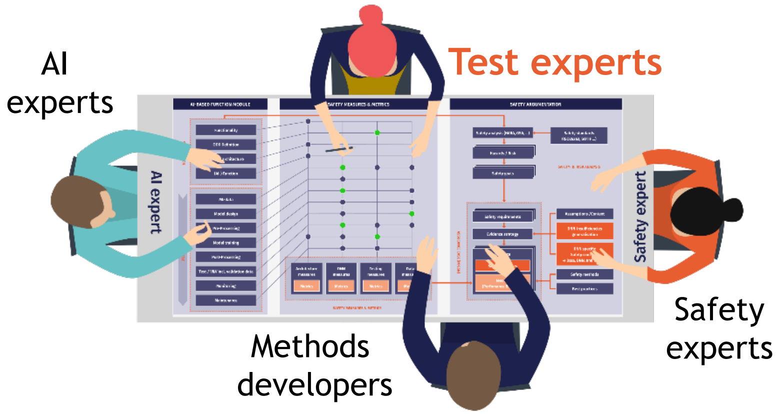
Safe AI for Automated Driving

Final Event | June 23rd 2022

Testing, Teststrategy & ML Lifecycle (AP4.4/AP4.5)

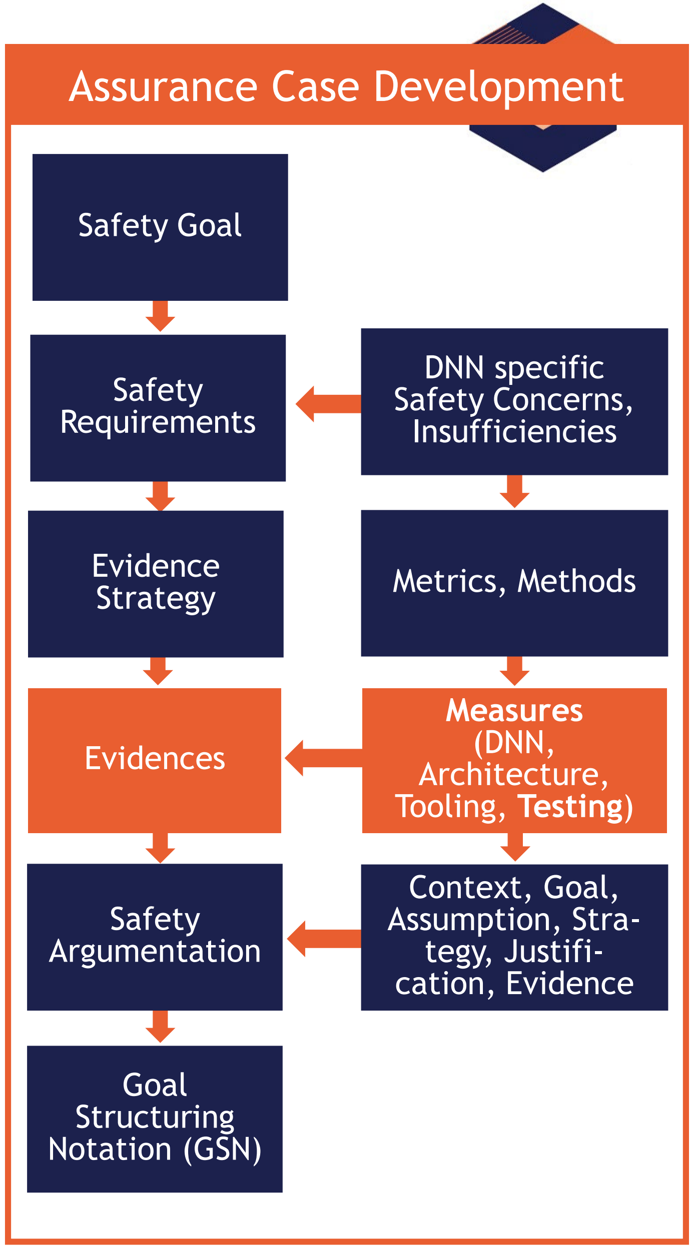
Frédérik Blank, Robert Bosch GmbH

Testing & KI-A test strategy



- Plays a major role in assuring safety of AI-based functions
- Results from newly developed test& test methods used as evidences in the safety argumentation
- Required: New approaches focusing on systematically testing the "AI-function" and "used data" in an iterative way
- KI-A test strategy
  - lists applicable test methods for specific test purposes
  - consists of 4 method classes:
    - Dataset Verification & Coverage Analysis
    - Neuronal Network Component Test
    - Data Pool Verification (dataset label quality analysis)
    - ML Integration & Qualification Test

