KI ABSICHERUNG
Safe AI for Automated Driving

KI Absicherung – Finale Ergebnissteckbriefe TP3

Version zur Veröffentlichung

Version	1.0
Editor	Dr. Fabian Hüger / Volkswagen AG Dr. Stephan Scholz / Volkswagen AG PD Dr. Michael Mock / Fraunhofer IAIS
Projektkoordination	Volkswagen AG / Fraunhofer IAIS
Fälligkeit	31.12.2022
Erstellungsdatum	03.11.2022

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

Dokumenteninformation

Autoren

Andreas Bär / TU Braunschweig

Fridolin Bauer / BMW Group

Andreas Blattmann / Universität Heidelberg

Dominik Brüggemann / Bergische Universität Wuppertal

Robin Chan / Bergische Universität Wuppertal

Philip-William Grassal / Universität Heidelberg

Christian Hellert / Continental AG

Alexander Hirsch / Robert Bosch GmbH

Fabian Hüger / Volkswagen AG

Fabian Küppers / Hochschule Ruhr West

Titus Leistner / Universität Heidelberg

Adrian Loy / Merantix Labs GmbH

Michael Mock/ Fraunhofer IAIS

Firas Mualla / ZF Friedrichshafen AG

Svetlana Pavlitskaya / Forschungszentrum Informatik FZI

Tobias Riedlinger / Bergische Universität Wuppertal

Carsten Rother / Universität Heidelberg

Timo Sämann / Valeo Schalter und Sensoren GmbH

Jan David Schneider / Volkswagen AG

Gesina Schwalbe / Continental AG

Deepthi Sreenivasaiah / Merantix Labs GmbH

Hanno Stage / Forschungszentrum Informatik FZI

Tom Thielo / e:fs TechHub GmbH

Christoph Thiem / Stellantis, Opel Automobile GmbH

Reviewer

Fabian Hüger / Volkswagen AG

Stephan Scholz / Volkswagen AG

Michael Mock / Fraunhofer IAIS

Christian Hellert/ Continental AG

Sebastian Gerres / Merantix Labs GmbH

Hanno Gottschalk/ Bergische Universität Wuppertal

Alexander Hirsch / Robert Bosch GmbH

Timo Sämann / Valeo Schalter und Sensoren GmbH

Jan David Schneider / Volkswagen AG

Jonas Schneider / e:fs TechHub GmbH

Kontakt

(in Vertretung für die Projektkoordination)

European Center for Information and Communication Technologies – EICT GmbH

EUREF-Campus Haus 13

Torgauer Straße 12-15

10829 Berlin

Germany

Email: projects@eict.de

Projektwebsite: https://www.ki-absicherung-projekt.de/

Revisionslog

Version	Datum	Kommentar	Autor	Partner
0.1	Bis 21.10.2022	Input auf Confluence	s.o.	s.o.
0.2	24.-25.10.2022	Ausspielen Word, Prüfung und Korrektur Übertrag, Strukturierung und Layout Dokument	Bert Hildebrandt	EICT
0.3	25.10.2022	Formelles Review und Layout	Dr. Nikos Papamichail	EICT
1.0	03.11.2022	Finalisierung	Dr. Nikos Papamichail	EICT

Inhaltsverzeichnis

1 AP3.1 Tracking State of Research 8

1.1 E3.1.1 Final: Kriterienkatalog und Gliederung zur Definition des relevanten Forschungsfelds 8

1.1.1 Formal Classification 8

1.1.2 Description of the result 8

1.1.3 Result 8

1.2 E3.1.2 Final: Initialer State-of-Research-Report 9

1.2.1 Formal Classification 9

1.2.2 Description of the result 9

1.2.3 Result 11

1.3 E3.1.3 Final: Auflistung und Kategorisierung relevanter State-of-Research (Repository) 12

1.3.1 Formal Classification 12

1.3.2 Description of the result 12

1.3.3 Result 13

1.4 E3.1.4 Final: Öffentlicher Zugang zu E3.1.3 14

1.4.1 Formal Classification 14

1.4.2 Description of the result 14

1.4.3 Result 15

1.5 E3.1.5 Final: Aktiver Austausch mit wissenschaftlichen Community 15

1.5.1 Formal Classification 15

1.5.2 Description of the result 15

1.5.3 Result 15

2 AP3.2 Höherwertige Funktion KPIs für KI Funktionen 17

2.1 E3.2.1 Final: nur projektintern für KI Absicherung verfügbar 17

2.2 E3.2.2 Final: nur projektintern für KI Absicherung verfügbar 17

2.3 E3.2.3 Final: nur projektintern für KI Absicherung verfügbar 17

2.4 E3.2.4 Final: nur projektintern für KI Absicherung verfügbar 17

2.5 E3.2.5 Final: nur projektintern für KI Absicherung verfügbar 17

3 AP3.3 Funktional verändernde Methoden und Maßnahmen 18

3.1 E3.3.1 Final: Algorithmische Implementierung und Dokumentation für optimierte Datensatz-Selektion 18

3.1.1 Formal Classification 18

3.1.2 Description of the result 18

3.1.3 Results 18

3.2 E3.3.2 Final: nur projektintern für KI Absicherung verfügbar 31
3.3 E3.3.3 Final: nur projektintern für KI Absicherung verfügbar 31
3.4 E3.3.4 Final: nur projektintern für KI Absicherung verfügbar 31
3.5 E3.3.5 Final: nur projektintern für KI Absicherung verfügbar 31
4 AP3.4 Introspektive Methoden und -Maßnahmen 32
4.1 E3.4.1 Final: Plausibilisierung 32
4.1.1 Formal Classification.....32
4.1.2 Description of the result 32
4.2 E3.4.2 Final: Unsicherheitsmodellierung 65
4.2.1 Formal Classification.....65
4.2.2 Description of the result 66
4.3 E3.4.3 Final: Robustheitsprüfung durch Manipulation 78
4.3.1 Formal Classification.....78
4.3.2 Description of the result 78
4.4 E3.4.4 Final: Online Überwachung.....80
4.4.1 Formal Classification.....80
4.4.2 Description of the result 80
4.5 E3.4.5 Final: Offline Validierung.....85
4.5.1 Formal Classification.....85
4.5.2 Description of the result 85
5 AP3.5 Externe Methoden und Maßnahmen 90
5.1 E3.5.1 Final: Surveillance and coverage of input data 90
5.1.1 Formal Classification.....90
5.1.2 Description of the result 90
5.2 E3.5.2 Final: Uncertainty estimation and calibration methods..... 108
5.2.1 Formal Classification..... 108
5.2.2 Brief description of the cluster content 108
5.2.3 Mechanisms..... 109
5.3 E3.5.3 Final: Adversarial attacks and teacher-student frameworks 127
5.3.1 Formal Classification..... 127
5.3.2 Description of the result 127
5.4 E3.5.4 Final: nur projektintern für KI Absicherung verfügbar 135
5.5 E3.5.5 Final: nur projektintern für KI Absicherung verfügbar 135
AP3.6 Aggregierte Methoden.....136
5.6 E3.6.1 Final: E3.6.2 Final: Auflösung von Bewertungsredundanzen und Synergien 136
5.6.1 Formal Classification..... 136

5.6.2 Description of the result 136
5.7 E3.6.3 Final: Implementierung von aggregierten Methoden und Maßnahmen und Bewertung hinsichtlich KPIs 140
5.7.1 Formal Classification 140
5.7.2 Description of the result 140
5.8 E3.6.4 Final: nur projektintern für KI Absicherung verfügbar 152
5.9 E3.6.5 Final: nur projektintern für KI Absicherung verfügbar 152

1 AP3.1 Tracking State of Research

1.1 E3.1.1 Final: Kriterienkatalog und Gliederung zur Definition des relevanten Forschungsfelds (zur Veröffentlichung)

1.1.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Document*
Group/Cluster	
Type of content	*Compilation/Overview/Catalogue/Classification*
Classification level	*PU*

1.1.2 Description of the result

1.1.2.1 Motivation

The state-of-the-art in the design and development of methods for improving safety-relevant aspects of neural networks includes a plentitude of different approaches and methods. Most of these approaches and methods are investigated for improvement in the scope of sub-project 3. The overall task of WP3.1 is to monitor and make available the state-of-the-art in a structured manner. Result 3.1.1 is defining the structure, under which the state-of-the art is clustered and an assignment to partners being responsible to lead the state-of-the-art monitoring in their assigned cluster.

1.1.2.2 Approach

Fraunhofer IAIS was organizing a 2 days face-to-face kick-off workshop with 30 participants for the WP 3.1. On this workshop, several moderation techniques (world-café, call-response techniques, moderated group discussion, prepared structured AP presentations) have been applied to come to a consistent and complete overview and clustering of the relevant state-of-the-art topics. This clustering has been aligned further with the KPI-structures being generated in WP 3.2. The result has been presented in follow-up TP-Lead and TP3-all workshops. An assigned of clusters to responsible lead partner for literature monitoring has been elaborated and agreed upon in iterative follow-up calls.

1.1.3 Result

The following table shows the resulting clustering of the related work, including sub-clusters. This clustering was used as basis for an assignment to responsible lead partners for literature monitoring:

Partner Assignements

Uncertainty	Generative Models		1			0					1
	MC Dropout					1					1
	BNN Approximations		0			1					1
	Markov Random Fields							1	0		1
	Gradient based Uncertainty		0	0	0	1			0		1
Interpretability	Visual Analytics					1	0				1
	Intermediate Representations							1	0		1
	Pixel Attributions	1	0					0	0	0	1
	Interpretable Proxies				1						1
	Generative Approaches									1	1
Compression	Pruning									1	1
	Quantization									1	1
	Distillation									1	1
Architectures	Building Blocks				0		1			0	1
	Architecture Search							1			1
	Multi-Task Networks								1		1
Dataset Optimisation	Outlier / Anomaly Detection					1				0	1
	Active Learning			1							1
	Domains		0		0	1					1
	Augmentation									1	1
	Corner Case Detection						0		1		1
Adversarial Attacks	Attacks and Defences		0				0	1		0	1
	More Realistic Attacks						1			0	1
Aggregation	Ensenmble Methods				0	1		0			1
	Temporal Methods			0	0	0	0	0	1	0	1
Verification	Formal testing		0		1	0				S	1
	Black Box methods					1					1
Robust Training	Modification of Loss			1							1
	(Domain) Generalization					0				1	1
	Hyperparameter Tuning	0						1			1
	Iterative Leaning							1			1
KPIs as such					1				1		1
		1	1	2	3	8	2	6	4	6	

1.2 E3.1.2 Final: Initialer State-of-Research-Report (zur Veröffentlichung)

1.2.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Document*
Group/Cluster	
Type of content	*State of the art*
Classification level	*PU*

1.2.2 Description of the result

1.2.2.1 Motivation

The goal of this result is to provide a comprehensive and almost complete state-of-the-art overview on methods for AI safety.

1.2.2.2 Approach

The clustering of the state-of-the-art has been refined to a build a section and sub-section chapter. Each section has an introduction into the cluster topic and each subsection is

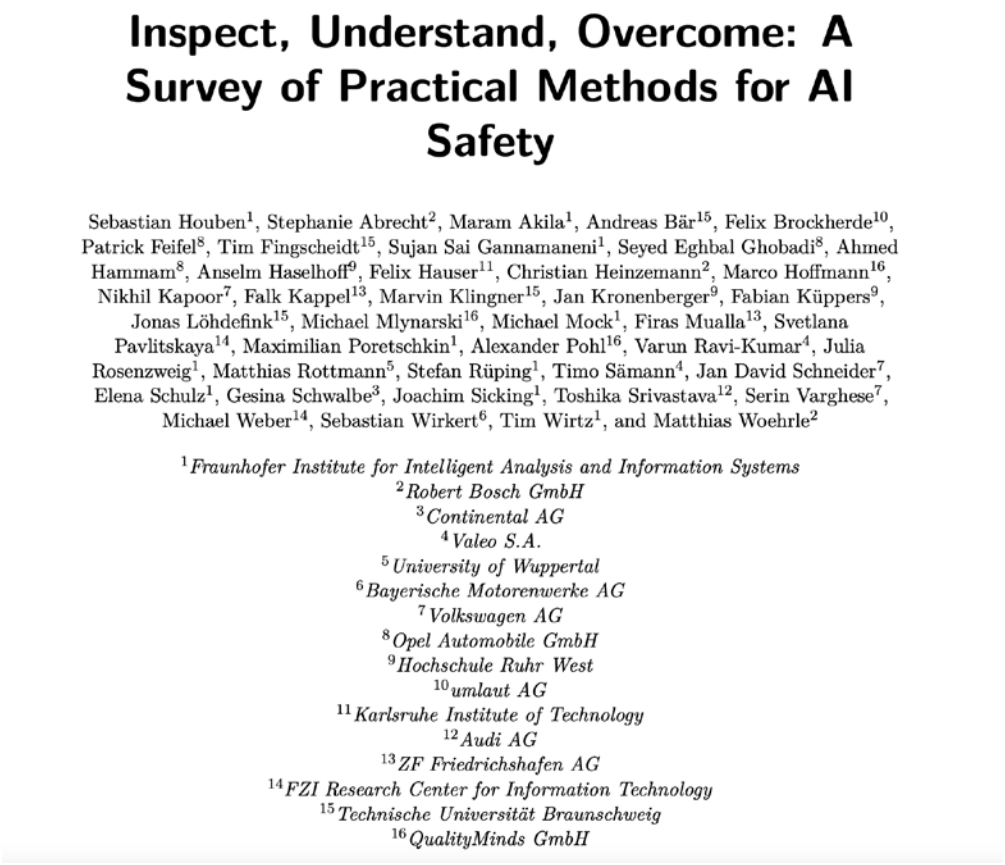
subdivided into the paragraphs "Definitions and Origins" showing basic work in this field and paragraphs "Challenges and Research Directions" showing the recent contributions and open questions in this field. The section and sub-section structure is the basis for the literature repository (E3.1.3 and E3.1.4) and the method taxonomy developed in WP3.2.

1. Introduction
2. Dataset Optimization
    a. Outlier/Anomaly Detection
    b. Active Learnig
    c. Domains
    d. Augmenation
    e. Corner Case Detection
3. Robust Training
    a. Hyperparameter Optimzation
    b. Modification of Loss
    c. Domain Generalization
4. Adverserial Attacks
    a. Adversarial Attacks and Defenses
    b. More Realistic Attacks
5. Interpretability
    a. Visual Analytics
    b. Intermediate Representations
    c. Pixel Attributions
    d. Interpretable Proxies
6. Uncertainty
    a. Generative Models
    b. Monte-Carlo Dropout
    c. Bayesian Neural Networks
    d. Uncertainty Metrics for DNNs in Frequentist Inference
    e. Markov Random Fields
    f. Confidence Calibration

7. Aggregation
    a. Ensemble Methods
    b. Temporal Consistency
8. Verification
    a. Formal Testing
    b. Block Box Methods
9. Architecture
    a. Building Blocks
    b. Multi-Tasks Networks
    c. Neural Architecture Search
10. Model Compression
    a. Pruning
    b. Quantization

1.2.3 Result

Although writing state-of-the-art survey paper has been designed and organized by the AP 3.1 Lead Fraunhofer IAIS, the participation has not been limited to partners formally involved in AP3.1. Partners from all sub-projects of KI-Absicherung contributed to the 93 pages overview that includes and describes more than 400 references.



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

Sebastian Houben[1], Stephanie Abrecht[2], Maram Akila[1], Andreas Bär[15], Felix Brockherde[10], Patrick Feifel[8], Tim Fingscheidt[15], Sujan Sai Gannamaneni[1], Seyed Eghbal Ghobadi[8], Ahmed Hammam[8], Anselm Haselhoff[9], Felix Hauser[11], Christian Heinzemann[2], Marco Hoffmann[16], Nikhil Kapoor[7], Falk Kappel[13], Marvin Klingner[15], Jan Kronenberger[9], Fabian Küppers[9], Jonas Löhdefink[15], Michael Mlynarski[16], Michael Mock[1], Firas Mualla[13], Svetlana Pavlitskaya[14], Maximilian Poretschkin[1], Alexander Pohl[16], Varun Ravi-Kumar[4], Julia Rosenzweig[1], Matthias Rottmann[5], Stefan Rüping[1], Timo Sämann[4], Jan David Schneider[7], Elena Schulz[1], Gesina Schwalbe[3], Joachim Sicking[1], Toshika Srivastava[12], Serin Varghese[7], Michael Weber[14], Sebastian Wirkert[6], Tim Wirtz[1], and Matthias Woehrle[2]

[1]*Fraunhofer Institute for Intelligent Analysis and Information Systems*
[2]*Robert Bosch GmbH*
[3]*Continental AG*
[4]*Valeo S.A.*
[5]*University of Wuppertal*
[6]*Bayerische Motorenwerke AG*
[7]*Volkswagen AG*
[8]*Opel Automobile GmbH*
[9]*Hochschule Ruhr West*
[10]*umlaut AG*
[11]*Karlsruhe Institute of Technology*
[12]*Audi AG*
[13]*ZF Friedrichshafen AG*
[14]*FZI Research Center for Information Technology*
[15]*Technische Universität Braunschweig*
[16]*QualityMinds GmbH*

The state-of-the-art overview is made available to the public under https://arxiv.org/abs/2104.14235 and is going to be published as introduction chapter in of Springer book on current research on safe AI.

1.3 E3.1.3 Final: Auflistung und Kategorisierung relevanter State-of-Research (Repository) (zur Veröffentlichung)

1.3.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Document*
Group/Cluster	
Type of content	*State of the art*
Classification level	*PU*

1.3.2 Description of the result

1.3.2.1 Motivation

In order to keep track of the state-of-the-art research after the initial coverage in result E3.1.2 (E3.1.2 Initialer State-of-Research-Report), a project internal literature repository for collaboration on related work should be set up.

1.3.2.2 Approach

The literature repository has been designed and set up on the basis of confluence by the AP 3.1 lead Fraunhofer IAIS, kindly supported by the project office EICT. The literature repository organizes entries in the structure identified in E3.2.1 (Clustering of the state-of-the-art) and provides a keyword based search facility, topic and keyword based indices, and timeline based overviews about latest additions. Templates for inserting new paper reviews in different predefined formats are provided as confluence macros. Entries made by partners are personalized to foster the immediate discussion between interested partners on specific research results on a personal level.

1.3.3 Result

The internal literature repository has been actively used by all project partners throughout the KI-Absicherung project.

Literature Repository

Information

Manual and Guideline: Brief introduction how to add papers.

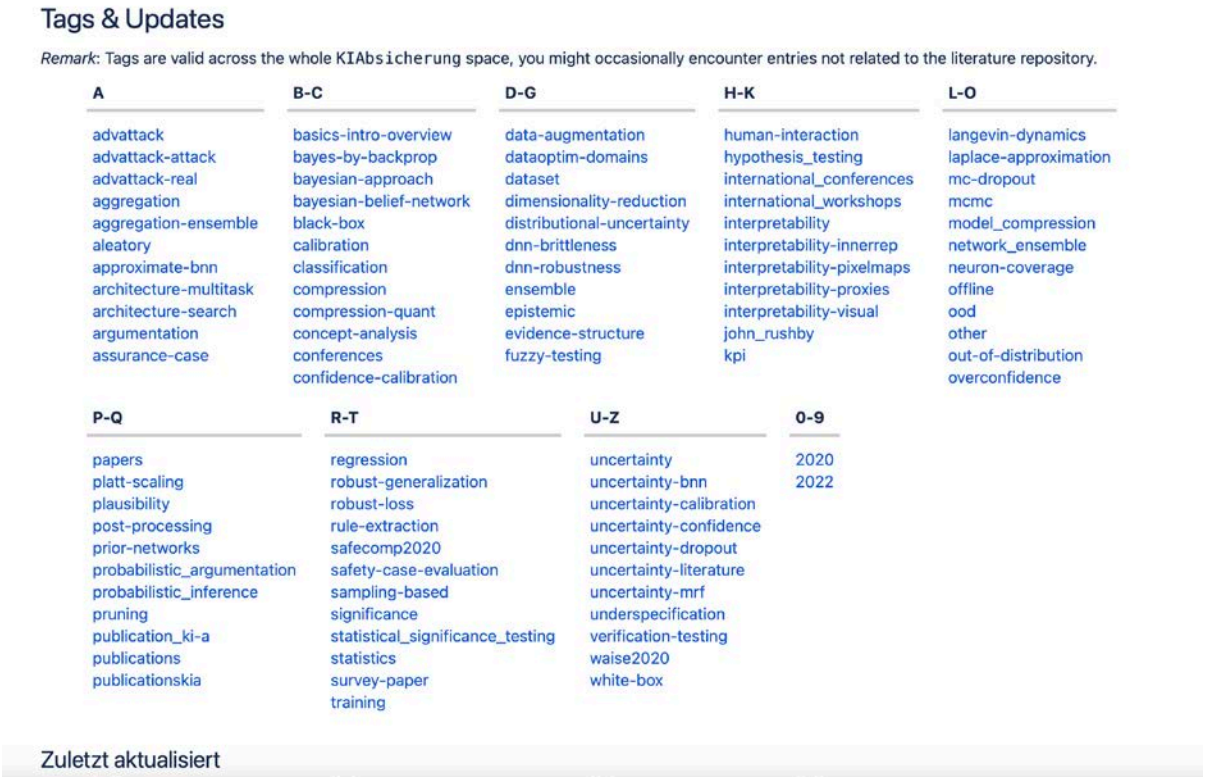
Responsible Person: AP3.1 Lead Michael Mock

The repository can be accessed either by the categorial list or filtered by the tags below.

People who create pages put some effort into it. Please let them know you appreciate their work by leaving a like or a comment if you found the page useful.

Categories

- Uncertainty
  - Generative Models
  - MC Dropout
  - Bayesian Neural Networks
  - Markov Random Fields
  - Confidence Calibration
- Interpretability
  - Visual Analytics
  - Intermediate Representations
  - Pixel Attributions
  - Interpretable Proxies
  - Generative Approaches
- Compression
  - Pruning
  - Quantization
  - Distillation
- Architectures
  - Building Blocks
  - Multi-Task Networks
  - Neural Architecture Search
- Dataset Optimization
  - Outlier/Anomaly Detection
  - Active Learning
  - Domains
  - Augmentation
  - Corner Case Detection
- Adversarial Attacks
  - Attacks and Defenses
  - More Realistic Attacks
- Aggregation
  - Ensemble Methods
  - Temporal Consitency
- Verification
  - Formal Testing
  - Black Box Methods
- Robust Training
  - Hyperparameter Optimization
  - Modification of Loss
  - (Domain) Generalization
- KPIs as such
- Others



Tags & Updates

Remark: Tags are valid across the whole `KIAbsicherung` space, you might occasionally encounter entries not related to the literature repository.

A	B-C	D-G	H-K	L-O
advattack	basics-intro-overview	data-augmentation	human-interaction	langevin-dynamics
advattack-attack	bayes-by-backprop	dataoptim-domains	hypothesis_testing	laplace-approximation
advattack-real	bayesian-approach	dataset	international_conferences	mc-dropout
aggregation	bayesian-belief-network	dimensionality-reduction	international_workshops	mcmc
aggregation-ensemble	black-box	distributional-uncertainty	interpretability	model_compression
aleatory	calibration	dnn-brittleness	interpretability-innerrep	network_ensemble
approximate-bnn	classification	dnn-robustness	interpretability-pixelmaps	neuron-coverage
architecture-multitask	compression	ensemble	interpretability-proxies	offline
architecture-search	compression-quant	epistemic	interpretability-visual	ood
argumentation	concept-analysis	evidence-structure	john_rushby	other
assurance-case	conferences	fuzzy-testing	kpi	out-of-distribution
	confidence-calibration			overconfidence

P-Q	R-T	U-Z	0-9
papers	regression	uncertainty	2020
platt-scaling	robust-generalization	uncertainty-bnn	2022
plausibility	robust-loss	uncertainty-calibration	
post-processing	rule-extraction	uncertainty-confidence	
prior-networks	safecomp2020	uncertainty-dropout	
probabilistic_argumentation	safety-case-evaluation	uncertainty-literature	
probabilistic_inference	sampling-based	uncertainty-mrf	
pruning	significance	underspecification	
publication_ki-a	statistical_significance_testing	verification-testing	
publications	statistics	waise2020	
publicationskia	survey-paper	white-box	
	training		

Zuletzt aktualisiert

1.4 E3.1.4 Final: Öffentlicher Zugang zu E3.1.3 (zur Veröffentlichung)

1.4.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*IT-infrastructure*
Group/Cluster	
Type of content	*Tool*
Classification level	*PU*

1.4.2 Description of the result

1.4.2.1 Motivation

The literature repository established in result E3.1.3 should be made available to the public.

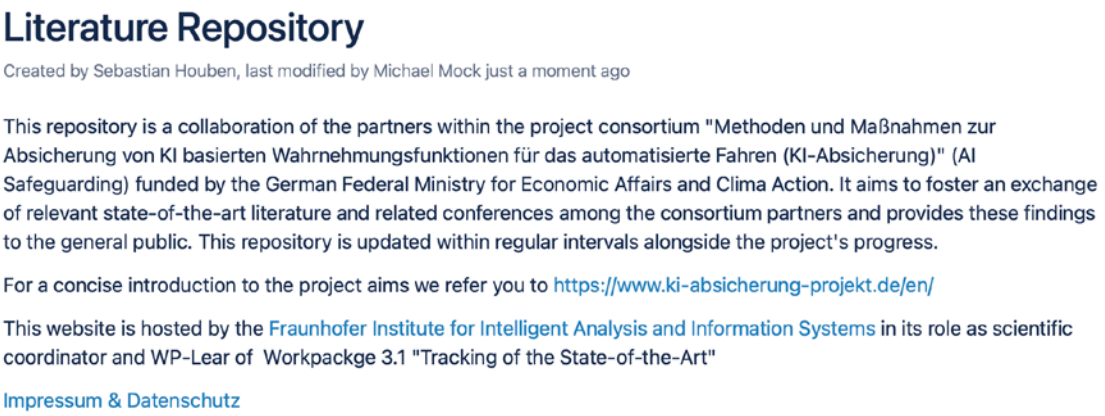
1.4.2.2 Approach

As the internal literature repository was intended to foster personal communication between members of KI-Absicherung and thus included personal names and comments, the consortium decided to make an anonymized derived version of this repository available to the public. All entries of the internal repository have been edited and possibly revised by the AP 3.1 lead Fraunhofer IAIS. Acknowledgments to the original authors have been added only in case they gave their consent to be mentioned in the public version. Also, the derived version has been approved following the standard publication process in KI-Absicherung.

1.4.3 Result

The literature repository is hosted by Fraunhofer IAIS and accessible under

https://jira.iais.fraunhofer.de/wiki/display/LiteratureRepositoryKIAbsicherung/Literature+Repository



Literature Repository

Created by Sebastian Houben, last modified by Michael Mock just a moment ago

This repository is a collaboration of the partners within the project consortium "Methoden und Maßnahmen zur Absicherung von KI basierten Wahrnehmungsfunktionen für das automatisierte Fahren (KI-Absicherung)" (AI Safeguarding) funded by the German Federal Ministry for Economic Affairs and Clima Action. It aims to foster an exchange of relevant state-of-the-art literature and related conferences among the consortium partners and provides these findings to the general public. This repository is updated within regular intervals alongside the project's progress.

For a concise introduction to the project aims we refer you to https://www.ki-absicherung-projekt.de/en/

This website is hosted by the Fraunhofer Institute for Intelligent Analysis and Information Systems in its role as scientific coordinator and WP-Lear of Workpackge 3.1 "Tracking of the State-of-the-Art"

Impressum & Datenschutz

1.5 E3.1.5 Final: Aktiver Austausch mit wissenschaftlichen Community (zur Veröffentlichung)

1.5.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Document*
Group/Cluster	
Type of content	*State of the art*
Classification level	*PU*

1.5.2 Description of the result

1.5.2.1 Motivation

This result intends to foster the scientific exchange and dissemination of the scientific results of the project KI-Absicherung.

1.5.3 Result

More than 50 scientific publications have emerged from the KI-Absicherung project, including publications on A ranked conferences such as ICCV, NeurIPS and CVPR. Best paper awards have been received for workshop publications on SAIAD (Safe AI for Autononous Driving) and WAISE (Workshop on AI Safety Engineering) workshops, which are co-located with CVPR and Safecomp conferences, respectively. The SAIAD workshop has been proposed, organized and held consecutively (3 editions during KI-Absicherung) by members of the KI-Absicherungs consortium. Scientists from academia and industry provided in total 15 chapters to the book project "Deep Neural Networks and Data for Automated Driving - Robustness, Uncertainty Quantification and

Insights Towards Saftety" with 450 pages edited by Tim Fingscheidt, Hanno Gottschalk and Sebastian Houben. The book proposal is currently under preparation for publication by Springer. A list of selected publications is available under https://www.ki-absicherung-projekt.de/en/publications.

2 AP3.2 Höherwertige Funktion KPIs für KI Funktionen

2.1 E3.2.1 Final: nur projektintern für KI Absicherung verfügbar

2.2 E3.2.2 Final: nur projektintern für KI Absicherung verfügbar

2.3 E3.2.3 Final: nur projektintern für KI Absicherung verfügbar

2.4 E3.2.4 Final: nur projektintern für KI Absicherung verfügbar

2.5 E3.2.5 Final: nur projektintern für KI Absicherung verfügbar

3 AP3.3 Funktional verändernde Methoden und Maßnahmen

3.1 E3.3.1 Final: Algorithmische Implementierung und Dokumentation für optimierte Datensatz-Selektion (zur Veröffentlichung)

3.1.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	Methods and Mechanisms
Classification level	*PU*

3.1.2 Description of the result

Motivation

This result focuses on methods and measures such as active or iterative learning, or uncertainty based methods for dataset optimization that aim at increasing the safety relevant metrics in particular by supporting the selection of relevant training data.

Approach

The individual development and implementation is left to the partners themselves. Synergies are identified by regular work package meetings with presentations and discussion of interim results.