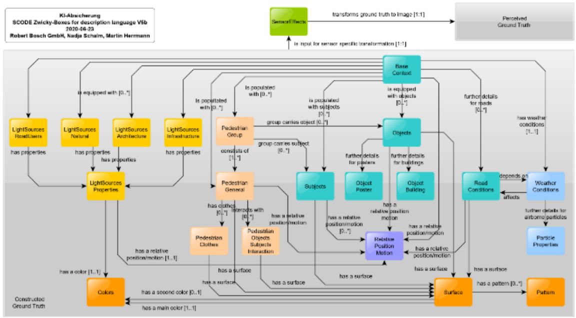
⇒ **Provide evidence on the quality of the system under test**



AI = Artificial Intelligence KI-A = KI Absicherung ML = Machine Learning

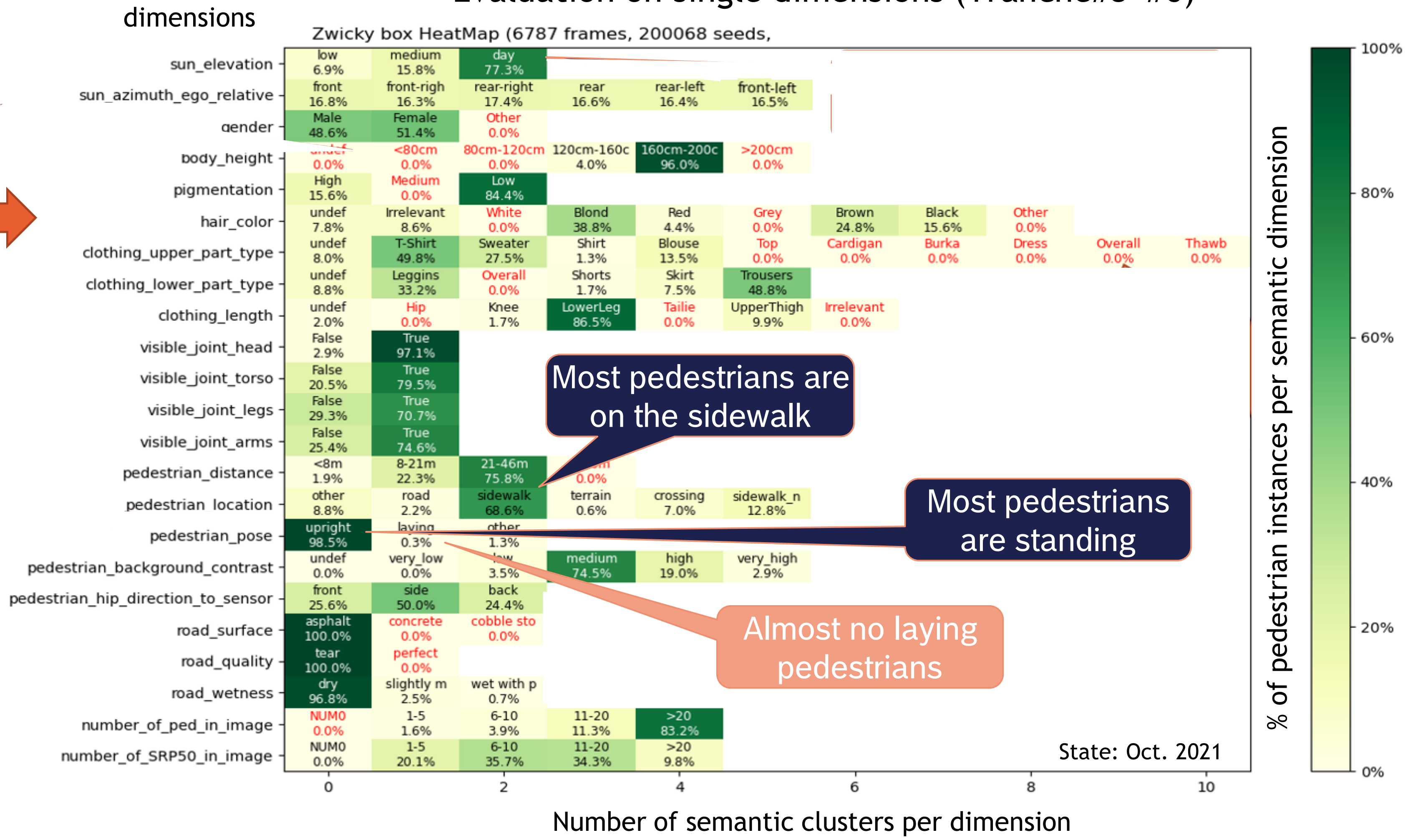
Dataset Verification & Coverage Analysis
Evaluating the Training Data Coverage with Heatmaps