3.1.3 Results

In particular, the following methods and measure are being developed:

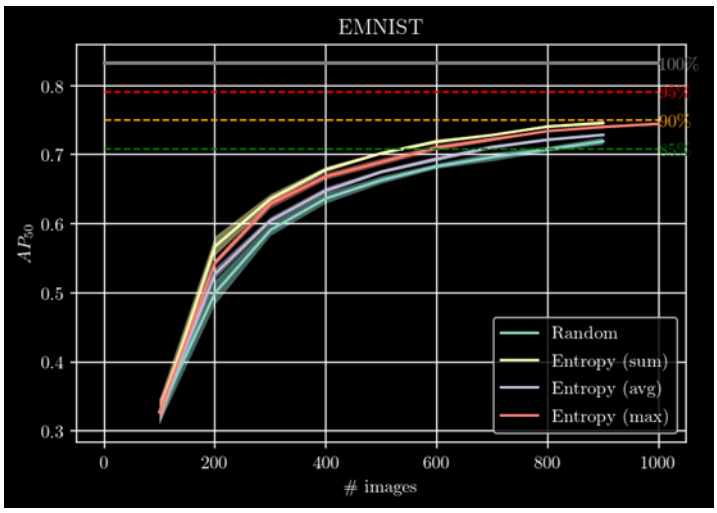
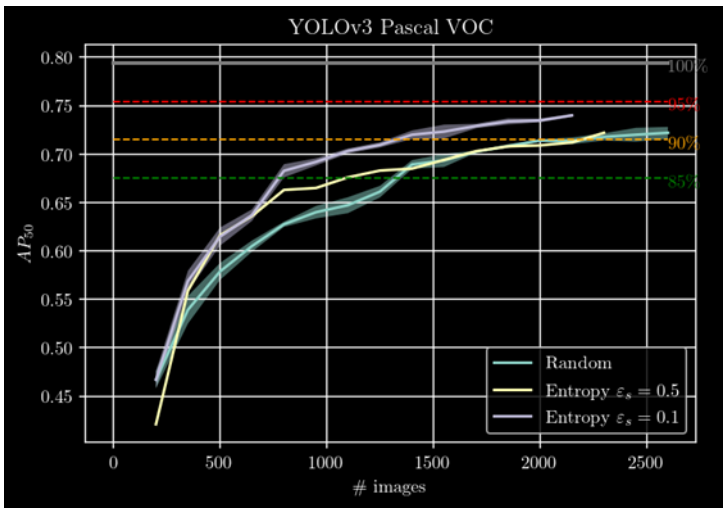
3.1.3.1 Bosch

Title (Mechanism ID)	Uncertainty Sampling (MECH-700701)
Leading and involved Partners (Name)	Robert Bosch GmbH
Mapping to Taxonomy Tree	Safe AI Mechanisms > Dataset Optimization > Active Learning > Uncertainty Sampling
Short Description of Mechanism	After training a neural network on an initial training set, the model will detect objects with a certain classification score. There are several possibilities (MC dropout, deep ensembles, SGLD, SVGD, snapshot ensembles) to assign a measure of uncertainty to this score. The model is now tested on an additional dataset. If the model is not performing well, one can pick the inputs the model was most uncertain about and add them to the training set. After retraining on the now extended training set, one would expect a

Title (Mechanism ID)	Uncertainty Sampling (MECH-700701)
	superior performance, as the most informative inputs were added. The procedure is repeated until the model performance saturates.
Used Data (train/val/test)	Synthetic dataset KIA-Tranche03 chosen due to availability
Used DNN / Task	SSD-r2-v2 chosen due to availability. The task is 2D object detection.
Main Safety Concern being adressed	Inferior model performance under limited labeling efforts • In reality, labeling input data is costly and therefore a minimal amount of data should be labeled. However, the model performance should not suffer from a smaller labeling budget. We therefore investigate if equal model performance can be reached using less data, and what selection criteria are most suitable for 'little labeled data' and 'high model performance'.
Summary of experiment results	Only 35% of the initial data are necessary to reach model performance convergence. None of the uncertainty based query functions improved over the performance obtained with random query
Summary of effectiveness compared to baseline / Level of Effectiveness	It seems that randomly adding more and more data simply until the model performance saturates is already a good (hard to beat) start. Beyond that, targeted improvements may be better than an uncertainty based query, where it is unclear from which part of the input space the new data is drawn.
Potential Evidences for an Safety Assurance Case	An optimized data selection avoids unnecessary data labeling, while retaining high model accuracy. Monitoring the training progress during the data selection ensures that data selection is not stopped before the model performance saturates. Beyond this, the method does not target specific performance limitations. A more targeted selection of images, or a weighting of the input images may be more beneficial

3.1.3.2 BUW

Title (Mechanism ID)	Active Learning for Object Detection (MECH-733147)
Leading and involved Partners (Name)	BUW
Mapping to Taxonomy Tree	Dataset Optimization > Active Learning > Uncertainty Sampling
Short Description of Mechanism	During the training of an Artificial Neural Network (NN) for Computer Vision tasks, a large amount of labelled data is required to reach satisfactory performance. Acquiring labelled data is a time-expensive process which needs a human to manually write ground truth labels for each single training image. However, not all images (samples) are equally useful to the NN to perform the task it is supposed to conduct. Active Learning aims to use a NN's current state (after having been trained on some amount of data) to choose (query) out of a pool U of unlabelled data those images to be labelled manually which will be of the most use (by some measure) to the NN. Unknown or lesser known (uncertain) concepts appearing in images from U will be queried with some preference by the NN. A ranking by uncertainty measures from inference has proven to provide a vital tool for Active Learning; uncertainty measures which can take on manifold forms, however, ranging from simple certainty scores provided by the direct output of the NN to sophisticated white-box methods taking into account information from all layers of the NN (therefore, providing more accurate uncertainty estimation). The application of meta classification in this project seeks to leverage the advanced confidence estimation developed in E3.4.2c and E3.4.4b together with prediction weighting strategies to optimize for information during each query step.
Used Data (train/val/test)	• MS COCO 2014, 2017 • PascalVOC 2007 • KITTI • BIT Tranche 3+4 • A2D2 • Internal toy model based on COCO-detection and (E)MNIST-classification
Used DNN / Task	• YOLOv3

Title (Mechanism ID)	Active Learning for Object Detection (MECH-733147)
	• Faster R-CNN@ResNet50 • RetinaNet@ResNet18 • TP1 Single Shot Detector (SSD)
Main Safety Concern being adressed	Active learning addresses data quality and rare critical situations by selecting data samples that are informative for the model under consideration. Under the assumption of correct model specification and optimization without heavy over-fitting, informativeness should correlate with uncertainty of the model and uncertain samples are hypothesized to include safety-critical samples.
Summary of experiment results	Experimental results include the refinement of different aggregation strategies of box-wise uncertainty scores over an image  as well, as the investigation of different prediction thresholds that can have a strong influence on the query result: 

Title (Mechanism ID)	Active Learning for Object Detection (MECH-733147)
Summary of effectiveness compared to baseline / Level of Effectiveness	The mechanism can, at present, not be compared with the TP2 baseline, as the active learning strategy is still under further development and will in the future be adapted to the project framework.
Potential Evidences for an Safety Assurance Case	Under certain conditions, active learning can be used to tune robustness of deep neural networks by active data selection. Currently, this project is aimed at reducing annotation cost through informed query strategies which can give rise to safety assurance-related evidences in the form of reliably reducing the required amount of data for the model to reach a certain performance. This hypothesized behavior of meta classification-based query strategies is planned to be tested on object detection benchmark data sets with a selection of standard object detection architectures. Validation can be achieved by repeating active learning runs with different random seeds, resulting in different starting and query distributions.

3.1.3.3 Continental

Title (Mechanism ID)	Hybrid Learning using Concept Enforcement (MECH-249938)
Leading and involved Partners (Name)	Continental
Mapping to Taxonomy Tree	Interpretability > Global Interpretation > Post-hoc Global Interpretation > Post-hoc White-box Inspection of Intermediate Representations Interpretability > Global Interpretation > Post-hoc Global Interpretation > Interpretable Surrogate > White- or Gray-box Monitoring > Model-independent Observer > Anomaly-detecting Observers > Output-only Observation > Multi-task Perception
Short Description of Mechanism	**Goal:** The goals of this method are: 1. Find **global representations of (binary) semantic concepts** in the DNN internal representation. 2. Use these to **extract per-sample information on the concepts** from DNN intermediate outputs (i.e. whether the DNN "saw" the concept or not). We here primarily consider **segmentation of concepts**.

Title (Mechanism ID)	Hybrid Learning using Concept Enforcement (MECH-249938)
	3. Use both global representations and per-sample information for **verification** purposes. 4. *(Potential future work enabled by this method)* Use the access to semantic concepts to improve internal representations and enforce requirements during training/fine-tuning. **Approach:** To find a vector representing a concept in a latent space, a **linear model** (convolution with 1 output channel) is **post-hoc attached to the layer output**. This *concept model* is trained **supervised** on few samples to **produce a segmentation mask for the concept**, i.e. to discriminate between concept and non-concept activation map pixels. After training, the **weights of the convolution represent the concept vector**, and the concept model predictions can be used as local concept information. Concretely, the so-chosen concept vector is the vector that *points into the direction of the concept* in the latent space. It is trained to have a small cosine distance (=high scalar product) with those activation map pixels of input samples in which the concept occurs. **Applications investigated here:** • Verification that **necessary concepts are "known"** to the DNN: Assess how successfully concept models can be trained. • Verification that internal representations of concepts encode **sensible concept similarities**: Cosine similarity of concept vectors can be compared with semantic similarity of concepts. • Verification that internal **representations are not biased**: Comparison of concept model performances for different training/test subsets. • Manual **inspection of internal logics**: Use concept outputs to post-hoc create interpretable proxies. • Verification and online-monitoring of compliance with (fuzzy) logic rules: Use concept outputs as (fuzzy) predicates in logic formulas. • *(Potential future work enabled by this method)* Improve insufficient internal representations of concepts by fine-tuning with a multi-task training scheme on concept embedding and rule compliance scores.
Used Data (train/val/test)	• First evaluation (cf. publication): • data: subset of German Traffic Signs Recognition dataset

Title (Mechanism ID)	Hybrid Learning using Concept Enforcement (MECH-249938)

	• concept segmentation labels: auto-generated (letters in speed limit signs)
	• Larger evaluation on pedestrians:
	  • data: MS COCO (restricted to images with commercial friendly license)
	  • concept segmentation labels: auto-generated from MS COCO 2014 keypoint annotations
	• KIA data Tranche 4 & 5 (see E3.3.1 Continental Hybrid Learning: Experiments)

Used DNN / Task	• Classifier architectures:
	  • First evaluation:
	  o 4-Conv-layer DNN (cf. publication)
	  • Object detection backend architectures:
	  o AlexNet (pytorch modelzoo)
	  o VGG16 (pytorch modelzoo)
	  o ResNeXt50 (pytorch modelzoo)
	• State-of-the-art object detection architectures:
	  • YOLOv3
	  • Mask R-CNN (pytorch modelzoo)
	  • EfficientDet D1 (commit 75e16c2f41, model ID tf_efficientdet_d1)
	  • KIA Opel SSD 300 releases 1, 3, and 3 with 512x512 resolution

Main Safety Concern being adressed	• SC-1.3 Incomprehensible behavior: Concept analysis can attach interpretable outputs and enable provision of interpretable proxies • SC-1.4 Insufficient Plausibility: Concept analysis can be used for diverse kinds of plausibility checks, including both offline and online ones. • SC-2.6 Unknown behavior in rare critical situations: Concept analysis applications to logical consistency checking can help in identifying symbolic descriptions/constraints of failure cases which can be used for data generation (see publication).

Title (Mechanism ID)	Hybrid Learning using Concept Enforcement (MECH-249938)
	• SC-3.1 Safety-aware metrics: Concept analysis can be used to define safety-related plausibility metrics like the logical consistency scores

Summary of experiment results	Summary of experiment results
	• A **concept embedding analysis** approach from literature was **adopted for object detection** via hyperparameter tuning and removal of pre-processing steps.
	• **Experiments** were conducted on a **variety of network architectures** (AlexNet, VGG16, ResNeXt50, Mask R-CNN, EfficientDet D1, KIA Opel SSD300, KIA Opel SSD512) to measure **how well** information on pre-defined semantic **concepts is embedded in the DNN** internal representations.
	• **Three applications** for post-hoc verification of DNNs were developed and demonstrated successfully that use concept models and additional concept outputs:
	  • verification of **DNN internal semantics** (similarity of concepts, absence of size bias),
	  • inspection via **interpretable proxy** models,
	  • monitoring of **compliance with fuzzy logic rules**.

Summary of effectiveness compared to baseline / Level of Effectiveness	**Statement on effectiveness:** • The method allows (for the first time) to access and verify symbolic background knowledge on DNN internal representations, both on single samples and globally. • The method does not introduce additional complexity: Considered representations are simply linear combinations of filters. **Statement on confidence:** • Data dependence: The quality of the concept analysis performance results is based on testing, and hence depends on an accurate test dataset. The results for logical consistency monitoring used a very simple ground truth definition of detection errors, hence have to be revised for practical application.
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis:** Verification of compliance with symbolic background knowledge based on visual semantic concepts (e.g. concept similarities or relations) requires access to the

Title (Mechanism ID)	Hybrid Learning using Concept Enforcement (MECH-249938)
	representation of semantic concepts within the DNN. This method provides such an access both globally and locally. 2. **Evidences:** Quantitative or qualitative verification, offline or online, whether • task-relevant concepts are embedded in DNN representations (and how well), and whether pre-defined semantic similarities are respected; • DNN internal representations are biased (i.e. better performing on specific test subsets); • DNN outputs and intermediate outputs are compliant with (fuzzy) logic rules (e.g. *arms* belong to *persons*). 3. **Further tests:** The results of the rule compliance checks could be strengthened by additional experiments on further rules, different ground truth (=detection errors) definitions, and potentially further networks. Especially the global rule compliance scores and cosine similarities should be observed for more models in comparison, including reference models which are known to be non-compliant. 4. **Reference to GSN tree:** see results of Work Stream 3: Incomprehensible Behaviour & Insufficient Plausibility
Link to papers	• Schwalbe, Gesina, and Martin Schels. 2020. "Concept Enforcement and Modularization as Methods for the ISO 26262 Safety Argumentation of Neural Networks." In *Proc. 10th European Congress Embedded Real Time Software and Systems*. Toulouse, France. https://hal.archives-ouvertes.fr/hal-02442796. • Rabold, Johannes, Gesina Schwalbe, and Ute Schmid. 2020. "Expressive Explanations of DNNs by Combining Concept Analysis with ILP." In *KI 2020: Advances in Artificial Intelligence*, edited by Ute Schmid, Franziska Klügl, and Diedrich Wolter, 148–62. Lecture Notes in Computer Science. Bamberg, Germany: Springer International Publishing. https://doi.org/10.1007/978-3-030-58285-2_11. • Schwalbe, Gesina. 2021. "Verification of Size Invariance in DNN Activations Using Concept Embeddings." In *Artificial Intelligence Applications and Innovations*, edited by Ilias Maglogiannis, John Macintyre, and Lazaros Iliadis, 374–86. IFIP Advances in Information and Communication Technology. Cham: Springer

Title (Mechanism ID)	Hybrid Learning using Concept Enforcement (MECH-249938)
	International Publishing. https://doi.org/10.1007/978-3-030-79150-6_30. • Schwalbe, Gesina, Christian Wirth, and Ute Schmid. 2022. "Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings." *CoRR* abs/2201.00572 (March). https://arxiv.org/abs/2201.00572.

3.1.3.4 Fraunhofer IAIS

Title (Mechanism ID)	Local Uncertainty Realism (MECH-030362)
Leading and involved Partners (Name)	Fraunhofer IAIS
Mapping to Taxonomy Tree	1) Safe AI Mechanisms > Uncertainty > (Approximate) Bayesian Neural Networks > Ensembles > Monte-Carlo Dropout 2) Safe AI Mechanisms > Uncertainty > Confidence Calibration > Calibration via Adaption of Training Process
Short Description of Mechanism	During training, we estimate the implicit uncertainty of a detection using the mismatch between prediction and Ground Truth as proxy for its variance. This estimate is included into the training and the dropout induced uncertainty adjusted to match the estimate. Uncertainty predictions are a crucial backbone of a neural networks "self-assessment" and find various applications, e.g. detection of out-of-domain input data, outliers or general flaws within the network prediction. To be of use for such purposes, uncertainty estimates need to be heteroskedastic, i.e. local, in the sense that they allow a statement on a specific given input and not only a global one on the confidence in general. The method proposed here tries to enhance such capabilities by directly estimating local data variances and explicitly including them into the training process of a well established uncertainty mechanism (MC Dropout).
Used Data (train/val/test)	• Release 4 used BTS data tranches 3-5, following otherwise the official train-test splits. • Release 5 follows the New, additional data split, decided 29.11.21 for SSD R3-v2 • Other datasets for extended evaluation (KITTI, A2D2, BDD100K, Nightowls, NuImages, SynScapes)

Title (Mechanism ID)	Local Uncertainty Realism (MECH-030362)
Used DNN / Task	• Release 4: Adapted Opel SSD (r1-v3) • Release 5: Adapted Opel SSD (r3-v2) • Other architectures for extended evaluation (SqueezeDet, ResNet) The task is 2D object detection.
Main Safety Concern being adressed	Unreliable Confidence Information
Summary of experiment results	The experiments clearly show that the mechanism leads to well calibrated regression uncertainty. However, transferring this approach to the regression component of an object detection task lead to noticeable performance decreases (up to 10%). Using our method we can significantly enhance the calibration of MC Dropout based uncertainty mechanisms. We tested this both for1D regression as well as for object detection (OD) tasks when considering "regression" of bounding boxes. Additionally, we performed out-of-distribution studies to confirm that the original properties of MC Dropout to react to epistemic uncertainty are still preserved. For this we either investigated artificial data splits (1D regression) or performed validation across unseen datasets (OD).
Summary of effectiveness compared to baseline / Level of Effectiveness	The deterministic baseline does not provide regression uncertainty and a direct comparison is therefore not possible. However, we can compare performance values and observe a degradation, which magnitude (up to 10%) depends on the used architecture, lower for SqueezeDet but noticeable for SSD or RetinaNet. As this did not occur for 1D regression we suspect this to be caused by the bounding box matching algorithm necessary to transfer the method to OD tasks. First experiments transferring the matching logic of BayesOD to our mechanism indicate that such degradation could be mitigated.
Potential Evidences for an Safety Assurance Case	There are two distinct cases where our method might be of use: 1. We argued that regression uncertainty may be helpful for downstream use. While not explored within the project MLSR20 states that DNN errors (in position) should be Gaussian distributed to facilitate temporal aggregation. Our method is designed to predict uncertainties such that the residual error should (ideally) follow a Gaussian. Including such predictions might therefore ease and robustify aggregation.

Title (Mechanism ID)	Local Uncertainty Realism (MECH-030362)
	2. We explored the robustness of our method under domain shift and noticed that uncertainty quality is much less affected than performance. The method might therefore also help in the handling of unforeseen scenarios or rare cases, e.g. new "types" of pedestrians or similar.
Link to papers	[1] Sicking, J.; Akila, M.; Wirtz, T.; Houben, S. & Fischer, A. "Characteristics of Monte Carlo Dropout in Wide Neural Networks", *ICML UDL Workshop,* **2020** [2] Sicking, J.; Akila, M.; Pintz, M.; Wirtz, T.; Fischer, A. & Wrobel, S., "A Novel Regression Loss for Non-Parametric Uncertainty Optimization", *Third Symposium on Advances in Approximate Bayesian Inference,* **2021** [3] Sicking, J.; Akila, M.; Pintz, M.; Wirtz, T.; Fischer, A. & Wrobel, S., "Wasserstein Dropout", arXiv:2012.12687v2, **2021**

3.1.3.5 Merantix

Title (Mechanism ID)	Active Learning for image segmentation (MECH-124694)
Leading and involved Partners (Name)	Merantix
Link to papers	2021-10-ERCVAD Virtual- Sreenivasaiah et al- MEAL: Manifold Embedding-based Active Learning
Main Safety Concern being adressed	Distributional shift over time (SC-2.5)
Mapping to Taxonomy Tree	Dataset Optimization > Active Learning > Uncertainty Sampling
Potential Evidences for an Safety Assurance Case	Model performance and hence the safety assurance can be achieved by improving the acquisition functions, i.e. by adding this objective to acquisition function during the acquisition phase, so that safety critical and informative images are selected by Active Learning and later used for training a robust model. Currently, the acquisition functions tested in this mechanism aim at reducing the generalization error of the model and increase diversity during the sample acquisition process. This can give rise to safety assurance-related evidences in the form of a training dataset which is highly informative and representative of the underlying distribution.

Title (Mechanism ID)	Active Learning for image segmentation (MECH-124694)
Short Description of Mechanism	We want to explore different techniques to decide what unlabeled data should be added to the dataset to maximize the performance of the resulting segmentation network. This is referred to as Active Learning. The common method from the literature is to choose those samples for which the model is uncertain about. The most common uncertainty measure used for the selection is Entropy calculated using the softmax outputs. We want to build a mechanism which effectively combines the informativeness and the representativeness of the unlabeled pool of data and compare the naive entropy methods to this approach. Since Active Learning is an iterative process meaning that the images are selected and dataset is incrementally increased, in real-world, this means that there will be a gap between subsequent acquisition steps as the selected images should be manually labeled.
Summary of effectiveness compared to baseline / Level of Effectiveness	• The baseline for this mechanism is a fully supervised model from TP1. The performance of the model with respect to mIoU when trained with full train dataset is higher compared to Active Learning. However, the effectiveness here is measured by the reduction in labeling required to achieve considerable performance. • In Camvid, with 5% of total data queried using AL, the mIoU is 81.6% of the fully supervised performance. • In Cityscapes, using just 1% of the total data, AL achieved 65% of the mIoU of the total dataset. • In KIA, using just 4.200 image patches, 70% of overall performance was achieved.
Summary of experiment results	• Experiments show that using minimal amount of carefully selected images, good performance of model can be achieved. • In experiments with CamVid, Cityscapes dataset, finetuning the backbone gives a marginal lift given the increased numbers of parameters that need to be trained. In both cases with and without finetuning, MEAL and MEAL-FT outperform the baselines given by Entropy Sampling and Random Sampling. • In case of KIA dataset, with finetuning the backbone, MEAL performs slightly better than other AL methods but the increase

Title (Mechanism ID)	Active Learning for image segmentation (MECH-124694)
	is not significant enough. One of the reasons could be because not enough tuning was done on hyper-parameters related to the acquisition function.

Used Data (train/val/test)	• **CamVid**: A dataset which is composed of street scene view images of size 360 × 480 with 11 semantic classes. It contains 367 (training), 101 (validation), 234 (test) images with labels.
	• **Cityscapes**: Composed of real-world street view images of dimension 2.048 × 1.024 with 19 semantic classes. The training set consists of 2.975 images with semantic labels and the validation set contains 500 images. Images are resized to 512 x 1.024.
	• **KIA dataset with segmentation labels** (BIT sequences from Tranche 3+4 train, val split from TP1):
	  • sequential data from the dataset ("`car-camera`")
	  • images resized to 300 x 300
	  • 34 classes

Used DNN / Task	This method uses semantic segmentation model provided by Intel in TP1. • DeepLabV3+ provided by Intel in TP1 with different backbone networks. We used MobilenetV2 as a backbone due to its resource constraint capabilities.

3.2 E3.3.2 Final: nur projektintern für KI Absicherung verfügbar

3.3 E3.3.3 Final: nur projektintern für KI Absicherung verfügbar

3.4 E3.3.4 Final: nur projektintern für KI Absicherung verfügbar

3.5 E3.3.5 Final: nur projektintern für KI Absicherung verfügbar

4 AP3.4 Introspektive Methoden und –Maßnahmen

4.1 E3.4.1 Final: Plausibilisierung (zur Veröffentlichung)

4.1.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Methods*
Classification level	*PU*

4.1.2 Description of the result

The work described here focused on saliency maps, which highlight the input regions that played a crucial role in DNN prediction. Numerous saliency maps methods were implemented and applied to the object recognition network SSD. A key result represents the transfer of saliency maps methods from the task of classification to detection. The application of saliency maps aims at increasing the plausibility. Validation of saliency map methods remains a major challenge.

4.1.2.1 Hochschule Ruhr West

Title (Mechanism ID)	Concept Validation (MECH-325047)
Leading and involved Partners (Name)	Fabian Küppers (Jan Kronenberger formerly)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Global Interpretation > Post-Hoc > Interpretable Surrogate > Black-Box
Short Description of Mechanism	The mechanism predicts the presence and contribution of visual concepts in the neural network in order to validate the classification of any detection (e.g.: How much does the *head* contribute to the prediction as *human* or the prediction as *human* is not valid due to the missing of the *torso*).

Used Data (train/val/test)	• Concept Model
	  • Cropped body part images as NumPy array with corresponding presence of the concepts and optional the body part segmentation.
	• Explainer
	  • Predictions of the teacher model. These consist of .json files with a list of all detections on a single image. The corresponding RGB Image has to be present as well. Each

Title (Mechanism ID)	Concept Validation (MECH-325047)
	detection is cropped from the original image and reshaped to 256x256x3. The confidence score is also needed.

Used DNN / Task	The approach consists of multiple models.
	• Teacher model: We used the detections made by the TP1 Mask RCNN E1.3.3d Implementierung der funktionalen Algorithmen: Instanz-Segmentierung
	old: *Teacher model: We used the detections made by the TP1 Opel SSD E1.3.3a Implementierung der funktionalen Algorithmen: 2D-Bounding Box*
	• Concept model: Multiple version of the concept model are used.
	  • Logits Only: ResNet50 from Torchvision
	  • Segmentations: DeepLabV3 from Torchvision
	• Explainer model:
	  • Encoder: ResNet50 from Torchvision
	  • Decoder: Gaussian Discriminant Analysis

Main Safety Concern being adressed	**Insufficient Plausibility (SC-1.4)**: The assumption of the method is that a prediction of a pedestrian detector (e. g. TP1 SSD) utilizes human interpretable visual concepts. The analysis has shown that the decisions of the TP1 SSD are not driven by these concepts. We found that the TP1 SSD detector is too underconfident in its predictions on the KI-A dataset tranche 3. This can be validated by D-ECE score and reliability diagram (see E3.5.2 Multivariate Confidence Calibration HRW MECH-043409). Additionally the TP1 SSD detector predicts only very few samples with a confidence score larger than 0.2. This problem was also detected in E3.5.2 Multivariate Confidence Calibration HRW MECH-043409.
Summary of experiment results	Our evaluation has shown, that the student (explainer) is able to reconstruct the teachers (TP1 SSD) confidence scores. The distributions were slightly modified to maintain the explainability. However, the results are not reliable because the data set is highly unbalanced. The TP1 SSD detections are also underconfident. Therefore, the evaluations have shown a low impact from the visual concepts on the confidence scores for the TP1 SSD.
Summary of effectiveness compared to	Unfortunately, the evaluations have shown a low impact from the visual concepts on the confidence scores for the TP1 SSD.

Title (Mechanism ID)	Concept Validation (MECH-325047)
baseline / Level of Effectiveness	
Potential Evidences for an Safety Assurance Case	The explainer network could be used to explain and to validate the predictions of a baseline object detection model. This could be useful especially for black-box models without any insights or out-of-distribution detection. However, in this setting, we have not been able to show sufficient evidences for our assumptions. Our results/experiments might be improved using a more representative network architecture.
Link to papers	CVPR2021 SAIAD Workshop https://openaccess.thecvf.com/content/CVPR2021W/SAIAD/papers/Haselhoff_Towards_Black-Box_Explainability_With_Gaussian_Discriminant_Knowledge_Distillation_CVPRW_2021_paper.pdf

4.1.2.2 Uni Heidelberg

Title (Mechanism ID)	E3.4.1c_HCI_Generative_Visualization_of_Learned_Representations
Leading and involved Partners (Name)	Andreas Blattmann
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Global Interpretation > Post-Hoc > WhiteBox
Short Description of Mechanism	The aim of this project is build a pipeline to visualize invariances of pose estimation networks. Such invariances often arise from discriminative training, Detection networks, such as bounding box detection and pose estimation, often develop invariances with respect to appearances of people, as their main goal is to identify the location and body configuration of persons, and appearance is not necessarily needed for this. Our solution is based on a Conditional Normalizing Flow Model, which enables efficient visualization of invariances directly in the pixel space. As conditional normalizing flows are known to disentangle their base representation from the conditioning (see also [1]), we condition our proposed model on the features of the investigated network and train the model to maximize the likelihood of the data. Thus the base representation of our Normalizing Flow will be independent of the DNN-features, and consequently contain all information which is not present in this conditioning. This enables visualizing in a given feature representation invariances taking the representation as a conditioning and drawing samples from the base distribution of the NF model, which is the standard normal. This is visualized in Fig.1 [1] Blattmann, A, Milbich, T, Dorkenwald, M and Ommer, B (2021). iPOKE: Poking a Still Image for Controlled Stochastic Video Synthesis. Proceedings of the International Conference on Computer Vision (ICCV) 2021
Used Data (train/val/test)	• Animals [2] train for training the models +val for validation • ImageNet train for training the models + val for validation • MV-Tranche 4 (train) + 5 (full) for training the models + Tranche 4 (val) for validation

Title (Mechanism ID)	E3.4.1c HCI Generative Visualization of Learned Representations
Used DNN / Task	• AlexNet, trained on the Animals Dataset [2] / Classification • FaceNet [3] / Face Recognition • ResNet-101 trained on ImageNet / Classification • MargiPose-V1, trained on KIA-Data MV Tranche 4+5 / Pose Estimation *[2] Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence* *[3] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2015*
Main Safety Concern being adressed	We address the safety concern of 'Incomprehensible Behavior' (SC-1.3). We validate this by visualizing specific feature layers of a given DNN in cases, where the model outputs an unexpected results, as e.g. when exposed to an adversarial attack, see Fig.3.
Summary of experiment results	**Generative Performance** To be able to make a principled statement on the obtained sample fidelity we compare our model to the recent latent space visualization method D&B [4]. For our aim of visualizing invariances, sample fidelity is of central importance, as artifacts and the like will hamper clear visualization. To this end, we evaluate FID scores for both methods on the Animals *[2]* dataset, when conditioned on different feature layers of AlexNet (each feature layer requires pretraining our model) trained on this dataset. We outperform the baseline for all of the investigated AlexNet feature layers. The figure shows corresponding qualitative results.

layer	conv5	fc6	fc7	fc8	output
ours	**23.6 ± 0.5**	**24.3 ± 0.7**	**24.9 ± 0.4**	**26.4 ± 0.4**	**27.4 ± 0.3**
D&B	25.2	**24.9**	27.2	36.1	352.6

Abbildung 1 Tab. 1: Quantitative evaluation and comparison with D&B

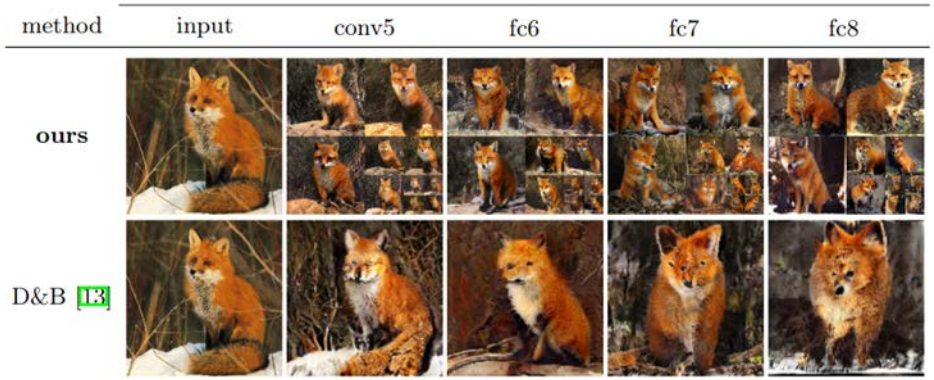


Fig. 1: Accompanying qualitative results for Tab. 1

Visualizing Invariances

The main goal of our method is to provide means to visualizing invariance of pretrained DNNs. Here, we present results of our model. To show the versatility of our proposed method and due to the late delivery of data with pose annotations we didn't only evaluate our model on pretrained pose estimation networks (as originally intended) but also on a variety of other models such as Classification networks (see AlexNet, in Tab.1 and Fig. 2 or Resnet101 in Fig. 4) and face detectors like FaceNet [3], for which we show visualizations in Fig, 2. As the base distribution of our NF model captures all information required to reconstruct the image beyond the conditioning , i.e. the DNN features, we can visualize the invariances of the feature representation by drawing different samples from the distribution for the same conditioning. We see that, for early layers, there are not many invariances in the network. After pooling, precise spatial information gets lost and invariances begin to increase. This is also indictated by the mean (second rightmost column) and variance of all the samples for a given conditioning feature representation.

Title (Mechanism ID)	E3.4.1c HCI Generative Visualization of Learned Representations

Fig. 2: Visualizing invariances of a pretrained FaceNet recongnition model

Moreover, our method allows for visualizations of adversarial attacks. We apply an FSGM [5] attack on a ResNet101 classifier pretrained on ImageNet and, by visualizing results from different feature layers, show how the attack propagates through the network.

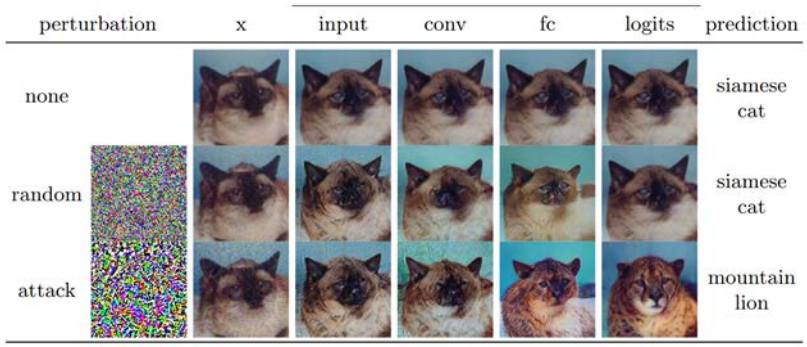


Fig. 3 Visualizing the spread of an adversarial attack on a ResNet101 classifier

Title (Mechanism ID)	E3.4.1c HCI Generative Visualization of Learned Representations

Finally we train the pose estimation network MargiPose-V1 on MV Tranche 4 and 5 and visualize the invariances in this representation in Fig. 4. We see that, as one would expect, the model becomes invariant to appearances of persons, as the appearance of the person can in most cases be neglected when intending to only predict keypoint locations. All samples are generated by a model conditioned on the bottleneck layer of the investigated network.

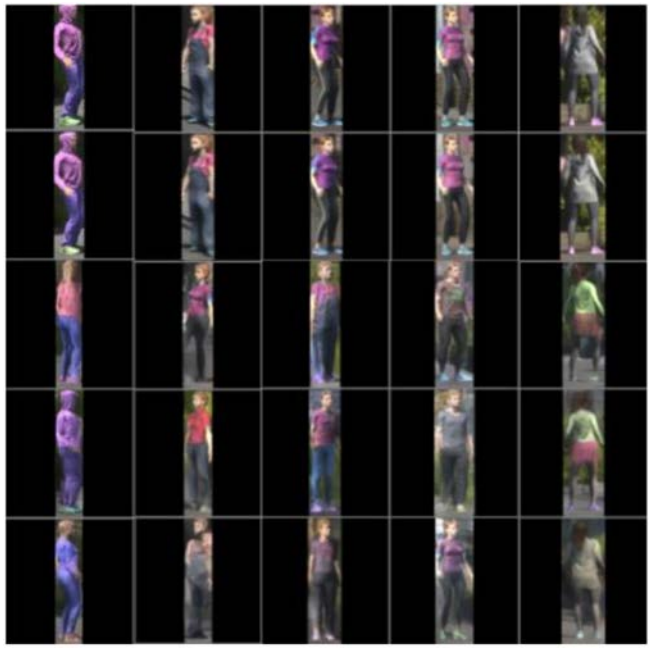


Fig. 4: Visualizations of invariances on the KIA dataset.

We conducted various other experiments to contribute a chapter in the joined book project of TP3. To review these and and obtain a concise overview of our method please check the corresponding chapter there.