Semantic domain model



- Input data coverage = degree a dataset covers the semantic domain model (training or testing)
- Extension: Combination of dimensions & semantic clusters with each other (e.g. pairs)

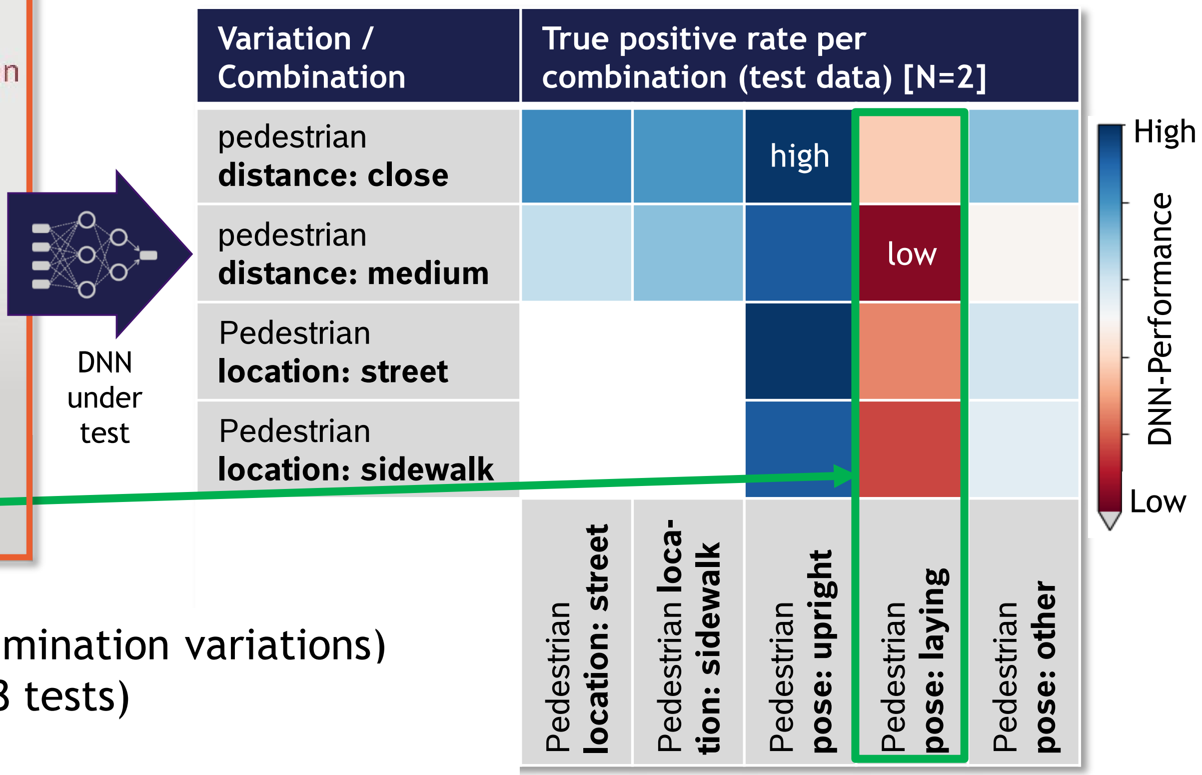
Evaluation on single dimensions (Tranche#5+#6)



Systematic generation of parametrizable safety critical scenarios (Euro-NCAP-like) using combinatorial testing

Source: Valeo, Bosch, ZF, Mackevision

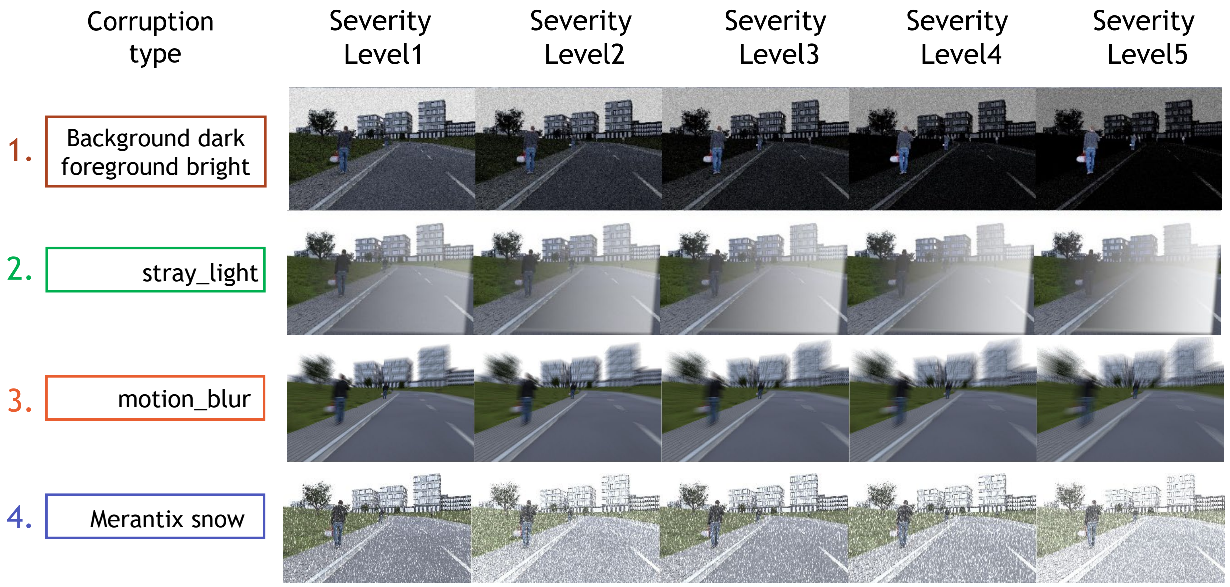
- Build safety-relevant and representative test datasets that systematically & efficiently cover ODD and fill data gaps (e.g. with **combinatorial testing**)
- Helps to identify safety-critical low performance data points & DNN insufficiencies