[2] Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence

Title (Mechanism ID)	E3.4.1c HCI Generative Visualization of Learned Representations
	[3] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2015 *[4] Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks, NIPS 2016* *[5] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)*
Summary of effectiveness compared to baseline / Level of Effectiveness	We tested against M&D where we saw clear improvements of our method. Although the achieved FIDs can be still improved, by further investigating means to improve the autoencoder, we are satisfied with the effectiveness of the method against baselines.
Potential Evidences for an Safety Assurance Case	Why could this mechanism be valuable for an safety assurance case? The ability to post-hoc visualize the details of a failure case is a valuable property, which we are sure could be valuable for an assurance case, since explaining the reasons for a failure is the first step towards preventing it from happening again. Which evidences for a safety assurance case might be derived? Our method can help to prove that the model reasons in the way we want it to reason and not for some other, unintended cause which results nonetheless results in an expected output.
Link to papers	Our ECCV 2020 paper about this model Chapter in Joint Book Project, where the method is explained in detail and many additional experiments are presented.

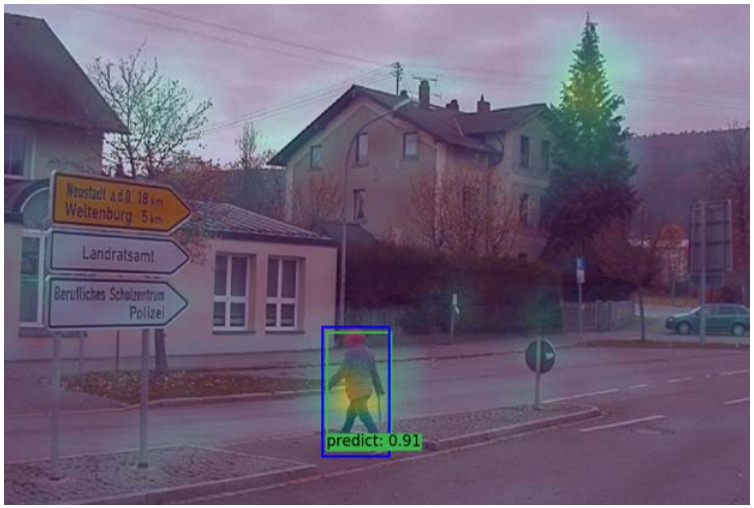
4.1.2.3 Bosch

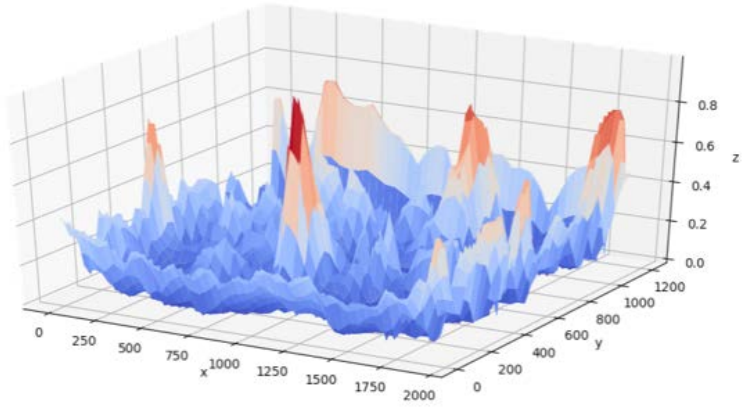
Title (Mechanism ID)	Comparing_Saliency_Maps_MECH-015039
Leading and involved Partners (Name)	Thomas Spieker
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Local Interpretation > Inspection of internal processing > Feature importance analysis > Saliency Maps > Gradient Based
Short Description of Mechanism	Saliency/attribution methods offer a way to explain the output of a neural network, by highlighting the important regions in the input image. Many different such attribution methods have been developed in the past. Usually they define a list of desiderata and analyze whether the new method fulfills these desiderata or not. However, it is unclear what a good saliency method is and what is not. It has been established that only rules, decision trees and linear models are truly explainable (see R. Guidotti et al.). Thus a saliency map needs to provide verifiable rules about the neural network, which can subsequently be proven (or disproven) in a quantitative analysis (e.g. to identify a car the model always checks for its wheels and/or license plate). We investigate different saliency methods, on if they qualify to provide these verifiable rules about the network. If a rule is identified from the saliency map, the data has to be relabeled, to include this rule. Afterwards the saliency maps of this relabeled test set can be calculated and compared to the labels.
Used Data (train/val/test)	images plus labels
Used DNN / Task	TP1.3 SSD-r1-v3
Main Safety Concern being adressed	Incomprehensible behavior (SC-1.3)

Title (Mechanism ID)	Comparing_Saliency_Maps_MECH-015039
Summary of experiment results	GradCAM clusters on all objects of the same class in an image. Integrated Gradients clusters on single detections. More fine-grained, which makes small deviations hard to interpret.
Summary of effectiveness compared to baseline / Level of Effectiveness	Compared to not knowing, which information the model uses, the method is quite effective. Obvious mistakes can be identified, but an automated, quantitative analysis is difficult/not feasible.
Potential Evidences for an Safety Assurance Case	Might help to clarify if the model bases its decisions on "reasonable" parts of the input. This will however be more of a debugging tool for development, since reliable/quantifiable and automated analysis is difficult
Link to papers	A Survey Of Methods For Explaining Black Box Models Towards Robust Interpretability with Self-Explaining Neural Networks Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization Axiomatic Attribution for Deep Networks

4.1.2.4 BMW

Title (Mechanism ID)	E3.4.1a_BMW_Heatmap_based_Attention_Consistency_Validation_MECH-156340
Leading and involved Partners (Name)	Fridolin Bauer (BMW) Tianming Qiu (fortiss)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Global > Post-hoc + inherently
Short Description of Mechanism	This mechanism proposes a self-explanatory model by inserting attention modules to existing pedestrian detection pipelines. The attention modules can generate heatmaps on pedestrians as visual interpretabilities. The mechanism also provides an off-line evaluation on the generated attention heatmap consistency.

Title (Mechanism ID)	E3.4.1a_BMW_Heatmap_based_Attention_Consistency_Validation_MECH-156340
Used Data (train/val/test)	KIA dataset Tranche 3+4
Used DNN / Task	n/a
Main Safety Concern being adressed	SC 1.3: Incomprehensible behavior SC 1.4: Insufficient Plausbility Heatmap can provide visual explanations for network detection behaviors.
Summary of experiment results	We show the heatmaps on the input images: 

Title (Mechanism ID)	E3.4.1a_BMW_Heatmap_based_Attention_Consistency_Validation_MECH-156340
	 • We don't improve the mAP results compared to the baseline model. But we provide self-explainary heatmap for visual interpretations. • The heatmap is shown as probability map on input images.
Summary of effectiveness compared to baseline / Level of Effectiveness	• mAP with attention module: 59.05% (BIT Tranche 3) • mAP without attention module: 61.55% (BIT Tranche 3)
Potential Evidences for an Safety Assurance Case	• Lower consistency distance index: higher alignment of attention distribution with ground truth bounding boxes. • The rate of 'ground-truth with attention and a predicted bounding box' demonstrate the interpretability of the model. The higher, the better. • Since the attention precision (TP/(TP+FP)) is around 60-70%, algorithm can trust more on these unsupervised learned attentions.

Title (Mechanism ID)	E3.4.1a_BMW_Heatmap_based_Attention_Consistency_Validation_MECH-156340
Link to papers	Papers: • HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis

4.1.2.5 EFS

Title (Mechanism ID)	Trainable_Concepts_MECH-852362
Leading and involved Partners (Name)	Tom Thielo Kai Fabi
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Local Interpretation > Visual Analytics
Short Description of Mechanism	This method uses a novel neural network architecture. The architecture is based on a conditional variational autoencoder (CVAE) and the gaussian discrimanant analysis (GDA) for generative classification. The GDA uses the latent representations of the CVAE's Encoder to perform its classification. The necessary distributions for GDA classification are learned by another network which is called Priorencoder. For clssification and reconstruction the same latent features are used. All components of this architecture are trained simultaneously. The typical CVAE loss terms are supplemented by a classification loss and an additional Kullback-Leibler-Divergence between the distributions of the Priorencoder and the latent representations of the input samples.
Used Data (train/val/test)	**Dataset** • German Traffic Sign Recognition Benchmark (GTSRB) **Reason / essential properties** • Real world mutli-category classification benchmark dataset.

Title (Mechanism ID)	Trainable_Concepts_MECH-852362
	• Efficient representation of real world data in latent space.
Used DNN / Task	Mechanism trains own model from scratch. The introduced mechanism can't be used to analyse existing models.
Main Safety Concern being adressed	**Adressed safety concern:** Incomprehensible behaviour (SC-1.3)
Summary of experiment results	The experiments showed that the model trained as expected and leanrs features suitable for classification and image reconstruction. Both classification and image reconstruction seem to work reasonably well. The hyperparameter tuning showed to be difficult, but can be handled by time consuming hyperparameter optimization approaches.
Summary of effectiveness compared to baseline / Level of Effectiveness	Method creates own models, so there is no baseline for comparison.
Potential Evidences for an Safety Assurance Case	1. Safety hypothesis: The generative part of the model allow partially human understandable classification decisions, especially in regards of the visualizations of the decoder. Our experimtens indicate human intuitive properties e.g. distance metrics between class distributions seem to correlate with human perceptive similarities. 2. Evidences for a safety assurance case: No guarantees can be derived. 3. Further tests:

Title (Mechanism ID)	Trainable_Concepts_MECH-852362
	Investigations for correlations between (modified) native trustwothiness scores and relevant safety KPI's have to be intensified.
Link to papers	Method represents a new concept. No external sources are available.

4.1.2.6 Umlaut

Title (Mechanism ID)	LRP_MECH-223169
Leading and involved Partners (Name)	Sabine Hug
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Local Interpretation > Inspection of internal processing > Feature Importance Analysis > Saliency maps > Back-propagation
Short Description of Mechanism	LRP is a method for the generation of attribution heatmaps for neural networks. The core idea is the backward propagation of so-called relevance $\mathbf{R_j}$ such that the sum of relevance is conserved layer by layer. Specific rules define how the relevance of the upper layer $\mathbf{R_j}$ is divided and propagated to the lower layer $\mathbf{R_i}$ and down to the input layer. In the basic rule (z-Rule) each input feature **i** is attributed relevance corresponding to the relative amounts that the input features $\mathbf{z_{ij}}$ contribute to the specific output features **j** of a layer. $$R_i = \sum_j \frac{z_{ij}}{\sum_j z_{ij}} R_j$$ There are several other rules suggested in the literature, a comprehensive overview (with heuristic use cases) can be found in Layer-Wise Relevance Propagation: An Overview.

Title (Mechanism ID)	LRP_MECH-223169
	The concept of layer-wise relevance conservation has to be modified in the context of residual connections as the notion of layer is not clearly defined and one uses the generalized concept of local relevance conservation (among connected neurons) $\sum_j R_{i\leftarrow j}^{(l,l+1)} = R_i^{(l)}, \qquad \sum_i R_{i\leftarrow j}^{(l,l+1)} = R_j^{(l+1)}$ cf. Opening the machine learning black box with Layer-wise Relevance Propagation.
Used Data (train/val/test)	Training not required. For evaluation we used TP2 data (BitTS Tranche 2 - 4) and A2D2 data.
Used DNN / Task	Training not required. For evaluation we used torchvision classifiers based on Resnet and VGG architecture as well as the TP1 SSD.
Main Safety Concern being adressed	SC 1.3: Incomprehensible behavior
Summary of experiment results	Pixel attributions generated via LRP with the specified rule show several different morphologies. A strong concentration of positive attribution values is not generally observed and we find several predictions for which a significant bulk of the associated pixel attributions lie outside the bounding box. The behavior that (according to the saliency map) important features lie outside the prediction is generally not problematic. The morphological clustering works, however to what extent it allows insights into model behavior is debatable. Attempts to leverage for understanding False Positives remains unsuccessful for Spray, whereas the Wasserstein k-means analysis is still ongoing.

Title (Mechanism ID)	LRP_MECH-223169
Summary of effectiveness compared to baseline / Level of Effectiveness	The method does not affect the baseline. We have serious concerns regarding our core assumption that LRP produces heatmap values, which reflect relevant areas or features in input space and the attributions exhibit a well-defined model and class-sensitive morphology. However, operating under the assumption that such a heatmap exists we believe that morphologically clustering the heatmaps is a good approach to obtain structural insights into the derived dataset of explanations.
Potential Evidences for an Safety Assurance Case	For DNNs global explainablity methods are not available (and possibly also not feasible). A meaningful aggregation of local explanations is therefore required to reduce human labor to reasonable levels. This can be achieved by clustering heatmaps by morphological similarity. The clustering provides groups of morphologically similar explanations, these can then be further analyzed by developers to possibly identify problematic behavior. The main obstacle to deriving any form of strong evidences is the lack of robust saliency map evaluation metrics.
Link to papers	Original research paper: https://doi.org/10.1371/journal.pone.0130140 LRP overview: https://link.springer.com/chapter/10.1007/978-3-030-28954-6_10 Lapuschkin thesis on LRP: http://dx.doi.org/10.14279/depositonce-7942

4.1.2.7 ZF

Title (Mechanism ID)	Heatmaps_DTD_ZF_MECH-918545
Leading and involved Partners (Name)	Firas Mualla
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Local Interpretation > Inspection of internal processing > Feature Importance Analysis > Saliency maps > Back-propagation
Short Description of Mechanism	The idea comes from the Taylor series in calculs: a function f(**x**) can be approximated as sum of the function's value at a reference point f($\mathbf{x}_0$) plus the inner product between the gradient of f at the reference point $\mathbf{x}_0$ and the direction from $\mathbf{x}_0$ to **x** (the vector **x** - $\mathbf{x}_0$). This linear approximation of f at **x** is valid as long as **x** is "not far" from $\mathbf{x}_0$. Translating this general mathematical concept to the domain of AI explainability: f(**x**) can be assumed to be the probability of class pedestrian as a function of an input image **x**. In the context of our project, f(**x**) is usually given as output of a deep neural network. The reference point $\mathbf{x}_0$ can be given as the same considered image **x,** but after (partially) removing the class evidence, for instance by smoothing part of the image pixels which contain the pedestrian object. As a result of this choice of $\mathbf{x}_0$, one can assume that f($\mathbf{x}_0$) is close to zero and $\mathbf{x}_0$ can be thus called a root of the function f(**x**). Consequently, the probability that the image **x** contains a pedestrian can be approximated as the inner product of the difference image **x** - $\mathbf{x}_0$ (each component is the difference between two intensity values) and the gradient of the model f evaluated at the reference image. This gradient is a vector, in which the i-th component is the derivative of f with respect to the intensity of the i-th pixel of $\mathbf{x}_0$. The concept sofar is not different from the linear approximation of Taylor series which is taught in secondary schools. The core idea is that the aforementioned inner product is kind of decomposition or redistribution of f(**x**) value on its variables (pixels) so that each pixel x has the following contribution: df/dx_0 (derivative) times x-x_0 (intensity difference to the corresponding pixel in the root image). The application of this technique directly to deep neural networks does not succeed due to at least two reasons: 1) it is difficult to estimate the gradient reliably in deep neural networks, 2) it is difficult to find a root point, since it

Title (Mechanism ID)	Heatmaps_DTD_ZF_MECH-918545
	has to be as close as possible to the original image, but at the same time a root of the model (yielding model's output close to zero). The idea of DTD is thus to apply the aforementioned "simple" Taylor decomposition in a deep network layerwise instead of the direct application to the input image, and hence the name deep Taylor decomposition.
Used Data (train/val/test)	No training is required
Used DNN / Task	No training is required. Networks are VGG classifiers and SSD based on VGG backbone.
Main Safety Concern being adressed	Incomprehensible behavior (SC-1.3)
Summary of experiment results	Experiments on the DTD for SSD show strong concentration of the heatmap energy inside the the bounding box (numerical results will come later). As mentioned above, not every feature inside the bounding box is stable (e.g. keychain), and not every feature outside it is unacceptable (e.g. sidewalk as auxiliary feature for pedestrians). The results of these experiments are thus to be considered not as a full-fledged proof, but as a kind of advanced sanity check.
Summary of effectiveness compared to baseline / Level of Effectiveness	
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis**: The method addresses the safety concern incomprehensible behavior (SC-1.3). It delivers some insights, as to what pixels contribute more to the final model's decision. Compared to other heatmap methods, it also claims kind of theoretical soundness due to the decomposition principle clarified above. 2. **Evidences for a safety assurance case**: Since the method gives a kind of "explanation", it increases the trust in the model. A special case of explanation is detecting cases when the deep learning model overfits to features that are not inherent in the object of concern but happen to cooccur with the object due to a bias in the

Title (Mechanism ID)	Heatmaps_DTD_ZF_MECH-918545
	training data. Detecting such cases and augmenting the training data accordingly increases the trust in the data and thus the model. 3. **Further tests**: More tests are needed for the robustness of the heatmap methods. For instance, some of these methods can be attacked to generate arbitrary heatmaps given almost the same model's input and output. So far no concrete implementation in safety argumentation in KI-Absicherung has been done.
Link to papers	• Method basis Explaining NonLinear Classification Decisions with Deep Taylor Decomposition • Tutrials website, e.g. to explain the difference between *decomposition* and *sensitivity analysis* (gradient): http://www.heatmapping.org/deeptaylor/ • Some explanations and details about the extension of DTD to SSD can be currently found in the ReadMe file in the repository

4.1.2.8 Block 1: General Information

Name of Mechanism	E3.4.1c_HCI_Generative_Visualization_of_Learned_Representations
Contact Person	Andreas Blattmann
Deliverable Number	E3.4.1
Version of Document	V1
Class of contribution (plausibility, robustness, white-box, …)	white box/ interpretability

Name of Mechanism	E3.4.1c HCI Generative Visualization of Learned Representations
Reason of choice of mechanism	Knowledge about on what a network focuses when making a prediction is of central importance for understanding the limit of applicability of deep methods. As our method enables visualizing this, it can help improve our understanding of NNS in plenty of ways.
Maturity level of the mechanism (e.g. new concept, known and acknowledged by the scientific community, already in use)?	Based on a peer-reviewed paper and results appreciated in scientific community.
Current state of development for the mechanism (e.g. under development, existing poof of concept/prototype, thoroughly tested)?	The following steps have been carried out: • Normalizing Flow model has been developed and trained on research datasets from • After testing we participated in the book project and wrote a chapter over this model • We also visualized invariances of pose estimation networks on the KIA data (see below)
Date of last code release	01/2021
Internal links (confluence, repo/commit, documentation of experiments with sacred)	Chapter in Joint Book Project, where the method is explained in detail and many additional experiments are presented.
Links to external sources (papers, data sets/labels)	Our ECCV 2020 paper about this model
Short description of functionality of mechanism to be understood by a	The aim of this project is build a pipeline to visualize invariances of pose estimation networks. Such invariances often arise from discriminative training, Detection networks, such as bounding box detection and pose estimation, often develop invariances with respect to appearances of people, as their main goal is to identify the location and body configuration of persons, and appearance is not necessarily needed

Name of Mechanism	E3.4.1c_HCI_Generative_Visualization_of_Learned_Representations
person with very limited knowledge about it.	for this. Our solution is based on a Conditional Normalizing Flow Model, which enables efficient visualization of invariances directly in the pixel space. As conditional normalizing flows are known to disentangle their base representation from the conditioning (see also [1]), we condition our proposed model on the features of the investigated network and train the model to maximize the likelihood of the data. Thus the base representation of our Normalizing Flow will be independent of the DNN-features, and consequently contain all information which is not present in this conditioning. This enables visualizing in a given feature representation invariances taking the representation as a conditioning and drawing samples from the base distribution of the NF model, which is the standard normal. [1] Blattmann, A, Milbich, T, Dorkenwald, M and Ommer, B (2021). iPOKE: Poking a Still Image for Controlled Stochastic Video Synthesis. Proceedings of the International Conference on Computer Vision (ICCV) 2021
State if the mechanism is to be used **online or offline**	offline

4.1.2.9 Block 2: Experiment Preparation

Prerequisites on model and data, State the **task**/s that is/are covered	Our model requires a pretrained detection network and a test dataset, where the model can be evaluated. In general, out approach is not limited to the exact type
Data sets and labels with respective version numbers which were used for training and/or evaluation. Reasons of choice and description of essential properties of data	We used MV Tranche 4 + 5 to train our model. For the investigated model, we waited for the final results on these datasets to be published, which are as of now (24.05.22) not present. See here

Prerequisites on model and data, State the **task**/s that is/are covered	Our model requires a pretrained detection network and a test dataset, where the model can be evaluated. In general, out approach is not limited to the exact type
Networks with respective version numbers which were used for training and/or evaluation. Reasons of choice and description of essential properties of data	The pose estimation networks MargiPose-V1, which is developed in E1.5.2. However since there were no pretrained weights for he KIA data avaialble until today, we trained MargiPose (see also the description here) on our own.
Input data format	Test dataset and pretrained DNN.
Output data format	Generated images showing the invariances of the pretrained DNN for given inputs.
Computational power needed for evaluation. Explain setup and amount of data which was used.	A single NVIDIA Titan 2080 Ti. For training the generative model, we used the train split of MV Tranche 4 and the entire MV Tranche 5 data. The resulting dataset consists of 50741 frames based on which we extract 148K crops, which serve as train inputs to our model.
Requirements on test data that are necessary to enable a thourough analysis of the mechanism and metrics (e.g. sensor noise, adversarial examples, occlusions, corner cases, outliers, etc.).	We assessed performance metrics of our generative normalizing flow model on the research dataset we tested the model on (see step 1 in the description of the current state in Block 1).

4.1.2.10 Block 3: Metrics and Evaluation

Description of metrics that were used for evaluation: metric name and full text description. Additionally: 1. Proof that there is correlation between the used metric and the effect which shall be measured (e.g. for robustification methods:	For the **generative model** we assess FID scores, which is the standard generative performance metric for measuring perceptual sample quality. It compares the real and the generated data distribution in the feature space of a pretrained classification network (inception v3). By assuming both real and generated to be normally distributed, the frechet distance can be calculated based on the estimated moments. As our proposed results does only aim at visualizing invariances and not actually improving performance we don't assess metrics for the investigated pose estimation network.

performance drop on corrupted images vs. normal images; e.g. for uncertainty methods: High uncertainty for False Positives, when standard DNN Output was confident) 2. Proper metric documentation, i.e. a. What does it intuitively mean? b. What are the extreme values of this metric? c. Which values are considered "bad, normal, good"? d. Is the scale of this metric linear? I.e. 2x value == 2x effect? Is the metric monotonic? e. Known limitations of the metric	
Description of metrics that were used for evaluation: diagram and/or formulas with units, parameters	see above.
Explanation of test design decision as well as description and explanation of **evaluation results**. Comparison of metric value before and after using the method, proof that value was improved on one or multiple test datasets by using this method.	Before visualizing the features of pretrained networks, we evaluate the generative performance of our proposed NF model and compare it with a recent method. Afterwards we visualize different layers of neural networks on research datasets before we also visualize invariances in the bottleneck layer of MargiPose-V1 trained on MV Tranche 4 and 5. **Generative Performance** To be able to make a principled statement on the obtained sample fidelity we compare our model to the recent latent space visualization method D&B [2]. For our aim of visualizing invariances, sample fidelity is of central importance, as artifacts and the like will hamper clear visualization. To this end, we evaluate FID scores for both methods on the Animals *[3]* dataset, when conditioned on different feature layers of AlexNet (each feature layer requires pretraining our model) trained on this dataset. We outperform the baseline for all

of the investigated AlexNet feature layers. The figure shows corresponding qualitative results.

layer	conv5	fc6	fc7	fc8	output
ours	$\mathbf{23.6 \pm 0.5}$	$\mathbf{24.3 \pm 0.7}$	$\mathbf{24.9 \pm 0.4}$	$\mathbf{26.4 \pm 0.4}$	$\mathbf{27.4 \pm 0.3}$
D&B	25.2	**24.9**	27.2	36.1	352.6

Abbildung 2 Tab. 1: Quantitative evaluation and comparison with D&B

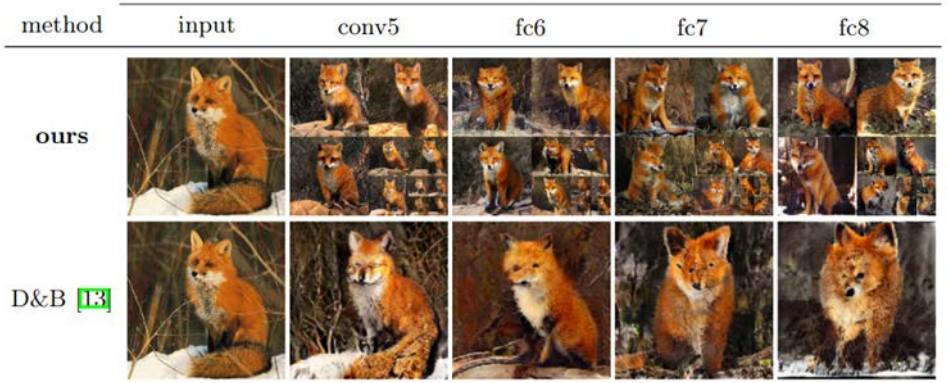


Fig. 1: Accompanying qualitative results for Tab. 1

Visualizing Invariances

The main goal of our method is to provide means to visualizing invariance of pretrained DNNs. Here, we present results of our model. To show the versatility of our proposed method and due to the late delivery of data with pose annotations we didn't only evaluate our model on pretrained pose estimation networks (as originally intended) but also on a variety of other models such as Classification networks (see AlexNet, in Tab.1 and Fig. 2 or Resnet101 in Fig. 4) and face detectors like FaceNet [4], for which we show visualizations in Fig, 2. As the base distribution of our NF model captures all information required to reconstruct the image beyond the conditioning , i.e. the DNN features, we can visualize the

invariances of the feature representation by drawing different samples from the distribution for the same conditioning. We see that, for early layers, there are not many invariances in the network. After pooling, precise spatial information gets lost and invariances begin to increase. This is also indictated by the mean (second rightmost column) and variance of all the samples for a given conditioning feature representation.

Fig. 2: Visualizing invariances of a pretrained FaceNet recongnition model

Moreover, our method allows for visualizations of adversarial attacks. We apply an FSGM [5] attack on a ResNet101 classifier pretrained on ImageNet and, by visualizing results from different feature layers, show how the attack propagates through the network.

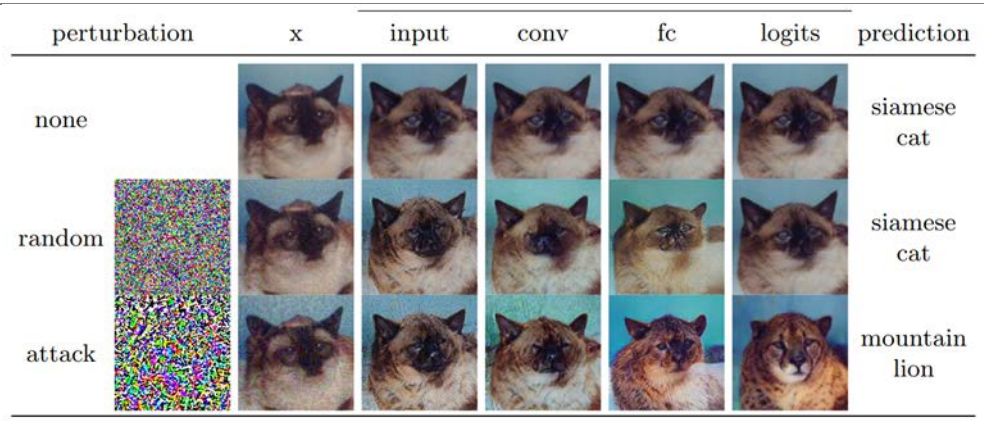


Fig. 3 Visualizing the spread of an adversarial attack on a ResNet101 classifier

Finally we train the pose estimation network MargiPose-V1 on MV Tranche 4 and 5 and visualize the invariances in this representation in Fig. 4. We see that, as one would expect, the model becomes invariant to appearances of persons, as the appearance of the person can in most cases be neglected when intending to only predict keypoint locations. All samples are generated by a model conditioned on the bottleneck layer of the investigated network.

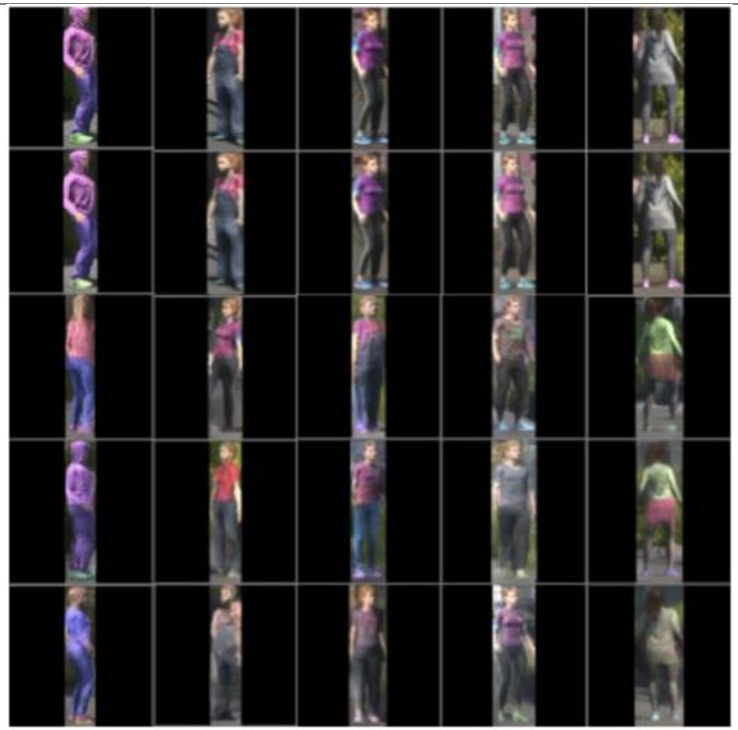


Fig. 4: Visualizations of invariances on the KIA dataset.

We conducted various other experiments to contribute a chapter in the joined book project of TP3. To review these and and obtain a concise overview of our method please check the corresponding chapter there.

[2] Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks, NIPS 2016

[3] Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence

	[4] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2015 *[5] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)*
Which safety concern is addressed and how is this validated?	We address the safety concern of 'Incomprehensible Behavior' (SC-1.3). We validate this by visualizing specific feature layers of a given DNN in cases, where the model outputs an unexpected results, as e.g. when exposed to an adversarial attack, see Fig.3.
Is there an **uncertainty** within the metric measurement or some kind of a **trustworthiness** score?	No
Benefits (i.e. Pros), drawbacks (i.e. Cons), limitations of using this mechanism and metric including assumptions that were made	Pros: • Our method enables efficient visualizations of the latent space of pretrained DNNs • By directly visualizing the features in the pixel space, we provide the most intuitive way of visualization for humans • Since we train the NF in the latent space of an pretrained Autoencoder, training does not take long Cons: • We have to retrain our model for every new layer we want to visualize. However since training does not take long this point is not that grave. • Our method does not provide a quantitative measure of performance for the investigated model.

Recommendations for **interactions with other methods** or **dependencies** on other methods.	No interactions with other methods
What needs to be done next in order to: (1) increase the effectiveness (2) gain more confidence in the results (besides "test on more data") and (3) bring this method into series production ? How much effort does this involve?	(1) To increase effectiveness, one could rework the implementation and use newest insights to build the MLP layers which are the backbones for each coupling block. Further scaling the train data and the number of parameters for the autoencoder would most likely lead to a stronger such model, which would have beneficial effects on the entire pipeline, as the overall achievable in general in bounded by the performance of the autoencoder. (2) Design a quantitative measure to measure the degree of invariances in an investigated representation. (3) The most difficult part of the method is training the autoencoder. Since NF models require a small representation as input to work well, we have to heaviliy compress the image by projecting it to a low dimensional feature space. This, on the other hand, hampers the overall generation quality.
State the biggest obstacles (e.g. convergence in GANs, usage of uncommon data points or additional labeling info, …) you had to tackle while implementation/evaluation. Could you solve them? If yes: How did you do it and how hard was it?	Training an autoencoder (see point (3), above) can be tricky due to the hight level of compression.

4.1.2.11 Block 4: Results, Effectiveness and Evidences

Ranking of mechanism regarding contribution to safety (1-5, 5 is highest)	4
Summary of experiment results	The aim of this project is build a pipeline to visualize invariances of pose estimation networks. Such invariances often arise from discriminative training, Detection networks, such as bounding box detection and

pose estimation, often develop invariances with respect to appearances of people, as their main goal is to identify the location and body configuration of persons, and appearance is not necessarily needed for this. Our solution is based on a Conditional Normalizing Flow Model, which enables efficient visualization of invariances directly in the pixel space.