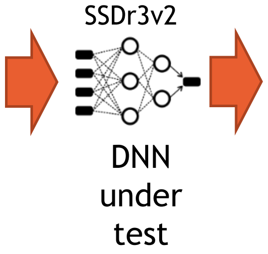


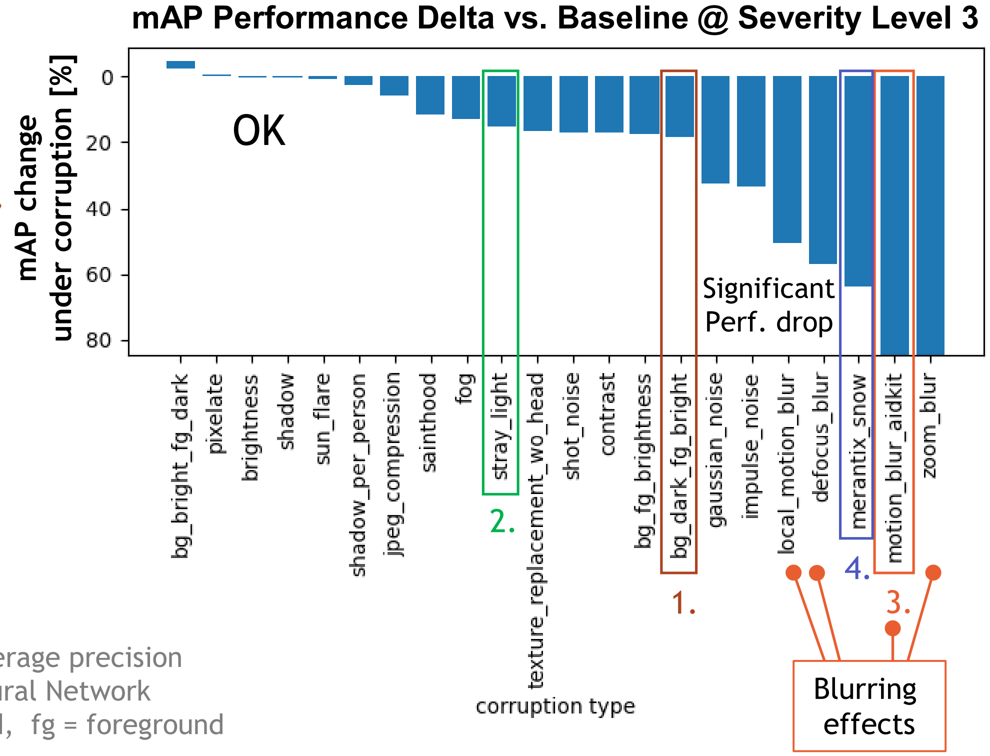
Testing all combination: **27.6 billion** (incl. vehicle color and illumination variations)
Combinatorial testing **3-wise: 6669 tests** or images (2-wise: 408 tests)

Neuronal Network Component Test -
Corruptions Testing to identify most critical corruption types

- DNN robust against natural corruptions (noise, weather & light effects, …)?
  - A robust DNN should ideally exhibit no performance drop when encountering natural corruptions
- Newly developed corruption types within KI-Absicherung revealed remaining robustness insufficiencies → Input for further robustification(s) and evidence to safety argumentation



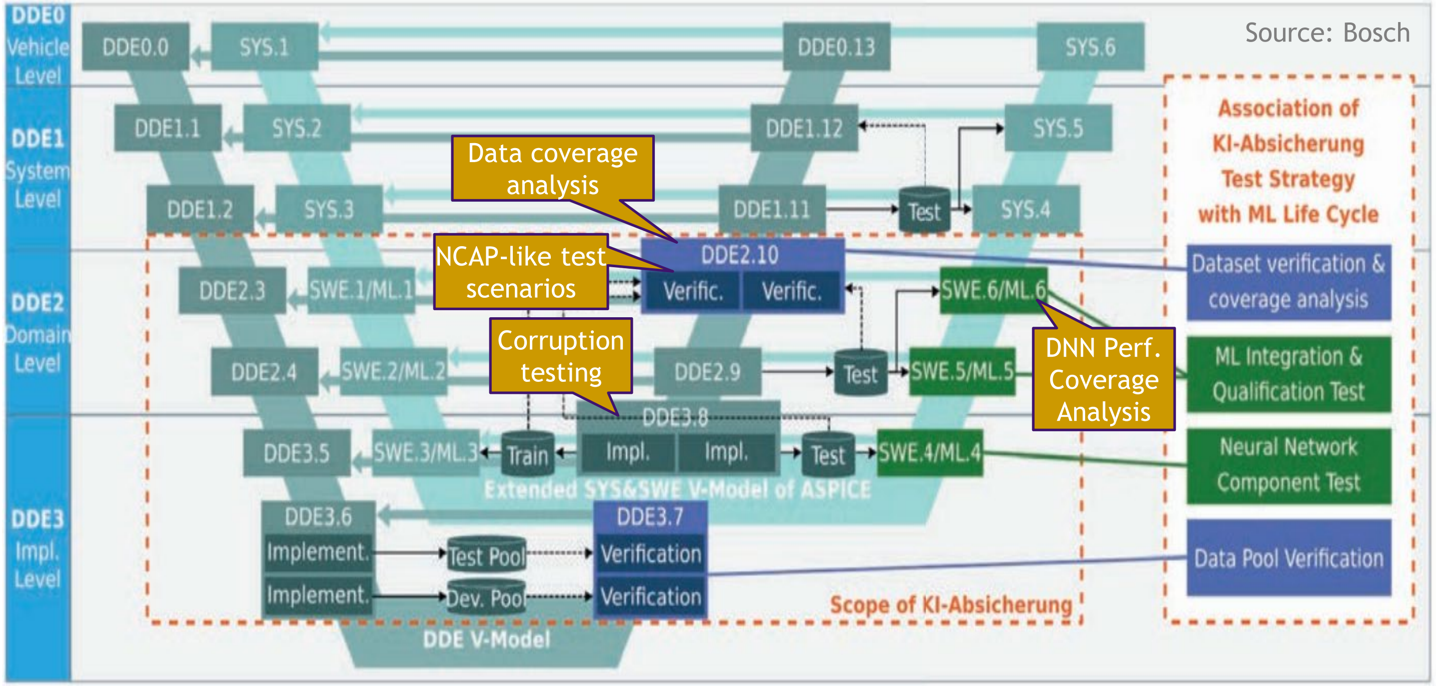




mAP = mean average precision
DNN = Deep Neural Network
bg = background, fg = foreground

ML Lifecylce model for ML development

- New consistent data-oriented ML Lifecycle model developed to
  - define systematic, structured ML data-driven development process
  - systematically specify, implement and verify training and testing data sets for SW with ML models
- Adds a second V-model for the data that collaborates with the SYS/SW V-model via defined datasets
- Links to KI Absicherung test strategy on implementation and domain level
- Planned as input for communication with ISO/PAS 8800



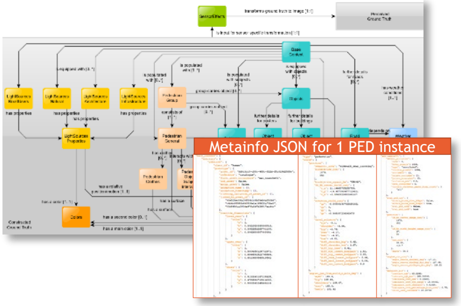
Source: Bosch

DDE: Data-Driven Engineering ML: Machine Learning SWE: Software Engineering SW: Software