As conditional normalizing flows are known to disentangle their base representation from the conditioning (see also [1]), we condition our proposed model on the features of the investigated network and train the model to maximize the likelihood of the data. Thus the base representation of our Normalizing Flow will be independent of the DNN-features, and consequently contain all information which is not present in this conditioning. This enables visualizing in a given feature representation invariances taking the representation as a conditioning and drawing samples from the base distribution of the NF model, which is the standard normal. This is visualized in Fig.1

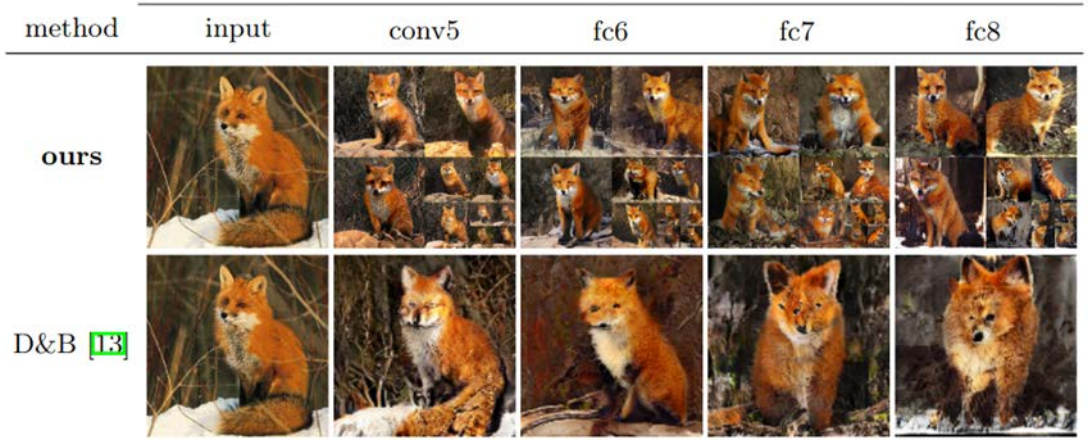
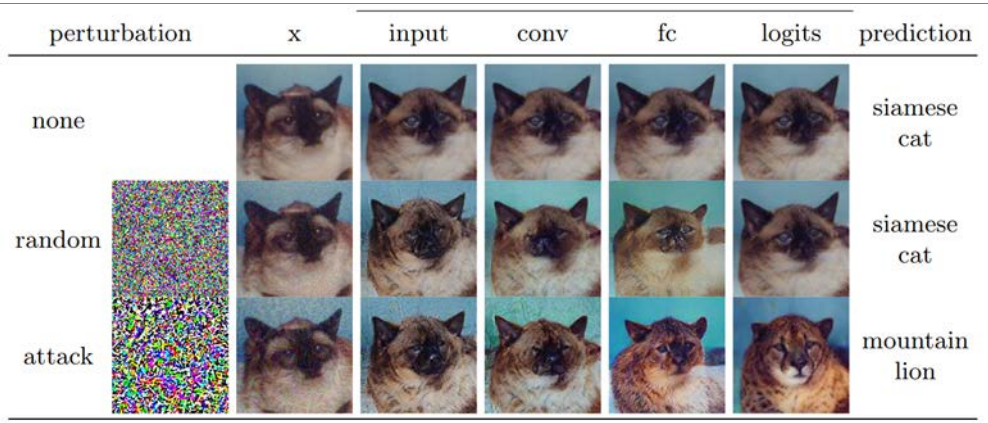


Fig. 1: Accompanying qualitative results for Tab. 1

Hence, in situations where a DNN fails, we can use our proposed framework to visualize different feature layers, and, hence, draw new conclusions. We visualize this in Fig. 2 for the case of an adversarial attack.

	 *Fig. 2 Visualizing the spread of an adversarial attack on a ResNet101 classifier* Due to its nice results, our proposed approach was also included as a chapter in the joint book project of TP3. [1] Blattmann, A, Milbich, T, Dorkenwald, M and Ommer, B (2021). iPOKE: Poking a Still Image for Controlled Stochastic Video Synthesis. Proceedings of the International Conference on Computer Vision (ICCV) 2021
Summary of gained findings/insights from experiments, including personal interpretation of results and statement on confidence/trust in the integrity of the experiments	We think that the provided model offers nice means to visualize specific feature layers of DNNs and, in doing so, can be used to search for explanations for failure cases. Since we conducted many experiments for different datasets and different discriminatively trained DNNs, we are certain that the method is applicable to wide range of feature extractors for various tasks.
Summary of effectiveness compared to baseline, including personal statement on	We tested against M&D where we saw clear improvements of our method. Although the achieved FIDs can be still improved, by further investigating means to improve the autoencoder, we are satisfied with the effectiveness of the method against baselines.

confidence/trust in the integrity of the effectiveness	
Potential Evidences for an Safety Assurance Case Please elaborate on: - Why could this mechanism be valuable for an safety assurance case? - Which evidences for a safety assurance case might be derived - What should be tested (in AP4.4) in order to derive a "strong" evidenc	Why could this mechanism be valuable for an safety assurance case? The ability to post-hoc visualize the details of a failure case is a valuable property, which we are sure could be valuable for an assurance case, since explaining the reasons for a failure is the first step towards preventing it from happening again. Which evidences for a safety assurance case might be derived? Our method can help to prove that the model reasons in the way we want it to reason and not for some other, unintended cause which results nonetheless results in an expected output.

4.2 E3.4.2 Final: Unsicherheitsmodellierung (zur Veröffentlichung)

4.2.1 Formal Classification

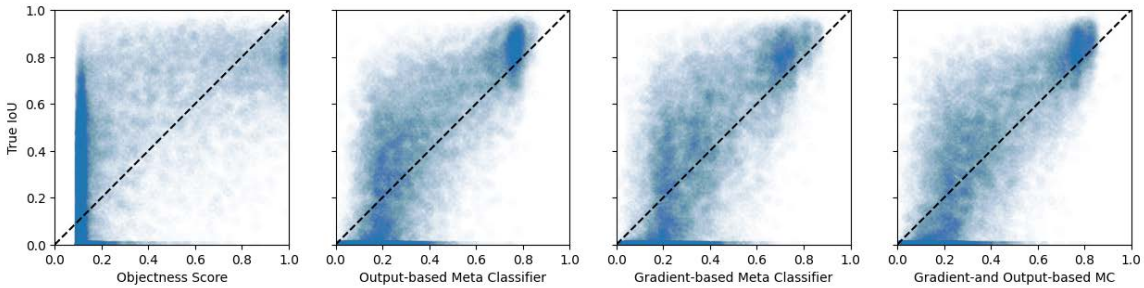
Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Mechanisms*
Classification level	*PU*

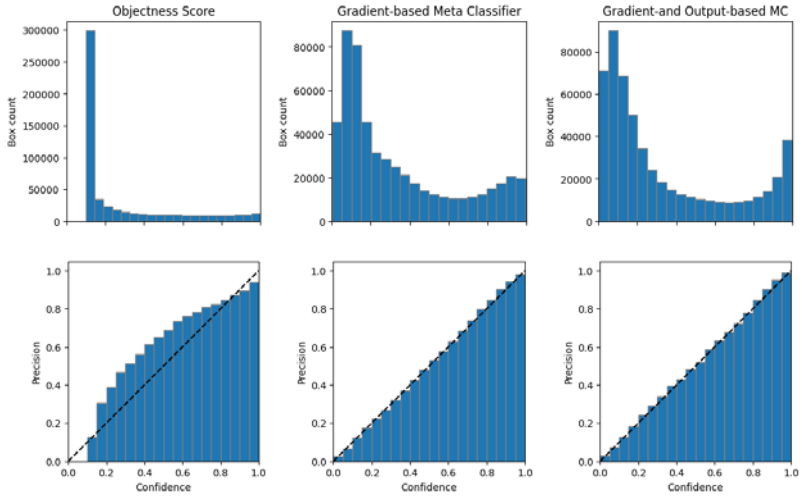
4.2.2 Description of the result

The implementation of reliable uncertainty estimation is an important component for safe AI. The work described here uses gradient metrics, additional layers, and the computation of Hessian matrices as part of its uncertainty estimation. Emphasis was placed on making the computation as efficient as possible so that it could be used for a real-time application. The effectiveness and comparison with other uncertainty methods has been demonstrated in numerous experiments.

4.2.2.1 Uni Wuppertal

Title (Mechanism ID)	BUW_Uncertainty_Quantification_by_Gradient_and_Activation_Information_MECH-173679 (MECH-173679)
Leading and involved Partners (Name)	Tobias Riedlinger (BUW)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Uncertainty > Frequentist Inference > Gradient-based Uncertainty
Short Description of Mechanism	Uncertainty measures are important not only for testing the performance of a network model under consideration but, moreover, for meta classification, meta segmentation and meta detection tasks, where an accurate and sensitive estimation of a model's uncertainty is vital. These approaches can then be used to trade uncertainty information and quantities computed on basis of those for improved model performance. Gradient uncertainty metrics use the magnitude of the self-learning gradient as a measure of confidence. This amounts to estimating the informativeness of a label coinciding with the network's prediction for the model parameters.
Used Data (train/val/test)	• MS COCO 2014, 2017 • PascalVOC 2007 • KITTI • MV+BIT Tranche 3+4

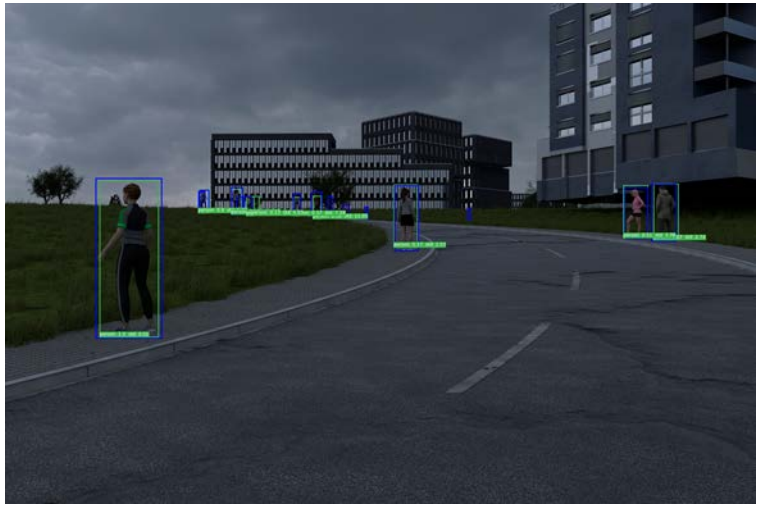
Title (Mechanism ID)	BUW_Uncertainty_Quantification_by_Gradient_and_Activation_Information_MECH-173679 (MECH-173679)
	• MV Tranche 5 test data • A2D2
Used DNN / Task	• YOLOv3 • Faster R-CNN • RetinaNet • Cascade R-CNN • TP1 Single Shot Detector (SSD)
Main Safety Concern being adressed	Gradient uncertainty metrics aim at mitigating unreliable confidence estimation in DNNs. We validate this in extensive experiments in the preprint, as well, as the meta classification and calibration experiments above. Depending on the application, uncertainty measures can be used to find labeling insufficiencies in the data set. We mention this as an outlook.
Summary of experiment results	• IoU estimates significantly better correlated with the true IoU of the prediction 

Title (Mechanism ID)	BUW_Uncertainty_Quantification_by_Gradient_and_Activation_Information_MECH-173679 (MECH-173679)
	• Confidence estimates based on gradient metrics is more reliable (better calibration) and confidence ranking of prediction is improved (AuROC, AP) over the score baseline  • Improved confidence ranking from meta classification carries over to object detection performance (so does the confidence calibration).
Summary of effectiveness compared to baseline / Level of Effectiveness	Compared with the objectness score baseline, meta classifiers based on gradient information show improved confidence ranking and mitigate calibration issues. Meta regression models allow for a reasonably reliable estimation of prediction IoU at inference time which the objectness score does not deliver.

Title (Mechanism ID)	BUW_Uncertainty_Quantification_by_Gradient_and_Activation_Information_MECH-173679 (MECH-173679)
Potential Evidences for an Safety Assurance Case	This mechanism was featured in EWS1 on Unreliable Confidence Estimation. The associated GSN includes the goals of improving confidence estimation performance in terms of AuROC and AP, improving confidence calibration in terms of ECE/MCE/ACE which have been investigated on some of the project data and shown to be improved by the use of gradient-based meta classification. Additional experiments improving the found evidences may include the improvement of object detection performance by meta classification, tests on other project data / versions of the SSD and other architectures.
Link to papers	Initial work conducted for classification tasks Preprint paper

4.2.2.2 BMW I

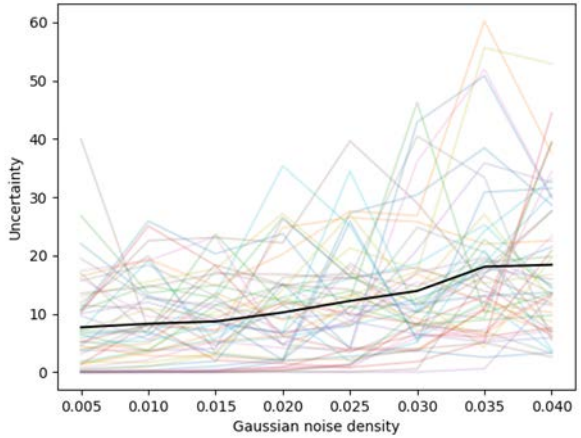
Title (Mechanism ID)	Bayesian_Uncertainty_with_Laplace_Approximation_MECH-962273 (MECH-191071)
Leading and involved Partners (Name)	Fridolin Bauer (BMW) Tianming Qiu (fortiss)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Uncertainty > (Approximate) Bayesian Neural Network
Short Description of Mechanism	Different from deterministic neural networks, Bayesian neural networks assume there is a probabilistic distribution on each weight. By calculating the distributions via Laplace approximation, distributions of neural network output can be derived and used for uncertainty estimation.
Used Data (train/val/test)	• KIA dataset tranches 3+4+5+6, with bounding box as annotations.

Title (Mechanism ID)	Bayesian_Uncertainty_with_Laplace_Approximation_MECH-962273 (MECH-191071)
	• Laplace approximation is an post-hoc method doesn't need to be trained. So we didn't really need to use a specific dataset to train the LA method, but any pre-trained model.
Used DNN / Task	n/a
Main Safety Concern being adressed	• Unreliable Confidence Information
Summary of experiment results	 Laplace approximation based Bayesian neural networks provide a post-hoc uncertainty estimation method. Thus it doesn't change prediction accuracy based on current training results. Apart from bounding box and the correspoding labels, BNN outputs the standard error of bounding box location as uncertainty estimation. In the above example image, the pedestrian far away from the vehicle shows higher uncertainty while the closer pedestrians detection shows less uncertainty.

Title (Mechanism ID)	Bayesian_Uncertainty_with_Laplace_Approximation_MECH-962273 (MECH-191071)

LA-BNNs provide an opportunity for real-time detetcion, each image takes 2-3 seconds for inference.

A positive correlation between Gaussian noise variance and uncertainty can be identified. The opaque black line is the average of all uncertainties of the bounding boxes from a randomly picked KIA dataset with 40 images with the same level of Gaussian noise.



Title (Mechanism ID)	Bayesian_Uncertainty_with_Laplace_Approximation_MECH-962273 (MECH-191071)
	The additive Gaussian noise comes from N(0, sigma^2), where sigma denotes the standard deviation. The Gaussian noise is added to the image and the output pixel value is restricted between 0 and 255 (The code implementation can be found in the bitbucket). On the synthetic image, the Gaussian distribution look unreal. The reason could be the rendering method is different for different part. Here we also put the real images from Kitti, where Gaussian noise looks more intuitive: 

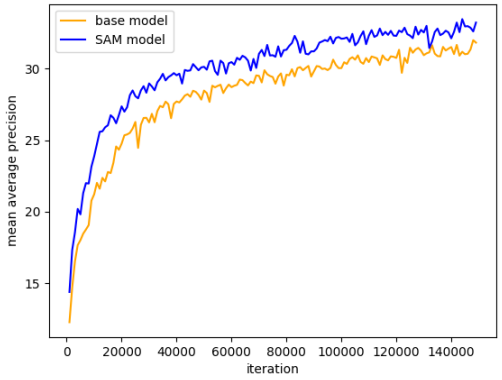
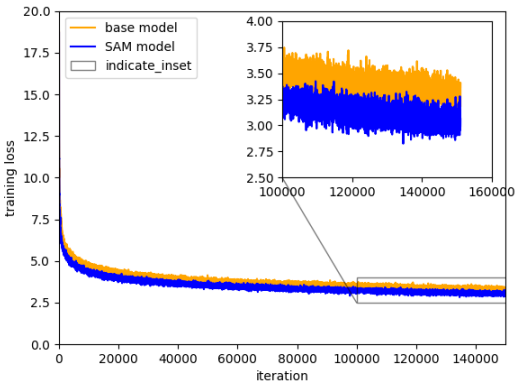
Title (Mechanism ID)	Bayesian_Uncertainty_with_Laplace_Approximation_MECH-962273 (MECH-191071)
Summary of effectiveness compared to baseline / Level of Effectiveness	• It is a relative new approach to show the uncertainty numerically, there is no comparable baseline.
Potential Evidences for an Safety Assurance Case	• Uncertainty estimation gives a confidencial indicator for predicted bounding box. • It provides the opportunity for human or down-stream decision making processes to evaluate the reliablity of current predicted results.

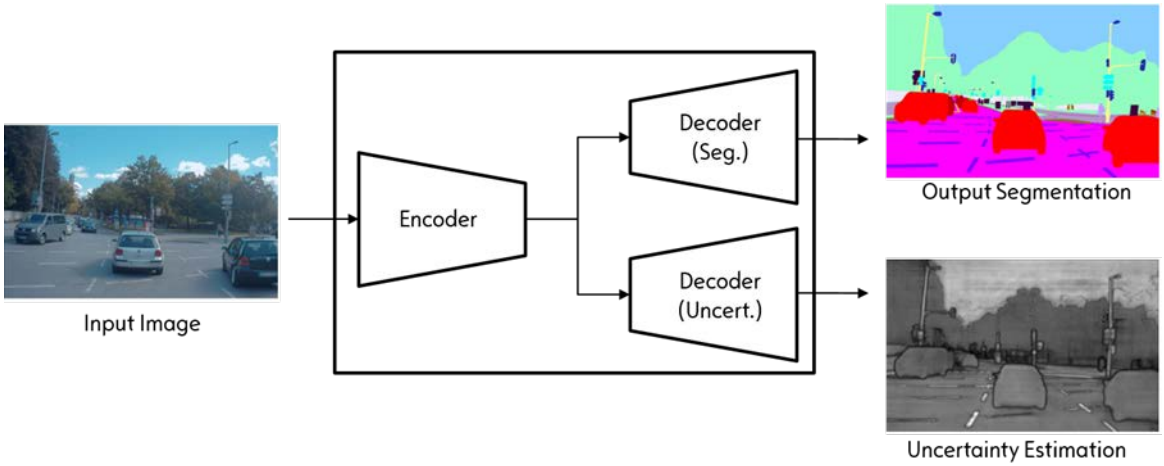
Link to papers	• Papers:
	  • Laplace Approximation with Diagonalized Hessian for Over-parameterized Neural Networks (accepted by NeurIPS workshop)
	  • 2022-10-ITSC Macau-Gui et al. -Laplace Approximation for Faster Uncertainty Estimation in Object Detection (submitted to ITSC22)

4.2.2.3 BMW II

Title (Mechanism ID)	BMW_Generalization_with_Gradient_and_Activation_MECH-963255
Leading and involved Partners (Name)	Fridolin Bauer (BMW) Tianming Qiu (fortiss)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Robustification > Corruption robustness > roboustness on model level > domain generalization

Title (Mechanism ID)	BMW_Generalization_with_Gradient_and_Activation_MECH-963255
Short Description of Mechanism	There is one valid assumption is that the loss function landscape of maximum a posteriori (MAP) point can affect the generalization ability. With gradient modification techinqiues during optimization, we seek a flatter landscape of MAP point.
Used Data (train/val/test)	KIA dataset Tranche 3 for training. Tranche 5 for testing. Since from Tranhce 3 to Tranche 5, there is a big domain gap, we don't use Tranche 5 at all during training.
Used DNN / Task	Opel SSD: ssd_300_kia_tp1_tranche_3+4_20210615_155652_97000
Main Safety Concern being adressed	As a metric for generalization ability description, it fits for SC-2.4: specification of the ODD. Insufficient generalization capability (FI-1)
Summary of experiment results	Results show that the sharpness of loss landscape may has relation to generalization performance of model. With the same training dataset, SAM model can achieve 2% improvement compared to the baseline model.

Title (Mechanism ID)	BMW_Generalization_with_Gradient_and_Activation_MECH-963255
	 
Summary of effectiveness compared to baseline / Level of Effectiveness	Within same training efforts, we get have 2% improvement on test dataset without putting target domain data into training procedures.
Potential Evidences for an Safety Assurance Case	• With proposed mechanism we can improve the generalization ability. • We also provide a metric to indicate the generalization ability. • Score of basin volumne shows consistent results as mAP on test dataset.
Link to papers	Papers: • On the importance of single directions for generalization • Sharpness-Aware Minimization for Efficiently Improving Generalization • Understanding Generalization through Visualizations

Title (Mechanism ID)	Opel_Lightweight_Aleatoric_Uncertainty_Estimation_MECH-687427
Leading and involved Partners (Name)	Ahmed Hammam (Opel)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Uncertainty > Aleatoric Uncertainty
Short Description of Mechanism	Given a pre-trained semantic segmentation deep neural network, this mechanism aims to produce aleatoric uncertainty estimations by adjusting the loss function of the DNN to incorporate Dirichlet distributions. 
Used Data (train/val/test)	• A2D2 Dataset • KIA Dataset
Used DNN / Task	DeepLab V3+ - Semantic Segmentation Intel ZF https://gitlab.com/kia2/tp1/ap1.3/semantic_segmentation/deeplabv3plus

Title (Mechanism ID)	Opel_Lightweight_Aleatoric_Uncertainty_Estimation_MECH-687427
Main Safety Concern being adressed	Unreliable confidence information (SC-1.1):
Summary of experiment results	This technique adds an extra part to the deep neural network in order to deliver the uncertainties it has. The uncertainties provided by this approach is aleatoric uncertainty, which is mainly due to the network itself.
Summary of effectiveness compared to baseline / Level of Effectiveness	The effectiveness, in this case, can't be quantified as the network is trained with an extra part added to the network to provide the uncertainty. However, we train the network to provide at least the same mean intersection over union provided by the original network.
Potential Evidences for an Safety Assurance Case	(1) This mechanism provides an uncertainty estimation of the network. this can be used as evidence to understand the deficiencies of the network. (2) With the use of the metrics correlating accuracy and certainty, a more quantitative result can be delivered providing evidence that can be used for safety assurance. (3) Wide testing with the network to fulfill and encompass a wide range of scenarios to properly test the reliability of the uncertainty estimation produced.
Link to papers	• Gast, Jochen, and Stefan Roth. "Lightweight probabilistic deep networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. • https://arxiv.org/pdf/1805.11327.pdf

4.3 E3.4.3 Final: Robustheitsprüfung durch Manipulation (zur Veröffentlichung)

4.3.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Mechanisms*
Classification level	*PU*

4.3.2 Description of the result

Title (Mechanism ID)	Photometric_Transformation_Robustness_MECH-418874
Leading and involved Partners (Name)	Tom Thielo
Mapping to Taxonomy Tree	Safe AI Mechanisms > Verification > Formal Testing > Proving Methods > Prove local properties
Short Description of Mechanism	The mechanism formally assesses the robustness of neural networks against photometric transformations such as scaling of luminosity or contrast. For this purpose the mechanism encodes the neural network, possible transformations and the relevant robustness properties as an equation system. Solutions for this equation system are counter examples. If such counter examples can be found the robustness property is refuted. Otherwise the property is proven for the given scenario.
Used Data (train/val/test)	• mechanism doesn't use labels (robustness is validated in respect to the original prediction) **GTSRB** (german traffic sign recognition benchmark): • real world images of german traffic signs • 43 classes, ~50.000 images, varying sizes, traffic signs are centered with a fixed border • Reason of choice: mechanism doesn't run on KI-A networks (run time issues). So separate, smaller networks were needed with according data complexity **GTSDB** (german traffic sign detection benchmark):

Title (Mechanism ID)	Photometric_Transformation_Robustness_MECH-418874
	• real world images of german streets with traffic signs in it • 43 classes, ~900 images, varying sizes, mutliple traffic signs per image • trained networks detect only one traffic sign at a time • Reason of choice: mechanism doesn't run on KI-A networks (run time issues). So separate, smaller networks were needed with according data complexity
Used DNN / Task	• no KI-A networks were used • Reason: to large complexity for the mechanism • exemplary networks are in the projects Repo
Main Safety Concern being adressed	Brittleness of DNNs (SC-1.2.1)
Summary of experiment results	Although in general adversarial examples are pretty easy to create, with the restriction to meaningful transformation it's not. I.e. we could observe robust behavior for simple neural networks for classification and object detection tasks on real world data. For state of the art neural networks the run time is not manageable. For smaller but still well performing networks this approach can perform formal verification. It is not trivial to give any recommendations in terms of a good trade of between complexity, the resulting run time and performance of the neural network.
Summary of effectiveness compared to baseline / Level of Effectiveness	The mechanism doesn't perform any model adjustments, consequently there can't be done any comparisons. The confidence in the integrity of the mechansim itself is high.
Potential Evidences for an Safety Assurance Case	1. The mechanism can prove the robustness for a set of properties (robustness against photometric transformations). Consequently, the mechanism adresses the safety concern 'Brittleness of DNNs (SC-1.2)'. 2. The mechanism provide formal proven robustness properties on specially designed test data sets (e.g. only test data with reasonable lumination if the robustness gainst luminosity

Title (Mechanism ID)	Photometric_Transformation_Robustness_MECH-418874
	transformations is verified) seem to be a strong evidence for the behavior of the neural network also for new data. 3. In order to perform meaningful formal verification with this method suitable datasets have to be derived depending on the property under test. So with a set of safety goals the related robustness properties can be derived and tested with suitable test data. (not applicable for KI-A data and networks!)
Link to papers	Paper *Formal Verification of CNN-based Perception Systems* https://arxiv.org/abs/1811.11373

4.4 E3.4.4 Final: Online Überwachung (zur Veröffentlichung)

4.4.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Mechanisms*
Classification level	*PU*

4.4.2 Description of the result

The main work in this UAP concerns anomaly detection, where objects or image regions in the input data are detected as anomalies. Anomalies are characterized by the fact that they were not preserved in the classes of training data or are very different in appearance. Major challenges of the approaches are a real time runtime as well as a good anomaly detection performance. The object recognition network SSD has been used in numerous experiments.

4.4.2.1 ZF

Title (Mechanism ID)	Semi_Supervised_White_Box_Anomaly_Detection_MECH-451461
Leading and involved Partners (Name)	Falk Kappel (ZF)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Robustification > Adverserial Robustness/ Adv. Attack Defenses > Robustification System Level > Out-of-Distribution Detection
Short Description of Mechanism	We investigate two approaches:

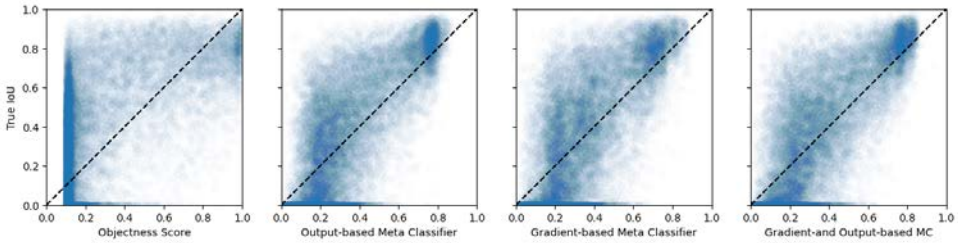
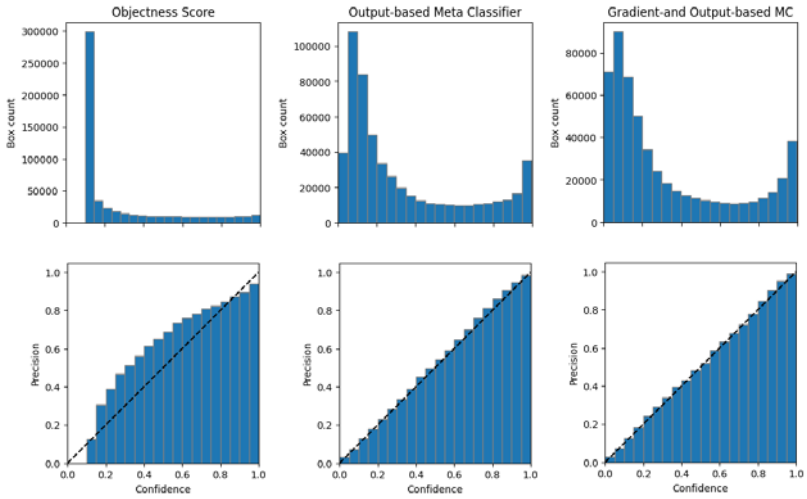
Title (Mechanism ID)	Semi_Supervised_White_Box_Anomaly_Detection_MECH-451461
	1. Generate activation statistics for several layers throughout the whole network. We choose all BatchNorm2d layers, as the activation statistics should stay more or less in the same range for the training data. Then try classification of anomalies based on these statistics by using different kinds of outlier detectors. 2. Train an Autoencoder to reproduce the features of an intermediate layer in the network. Then try to use either the reconstruction error or the output of the encoder part for further classification into anomaly/normal data with different kinds of outlier detectors. The outlier detectors used include One-Class SVM, Isolation Forests, and something we call "Dynamic Threshold" which is a simple check if each feature lies within a parametrizable multiple of the standard deviation. The outlier detectors only use the training data for fitting.
Used Data (train/val/test)	KIA Dataset Tranche 3+4+5+6 The official train data split is used for training the Autoencoders and Classifiers in order to not introduce bias. This data is considered as normal data that doesn't include anomalies. The official validation data split is used for evaluation and also considered as normal data.
Used DNN / Task	ssd_300_kia_tp1_tranche_3+4+5+6_20211220_083144_123000 was used for white box testing. The trained Autoencoder is an own design based on several blocks consisting of convolutional layers and ResNet-like skip connections. The convolutions often have a stride in the encoder part or are replaced by strided transposed convolutions in the decoder part. We make use of both ReLU activations and Concatenated ReLU activations. The Concatenated ReLU activation turned out to be particularly useful on the relatively small number of features in the encoding as opposed to a regular ReLU.
Main Safety Concern being adressed	Unreliable confidence information (SC-1.1) Unknown behavior in rare critical situations (SC-2.6)
Summary of experiment results	• Outlier detection using layerwise statistics of BatchNorm layers of a model under test, is already a good basis for classifying anomalies for that model. • The channel-wise calculated root mean square (reconstruction) error of an Autoencoder, trained to reproduce the features of an intermediate layer of the model under test, with an outlier detector

Title (Mechanism ID)	Semi_Supervised_White_Box_Anomaly_Detection_MECH-451461
	on top, can potentially work very well on anomaly detection and is the best method we found.
Summary of effectiveness compared to baseline / Level of Effectiveness	There is no comparable baseline.
Potential Evidences for an Safety Assurance Case	With anomaly detection we can identify data that differs greatly from the training data (out of distribution). We've proven good classification results at least for hand crafted anomalous samples. Even if the model under test were to perform good in some of those cases, it can never be predicted how well it performs in general when presented with anomalous data. Therefore, anomaly detection can be a reasonable addition to improve the measurement of uncertainty for predictions. 1. **Safety hypothesis**: The method addresses the safety concern 'Unreliable Confidence Information' (SC-1.1). It informs about the presence of anomalies in the input data or the internal state of the model under test. 2. **Evidences for a safety assurance case**: Since the method informs about anomalous situations, it can be used to adapt the uncertainty of the model. 3. **Further tests**: An evaluation for more possible anomalies of different magnitude has to be carried out. So far no concrete implementation in safety argumentation in KI-Absicherung has been done.
Link to papers	no

4.4.2.2 Uni Wuppertal

Title (Mechanism ID)	Uncertainty_Metrics_for_Meta_Classifiers_MECH-593230
Leading and involved Partners (Name)	Tobias Riedlinger (BUW)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Uncertainty > Frequentist Inference > Estimation based on Model (Softmax) Output

Title (Mechanism ID)	Uncertainty_Metrics_for_Meta_Classifiers_MECH-593230
Short Description of Mechanism	Uncertainty measures are important not only for testing the performance of a network model under consideration but, moreover, for meta classification, meta segmentation and meta detection tasks, where an accurate and sensitive estimation of a model's uncertainty is vital. These approaches can then be used to trade uncertainty information and quantities computed on basis of those for improved model performance. Output uncertainty metrics reflect the model confidence, class probability distribution and the prediction geometry. These quantities can be used to great effect in order to base advanced confidence estimates upon.
Used Data (train/val/test)	• MS COCO 2014, 2017 • PascalVOC 2007 • KITTI • MV+BIT Tranche 3+4 • MV Tranche 5 test data • A2D2
Used DNN / Task	• YOLOv3 • Faster R-CNN • RetinaNet • Cascade R-CNN • TP1 Single Shot Detector (SSD)
Main Safety Concern being adressed	Output uncertainty metrics aim at mitigating unreliable confidence estimation in DNNs. This was validated in extensive experiments in the external publication, as well, as the meta classification and calibration experiments above. Depending on the application, uncertainty measures can be used to find labeling insufficiencies in the data set. We mention this as an outlook.
Summary of experiment results	• IoU estimates significantly better correlated with the true IoU of the prediction

Title (Mechanism ID)	Uncertainty_Metrics_for_Meta_Classifiers_MECH-593230
	 • Confidence estimates based on gradient metrics is more reliable (better calibration) and confidence ranking of prediction is improved (AuROC, AP) over the score baseline  • Improved confidence ranking from meta classification carries over to object detection performance (so does the confidence calibration).
Summary of effectiveness compared to baseline / Level of Effectiveness	Compared with the objectness score baseline, meta classifiers based on entire output information show improved confidence ranking and mitigate calibration issues. Meta regression models allow for a reasonably reliable estimation of prediction IoU at inference time which the objectness score does not deliver.
Potential Evidences for an Safety Assurance Case	Evidences for this mechanism can be generated similarly to MECH-173679. The associated GSN includes the goals of improving confidence estimation performance in terms of AuROC and AP, improving confidence calibration in terms of ECE/MCE/ACE which have been investigated on some of the project data and shown to be improved by the use of output-based meta classification. Additional experiments improving the found evidences may include the improvement of object detection performance by meta classification, tests on other project data / versions of the SSD and other architectures.

Title (Mechanism ID)	Uncertainty_Metrics_for_Meta_Classifiers_MECH-593230
Link to papers	External publication

4.5 E3.4.5 Final: Offline Validierung (zur Veröffentlichung)

4.5.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Mechanisms*
Classification level	*PU*

4.5.2 Description of the result

4.5.2.1 BMW

Title (Mechanism ID)	Formale_Verifikation_von_Robustheitseigenschaften_Neuronaler_Netze_MECH-936804
Leading and involved Partners (Name)	Fridolin Bauer
Mapping to Taxonomy Tree	Safe AI Mechanisms > Robustness testing > Natural corruptions > Image artifacts
Short Description of Mechanism	We are interested in developing techniques that help us to provide robustness metrics for neural networks in the context of pedestrian detection.
Used Data (train/val/test)	KIA-Tranche3-BIT-MV
Used DNN / Task	Trained SSD (E1.3.3a, https://gitlab.com/kia2/tp1/ap1.3/2d-bounding-box/ssd)
Main Safety Concern being adressed	*Brittleness of DNNs. It is evaluated perturbing the input image with impulse noise, fgsm and Gaussian noise, computing the degree of perturbation and*

Title (Mechanism ID)	Formale_Verifikation_von_Robustheitseigenschaften_Neuronaler_Netze_MECH-936804
	then validating the result but computing the bounding boxes on the input image with the found degree of perturbation.
Summary of experiment results	Experiments with the number of images in the dataset, the epsilon values, global epsilon values, the different analysis for fgsm, impulse noise and Gaussian
Summary of effectiveness compared to baseline / Level of Effectiveness	The effectiveness was compared against different attacks used in the experiments. Also by taking the computed degree of perturbation for a given image in the dataset, taking the ground truth of that image and the adequate perturbation method. Perturb the image with the degree of perturbation found. Check if the ground truth was covered by the perturbed image. Then, decrease the degree of perturbation (epsilon) value, take the input image and its ground truth, perturb the image with the decreased epsilon. Check if the ground truth is covered by the perturbed image. The image should not be covered by the perturbed imaged with the decreased epsilon.
Potential Evidences for an Safety Assurance Case	For design time, to detect how robust is the network against the perturbations used.
Link to papers	Verifying neural networks : (https://arxiv.org/ftp/arxiv/papers/1903/1903.06758.pdf) Global Robustness : (https://www.ijcai.org/proceedings/2019/0824.pdf) Maximum resilience : (https://arxiv.org/abs/1705.01040) Framework : NNDK (https://github.com/dependable-ai/nn-dependability-kit)

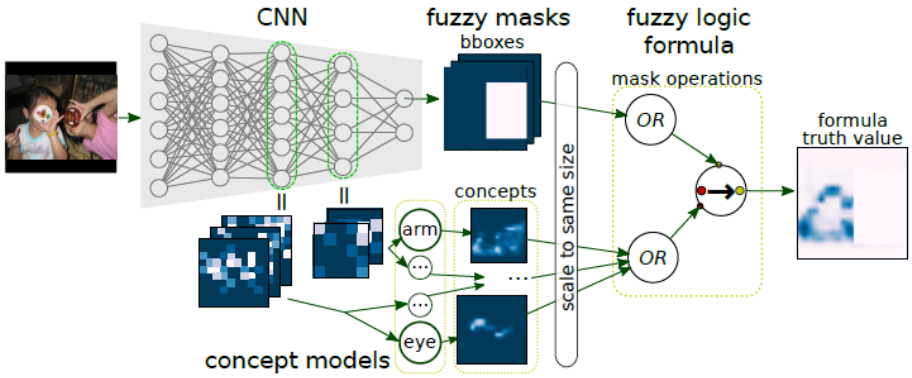
4.5.2.2 Continental

Title (Mechanism ID)	Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings (part of MECH-249938)
Leading and involved Partners (Name)	Continental

Title (Mechanism ID)	Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings (part of MECH-249938)
Mapping to Taxonomy Tree	Verification > Incomplete / Sampling-based Test Methods > Model Analysis / Review Monitoring > Model-independent Observer > Anomaly-detecting Observers > Output-only Observation > Multi-task Perception

Short Description of Mechanism	**Goal:** The goal of this method is verify in how far a trained DNN complies with given symbolic requirements that can be formulated as (fuzzy) first-order logic rules. An example for such a rule is the occlusion robustness rule "A head usually belongs to a person".
	
	Approach: The rules are formulated as (fuzzy) logic constraints on DNN outputs. The continuous truth value of these constraints then serves as a score in how far the DNN complies with the rule per sample or on a test set. In case the rule relates symbolic information that is not directly available from the DNN outputs (e.g. "is there a head"), this information is post-hoc attached as additional DNN output. This is done using the concept analysis method developed in E3.3.1 (MECH-249938). The method trains and attaches small linear models to DNN intermediate layers in a supervised fashion, given labels for the symbolic concepts of interest.
	
	Applications investigated here: The logical consistency score can be for
	• Offline verification of rule compliance, e.g. for comparison of DNNs;
	• For meta-classification of samples:
	  • Pre-selection of anomalous samples for manual inspection;
	  • Self-supervised online-monitoring of rule compliance.

Used Data (train/val/test)	• Data: MS COCO • Concept segmentation labels: auto-generated from MS COCO 2014 keypoint annotations
Used DNN / Task	State-of-the-art object detection architectures: • Mask R-CNN (pytorch modelzoo) • EfficientDet D1 (commit 75e16c2f41, model ID tf_efficientdet_d1)
Main Safety Concern being adressed	• SC-3.1 Safety-aware metrics: The logical consistency scores can serve as safety-related plausibility metrics. • SC-2.6 Unknown behavior in rare critical situations: Logical consistency checking can help in identifying symbolic descriptions/constraints of failure cases which can be used for data generation.

Title (Mechanism ID)	Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings (part of MECH-249938)
	• SC-1.4 Insufficient Plausibility: The rule compliance checking framework can be used for plausibility checks with respect to logical rules, both offline and online.
Summary of experiment results	Illustration of the developed framework for the rule *"IF arm OR … OR eye THEN person"*:  • A good setup for obtaining well-calibrated and -performing concept models is to use **BCE** for optimization followed by **threshold tuning**. • **A substantial amount of false negatives can be found**, as well as a couple of the few false positives, using the image level monitor and the occlusion robustness rule "A bodypart usually belongs to a person". • Pre-selection of corner cases wrt. logical consistency **helped uncover some symbolic logical issues** of Mask R-CNN and EfficientDet D1 (e.g. strong dependence on facial features, confusion of person with animal features). • **Fuzziness and calibration bring** slight **performance benefits** for the compliance meta-classifier; in case threshold tuning cannot be applied, they give a substantial boost compared to a Boolean uncalibrated baseline (which simply are mask unions and intersections).
Summary of effectiveness compared to baseline / Level of Effectiveness	**Statement on effectiveness:** • The method allows (for the first time) to access and verify symbolic background knowledge on DNN internal representations. • The method does not introduce additional complexity: Simple implementations of the rules are chosen, and the concept models are linear. **Statement on confidence:** • Data dependence: Both the preliminary concept analysis as well as the actual logical consistency scoring are sample-based. Hence, they

Title (Mechanism ID)	Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings (part of MECH-249938)
	depend on accurate training/test datasets. The results for logical consistency monitoring used a very simple ground truth definition of detection errors, hence have to be revised for practical application.
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis:** The DNN should comply with prior knowledge in the form of symbolic rules in order to ensure absence of logical issues (i.e., bias that can be described in natural language). This method allows to measure and compare this compliance. 2. **Evidences:** Logical consistency scores tell whether DNN outputs and intermediate outputs are compliant with (fuzzy) logic rules (e.g. *arms* belong to *persons*). 3. **Further tests:** The results of the rule compliance checks could be strengthened by additional experiments on further rules, different ground truth (=detection errors) definitions, and potentially further networks. 4. **Reference to GSN tree:** see Work Stream 3: Incomprehensible Behaviour & Insufficient Plausibility
Link to papers	• Schwalbe, Gesina, Christian Wirth, and Ute Schmid. 2022. "Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings." *CoRR* abs/2201.00572 (March). https://arxiv.org/abs/2201.00572.

5 AP3.5 Externe Methoden und Maßnahmen

5.1 E3.5.1 Final: Surveillance and coverage of input data (zur Veröffentlichung)

5.1.1 Formal Classification

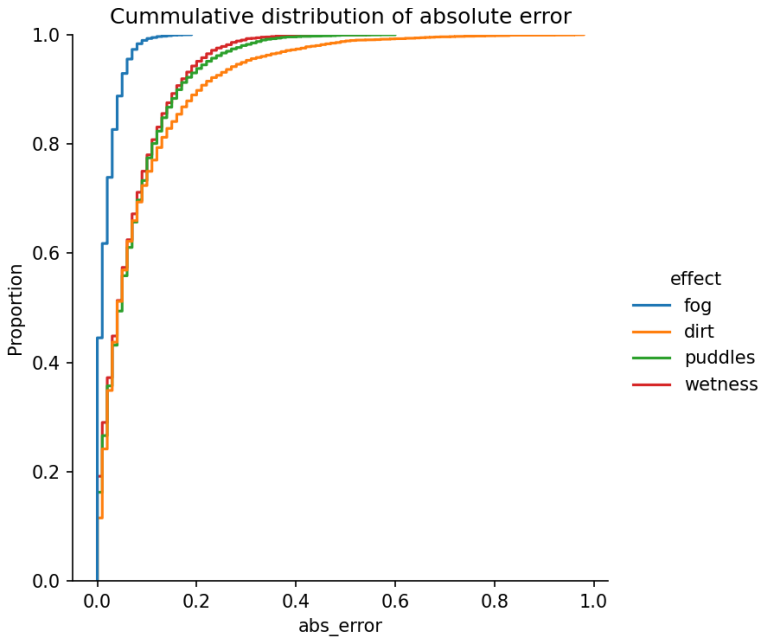
Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Methods and Mechanisms*
Classification level	*PU*

5.1.2 Description of the result

Title (Mechanism ID)	Input Monitoring (MECH-813975)
Leading and involved Partners (Name)	ZF (Falk Kappel)
Mapping to Taxonomy Tree	Monitoring > (model-independent) Observer > Anomaly Detecting Observers for Perf. Estimation > Input & Output Observation Seperately > Measure Distributional Shift Train / Test on input or output
Short Description of Mechanism	This method aims at detecting environmental effects on the input camera data. Those effects might lead to a decrease of accuracy of the networks output, therefore, knowledge of situations where they occur is of advantage. The effects to be detected have been chosen based on the availability within tranche 7 (fog, dirt, puddles, wetness). Resnet architectures of different depth (18 and 50) are used and compared against each other. The models are trained on Tranche 7 data to output 4 values corresponding to the predicted magnitude per environmental effect. The magnitude thereby is a value between 0 and 1 corresponding to the usage in synthetic data generation, with 0 signalizing that the effect is not present. A value of 1 corresponds to the effect occuring with maximum magnitude. The method can be used to contribute to an uncertainty value of the model under test (online usecase). For this a calibration per model under test and set of weights has to be done first. In an offline use case, the method can be used to filter a dataset and specifically search for images with the environmental effects covered here. This is useful when wanting to gather more data with such effects for training and improving the performance of the model under test in such scenarios. If on the other

Title (Mechanism ID)	Input Monitoring (MECH-813975)
	hand it is clear that the system does not have to operate under such conditions, the data can be excluded from the test set for a more meaningful evaluation.
Used Data (train/val/test)	Input Monitoring Data from KIA **Tranche 7** has been used, as this is the only Tranche of KIA data containing environmental effects. Only such 7 sequences of Tranche 7 that contain effects went into training validating and testing the model. Those sequences are the following: 'mv_results_sequence_0084_33190a04594547f3b126ec5d7be1ac8d' 'mv_results_sequence_0088_60dae98803fc4ad7bc9f51e023c6a1e6' 'mv_results_sequence_0089_6c1eeba5f5b84791a56e560bf27e86b2' 'mv_results_sequence_0090_d451639322d144a7b7d3b8bcfc4b681d' 'mv_results_sequence_0091_5b55471851cb441091578854dfa9da56' 'mv_results_sequence_0093_f377cafae31a450d883d6b0ea860dbdb' 'mv_results_sequence_0095_d26cfb610d064747b4599a1f2e150aa2' 'mv_results_sequence_0096_86a1c4741e7c49ef9286db7f5a4413bb' Correlation Analysis For the evaluation of the influence of environmental effects on the SSD performance all Tranche 7 sequences are used.
Used DNN / Task	Input Monitoring Resnet architectures of different depths (resnet18 and resnet50) have been used to predict the environmental effects and their magnitudes for input images. The output dense layer has 4 neurons corresponding to the 4 environmental effects present in Tranche 7 (fog, dirt, puddles, wetness). The activation function for the output is a Simgoid function, guaranteeing the outputs to be in a range from 0 to 1 signifying the magnitude of an effect as is the case for the annotations. During training, Binary Cross Entropy has been used as a loss function. For numerical stability the implementation uses torchs implementation of Binary Cross Entropy including the sigmoid activation rather than applying the simgoid function before. Correlation Analysis SSD (ssd_300_kia_tp1_tranche_3+4+5+6_20211220_083144_123000)

Title (Mechanism ID)	Input Monitoring (MECH-813975)
Main Safety Concern being adressed	Unreliable uncertainty information (SC-1.1)
Summary of experiment results	The graph below shows the cummulative distribution of the absolute error per class after 30 epochs of training. Looking at the distribution it can be seen what percentage of the predictions for a class are within what range of absolute errors. It becomes clear that fog is the effect that can be predicted most accurately with the other effects having a more similar cummulative distribution.



	mean error	std	max abs error
fog	-0.0073	0.0274	0.19
dirt	-0.0039	0.1386	0.98
puddles	-0.0028	0.1023	0.6
wetness	-0.0208	0.0922	0.55

Comparing the results achieved with the resnet50 and resnet18 architecture it becomes clear, that there is no big difference between them. As expected the resnet18 model can be trained more quickly give the smaller number of parameters. The bigger depth of the resnet50 architecture does not seem to increase the ability to tackle the task at hand on this dataset. It has to be kept in mind that this could behave differently on more complex real world data, however. Given the

Title (Mechanism ID)	Input Monitoring (MECH-813975)
	comparable performance as well as lower memory needs and better runtime, resnet18 is preferable here. In general the magnitude of environmental effect fog can be predicted with highest accuracy compared to the other effects. The highest error in prediction for class fog using the resnet18 architecture is 0.19 and the standard deviation 0.0274. Fog is the effect that can potentially have the biggest effect on the performance of the model under test, rendering camera algorithms in general useless from a certain density onwards. This method can be used to inform about those cases, allowing for an adaption of the uncertainty of the model under test in those scenarios.
Summary of effectiveness compared to baseline / Level of Effectiveness	Analyzing the correlation of environmental effects and the performance of the SSD show, that the performance decreases with increasing magnitude of those effects. Especially the presence of fog has a big impact on the performance. Given the decrease in performance of the SSD it is crucial to inform about such scenarios with high magnitude of environmental effects. The correlation of performance of the model under test (here SSD) and the effects present is different for each model architecture and training state, making an analysis necessary for every model under test.
Potential Evidences for an Safety Assurance Case	1: The performance of the model under test depends on the enironmental situation. Knowlegde about environmental effects can contribute to an uncertainty score of the model under test. Even the best model under test using images as input will suffer from physically not being able to see through fog from a certain density onwards. 2: The detection of environmental effects works especially well for fog. For other effects (dirt, puddle, wetness) roughly 80% of the predictions are within an absolute error of 0.2. The method, therefore, could be used in combination with other methods to indicate uncertainties. 3: The method should be tested on real world data. Further an investigation of the exact dependence of the SSDs perfomance on the single effects has to be done to derive concrete uncertainties.

5.1.2.1 Bosch

Title (Mechanism ID)	VAE Reconstruction Error Likelihood (MECH-356848)
Leading and involved Partners (Name)	Bosch (Laureen Lake)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Dataset Optimization > Outlier Detection/Anomaly Detection/Corner Case Selection > End-to-End (Anomaly) Classification > Input & Output Observation Seperately (e.g. measure distributional shifts) > Binary Classification OR Uncertainty > MECH-356848 also: Safe AI Mechanisms > Dataset Optimization > Outlier Detection/Anomaly Detection/Corner Case Selection > Learning Feature Representatons for Normality > Generative Models > MECH-356848
Short Description of Mechanism	The Variational Auto-Encoder (VAE) consists of two subnetworks, the encoder and the decoder. The encoder maps the input data (here an image) to a lower-dimensional feature space, also called latent space. The VAE predicts the parameters specifying the distribution of the latent variable, which is typically (and here assumed to be) multi-dimensional normally distributed. Accordingly, the output of the encoder network comprises the mean and the standard deviation of the learned normal distribution in the latent space. In contrast, the decoder network is the mirrored encoder network. It reconstructs the high-dimensional input from the latent space resulting in a reconstruction of the original input image. The VAE is trained with several images with the aim to improve regarding the loss function: MSE + beta * KLD The Mean Squared Error (MSE) is typically used to calculate the Reconstruction Error (RE) between the original and the reconstructed images while the Kullback-Leibler-Divergence (KLD) quantifies the difference between the predicted latent space distribution and its prior distribution which is assumed to be multi-dimensional standard normal.

Title (Mechanism ID)	VAE Reconstruction Error Likelihood (MECH-356848)
	Assuming that the VAE is less capable of reconstructing abnormal samples well and that abnormal samples differ from normal samples not only in input feature space, but also in the VAE latent space, we can use metrics to obtain input uncertainties for test images
Used Data (train/val/test)	BIT-TS Tranche 3, MV Tranche 2 (PoPC split)
Used DNN / Task	Performance on image level based on the following models have been used for correlation analyses: - TP1 DeepLabv3+ (trained from scratch on PoPC split by Andreas Baer) - TP1 Opel SSD
Main Safety Concern being adressed	The mechanism was developed to address following Safety Concerns: • Unreliable confidence information (SC-1.1) → VAE metric (e.g. Reconstruction Error) as input uncertainty correlating with model performance • Data distribution is not a valid approximation of real world (SC-2.1) → VAE metric as indicator for novelty of data point w.r.t. already collected training data (Active Learning) • Distributional shift over time (SC-2.5) → VAE metric as online monitor to detect when training distribution is left
Summary of experiment results	The PoPC experiments showed that the distribution of Reconstruction Errors of the out-of-ODD test data are more shifted to higher Reconstruction Errors compared to the distribution of Reconstruction Errors of the in-ODD test data. Also the average Reconstruction Error of out-of ODD test data is significantly higher than that of in-ODD test data. However, the two distributions of Reconstruction Errors for in-ODD and out-of-ODD test data still overlap, revealing that an out-of-distribution detection with the VAE on the basis of a single image alone is not possible.

Title (Mechanism ID)	VAE Reconstruction Error Likelihood (MECH-356848)

PoPC Data split (TP3 only)	Avg. Reconstruction Error
Train (in-ODD)	0.00732
Dev (Val) (in-ODD)	0.00755
TestA (in-ODD)	0.00922
TestB (out-ODD)	**0.01079**

Table 1: Average Reconstuction Error on different datasplits of PoPC data

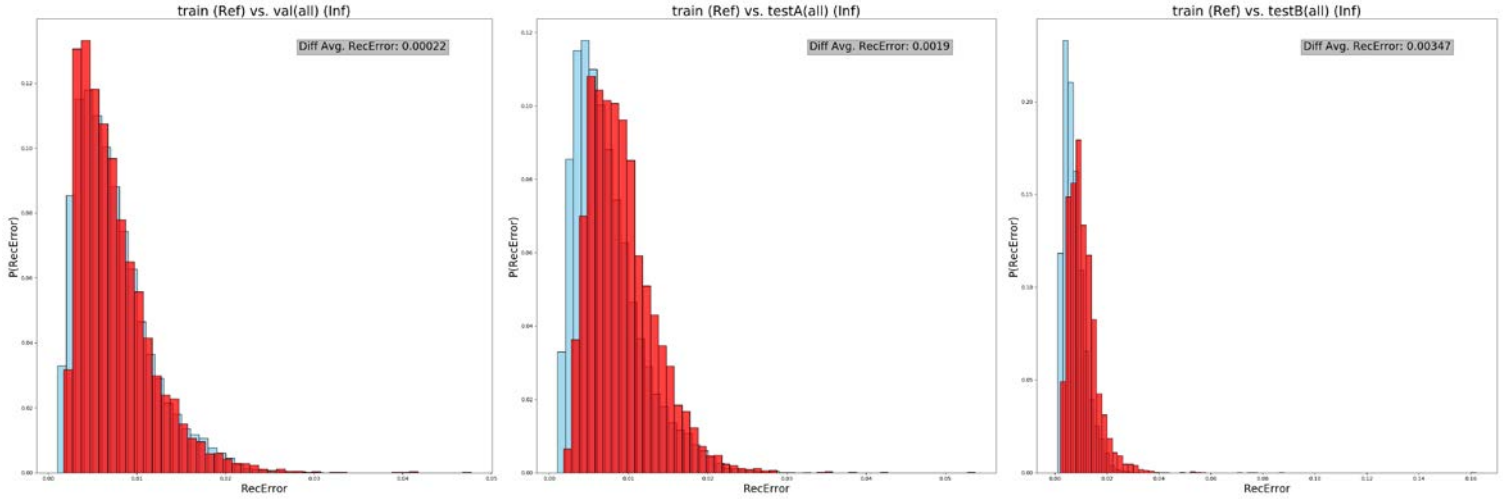


Figure 1: Histograms of VAE reconstructions errors for train (in-ODD; blue) vs. val (in-ODD; red), TestA (in-ODD; red) and TestB (out-of-ODD; red)

Title (Mechanism ID)	VAE Reconstruction Error Likelihood (MECH-356848)
	The sequence-based analyses showed different results than the analyses on average reconstruction errors on dataset level. They revealed that the average Reconstruction Error for some out-of-ODD sequences is comparable to that of in-ODD sequences and thus these sequences would be wrongly classified as being in-ODD. For the correlation analysis, Bosch considered scatterplots with regression lines and calculated different correlation statistics to investigate the relation between the mIoU (mean Intersection over Union) and the VAE Reconstruction Error (or mAP (mean Average Precision) vs. Reconstruction Error for SSD). The asumption was that there exists a negative correlation between these two metrics, i.e. bad performance, high Reconstruction Error and vice versa. However, the experiments showed a similar picture as already observed for the other sequence-based analyses: Some sequences show a significant negative correlation with the DeepLabv3+ & SSD performance, other sequences however revealed an insignificant negative or even a significant positive correlation. Apart from the experiments conducted for the PoPC, Bosch also performed the same analyses on the second implemented metric, i.e. the latent likelihood, and for the complete VAE loss metric (Rec. Err. + beta * Kullback-Leibler-Divergence) showing similar results.
Summary of effectiveness compared to baseline / Level of Effectiveness	As the mechanism is standalone and thus has no effect on the baseline performance, no effectiveness statement can be done in this regard. However, based on the quantitative evaluation on the PoPC split, the desired results that support mitigation of considered Safety concerns were not observed and thus the mechanism is - as is - not deemed to be effective for out-of distribution detection.
Potential Evidences for an Safety Assurance Case	This mechanism might be used at run-time to detect when we are leaving the training distribution (e.g. in- and out-of-ODD Detection). This is based on a metric value that indicates whether we are in or out of training distribution based on a threshold for the metric. The mechanism might also be used for Active Learning, i.e. to identify POIs in the field for complementing the training/test data. These applications are in scope of following saftey concerns

Title (Mechanism ID)	VAE Reconstruction Error Likelihood (MECH-356848)
	• Unreliable confidence information (SC-1.1) → VAE metric (e.g. Reconstruction Error) as input uncertainty correlating with model performance (for Active Learning OR onitoring) • Data distribution is not a valid approximation of real world (SC-2.1) → VAE metric as indicator for novelty of data point w.r.t. already collected training data (Active Learning) • Distributional shift over time (SC-2.5) → VAE metric as online monitor to detect when training distribution is left One might (in this project, we were not getting desried results in experments!) derive the evidence, that the novelty metric (here e.g. Reconstruction error) correlates with the model's performance in that it shows high reconstruction error in cases where the model performance is bad and vice versa. Also one might (in this project, we were not getting desried results in experments!) derive the evidence, that the mechanism detects when we are out-of distribution by showing a high reconstruction error for those samples. Tests should be made for the above mentioned evidences with target software (target detection algorithm) and several different datasets to get a stronger evidence. Also, we need a clear cut between in- and out-of-distribution in the experiment data as otherwise results might be misleading. It should also be observed, that VAE recontructions are good enough in order to be sure that it learned features from data correctly. Link to GSN file: https://confluence.vdali.de/pages/viewpage.action?pageId=14518653&preview=/14518653/16188450/kia-vae-minigsn.pdf

5.1.2.2 TU Braunschweig

Title (Mechanism ID)	Domain Mismatch Estimation (MECH-289269)
Leading and involved Partners (Name)	TU BS (Andreas Bär)

Title (Mechanism ID)	Domain Mismatch Estimation (MECH-289269)
Mapping to Taxonomy Tree	Monitoring ==> (model-independent) Observer ==> Anomaly-detecting Obervers for Perf. Estimation ==> Input & Output Observation seperately ==> Measure Distributional Shift Train/Test on input or output
Short Description of Mechanism	The semantic segmentation produces a pixel-wise semantic class prediction of the original input image. The goal is to detect performance drops caused by domain shifts. One way to do this, is to analyze the input space. The concept of the method is to learn an autoencoder to compress and resconstruct the image being fed to the deep neural network to be monitored. If the autoencoder is learnt with the same data as the semantic segmentation, it can be expected to also have performance drops when being confronted with the same domain shift (see the following figure).

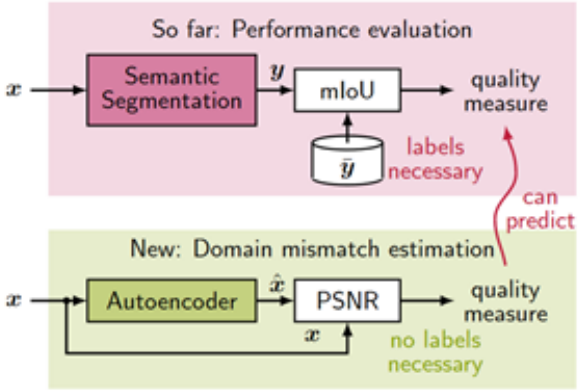
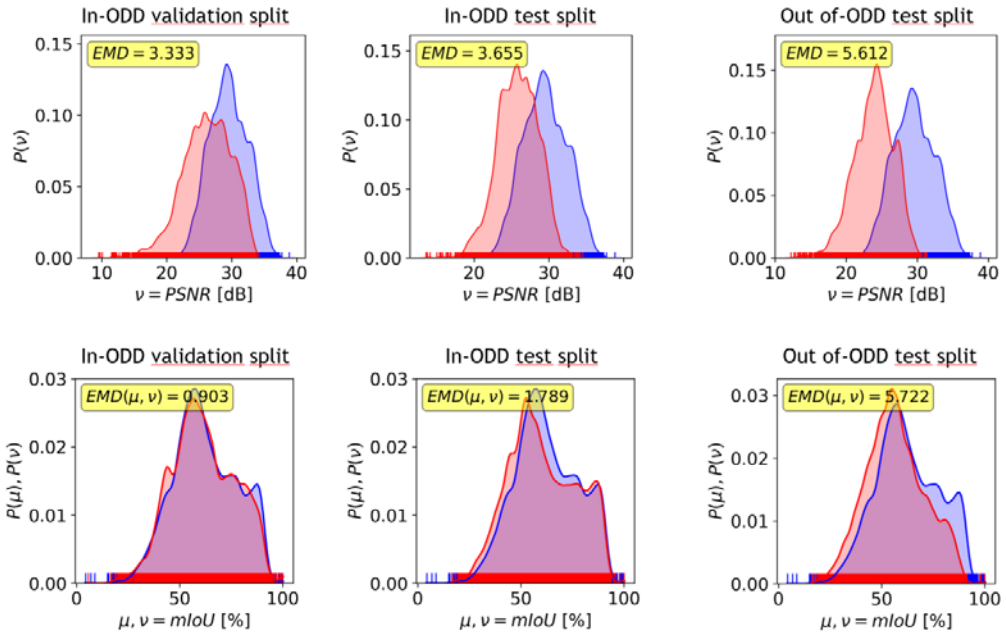


Figure 1: **Performance evaluation of semantic segmentation** (simplified sketch). Evaluation of the mean intersection over union (mIoU) requires ground truth segmentation labels $\bar{y}$, while the proposed domain mismatch estimation is performed on the basis of the PSNR of an autoencoder, trained and evaluated without labels.

The advantage of autoencoders is their ability of self-supervised training and thus they can be evaluated without the need of labels. So we use the reconstruction quality as an out of domain indicator. To yield more reliable results, the function does not operate on single images but on entire data (sub-) sets instead. The quality measure scores of

Title (Mechanism ID)	Domain Mismatch Estimation (MECH-289269)
	various data sets are compared using the earth mover's distance, also often referred to as Wasserstein-1 distance. This measure provides an estimate of the domain mismatch.
Used Data (train/val/test)	Cityscapes • training: 2975 training • validation: 500 Berkeley DeepDrive (7000 training + 1000 validation) • training: 7000 • validation: 1000 KITTI • training: 200 Release #2: • Mackevision 1st tranche (variations) • BIT TS 1st tranche (variations) Release #3: • BIT TS tranche 3 • Proof-of-project concept PoPC data Release #4 (BIT TS tranche 3 +4): • training: 31.757

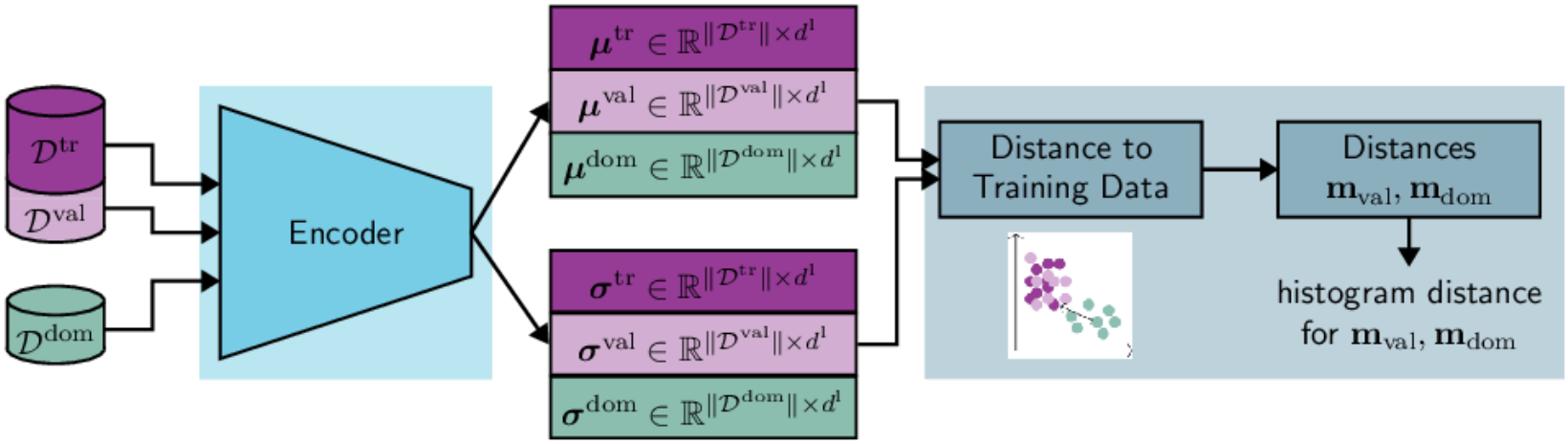
Title (Mechanism ID)	Domain Mismatch Estimation (MECH-289269)
	• validation: 7.419 • test: 9.903
Used DNN / Task	DeepLabv3+ (semantic segmentation): Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (ECCV 2018), project partner (Intel) implementation based on a GitHub code base. ERFNet (semantic segmentation): ERFNet: Efficient Residual Factorized ConvNet for Real-time Semantic Segmentation (IEEE Trans. on Int. Trans., vol. 19, no. 1, 2018), own implementation based on a GitHub code base. variant of Least Squares GAN (image reconstruction): relevant papers: Least Squares Generative Adversarial Networks (ICCV 2017), Generative Adversarial Networks for Extreme Learned Image Compression (ICCV 2019), High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs (CVPR 2018), Perceptual Losses for Real-Time Style Transfer and Super-Resolution (ECCV 2016), own implementation based on a GitHub code base.
Main Safety Concern being adressed	Brittleness of DNNs (SC-1.2), **Inadequate separation of test and training data (SC-2.2),** Specification of the ODD (SC-2.4), Distributional shift over time (SC-2.5), Safety-aware metrics (SC-3.1)
Summary of experiment results	In the course of the proof of project concept we provided an in-depth analysis about this method. Some results are depicted in the following

Title (Mechanism ID)	Domain Mismatch Estimation (MECH-289269)
	Reference (blue): In-ODD training split Inference (red): Tested on **in-ODD data and on out-of-ODD data (320x480)**  One can see that both the PSNR and the mIoU behave similar. We see increased numbers for **mtrc01000** (here termed as EMD) on the In-ODD test split and the Out-of-ODD test split. Thus a correlation of some kind is there on a dataset level. On a subsplit-level (so a small amount of images) and on an image level we could not observe a correlation limiting this method to larger sets of images.
Summary of effectiveness compared	The plain baseline does not offer any indicators for a domain shift. On the other hand, the additional use of our method indicates a domain shift based on the reconstruction error on data collections.