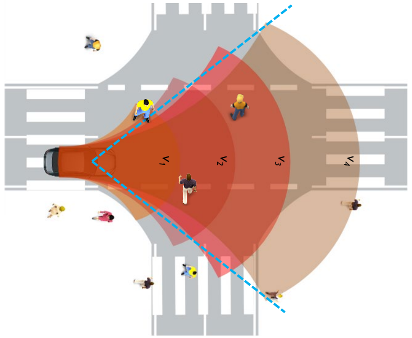
INTERNAL

To summarize the work and some of the highlights of TP4…

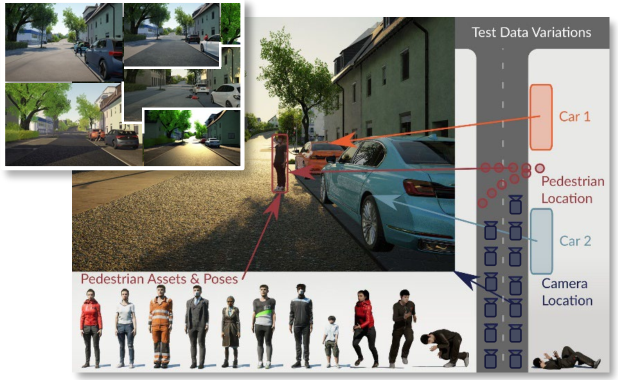
ODD, Ontology & enriched metadata



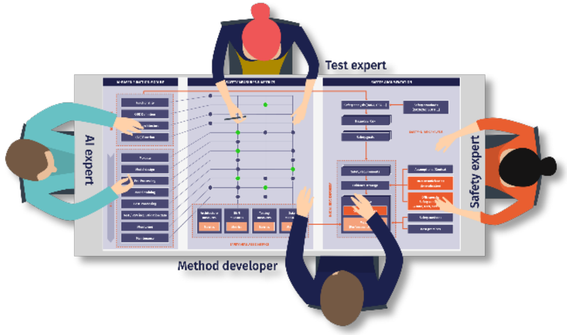
Safety relevant Pedestrians



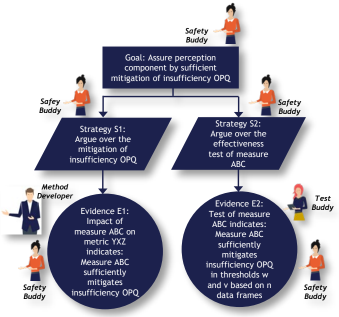
Systematic parametrized NCAP-like safety scenarios



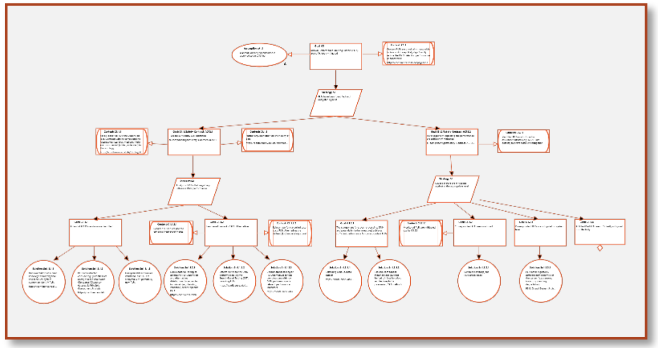
Participation of Safety & Test experts @ Evidence Workstreams



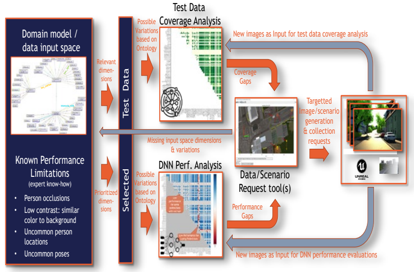
Evidence-based safety Argumentation



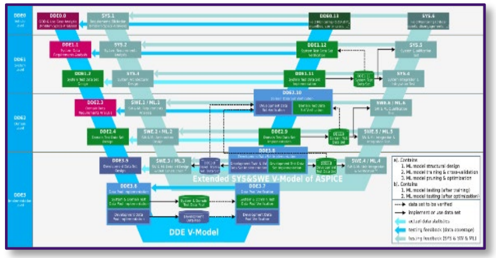
GSN-Fragments from EWS



Test methods & testing with closed-data loop



ML-Lifecycle & Data-driven Engineering Process



ODD = Operational Design Domain; EWS = Evidence Work Streams
ML = Machine Learning; GSN = Goal Structuring Notation
DDE = Data-driven Engineering

KI ABSICHERUNG

Safe AI for Automated Driving

Frédérik Blank, Robert Bosch GmbH

Frederik.Blank@de.bosch.com

KI Absicherung ist ein Projekt der KI Familie und wurde aus der VDA Leitinitiative autonomes und vernetztes Fahren heraus entwickelt.

KI FAMILIE

Gefördert durch:

Bundesministerium für Wirtschaft und Energie

aufgrund eines Beschlusses des Deutschen Bundestages

www.ki-absicherung.vdali.de @KI_Familie KI Familie

13:00-14:00 Mittagspause mit paralleler Postersession

14:00-15:00 Postersession

15:00-15:30 drei parallele Highlightvorträge