Title (Mechanism ID)	Domain Mismatch Estimation (MECH-289269)
to baseline / Level of Effectiveness	
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis**: The method mainly addresses the safety concern *inadequate separation of test and training data (SC-2.2)*. It measures, if a monitored DNN is exposed to a domain shift or not. 2. **Evidences for a safety assurance case**: As demonstrated in our experiments, the method shows correlation between image reconstruction performance and semantic segmentation performance on a dataset split level, i.e., sets of images >> 1. On a per-image level, no correlation is observed. 3. **Further tests**: Stronger evidences can be derived by determing the exact number of images that are at least needed to make a statement of a possible domain shift.
Link to papers	Self-Supervised Domain Mismatch Estimation for Autonomous Perception (CVPR SAIAD 2019) Performance Prediction for Semantic Segmentation by a Self-Supervised Image Reconstruction Decoder (to appear in CVPR WAD 2022)

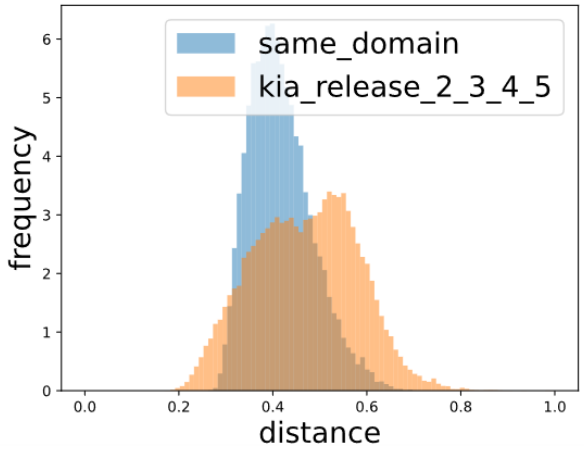
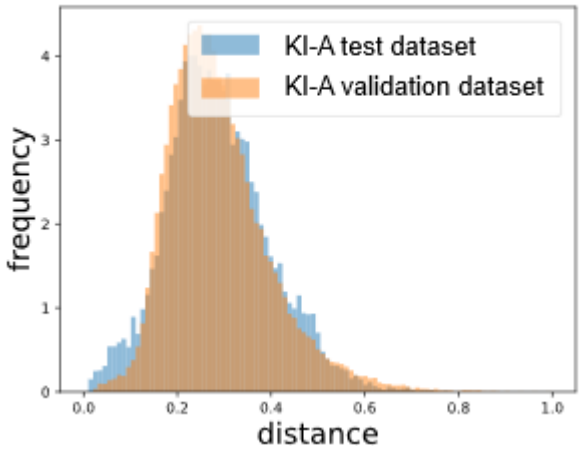
5.1.2.3 FZI

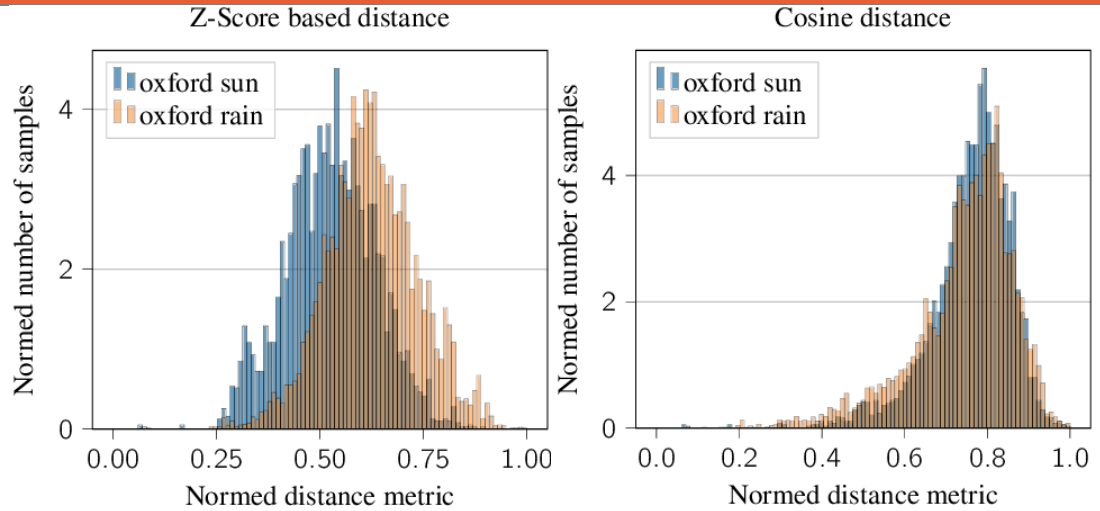
Title (Mechanism ID)	Object Centric Domain Shift (MECH-638536)
Leading and involved Partners (Name)	FZI (Hanno Stage)
Mapping to Taxonomy Tree	Dataset Optimization -> Inspection of Domain Mismetch -> VAE-Latent Space investigation -> VAE (MECH-638536)

Title (Mechanism ID)	Object Centric Domain Shift (MECH-638536)
Short Description of Mechanism	Variational Autoencoder (VAE) consist of an encoder network and a decoder network with a bottleneck between them. The goal of the autoencoder is to reconstruct the input image at the output. The encoder network encodes the input image and stores a representation of the image in the so-called latent space (the bottleneck). From this representation the decoder net tries to reconstruct the image. The underlying assumption of the mechanism, is that important features of the input data are stored inside the latent space. Two datasets from the same domain would have a very similar distribution inside the latent space, whereas other domain data would be encoded differently. By analyzing distances in latent space, a statement on dataset similarity can be given.  In the above image, the overall functionality of the mechanism is explained. A given Dataset A is split in a Training and validation dataset. After training a VAE with the given split, the latent space distances to the training data can be compared. This leads to a measure, how new or different a Dataset B is. Advantage of this approach is the comparison of data without further specification of features or parameters to compare. The mechanism focuses on the labeled objects in the training data, which is achieved by training only with the extracted bounding boxes, which are scaled to a fixed size.

Title (Mechanism ID)	Object Centric Domain Shift (MECH-638536)
	As a side benefit of the mechanism, the images with very high distances in latent space can be investigated to find "anomaly objects".
Used Data (train/val/test)	KI-A Release #2, #3, #4 #5 Dasplit Oxford RobotCar Dataset nuScenes A2D2
Used DNN / Task	The project SSD from Opel was used for the correlation analysis.
Main Safety Concern being adressed	Inadequate separation of test and training data (SC-2.2)

Summary of experiment results	• Experiments with real world data (NuScenes, A2D2, Oxford) have shown that the method can work for domain shifts
	  • Results with the KI-A dataset not quite as good
	  • Global shifts are detected
	• The best combination of VAE architecture and latent space distance is JointVAE with the Z-score
	• It is difficult to define a suitable threshold for the method to be used online
	KI-A Results

Title (Mechanism ID)	Object Centric Domain Shift (MECH-638536)
	  On the left side is the histogram with the reconstruction error and on the right side the histogram with the latent spatial distance. It is clearly visible that there is no domain shift in the KI-A dataset. This is expected, because e.g. in the split for tranche 3 all different domains occur in all data splits and so it is prevented that there is a domain shift. It can also be seen that the latent space defines the data sets as more similar compared to the reconstruction error. The ideal image for a domain shift is achieved when both histograms have no overlap. This means that there is a domain shift between both data splits.
Summary of effectiveness compared to baseline / Level of Effectiveness	• As a baseline for our mechanism, we could take the reconstruction quality • Here it could be shown in experiments that the latent space is more robust at different domain shifts Examples of the generated histograms for the Oxford dataset. VAE was trained with sunny image data and evaluated with rainy image data:

Title (Mechanism ID)	Object Centric Domain Shift (MECH-638536)
	 Compared to the reconstruction error (baseline), our method works more stable. This means that small and large domain shifts are recognised equally.
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis**: The method deals with the monitoring of the input data. It can prevent that there is a domain shift between different data splits. (SC-1.2; SC-2.1; SC-2.2; SC-2.4; SC-2.5) 2. **Evidences for a safety assurance case**: When training ML models, it is important to know that the different data splits are in the same domain or that they are not in the same domain. Our experiments have shown that domain shifts between data sets are found with the method. 3. **Further tests**: The hyperparameters of the current model could still be optimized. In addition, the latent space could be better investigated in order to understand it better and to make the domain shift even more explainable. Stronger evidences, could also be achieved by even more tests with different data sets.

Title (Mechanism ID)	Object Centric Domain Shift (MECH-638536)
Link to papers	To be appear in the KI-A Book: Stage, H. et al. (2022). Analysis and Comparison of Datasets by Leveraging Data Distributions in Latent Spaces. In T. Fingscheidt, H. Gottschalk, S. Houben et al.. Deep Neural Networks and Data for Automated Driving. Springer. doi: 10.1007/978-3-031-01233-4_3

5.2 E3.5.2 Final: Uncertainty estimation and calibration methods (zur Veröffentlichung)

5.2.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Document*
Group/Cluster	
Type of content	*Requirement*
Classification level	*PU*

5.2.2 Brief description of the cluster content

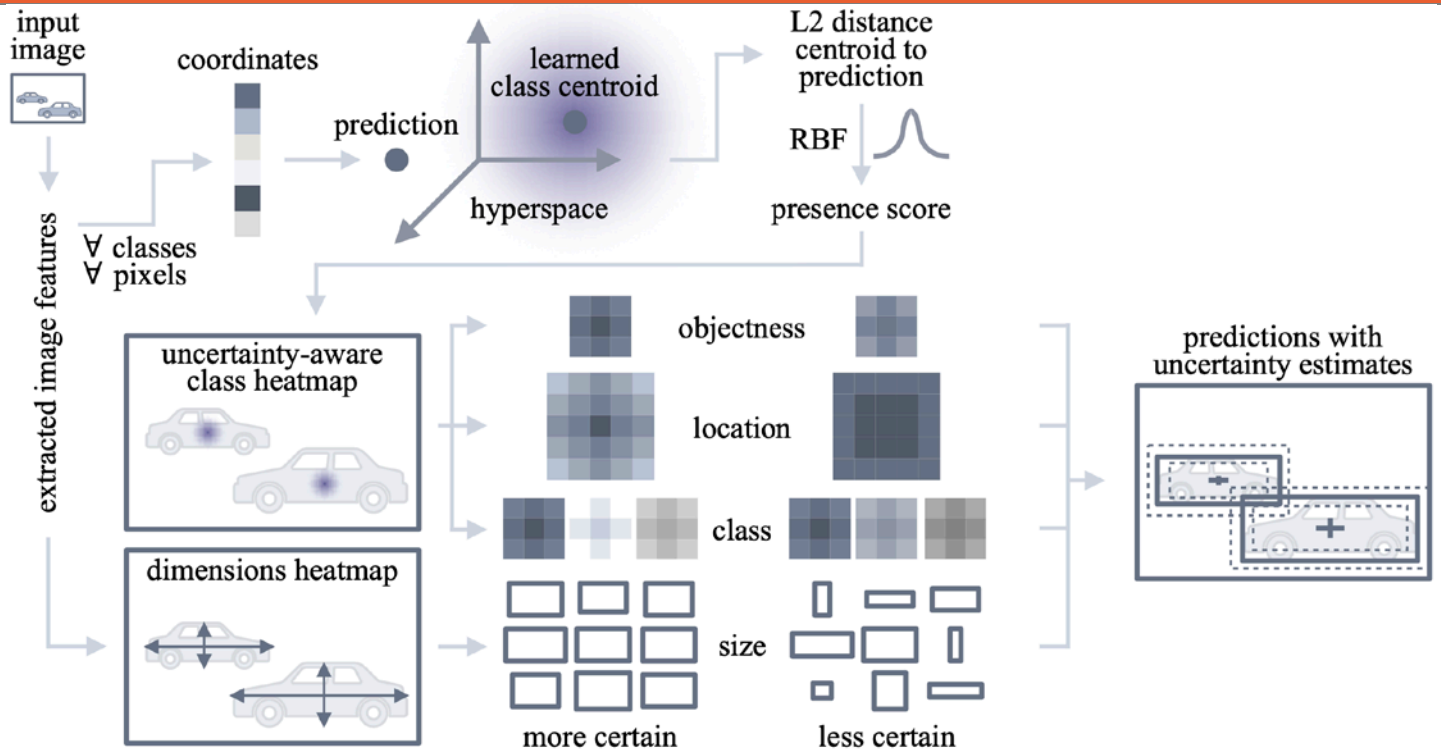
The target of the mechanisms developed within this cluster is to estimate and evaluate the predictive uncertainty that is obtained by black-box models in the scope of object detection. In the context of this cluster, one part of the methods deals with the uncertainty estimation. In this scope, we distinguish between aleatoric (data) uncertainty and epistemic (model) uncertainty. The other part of the methods deals with uncertainty evaluation and calibration, where the obtained uncertainty is compared to the observed prediction error and evaluated for consistency.

5.2.3 Mechanisms

5.2.3.1 BMW

Title (Mechanism ID)	Uncertainty Box (MECH-279269)
Leading and involved Partners (Name)	BMW
Mapping to Taxonomy Tree	Safe AI Mechanisms > Uncertainty > Confidence Calibration > Calibration via Adaption of Training Process and Safe AI Mechanisms > Uncertainty > Confidence Calibration > Calibration via Post-Processing
Short Description of Mechanism	Our work estimates the uncertainty in a sampling-free fashion independently for every aspect of object detection (i.e. object presence (also called objectness), location, size, and class). We achieve this in different ways at inference time: for the objectness we are computing the features distance to class representatives learned during training (as shown in the top of the figure below) and using that as uncertainty measure to create an uncertainty-aware class heatmap; for the other 3 output signals (i.e. location, size and class), we are comparing the discarded predictions surrounding the final predictions. In particular, as we built on top of a CenterNet architecture (anchor-less, or keypoint-based), we have the guarantee of having a bounding box predicted for each pixel. Therefore, we can compare this multitude of outputs with the final model predictions, thereby estimating the uncertainty depending on the disagreement of these outputs.

Title (Mechanism ID)	Uncertainty Box (MECH-279269)



Specifically, for the objectness uncertainty, we exploit the idea of a prior paper, namely Deterministic Uncertainty Quantification (DUQ), which was proposed for out of distribution prediction in whole-image classification settings. We thoroughly extended DUQ to a dense task (the prediction of the uncertainty-aware class heatmap) for object detection, while majorly improving its convergence. This works by computing the Euclidean distance in a feature hyperspace w.r.t. learned class representatives, which are trailing averages of positive samples seen during training. This distance is then passed through a radial basis function (RBF), in the form of a Gaussian kernel, providing a percentage value for the objectness uncertainty.

Title (Mechanism ID)	Uncertainty Box (MECH-279269)
	For the other output signals (location, size and class), the uncertainty computation is different. Location and class are also computed on the same uncertainty-aware class heatmap described above, rendering them uncertainty-aware as well, while the object dimensions (i.e. size) are extracted from a different heatmap, namely the dimensions heatmap. As can be seen in the figure above, intuitively, for the dimensions, a higher degree of agreement between neighboring outputs corresponds to a more certain size prediction, while if the outputs are very different from one another, the model is less certain on the dimensions. Instead, the location uncertainty is computed on the uncertainty-aware class heatmap, depending on the heatmap peaks around the object center: intuitively a sharper and more identifiable peak corresponds to a more certain location prediction, while a flatter and more diffused peak is linked to a more uncertain location prediction. Lastly, the class uncertainty is computed according to the different peaks in the various class heatmaps (each channel corresponds to a different class), so if only a class is activated at the object center location, it corresponds to a relatively certain prediction for that class, while if other classes are also activated at the same location, then the uncertainty on the class is higher. To these intuitive explanations correspond actual quantitative measurements which are described in the paper we published at IEEE RA-L.
Used Data (train/val/test)	While training was done on the third tranche of the BIT KI-A dataset, extensive experiments were carried only on public datasets (KITTI, BDD100K, nuImages) due to the readily available different conditions (country, vehicles, weather, illumination, aspect ratio, mounting position, ...).
Used DNN / Task	DLA-34 architecture, also used in CenterNet. This was chosen as it provides a good trade-off between speed and accuracy.
Main Safety Concern being adressed	This work mainly addresses the concerns of unreliable confidence information and poor generalization to unseen domains and conditions. We found uncertainty estimation to not only provide a good estimate of the domain gap, but also bridging this gap, significantly improving the overall performance.

Title (Mechanism ID)	Uncertainty Box (MECH-279269)
Summary of experiment results	According to the main results, the proposed method generalizes substantially better to out-of-domain data and unseen conditions, which is a highly wanted characteristic for any autonomous driving system. Just by making it uncertainty-aware, our model, trained on images recorded in Germany in sunny or cloudy conditions (KITTI), outperformed the model it is based on (CenterNet) by a large margin on other data recorded in USA and Asia (BDD100K and nuImages) in different weather and illumination conditions. This shows the added benefit of uncertainty estimation, which allows to better estimate the domain shift, thus delivering better predictions. This can be seen in the quantitative (table) and qualitative results (figure) below. Uncertainty quality comparison with related works on the validation sets of KITTI, BDD100K, and nuImages (nuIm.). All models were trained on KITTI, and transferred (→) to the other datasets. All results are for *car* (moderate for KITTI). on CenterNet‡ we applied our location and dimensions uncertainty estimations as post-processing (so each value under the columns location and dimensions refers to CenterNet + ours).
Summary of effectiveness compared to	In this work, one could identify two different baselines, namely CenterNet and DUQ, which we combined and extended. The experiments show a clear edge from each of them, proving the effectiveness of our modifications, especially in terms of the significantly improved generalization to unseen domains (BDD100K and nuImages). On top of the higher detection

	Method	AP	objectness AUPR-In	objectness AUPR-Out	objectness AUROC	objectness ECE ↓	objectness UE ↓	location CE ↓	location UBQ	location BR	dimensions CE ↓	dimensions UBQ	dimensions BR
KITTI	GaussianYOLO [3]	84.36	74.93	91.24	86.23	23.37	20.02	11.58	60.01	60.56	6.18	75.34	77.47
	CenterNet‡ [18]	89.15	72.76	99.56	96.65	10.49	6.00	7.17	68.27	69.80	**4.23**	85.84	89.58
	5-Ensemble [10]	**89.75**	73.79	93.61	87.13	16.93	17.72	4.59	**86.71**	**90.43**	5.01	84.28	87.54
	CertainNet [ours]	89.36	**78.16**	**99.81**	**98.00**	**4.60**	**4.98**	**4.54**	75.92	77.81	4.47	**86.76**	**91.05**
→ BDD	GaussianYOLO [3]	31.03	88.73	79.72	85.84	22.80	20.68	9.48	57.32	58.49	**7.67**	57.94	59.19
	CenterNet‡ [18]	34.51	81.35	**98.06**	91.74	7.78	**14.03**	25.72	35.90	36.76	10.53	73.43	79.55
	5-Ensemble [10]	26.75	**96.42**	81.84	**91.99**	25.64	15.26	**3.31**	**84.83**	**93.54**	9.40	**84.81**	**93.50**
	CertainNet [ours]	**40.93**	78.77	97.82	91.17	**4.89**	14.85	14.79	44.71	46.03	9.68	72.91	78.89
→ nuIm.	GaussianYOLO [3]	30.38	86.24	87.95	87.78	17.05	17.74	10.42	56.20	57.62	8.74	58.38	60.14
	CenterNet‡ [18]	43.93	75.94	98.93	91.58	**5.51**	15.70	27.07	35.69	36.21	8.39	74.79	79.21
	5-Ensemble [10]	35.31	**93.93**	90.90	93.04	17.74	13.98	**3.37**	**89.48**	**98.51**	8.36	**89.31**	**98.21**
	CertainNet [ours]	**53.14**	79.97	**99.28**	**94.24**	7.80	**11.90**	16.21	43.57	44.46	**7.73**	75.36	80.25

Title (Mechanism ID)	Uncertainty Box (MECH-279269)
baseline / Level of Effectiveness	performance (AP), in the context of autonomous driving, our added uncertainty estimates are valuable for downstream tasks (e.g., path planning).
Potential Evidences for an Safety Assurance Case	This mechanism would be valuable in this context as it improves in 2 directions: confidence calibration and generalization to out-of-domain data. For safety critical systems, these two are fundamental properties. The evidences can be given by the extensive experiments which can be found in the paper, where a detailed ablation study shows the added benefits of our method, a plurality of metrics test the improved calibration, and multiple transfers to challenging and highly different out-of-domain data support the generalization claim. A "strong" evidence could be derived by making use of the added uncertainty estimates in a downstream task (e.g., path planning). Proper safety derivations were done by Andre Rossbach.
Link to papers	IEEE RA-L publication and presentation at IEEE ICRA 2022: https://ieeexplore.ieee.org/abstract/document/9627771

5.2.3.2 Hochschule Ruhr West, EFS I

Title (Mechanism ID)	Detection Expected Calibration Error (MECH-280456)
Leading and involved Partners (Name)	Hochschule Ruhr West, EFS
Mapping to Taxonomy Tree	Uncertainty → Confidence Calibration → Calibration via Post-Processing
Short Description of Mechanism	Miscalibration in the scope of neural networks is defined as the deviation between estimated confidence and observed accuracy of a model on a specific target dataset. The miscalibration is measured by dividing the confidence space into several bins in order to measure the average accuracy as well as the average confidence in each bin. The Expected Calibration Error (ECE) as the common miscalibration score is computed as the weighted sum of the gaps between accuracy and confidence over all bins.
Used Data (train/val/test)	We measure miscalibration on the KI-A tranche 3+4+5+6 dataset using the official TP1 dataset split: New, additional data split, decided 29.11.21 for SSD R3-v2. The validation and test sets are also used in E3.5.2 Multivariate Confidence Calibration HRW MECH-043409 and E3.5.2 Uncertainty in Confidence Calibration HRW MECH-796456 for calibration training and testing. Thus, we provide calibration evaluation on both dataset using this mechanism.
Used DNN / Task	KI-A data and networks: TP1 Opel SSD E1.3.3a Implementierung der funktionalen Algorithmen: 2D-Bounding Box using the inference on the validation and test set on tranche 3+4+5. For calibration, we need the confidence scores as well as the bounding box information (position/scale) to perform confidence calibration. These information are provided by the SSD detector.
Main Safety Concern being adressed	SC-1.1: Unreliable Confidence Information
Summary of experiment results	Similar to our results in our research paper, we could find a (minor) correlation between position information (i.e., data distribution) and miscalibration. Most of the pedestrians are located at the boundaries of the horizontal axis and in the center region facing the vertical axis. In these cases, the gain in precision of the detection model exceeds the gain of confidence. Therefore, we find that the TP1 Opel SSD r3 v2 is consistently **underconfident** in its prediction which is only known by a RetinaNet (trained with focal loss) so far.

Title (Mechanism ID)	Detection Expected Calibration Error (MECH-280456)
Summary of effectiveness compared to baseline / Level of Effectiveness	The D-ECE is the baseline to evaluate the miscalibration for object detection in conjunction with Brier and NLL scores.
Potential Evidences for an Safety Assurance Case	This metric is necessary to assess the reliability of the confidence information provided by an object detection model. We derive the evidence for this statement from the common literature in the field of confidence calibration (especially for the standard ECE) as well as from our last publications "Küppers et al.: *Multivariate Confidence Calibration for Object Detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, 326-327"* and "Küppers et al.: Bayesian Confidence Calibration for Epistemic Uncertainty Modelling, *2021 IEEE Intelligent Vehicles Symposium (IV),* 2021". **Further experiments**: different detection models, different data sets/data distributions
Link to papers	Published paper: https://openaccess.thecvf.com/content_CVPRW_2020/papers/w20/Kuppers_Multivariate_Confidence_Calibration_for_Object_Detection_CVPRW_2020_paper.pdf Published paper: Küppers et al. (in press) Confidence Calibration for Object Detection and Segmentation. In: Tim Fingscheidt, Hanno Gottschalk and Sebastian Houben (eds.) Deep Neural Networks and Data for Automated Driving - Robustness, Uncertainty Quantification, and Insights Towards Safety. Springer Nature Switzerland, Cham, Switzerland.

5.2.3.3 Hochschule Ruhr West, EFS II

Title (Mechanism ID)	Multivariate Confidence Calibration (MECH-043409)
Leading and involved Partners (Name)	Hochschule Ruhr West, EFS
Mapping to Taxonomy Tree	Uncertainty → Confidence Calibration → Calibration via Post-Processing
Short Description of Mechanism	Miscalibration in the scope of neural networks is defined as the deviation between estimated confidence and observed accuracy of a model on a specific target dataset. The Expected Calibration Error

Title (Mechanism ID)	Multivariate Confidence Calibration (MECH-043409)
	(ECE) is the common miscalibration score in the scope of classification. There are several methods to correct biased confidence estimates. This is done by assigning predictions with a certain confidence new confidence values that match the observed accuracy. However, common literature for confidence calibration is limited to the classification domain. On object detection tasks, a neural network also estimates a bounding box assigned to each prediction. Thus, we examined if the miscalibration of object detectors also depends on the location and/or the scale of the predicted box. We further provide a framework for confidence calibration by also taking the position and/or scale into account. On the one hand, we extend the well-known histogram binning calibration method from classification to a multivariate histogram binning that is also able to use additional features for the binning. On the other hand, we extend logistic calibration (also known as Platt scaling) and beta calibration to use multivariate normal/beta distributions in order to keep track of correlations between all given features. This can be useful e.g. on an pedestrian detectioin system. In road traffic it often occurs that a pedestrian suddenly appears at the roadside to cross the road and is therefore located at the image boundaries. If we observe a miscalibration especially at the image boundaries, we cannot trust in the confidence score. With our proposed methods, we can assign properly confidence estimates to predictions that match the observed precision in those cases.
Used Data (train/val/test)	We performed calibration on the KI-A tranche 3+4+5+6 dataset using the official TP1 dataset split: New, additional data split, decided 29.11.21 for SSD R3-v2 The validation set is used for building the calibration mapping and the test set is used for calibration evaluation.
Used DNN / Task	KI-A data and networks: TP1 Opel SSD E1.3.3a Implementierung der funktionalen Algorithmen: 2D-Bounding Box. Version: Release 3 v2: E1.3.3a Release of SSD-r3-v2 For calibration, we need the confidence scores as well as the bounding box information (position/scale) to perform confidence calibration. These information are provided by the SSD detector.
Main Safety Concern being adressed	SC-1.1: Unreliable Confidence Information

Title (Mechanism ID)	Multivariate Confidence Calibration (MECH-043409)
Summary of experiment results	Similar to our results in our research paper, we could find a (minor) correlation between position information (i.e., data distribution) and miscalibration. Most of the pedestrians are located at the boundaries of the horizontal axis and in the center region facing the vertical axis. In these cases, the gain in precision of the detection model exceeds the gain of confidence. Therefore, we find that the TP1 Opel SSD r3 v2 is consistently **underconfident** in its predictions which is only known by a RetinaNet (trained with focal loss) so far.
Summary of effectiveness compared to baseline / Level of Effectiveness	Our experiments show that especially the Beta Calibration is a very effective method for confidence recalibration as the output distribution of the Opel SSD predictions is represented best by this calibration method. The Beta calibration method achieves a reduction of the D-ECE of 15% from ~21% to ~6% for IoU @ 0.5 which is very effective in our understanding. This is underlined by the complementary metrics Brier score and NLL which also show an improvement in calibration while not degrading the baseline prediction performance.
Potential Evidences for an Safety Assurance Case	Our hypothesis is that a detection model should produce reliable confidence information about its prediction to assess its confidence in its predictions. We show in our experiments that this mechanism is able to successfully recalibrate the confidence information so that meaningful confidence information are available. **Evidences**: • It is not sufficient to only evaluate the D-ECE score. All calibration methods show an improvement in the D-ECE. It is necessary to jointly inspect D-ECE, Brier, and NLL scores to evaluate the calibration properties. • Scaling methods like Logistic Calibration or Beta Calibration work best for object detection in general as these methods do not require a large amount of data (in contrast to Histogram Binning). Previous experiments already revealed that Histogram Binning does not yield a sufficient recalibrated data distribution. • It is convenient to use Logistic calibration as well as Beta calibration during calibration training and determine the superior model afterwards as previous experiments have shown that also Logistic calibration yields a good calibration performance, depending on the data distribution. **Further experiments**: different detection models, different data sets/data distributions

Title (Mechanism ID)	Multivariate Confidence Calibration (MECH-043409)
Link to papers	Published paper: https://openaccess.thecvf.com/content_CVPRW_2020/papers/w20/Kuppers_Multivariate_Confidence_Calibration_for_Object_Detection_CVPRW_2020_paper.pdf Published paper: Küppers et al. (in press) Confidence Calibration for Object Detection and Segmentation. In: Tim Fingscheidt, Hanno Gottschalk and Sebastian Houben (eds.) Deep Neural Networks and Data for Automated Driving - Robustness, Uncertainty Quantification, and Insights Towards Safety. Springer Nature Switzerland, Cham, Switzerland.

5.2.3.4 Hochschule Ruhr West, EFS III

Title (Mechanism ID)	Uncertainty in Confidence Calibration (MECH-796456)
Leading and involved Partners (Name)	Hochschule Ruhr West, EFS
Mapping to Taxonomy Tree	Uncertainty → Confidence Calibration → Calibration via Post-Processing
Short Description of Mechanism	It is possible with common Bayesian neural networks (BNN) to get a quantification of epistemic/intrinsic uncertainty about a prediction. For example, using MC dropout or variational inference, it is possible to obtain distributions for a prediction rather than a single point estimate. We transfer this technique to common confidence calibration mappings (logistic calibration, beta calibration, etc.) to obtain calibrated confidence estimates **with epistemic uncertainty attached** to each prediction. Instead of classical maximum likelihood estimation (MLE), we use stochastic variational inference (SVI) to estimate a variational distribution for each parameter instead of a point estimate. Therefore, it is possible to obtain an uncertainty quantification even for standard non-BNNs. We show that it is possible to even neglect calibrated estimates if a calibration mapping is uncertain in a specific region.
Used Data (train/val/test)	We performed calibration on the KI-A tranche 3+4+5+6 dataset using the official TP1 dataset split: New, additional data split, decided 29.11.21 for SSD R3-v2 The validation set is used for building the calibration mapping and the test set is used for calibration evaluation.
Used DNN / Task	KI-A data and networks: TP1 Opel SSD E1.3.3a Implementierung der funktionalen Algorithmen: 2D-Bounding Box using the inference on the validation and test set on tranche 3+4+5. For calibration, we need the

Title (Mechanism ID)	Uncertainty in Confidence Calibration (MECH-796456)
	confidence scores as well as the bounding box information (position/scale) to perform confidence calibration. These information are provided by the SSD detector.
Main Safety Concern being adressed	SC-1.1: Unreliable Confidence Information
Summary of experiment results	**Evaluation results:** we observe that the uncertainty correlates with the amount of available data. Therefore, we suggest to use this additional uncertainty quantification as a **sufficient** criterion to detect a possible covariate shift. Furthermore, we found that the TP1 SSD detection model is too **underconfident** in its predictions. That is orthogonal to current research as most of the related work founds modern neural networks to be too overconfident (except models using special regularization or focal loss).
Summary of effectiveness compared to baseline / Level of Effectiveness	Our experiments show that especially the Beta Calibration is a very effective method for confidence recalibration. The Beta calibration method achieves a reduction of the D-ECE of 14% from ~21% to ~7% for IoU @ 0.5 which is very effective in our understanding. This is underlined by the complementary metrics Brier score and NLL which also show an improvement in calibration while not degrading the baseline prediction performance. Furthermore, the metrics for the evaluation of the epistemic uncertainty of the calibration methods (PICP and MPIW) show a reliable and thus informative epistemic uncertainty quantification for the calibration methods itself. In addition, these metrics show that the Logistic calibration has a high epistemic uncertainty which is an additional hint towards a bad calibration performance. This behavior is beneficial to gain evidences and trust into these calibration methods as they are now able to indicate a possible failure mode.
Potential Evidences for an Safety Assurance Case	Our hypothesis is that a detection model should produce reliable confidence information about its prediction (see MECH-043409 Multivariate Confidence Calibration). Furthermore, a calibration model should also indicate if its uncertain about its recalibration estimate which is the focus of this mechanism. In our experiments, we can show that the Logistic calibration method does not offer a good calibration performance which is, in turn, also indicated by its epistemic uncertainty. This gives us the evidence that this mechanism is a very effective approach to evaluate the calibration methods itself. This is beneficial for the overall assurance case of reliable confidence information as allows to further evaluate the recalibrated confidences.

Title (Mechanism ID)	Uncertainty in Confidence Calibration (MECH-796456)
	Further experiments: different detection models, different data sets/data distributions
Link to papers	Published paper: https://ieeexplore.ieee.org/document/9575841

5.2.3.5 Hochschule Ruhr West, EFS IV

Title (Mechanism ID)	Instance Segmentation Calibration (MECH-250133)
Leading and involved Partners (Name)	Hochschule Ruhr West, EFS
Mapping to Taxonomy Tree	Uncertainty → Confidence Calibration → Calibration via Post-Processing
Short Description of Mechanism	Based on previous works regarding confidence calibration for object detection (see E3.5.2_Detection_ECE_HRW_MECH-280456 and E3.5.2_Multivariate_Confidence_Calibration_HRW_MECH-043409), we adapt those methods for pixel-wise confidence calibration on instance segmentation masks. We use methods that inlude not only the confidence but also the position of each pixel into a calibration mapping. We aim to improve the confidence scores of each mask pixel in order to reliably reflect the expected accuracy.
Used Data (train/val/test)	We performed calibration on the KI-A tranche 3 dataset using the dataset split. The validation set is used for building the calibration mapping and the test set is used for calibration evaluation.
Used DNN / Task	KI-A data and networks: TP1 Intel Detectron Mask-RCNN E1.3.3d Implementierung der funktionalen Algorithmen: Instanz-Segmentierung using the inference on the validation and test set on tranche 3. For calibration, we need the pixel confidence scores, the shortest distance of each pixel to the next segment boundary as well as the relative position of each pixel within the detected bounding box. These information are provided by the Mask-RCNN detector.
Main Safety Concern being adressed	SC-1.1: Unreliable Confidence Information
Summary of experiment results	We found that the TP1 instance segmentation model already provides a good calibration performance. We have been able to successfully decrease potential miscalibration by using the multivariate extension of the histogram binning method. The confidence calibration is applied

Title (Mechanism ID)	Instance Segmentation Calibration (MECH-250133)
	before thresholding the confidence scores of each pixel within an instance mask. This leads to an improvement in calibration.
Summary of effectiveness compared to baseline / Level of Effectiveness	In our experiments, we can show that especially the histogram binning is able to reduce miscalibration from ~2,6% to 1,7%. The effect on calibration is not as big as within object detection. However, compared to the amount of available data, this is still a feasible calibration result.
Potential Evidences for an Safety Assurance Case	Our hypothesis is that an instance segmentation model should produce reliable confidence information about its prediction to assess its confidence in the predicted instance segmentation masks. We show in our experiments that this mechanism is able to successfully recalibrate the pixel confidence information so that meaningful confidence information are available. **Evidences**: • It is not sufficient to only evaluate the D-ECE score. All calibration methods show an improvement in the D-ECE. It is necessary to jointly inspect D-ECE, Brier, and NLL scores to evaluate the calibration properties. • Binning methods like histogram binning work best if a sufficient amount of data is available. This is the case within the context of segmentation calibration as each pixel can be used as an own sample. **Further experiments**: different detection models, different data sets/data distributions
Link to papers	Published paper: Küppers et al. (in press) Confidence Calibration for Object Detection and Segmentation. In: Tim Fingscheidt, Hanno Gottschalk and Sebastian Houben (eds.) Deep Neural Networks and Data for Automated Driving - Robustness, Uncertainty Quantification, and Insights Towards Safety. Springer Nature Switzerland, Cham, Switzerland.

5.2.3.6 Hochschule Ruhr West, EFS V

Title (Mechanism ID)	Regression Calibration (MECH-432534)
Leading and involved Partners (Name)	Hochschule Ruhr West, EFS

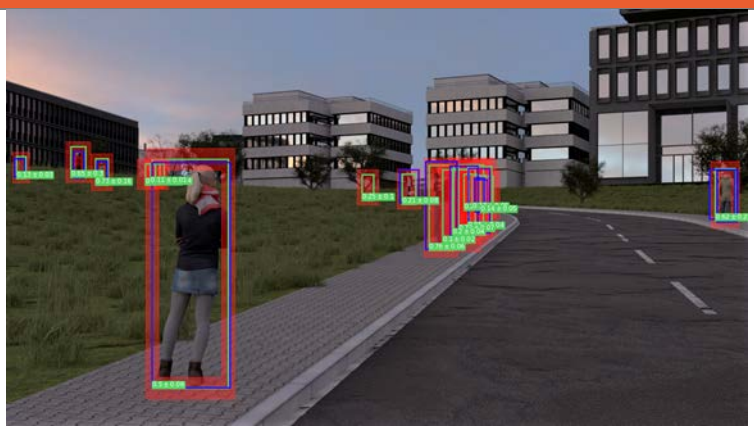
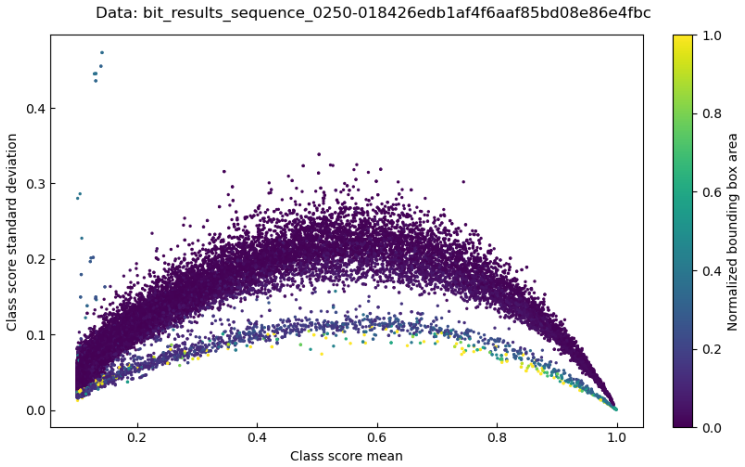
Title (Mechanism ID)	Regression Calibration (MECH-432534)
Mapping to Taxonomy Tree	Uncertainty → Confidence Calibration → Calibration via Post-Processing
Short Description of Mechanism	A probabilistic regression model outputs an uncertainty score for each prediction which is commonly the estimated variance of the prediction. This variance can be interpreted as the aleatoric uncertainty of such a model. Similar to the task of confidence calibration for the objectness score (cf. E3.5.2 Multivariate Confidence Calibration HRW MECH-043409), we expect the predicted uncertainty to match the observed error. If we observe a deviation, a regression model is called miscalibrated. Common calibration methods such as Isotonic Regression (Kuleshov et al., 2018, see: https://arxiv.org/abs/1807.00263), Variance Scaling (Levi et al., 2019, see: https://arxiv.org/pdf/1910.03127.pdf) or GP-Beta (Song et al., 2019, see: https://arxiv.org/abs/1905.06023) recalibrate the uncertainty by rescaling the predicted cumulative density function (CDF) to match the observed (error) distribution. However, this recalibration yields a non-parametric probability distribution which has no closed-form representation and thus might be difficult to integrate in subsequent processes, e.g., in Kalman filtering. Therefore, we extend the GP-Beta calibration framework by (Song et al., 2019) and propose the GP-Normal method, which expects a normal distribution as input and also outputs a parametric Gaussian distribution as calibration result. This allows for an easier integration into process chains that assert single scores such as standard deviation or variance as input.
Used Data (train/val/test)	We performed calibration on the KI-A tranche 3+4+5+6 dataset using the official TP1 dataset split: New, additional data split, decided 29.11.21 for SSD R3-v2 The validation set is used for building the calibration mapping and the test set is used for calibration evaluation.
Used DNN / Task	KI-A data and networks: Extension (!) of the TP1 Opel SSD E1.3.3a Implementierung der funktionalen Algorithmen: 2D-Bounding Box, Version: Release 3 v2: E1.3.3a Release of SSD-r3-v2, by a probabilistic loss to obtain positional estimates as well as the according uncertainty (represented by a Gaussian variance). The JSON prediction files for the extended network can be found on the Fraunhofer DSP: https://kip.gpu-cluster.itwm.fraunhofer.de/minio/lfs-objects/TP3/AP3.5/E3.5.2_Regression_Calibration/AleatoricSSD/

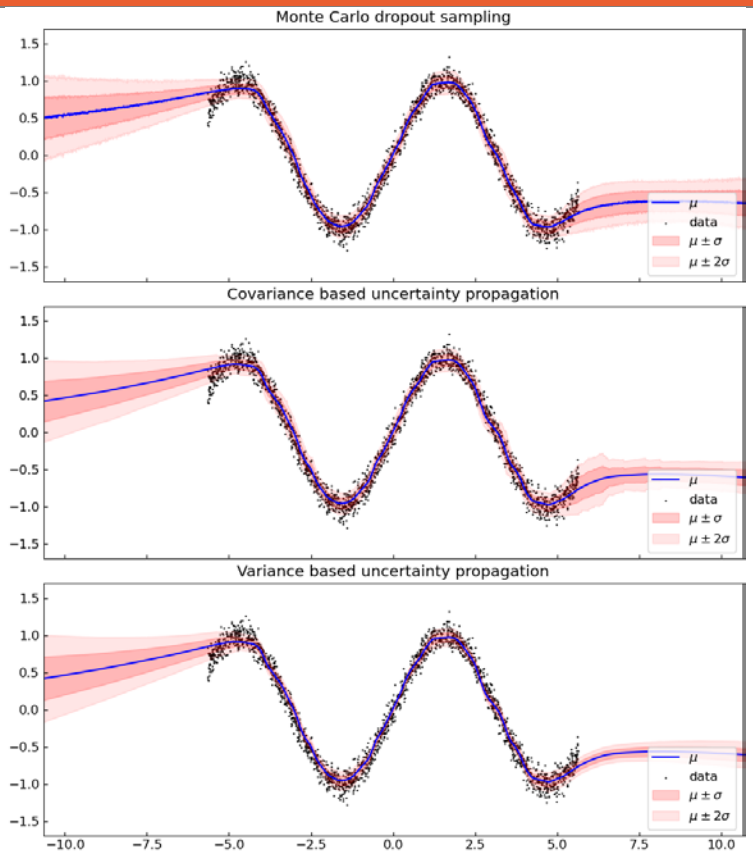
Title (Mechanism ID)	Regression Calibration (MECH-432534)
	For calibration, we need the position information as well as the according uncertainty to perform uncertainty calibration. These information are provided by the extended SSD detector.
Main Safety Concern being adressed	SC-1.1: Unreliable Confidence Information
Summary of experiment results	We investigate the calibration properties of an object detector that outputs probabilistic forecasts for the object's position in terms of • Quantile calibration: the predicted quantiles should matched the observed quantile coverage of the ground-truth samples • Variance calibration: the predicted standard deviation/variance should match the observed (root) mean squared error We observe a high miscalibration of the predicted uncertainty by default which is indicated by the metrics used for uncertainty evaluation. This miscalibration is successfully mitigated using the calibration methods, especially the non-parametric Isotonic Regression method. In this case, our GP-Normal does not lead to an improvement in calibration compared to the standard Variance Scaling method. Therefore, we conclude that the non-parametric Isotonic Regression achieves best results for quantile calibration, whereas the GP-Beta is able to achieve good results for quantile calibration as well as for variance calibration.
Summary of effectiveness compared to baseline / Level of Effectiveness	We assume a high effectiveness as all of the used calibration metrics indicate an improvement in uncertainty calibration, especially for the non-parametric Isotonic Regression and the GP-Beta methods.
Potential Evidences for an Safety Assurance Case	Our hypothesis is that a probabilistic regression model should produce reliable uncertainty estimates about its predictions to assess its confidence in its predictions. We show in our experiments that this mechanism is able to successfully recalibrate the uncertainty information so that meaningful uncertainty information are available. **Evidences**: • It is not sufficient to only evaluate a single score such as NLL or Pinball. For a reasonable comparison, it is necessary to inspect all previously mentioned metrics for uncertainty evaluation.

Title (Mechanism ID)	Regression Calibration (MECH-432534)
	• Non-parametric calibration methods are beneficial in most cases and very flexible. However, it might also be a challenging task to include such non-parametric distributions into subsequent processes **Further experiments**: different detection models, different data sets/data distributions
Link to papers	Publication under peer-review, nothing published so far

5.2.3.7 EFS

Title (Mechanism ID)	Sampling-free Epistemic Uncertainty Estimation (MECH-600492)
Leading and involved Partners (Name)	EFS
Mapping to Taxonomy Tree	Uncertainty → Approximate Bayesian Neural Networks → Variational Inference
Short Description of Mechanism	This mechanism represents a sampling-free method for epistemic uncertainty estimation of neural networks that is based on error propagation. Quantifying the model's intrinsic uncertainty is important for safety-critical applications as it allows to identify where the model lacks knowledge. This information can then be used to make better statements about the model's performance and the plausibility of its predictions.
Used Data (train/val/test)	Experiments are performed on KIA test datasets of Tranche 3+4+5+6 according to data split.
Used DNN / Task	TP1 Opel-SSD (E1.3.3a Implementierung der funktionalen Algorithmen: 2D-Bounding Box)
Main Safety Concern being adressed	SC-1.1: Unreliable Confidence Information
Summary of experiment results	Despite the constraints of an insufficiently trained detection network and time limits that do not allow further experiments to be conducted with the newly trained SSD, the experiments show promising qualitative results that model uncertainty estimation can reveal network flaws.

Title (Mechanism ID)	Sampling-free Epistemic Uncertainty Estimation (MECH-600492)
	 We found that estimates of the epistemic uncertainty gives better results for the classification task than for the prediction of bounding boxes. This is probably also related to the poor performance of the network. Overall, however, the results are promising. 
Summary of effectiveness compared to baseline / Level of Effectiveness	• Sampling-free epistemic uncertainty can serve as a fast and reliable uncertainty estimation method. • The example below shows sampling-free uncertainty estimation (variance- and covariance-based) compared to Monte Carlo dropout sampling (1000 samples) as a baseline for a toy example. As one can see, the qualitative results of both methods look very similar. However, the sampling-free variance-based approach is in this case are faster by a factor of ~500.

Title (Mechanism ID)	Sampling-free Epistemic Uncertainty Estimation (MECH-600492)
	
Potential Evidences for an Safety Assurance Case	1. Safety hypothesis: The method addresses the safety concern of Unreliable Confidence Information (SC-1.1). Quantifying epistemic uncertainty helps to identify potentially critical inputs the network has not enough knowledge about. Information about high model uncertainty can be used to detect data points the model was not sufficiently or representatively exposed to during training. 2. Evidences for a safety assurance case: Data points with high model uncertainty can be used to trigger safety mechanisms further downstream. Adding such detections that represent underrepresented data points to the training set can potentially reduce the model's ignorance in these domains. Experiments have shown that the model's uncertainty correlates with bounding box positions and class predictions. 3. Further tests: Stronger evidences can be derived by performing further experiments with well-trained models and data covering corner cases, outliers, sensor noise, and occlusion. To derive unbiased and strong evidences, the use of fully disjoint training and test datasets should be considered.

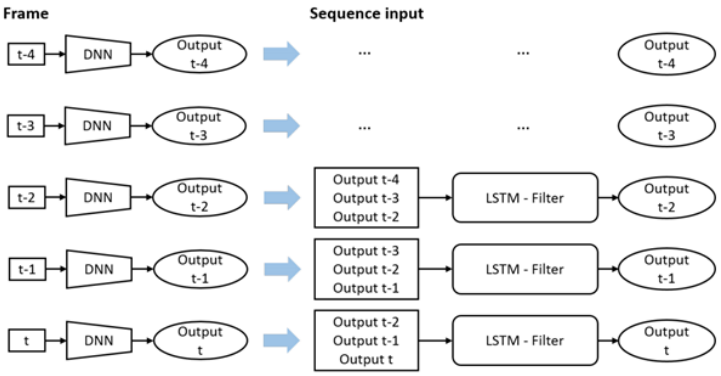
5.3 E3.5.3 Final: Adversarial attacks and teacher-student frameworks (zur Veröffentlichung)

5.3.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Document*
Group/Cluster	
Type of content	*Specification*
Classification level	*PU*

5.3.2 Description of the result

Title (Mechanism ID)	E3.5.3_Opel_Adversarial_Attacks_Assessment_MECH-232004
Leading and involved Partners (Name)	Ahmed Hammam
Mapping to Taxonomy Tree	Robustification ==> Adversarial Robustness/ Adversarial Defenses ==> Robustification on Model Level ==> Modification of Architecture ==> Temporal Consistency Architecture ==> Recurrent Nerual Networks (RNNs)
Short Description of Mechanism	Given a pre-trained model that has already been trained for a special task, an LSTM-Filter is introduced that works on the outputs of this DNN in order to improve its robustness against adversarial attacks. With the use of video sequences, frames are passed into a DNN to produce respective predictions. The DNN's outputs are then passed to a consecutive LSTM which correlates between attacked and unattacked frames in order to increase the DNN's performance. In order to decrease the influence of adversarial perturbations, the LSTM-Filter is used to filter the DNN's outputs such that the LSTM-filtered output can provide not only better results but also higher robustness. The predictions or decoding process can be used directly, based on the outputs from the LSTM-Filter.

Title (Mechanism ID)	E3.5.3_Opel_Adversarial_Attacks_Assessment_MECH-232004
	
Used Data (train/val/test)	KITTI Tracking Dataset (2D Bounding Box Detection) KIA Dataset
Used DNN / Task	Yolo V3 - (https://arxiv.org/pdf/1804.02767.pdf) 2D-Bounding Box HCI Opel ZF - SSD (https://gitlab.com/kia2/tp1/ap1.3/2d-bounding-box)
Main Safety Concern being adressed	Brittleness of DNNs (SC-1.2)
Summary of experiment results	The incorporation of an LSTM module to a pretrained model improves the performance of the network against adversarial attacks. The approach improves the mAP by around 5% and decreases the LAMR by also around 5% (at a sparsity level of 50%) for both the class car and the class pedestrian.
Summary of effectiveness compared to baseline / Level of Effectiveness	• Results show the improvement when using temporal sequences over using single images. The incorporation of more than one frame helps in robustifying the network.
Potential Evidences for an Safety Assurance Case	• The effects of adversarial attacks on normal networks not trained against adversarial attacks deteriorate heavily. Therefore, some measures should be taken to decrease this deterioration. • The results from the new network show that taking some measures might somehow retain the performance of the original network.
Link to papers	• Lu, Yongyi, Cewu Lu, and Chi-Keung Tang. "Online video object detection using association LSTM." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

Title (Mechanism ID)	E3.5.3_Opel_Adversarial_Attacks_Assessment_MECH-232004
	https://openaccess.thecvf.com/content_ICCV_2017/papers/Lu_Online_Video_ICCV_2017_paper.pdf • Wei, Xingxing, et al. "Sparse adversarial perturbations for videos." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019. https://arxiv.org/pdf/1803.02536.pdf

5.3.2.1 Opel /Stellantis

Title (Mechanism ID)	E3.5.3_Opel_OcclusionSensitivity_MECH-301022 (will be removed, 10.05.2022, as no sufficient information is given)
Leading and involved Partners (Name)	(Ahmed Hammam), Patrick Feifel (acc. to Method Catalog)
Mapping to Taxonomy Tree	Safe AI Mechanisms > Robustness Testing > Natural Corruptions > Image Artifacts > Other Sensor Effects
Short Description of Mechanism	We are focused on the task of 2d Object Detection for pedestrians. Thereby it is well known that the occlusion of objects might cause serious misclassifications or -localizations and False Negatives. To address the individual sensitivity for occluded objects for a given model, we run different shapes and sizes of an occluding patch in a systemical order through the original image. This setting seems to be reasonable to consider because the receptive field of a specific anchor box isn't perfectly aligned with the generated 2d BB. The final classification or localization cell sees more of the original image than the bounding box suggests. For each image in the test dataset with True Positives, we generate an augmented image set with occluded patches. We conduct the inference and measure deviations to the reference True Positives.
Used Data (train/val/test)	A2D2 and KI-A Tranche02 BIT
Used DNN / Task	SSD-r1-v3
Main Safety Concern being adressed	Unreliable confidence information (SC-1.1)
Summary of experiment results	Although a metric was defined, the method was abandoned due to high computational cost.

Title (Mechanism ID)	E3.5.3_Opel_OcclusionSensitivity_MECH-301022 (will be removed, 10.05.2022, as no sufficient information is given)
Summary of effectiveness compared to baseline / Level of Effectiveness	Method and metric are limited to a relative evaluation for a set of DNNs.
Potential Evidences for an Safety Assurance Case	Relative guarantees: One DNN performs better than another with respect to their occlusion sensitivity.
Link to papers	Zeiler, Fergus - Visualizing and Understanding Convolutional Networks

5.3.2.2 Fraunhofer IAIS

Title (Mechanism ID)	E3.5.3_Interpretable_student_networks_for_introspection_of_teacher_via_knowledge_distillation_Fraunhofer_MECH-050881
Leading and involved Partners (Name)	Julia Rosenzweig
Mapping to Taxonomy Tree	Safe AI Mechanisms ==> Interpretability ==> Global Interpretation ==> post-hoc Global Interpretation ==> interpretable Surrogates ==> Black-box; model agnostic
Short Description of Mechanism	After training an interpretable-by-design student network to be "close" to the teacher network (trained on the same dataset), we develop insights on/for it by searching for shortcut patches (i.e., image patches that are highly predictive of the concept of interest - target class - although not containing any pixels of that concept). In a next step, we test these insights on the teacher. We check whether the image augmented with an additional copy of the shortcuts patch is misclassified by the teacher (as belonging to the target class). Thus, we get an introspective view on a (black box) teacher network and can check its decision making and detect possible failure modes via transparent and interpretable students. Finally, we use this approach to also compare datasets w.r.t. their feature diversity (assumption: More diverse features allow for more shortcuts to be learned which are then found with using our approach).
Used Data (train/val/test)	As dataset we have used: • real: • A2D2

Title (Mechanism ID)	E3.5.3 Interpretable student networks for introspection of teacher via knowledge distillation Fraunhofer MECH-050881

	• Cityscapes
	• synthetic:
	  • GTA5
	  • Synscapes
	  • BIT-TS Tranche 4
	The training dataset of the teacher is at the same time also the dataset for training the student. However in case of knowledge distillation, labels for training the student are extracted from the teacher's output on the training dataset (potentially also initial labels of the training dataset).

Used DNN / Task	Task: Classification , DNNs: ResNet, BagNet
Main Safety Concern being adressed	Incomprehensible behavior (SC-1.3), Insufficient Plausibility (SC-1.4), Data distribution is not a valid approximation of real world (SC-2.1).
Summary of experiment results	We successfully use BagNets as interpretable-by-design networks for the analysis of black box classification networks. Our method effectively identifies shortcut patches that provoke misclassification for the black box teacher network under investigation - indicating possible risks due to non-generalizing concepts learned. We published the results on A2D2 in our publication (Rosenzweig et al.: Patch Shortcuts: Interpretable Proxy Models Efficiently Find Black-Box Vulnerabilities, CVPR-SAIAD Workshop 2021) and extended the evaluations onto multiple real and synthetic datasets, finding relevant patch shortcuts.
Summary of effectiveness compared to baseline / Level of Effectiveness	While the method can definitely identify relevant patch shortcuts of the black box network and thus point to weaknesses that have to be mitigated, there are several aspects one needs to consider: • It seems that the results (and the diversity of identified shortcut patches and with this the identified weaknesses of the teacher) depend on the dataset diversity itself: In a dataset that has rather low variety of scenes, it can happen that those scenes already contain shortctuts so that the relative success of the method is high although the shortcut patches themselves might be quite similar and one is able to identify only a small part of vulnerabilities. • It seems that the concept for which shortcuts are searched needs to reasonably large (e.g. covering enough pixels realtive to image size), as the method seems to work well for finding shortcuts for

Title (Mechanism ID)	E3.5.3 Interpretable student networks for introspection of teacher via knowledge distillation Fraunhofer MECH-050881
	class car but is not really effective for pedestrians (on KI-A data). Concerning the use as a dataset comparison method, the evaluation results are unclear: on some synthetic datasets, there are many shortcuts identified, on others rather few. So, the effectiveness inside real or synthetic datasets varies too much and leads to inconclusive results.
Potential Evidences for an Safety Assurance Case	Our method effectively finds patch shortcuts - i.e. image patches that the black box bases its decision on although they do not contain pixels of the class of interest. A network that relies on such patches poses a potential safety threat (shortcuts do not generalize) and such vulnerabilities should be addressed. On the flipside - if no patch shortcuts are identified with our method, this provides evidence that the network under inspection works in a plausible way and likely does not contain vulnerabilities in form of learned patch shortcuts.
Link to papers	Our publication on SAIAD Workshop 2021: Rosenzweig et al.: Patch Shortcuts: Interpretable Proxy Models Efficiently Find Black-Box Vulnerabilities, CVPR-SAIAD Workshop 2021

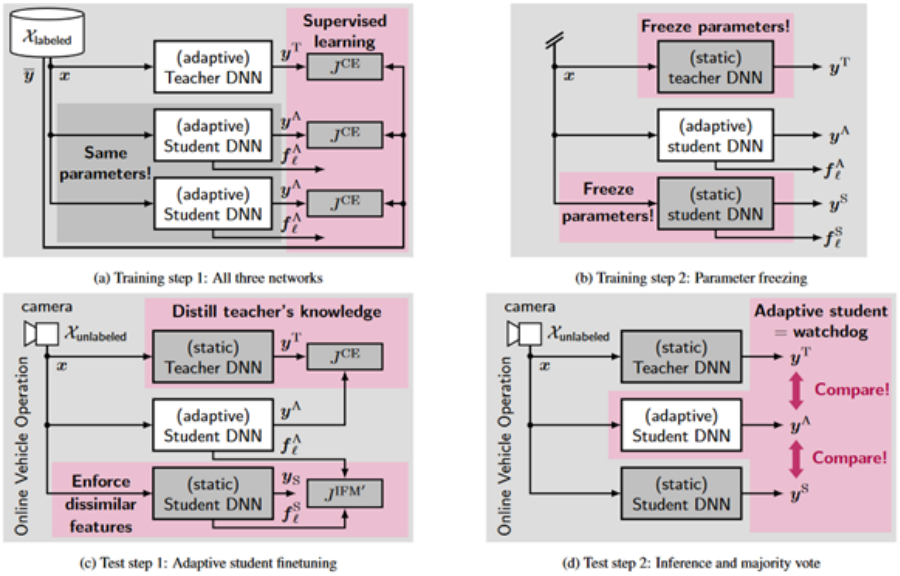
5.3.2.3 TU Braunschweig

Title (Mechanism ID)	E3.5.3 VW TUBS OutputOverlap MECH-292206
Leading and involved Partners (Name)	Andreas Bär
Mapping to Taxonomy Tree	Robustification ==> Adversarial Robustness ==> Robustification on Model Level ==> Modification of Architecture ==> Teacher Student Redundancy
Short Description of Mechanism	The semantic segmentation produces a pixel-wise semantic class prediction of the original input image. Here, a teacher-student triplet is used, with two architecturally identical students, one being adaptive (= parameters are optimized) and the other one being static (=parameters are frozen), and one static (=parameters are frozen) teacher. The teacher here is considered as the black box model to be monitored. At first, all three networks are pretrained on the same dataset (e.g., Citycapes). Second, the teacher as well as one of the students are frozen to obtain a static teacher, a static student, and an adaptive student. Then, the adaptive student is further optimized using the output of the teacher on unseen data (of the same domain!) as well as the feature representations (i.e., feature maps) of the static student. The aim is to robustify the adaptive student against adversarial attacks computed for both static networks. This way, if one

Title (Mechanism ID)	E3.5.3_VW_TUBS_OutputOverlap_MECH-292206

of the static networks is attacked, always two networks align with their predictions and thus have a high output overlap.

An overview of the training and test steps can be seen in the following figure with explaining text inside the figure caption.



(a) Training step 1: All three networks

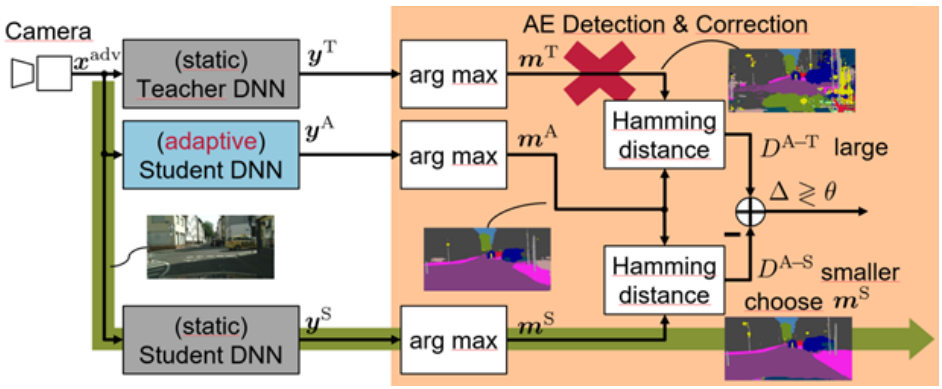
(b) Training step 2: Parameter freezing

(c) Test step 1: Adaptive student finetuning

(d) Test step 2: Inference and majority vote

Figure 2: General concept of the **proposed teacher-student framework**. **Training step 1**: Pretrain all deep neural networks (DNNs) on labeled data $x \in \mathcal{X}_{\text{labeled}}$, with $\overline{y}$ being the respective ground truth label of x, using the cross-entropy (CE) loss J^{CE} between the ground truth $\overline{y}$ and the respective DNNs' outputs y^{T}, y^{A}. Note that both student DNNs have the exact same architecture and parameters. **Training step 2**: Freeze the parameters of the teacher DNN and one student DNN. Both are now considered being static. **Test step 1**: Finetune the adaptive student DNN on unlabeled data $x \in \mathcal{X}_{\text{unlabeled}}$ using the CE loss J^{CE} between the static teacher's soft outputs y^{T} and the adaptive student's soft outputs y^{A}, and the layer-dependent inverse feature matching (IFM′) loss $J^{\text{IFM}'}$ between the static student's feature representations f_ℓ^{S} and the adaptive student's feature representations f_ℓ^{A} at layer ℓ. **Test step 2**: During inference, use the adaptive student DNN as a watchdog by comparing its output y^{A} with the static DNNs' outputs y^{T}, y^{S}. Whenever there is more similarity, the respective static network output is taken as the corrected semantic segmentation output. Test steps 1 and 2 are alternatingly executed in the online watchdog application of the adaptive student DNN, *e.g.*, in a vehicle (no labels required!). For the experimental validation in this paper, test step 1 is cast to the training (training step 3) and the alternating call schedule is omitted.

The following figure illustrates the overall workflow of the adversarial example detection and correction algorithm based on the output overlap. It gives a more detailed overview of Figure 2 (d).



The outputs of the static networks are compared to the output of the adaptive network using the Hamming distance. Whenever we measure a high Hamming distance, we have a big misalignment in the semantic

Title (Mechanism ID)	E3.5.3_VW_TUBS_OutputOverlap_MECH-292206
	segmentation output between the adaptive student and one of the static networks.

Used Data (train/val/test)	• Cityscapes
	  • training: 2.975
	  • training coarse: 19.998
	  • validation: 500
	• Release #4: BIT TS tranche 3 + 4:
	  • training: 31.757
	  • validation: 7.419
	  • test: 9.903

Used DNN / Task	ENet (semantic segmentation): ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation (arXiv, June 2016), GitHub code base ERFNet (semantic segmentation): ERFNet: Efficient Residual Factorized ConvNet for Real-time Semantic Segmentation (IEEE Trans. on Int. Trans., vol. 19, no. 1, 2018), own implementation based on a GitHub code base. DeepLabv3+ (semantic segmentation): Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (ECCV 2018), project partner (Intel) implementation based on a GitHub code base.
Main Safety Concern being adressed	**Brittleness of DNNs (SC-1.2),** Unknown behaviour in rare critical situations (SC-2.6), Safety-aware metrics (SC-3.1)
Summary of experiment results	In our experiments, we tested different Teacher-Student-Setups (DeepLabv3+, ERFNet, and ENet). We conclude that using the DeepLabv3+ as a teacher and the ENet as a student yield the best results in terms of adversarial attack detection and decoupled behaviour of teacher attacks and student attacks. Currently, we still see an issue with clean images, where the system is biased towards attack detection on the teacher (DeepLabv3+). However, as both networks, i.e., teacher and student, perform well on clean images, we only see a slight drop in clean performance when we use our detection scheme. However, we did not observe this behaviour with data from Cityscapes.

Title (Mechanism ID)	E3.5.3_VW_TUBS_OutputOverlap_MECH-292206
Summary of effectiveness compared to baseline / Level of Effectiveness	The plain baseline does not offer any indicators for adversarial attack exposure. On the other hand, the additional use of our method indicates a possible adversarial attack on a per-image basis. However, this comes with a computational overhead in form of additional models in a teacher-student setup and a performance trade-off due to model averaging.
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis**: The method addresses the safety concern *Brittleness of DNNs (SC-1.2)*. It enhances the performance of a DNN under attack and/or can be used for adversarial attack detection. 2. **Evidences for a safety assurance case**: In our experiments, we observed that an adversarial attack detection is possible. However, with the current setup, we also observe that we falsely detect adversarial attacks on clean images. Thus, on clean images we would expect a smaller performance. However, usually DNNs perform well on clean images, and we only see a slight performance drop. 3. **Further tests**: Stronger evidences can be derived by further experiments on clean images to mitigate the effect of false adversarial attack detection on clean images.
Link to papers	On the Robustness of Redundant Teacher-Student Frameworks for Semantic Segmentation (CVPR SAIAD 2019) Robust Semantic Segmentation by Redundant Networks With a Layer-Specific Loss Contribution and Majority Vote (CVPR SAIAD 2020)

5.4 E3.5.4 Final: nur projektintern für KI Absicherung verfügbar

5.5 E3.5.5 Final: nur projektintern für KI Absicherung verfügbar

AP3.6 Aggregierte Methoden

5.6 E3.6.1 Final: E3.6.2 Final: Auflösung von Bewertungsredundanzen und Synergien

(zur Veröffentlichung)

5.6.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Methods and Mechanisms*
Classification level	*PU*

5.6.2 Description of the result

3.6.2 encompasses a set of aggregation and ensembling techniques for providing enhanced self-assessment (e.g. uncertainty measures, confidence values). Redundancies are created by combining multiple algorithms from TP3 to improve regarding safety requirements.

5.6.2.1 FZI

Title (Mechanism ID)	Mixture-of-Experts Layers Embedded in CNNs_MECH-741255
Leading and involved Partners (Name)	FZI
Mapping to Taxonomy Tree	
Short Description of Mechanism	With hard constraints, the weights of certain experts are allowed to become zero, while soft constraints balance the contribution of experts with an additional auxiliary loss. As a result, soft constraints handle expert utilization better and support the expert specialization process, hard constraints mostly maintain generalized experts and increase the model performance for many applications. Our findings demonstrate that even with a single dataset and end-to-end training, experts can implicitly focus on individual sub-domains of the input space. Experts in the proposed models with MoE embeddings implicitly focus on distinct domains, even without suitable predefined datasets. As an example, experts trained for CIFAR-100 image classification specialize in recognizing different domains such as sea animals or flowers without previous data clustering. Experiments with RetinaNet and the COCO dataset further indicate that object detection experts can also specialize in detecting objects of distinct sizes.

Title (Mechanism ID)	Mixture-of-Experts Layers Embedded in CNNs_MECH-741255

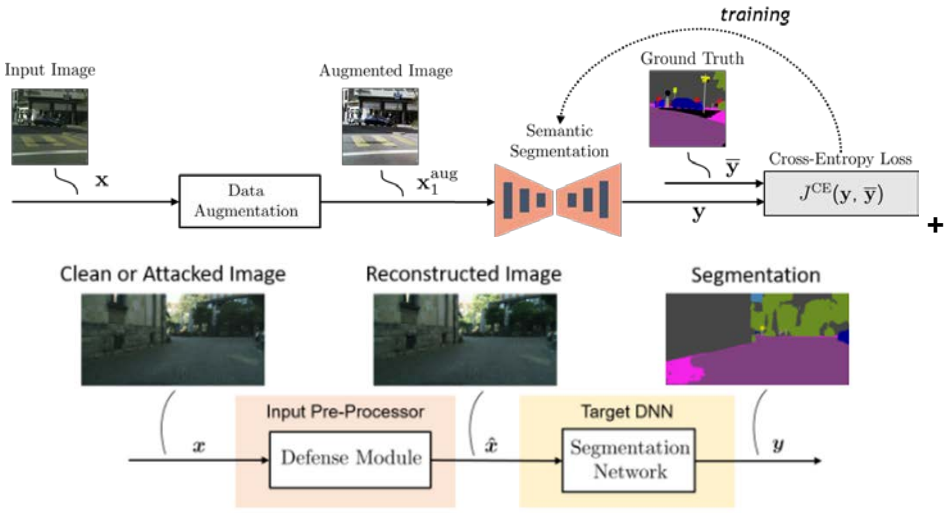
Used Data (train/val/test)	• **CIFAR-100 image classification**:
	  • official train/test split is used
	• **COCO object detection**:
	  • 2017 train and test-dev split.
	  • Metrics are computed on the 2017 validation set containing 5,000 images and also by uploading detection results onto an evaluation server.

Used DNN / Task	For the **image classification** experiments, we use ResNet-18. To match the input dimension of CIFAR-100, we reduce the filter size of the input layer to 3 × 3 with stride 1 and remove max-pooling. We use the official PyTorch implementation available at https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py For the **object detection** experiments, we use a pretrained RetinaNet as a baseline. It uses ResNet-50 backbone and an image scale of 600. We train all models using $\gamma = 2$ and $\alpha = 0.25$ for focal loss. All models are based on an unofficial PyTorch reimplementation for RetinaNet, available at https://github.com/yhenon/pytorch-retinanet (Apache License 2.0). F
Main Safety Concern being adressed	SC-1.3 Incomprehensible Behavior
Summary of experiment results	• MoE embedding allowed us to increase the model capacity extensively while keeping an appropriate inference time by determining the number of active experts per forward pass. In our experiments both soft- and hard-constrained models achieved comparable test results, some MoE models even outperformed the baseline and showed accuracy improvements. The prediction performance of MoE models can be even further improved by increasing the number of active experts. This way, we can balance the trade-off between model accuracy and computational complexity with a single parameter. This helps in training models with many experts and parameters on powerful hardware and then scale the runtime complexity based on the deployment device. • Experts trained end-to-end without predefined dataset splits were still able to specialize in distinct subdomains. For image classification task, experts focused on specific class groups,

Title (Mechanism ID)	Mixture-of-Experts Layers Embedded in CNNs_MECH-741255
	whereas for object detection, they specialized on objects of distinct sizes.
Summary of effectiveness compared to baseline / Level of Effectiveness	• Models with embedded MoE layers outperform the baseline, especially when a large number of activated experts is involved • The effect was demonstrated on two different computer vision tasks
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis**: The method addresses the safety concern "Incomprehensible Behaviour" of DNNs 2. **Evidences for a safety assurance case**: The proposed constraints tackle the problem from different angles and lead to different MoE behavior. Hard constraints result in a better overall performance and generalized experts, although the mean importance constraint is particularly prone to the dying experts problem. Soft constraints, on the other hand, lead to better expert specialization. 3. **Further tests**: Stronger evidences can be derived by further expeiments with a broader spectrum of architectures So far no concrete implementation in safety argumentation in KI-Absicherung has been done.
Link to papers	Submitted to ECCV 2022: 2022-10-ECCV Tel Aviv - Pavlitskaya et al. - Balancing Expert Utilization in Mixture-of-Experts Layers Embedded in CNNs

5.6.2.2 Volkswagen

Title (Mechanism ID)	Robustness via Data Augmentation & Pre-Processors
Leading and involved Partners (Name)	Volkswagen
Mapping to Taxonomy Tree	Robustification → Robustification on model level → Modification of Training process → Robustness-oriented Loss Functions Robustification → Robustification on system level → Pre-processing strategies
Short Description of Mechanism	This mechanism is combining online and offline mechanisms for improving robustness of pedestrian detection. We combine both, data augmentation during training (standard data augmentation and AugMix - see MECH-093532) and pre-processors (denoising autoencoders and Wiener filtering - see MECH-088167). Various combinations of

Title (Mechanism ID)	Robustness via Data Augmentation & Pre-Processors
	mechanisms are implemented and evaluated for their overall robustness improvement compared to the baseline. Extensive ablation studies with e.g. different loss types are performed. 
Used Data (train/val/test)	Cityscapes, Cityscapes-c, ACDC
Used DNN / Task	ICNet, Semantic Segmentation
Main Safety Concern being adressed	SC-1.2 Brittleness of DNNs
Summary of experiment results	• AugMix-JSD is a very strong data augmentation method that helps improve robustness to a variety of unseen real-world distributions → improved generalization. • Robustness tests done on simulated distributions scales to real-world distributions. Safe to assume that if robustness is improved on simulated noise, might also be true in real noise/adverse situations. • VQVAE trained on random gaussian noise (RN) helps improve adversarial robustness significantly when compared to comparable denoising autoencoders.
Summary of effectiveness compared to baseline / Level of Effectiveness	• The combination of both data augmentation during training and pre-processors at runtime significantly outperforms all baselines (no robustification, only augmentation, only filtering)

Title (Mechanism ID)	Robustness via Data Augmentation & Pre-Processors
Potential Evidences for an Safety Assurance Case	• Improved robustness to a variety of unseen real-world distributions • Robustness tests done on simulated distributions scales to real-world distributions
Link to papers	Augmix https://arxiv.org/pdf/1912.02781.pdf Wiener filter https://arxiv.org/pdf/2012.01558.pdf

5.7 E3.6.3 Final: Implementierung von aggregierten Methoden und Maßnahmen und Bewertung hinsichtlich KPIs (zur Veröffentlichung)

5.7.1 Formal Classification

Criteria	Classification according to VHB
Type of result	*Code*
Group/Cluster	
Type of content	*Methods and Mechanisms*
Classification level	*PU*

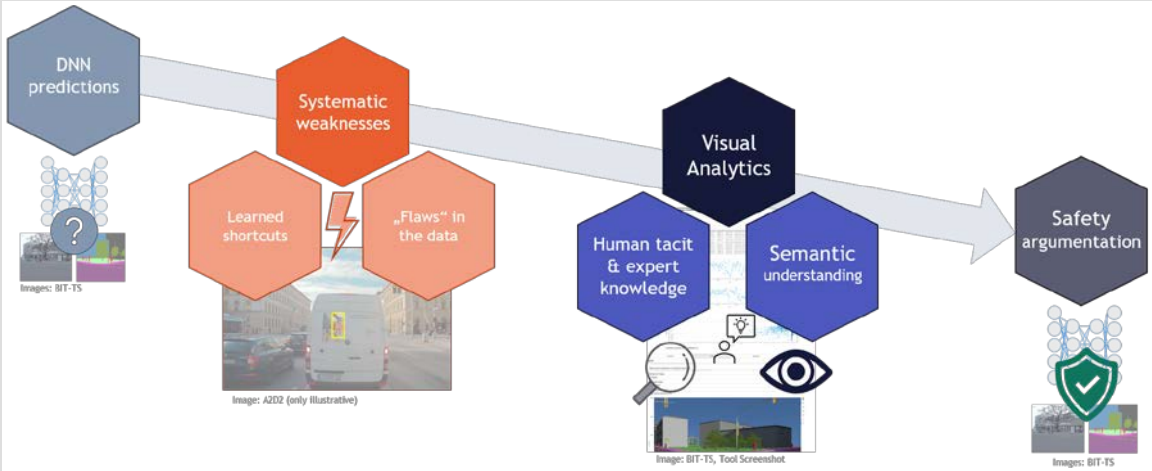
5.7.2 Description of the result

E3.6.3 provides a collection of aggregated Methods and measures. One or more mechanism from AP3.3, AP3.4 and AP3.5 are combined in order to further improve regarding safety requirements.

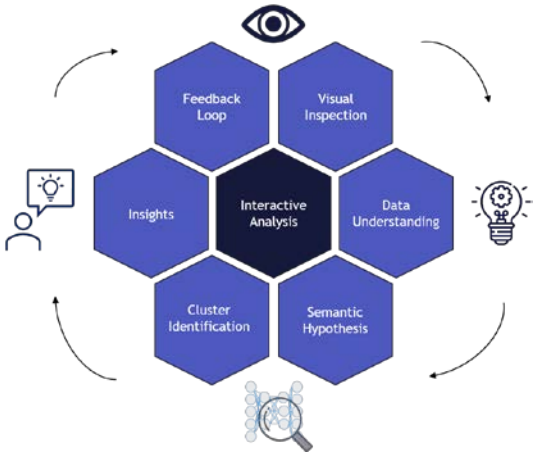
5.7.2.1 Fraunhofer IAIS

Title (Mechanism ID)	Aggregation based dependency analysis of neural networks with Visual Analytics (MECH-116617)
Leading and involved Partners (Name)	Fraunhofer IAIS
Mapping to Taxonomy Tree	Safe AI Mechanisms > Interpretability > Local Interpretation > Visual Analytics
Short Description of Mechanism	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

Title (Mechanism ID)	Aggregation based dependency analysis of neural networks with Visual Analytics (MECH-116617)
	of networks. This can be achieved by understanding the semantic concepts of the data. As an automated analysis of semantics is difficult, we utilize 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 • **Understand semantic concepts** of DNN predictions and data (similar & different failure patterns) • **Aggregate methods and metrics** in an interactive analyzation process • **Identify semantic clusters** to gain insights • **Visually compare DNN performance** with/without usage of safety mechanisms (w.r.t. the identified semantic concepts) All in all, this enables a stringent safety argumentation that can be built upon human understandable arguments.

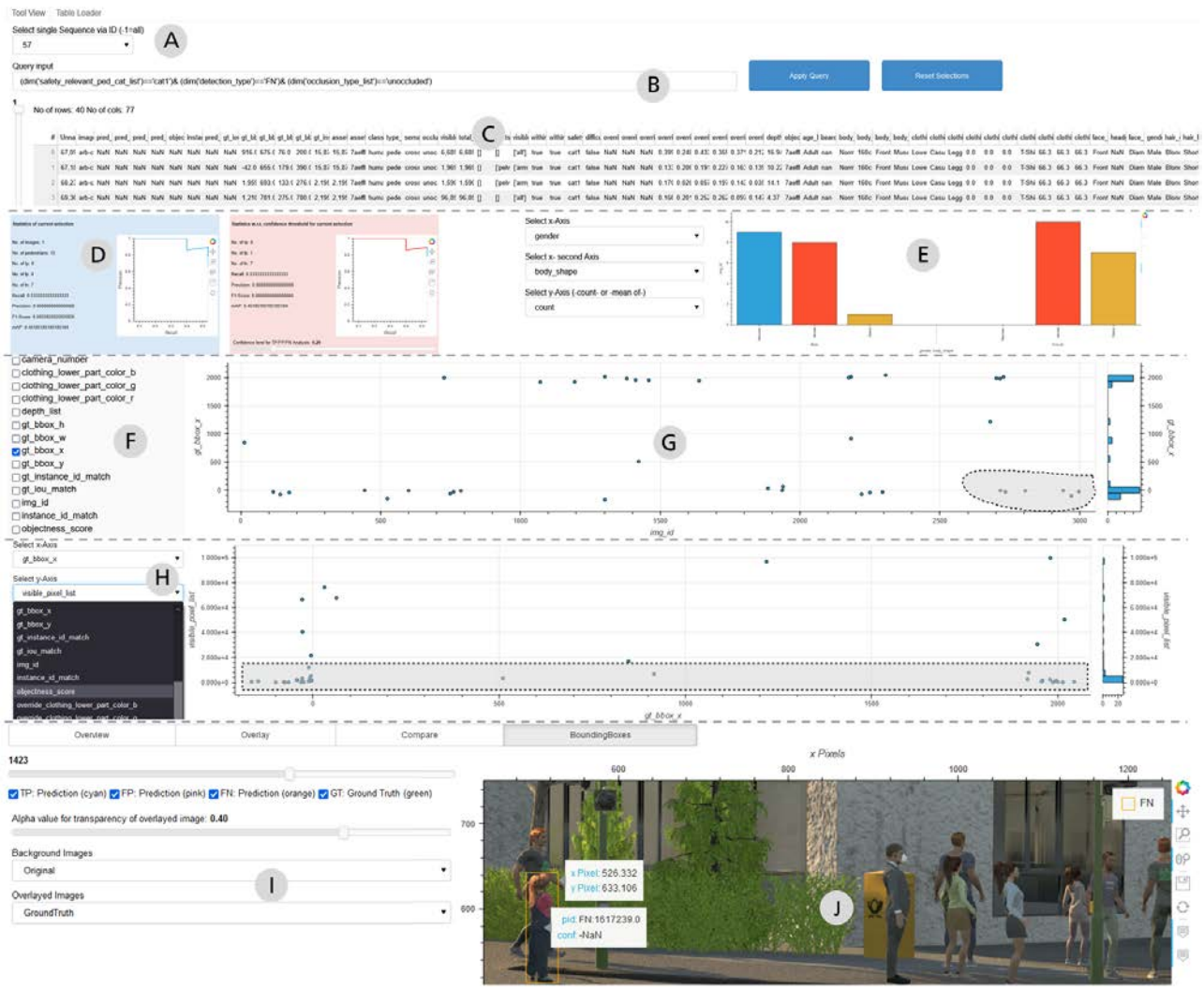


Understandig semantic concepts of the data with Visual Analytics and human knowledge



Building blocks of the interactive analysis process

Screenshots of features of current version (the widgets are stacked/truncated to provide better overview):



A: Dropdown to select sequence

B: Input for arbitrary textual queries

C: Table of metadata

D: Observe statistics of current selection and influence of adjustments of specific parameters (here: confidence threshold)

E: Histogram for categorical and numerical data

F: Selection for plots to be displayed (based on available numerical metadata)

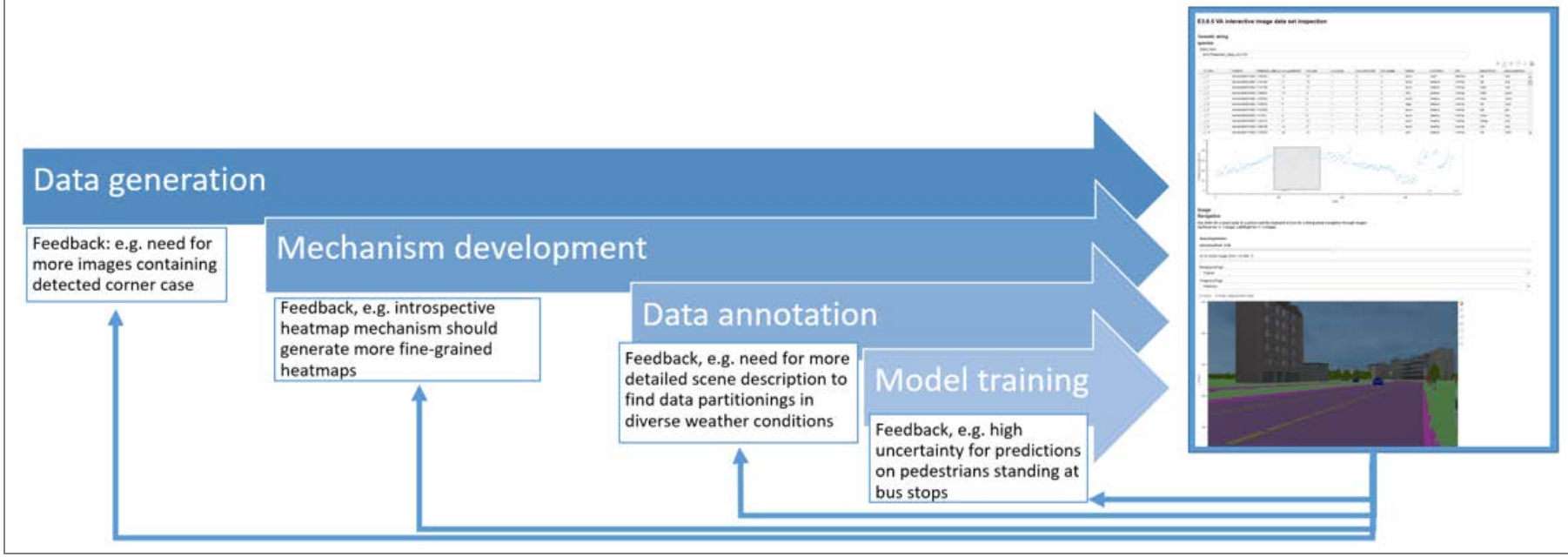
G: Plots for selected metadata

H: Dropdown for X- and Y-axis for a correlation plot

I: Slider for the transparency of the overlayed images and dropdown menus to select the foreground and background image source (e.g. ground truth, semantic segmentation, VAE images etc)

J: Current image with panning, zooming and mouse-hover information

By combining all those visual analyzation methods for neural networks/data sets into one interactive tool, the trained networks with their corresponding data sets can be thoroughly examined. This enables the user to interactively analyse the KPIs with respect to safety and e.g. look for corner cases. These insights can then in turn be used to enhance the training methods of the neural networks or the data set generation. By this, a feedback loop could be established between data generation, neural network training and analyses of both, as sketched below:



Title (Mechanism ID)	Aggregation based dependency analysis of neural networks with Visual Analytics (MECH-116617)
Used Data (train/val/test)	Bit-TS Tranche 3, Meta data Mackevision Tranche 3, Meta data
Used DNN / Task	Intel DeepLabV3+, Opel SSD
Main Safety Concern being adressed	Incomprehensible behavior (SC-1.3)

Summary of experiment results	**Experiments with BIT-TS Tranche 3 and Intel Deeplab V3+ and VAE**
	• Experiments:
	  • general exploration of segmentation performance
	  • exploration of VAE results
	  • focus on data and labels
	• Results:
	  • Pedestrians standing behind transparent bus stop are recognized by model but not labeled in the ground truth (fully occluded by bus stop)
	  • VAE has high error on scenes with large sky/horizon and few buildings
	
	Experiments with Mackevision Tranche 5 and OpelSSD
	• Experiments:
	  • investigated performance of pedestrian detection with a focus on safety categories
	  • focus on pedestrians that are not recognized
	• Results:
	  • A lot of FN are truncated Pedestrians at left/right edges -> remove from evaluation
	  • Investigation of False Negatives of Cat1 pedestrians
	  o FN Cat1 Pedestrians are mostly located on the left and right edges of the images
	  o a lot are truncated (though metadata indicates „truncated=FALSE“, unoccluded=TRUE)

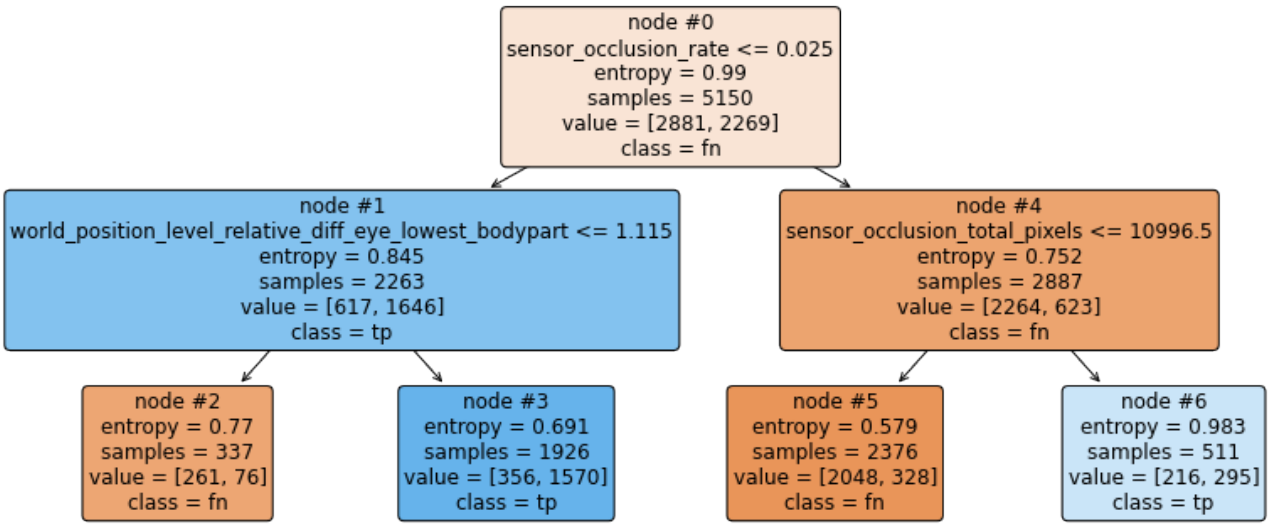
Title (Mechanism ID)	Aggregation based dependency analysis of neural networks with Visual Analytics (MECH-116617)
	o A lot of big pedestrians in the middle of the images are not recognized (visible_pixel) • Set uncertainty threshold to 0.2 for better F1 score with same mAP • Focus on analysis of „relevant FNs and FPs“ -> proposal for area based metric • Faulty medata data cat5 pedestrians („semantic_area=NaN“ or „ semantic_area=other) + bad detction of „huge“ pedestrians (probably need more training data) • Faulty TP for fully occluded assets -> network learned something wrong? • Preliminary Analaysis NCAP Scenario (based on results provided by ZF) o Pose4 recognition worse than others, corruption severity irrelevant,..
Summary of effectiveness compared to baseline / Level of Effectiveness	Not applicable (The VA Tool does not directly improve the DNN model, thus results are not directly comparable to baseline. However, with the interesting cases and evidences found, feedback to the method/DNN/data developer can be given to then mitigate the weaknesses/improve the safety of the DNN.).
Potential Evidences for an Safety Assurance Case	• Valuable for ***SC-1.3: Incomprehensible Behavior.*** The VA Tool reveals showing interesting cases/flaws in data and NN decisions. Literature is proving that VA is a well established approach. • Several evidences for interesting cases/flaws in data or NN decisions have been found, e.g. • a lot of big pedestrians in the middle of the images are not recognized (visible_pixel) • Faulty medata data cat5 pedestrians („semantic_area=NaN“ or „ semantic_area=other) + bad detction of „huge“ pedestrians (probably need more training data • Further investigate the data and NN decisions based on the hypothesis/insights gained so far. Especially focus on pedestrians that are not recognized and investigate whether the model has learnt something spurious on more data/image sequences.

Title (Mechanism ID)	Aggregation based dependency analysis of neural networks with Visual Analytics (MECH-116617)
Link to papers	• **External papers**: Papers 1 to 4 focus on a VA tool for a specific deep learning application, paper 5 is a survey paper on VA in deep learning. a. Spinner et al: "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning", 2019 b. Kwon et al: "RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records", 2019 c. Wexler et al: "The What-If Tool: Interactive Probing of Machine Learning Models", 2019 d. Pezzotti et al: "DeepEyes: Progressive Visual Analytics for Designing Deep Neural Networks", 2018 e. Hohman et al: "Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers", 2018 f. Choo, Liu: "Visual Analytics for Explainable Deep Learning", 2018 g. Cashman et al: "A User-based Visual Analytics Workflow for Exploratory Model Analysis", 2018 h. Liu et al: "Towards a better analysis of Deep Convolutional Neural Networks", 2019 i. Jiang et al: "Recent Research Advances on Interactive Machine Learning", 2018 j. Keim et al: " Visual Analytics: Definition, Process, and Challenges", 2008

5.7.2.2 EFS

Title (Mechanism ID)	Analysis of the Confidence Measures in the Context of ISO 21448 SOTIF (MECH-280115)
Leading and involved Partners (Name)	EFS
Mapping to Taxonomy Tree	Safe AI Mechanisms > Dataset Optimization > Corner Case Selection > Learning Feature Representations for Normality > Use discriminative Model in Representation Space
Short Description of Mechanism	According to ISO: 21448: "*Road vehicles — Safety of the intended functionality" (SOTIF)*, functional insufficiencies (FI) are deficiencies in

Title (Mechanism ID)	Analysis of the Confidence Measures in the Context of ISO 21448 SOTIF (MECH-280115)
	the functionality of a system, leading to erroneous and hazardous behavior when triggered. The particular set of conditions that is able to unveil the misbehavior is called triggering conditions (TC). For the identification and evaluation of potential functional insufficiencies and triggering conditions, clause 7 of the SOTIF suggests establishing a systematic method, which should consider knowledge gained from similar projects, experts and field experience. As further detailed in Annex B, exploratory methods (like exploratory analysis and exploratory simulation) are regarded as useful bottom up tools for identifying triggering conditions. To enable learning from older projects and generate the adequate work products (safety artifacts) in accordance to ISO: 21448 Clause 7 and Annex B, we need a container for collecting information from exploratory analysis. With this method, we provide templates for a SOTIF Failure Mode and Effects Analysis (S-FMEA). Using the processing capabilities of the metric benchmark tool, we evaluate 2d bounding-box predictions of a DNN for pedestrian detection. As extension, we build decision trees models, trained to discriminate between true positive and false negative instances based on meta-information features. We then parse the decision tree structure to part-automatically fill in the S-FMEA template, regarding triggering conditions and corner case descriptions.



Title (Mechanism ID)	Analysis of the Confidence Measures in the Context of ISO 21448 SOTIF (MECH-280115)
Used Data (train/val/test)	MV Tranche 4, 5, 6
Used DNN / Task	TP1 Opel SSD
Main Safety Concern being adressed	Specification of the ODD (SC-2.4)
Summary of experiment results	• Decision tree classifiers can predict true positive and false negative pedestrian detections with a test accuracy of 73-76% for safety relevant instances within KIA dataset using meta-information as input • Balanced datasets between true positive and false negative pedestrian instances are desirable, which may pose a restriction to the applicability of the proposed mechanism in its current state • The resulting decision tree structure depends strongly on the input data distribution
Summary of effectiveness compared to baseline / Level of Effectiveness	The method provides a tool to analyze 2d bounding-box detections for pedestrians and part-automatically fill in S-FMEA templates. Therefore, there is no baseline method that could be used to compare performance to.
Potential Evidences for an Safety Assurance Case	1. **Safety hypothesis:** The method addresses the safety concern *Specification of the ODD (SC-2.4)*. We use a decision tree algorithm trained to discriminate between true positive and false negative pedestrian instances, based on features in the available meta-information. Processing the tree structure we derive corner case descriptions. 2. **Evidences for a safety assurance case:** As demonstrated in our proof-of-concept experiments, training and parsing of decision trees can in principle be used to identify sub-domains in the meta-information where the regarded 2d bounding-box detection performs comparatively poor with regards to the recall, i.e. the fraction of pedestrians that were detected. 3. **Further tests:** Stronger evidences can be derived by filling in the missing parts of the S-FMEA analysis. In particular the assessment of occurrence and severity as well as design strategy should lead to

Title (Mechanism ID)	Analysis of the Confidence Measures in the Context of ISO 21448 SOTIF (MECH-280115)
	explicit measures to improve the SOTIF and ultimately reach a rationale of acceptance. So far no concrete implementation into safety argumentation in KI-Absicherung has been done.
Link to papers	No external sources available

5.7.2.3 Valeo

Title (Mechanism ID)	Improving Predictive Performance and Calibration by Weight Fusion (MECH-109816)
Leading and involved Partners (Name)	Valeo
Mapping to Taxonomy Tree	Safe AI Mechanisms > Aggregation > Weight Fusion
Short Description of Mechanism	The goal is to create a deep ensemble for the task of semantic segmentation in which the members have different failure modes. Through the different failure modes, the members should complement each other, so that finally a higher accuracy is achieved. In order to implement a different error mode among the members as much as possible, uncertainty modeling (aleatory and epistemic) is integrated into the training process of the deep ensemble. At runtime, the outputs of the deep ensemble members are fused at feature map level using uncertainty modeling. The Deeplabv3+ for semantic segmentation and the BDD100k dataset are used to perform the experiments.
Used Data (train/val/test)	Cityscapes, BDD100K, ACDC
Used DNN / Task	DeeplabV3+
Main Safety Concern being adressed	Insufficient generalization capability (FI-1)
Summary of experiment results	Averaging predictions of a DNN ensemble (deep ensemble) is a popular method to improve predictive performance and calibration in a variety of benchmarks and Kaggle competitions. The runtime and computational cost of deep ensembles grow linearly with the number of deep ensemble members, making them unsuitable for many applications. The averaging of weights circumvents this disadvantage

Title (Mechanism ID)	Improving Predictive Performance and Calibration by Weight Fusion (MECH-109816)
	and leads, if at all, to a higher computational effort in generating the weights for averaging. We show that weight fusion (WF) can lead to a significantly improved performance and calibration. We demonstrate this using state of the art segmentation CNNs and Transformer as well as real world datasets such as BDD100K and Cityscapes. We describe what prerequisites the weights must meet in terms of weight space, functional space and loss. We present a new test method (called oracle test) to measure the functional space between weights. Furthermore, we compare WF with similar approaches and show our superiority for in- and out-of-distribution data in terms of predictive performance and calibration. (Taken from abstract: Saemann et al.: "Improving Predictive Performance and Calibration by Weight Fusion in Semantic Segmentation")
Summary of effectiveness compared to baseline / Level of Effectiveness	We have presented a strategy to fuse two or more weight files into a single weight file. The creation of the weight files can be done by training from the scratch or by a finetuning process. We have derived from extensive experiments that the similarity in weight space between the weights must be as large as possible and at the same time the similarity in functional space must be as small as possible. To measure the functional space, we introduced a new testing method called oralce testing. For the creation of the weights through a finetuning process, the validation loss plays a decisive role, which should be as close as possible to the local minimum. Based on intensive studies with SOTA architectures (CNNs and Transformer) we presented the improvements in predictive performance and calibration. We have shown that equal-weighted fusion as performed in SWA does not usually lead to the best results and can even be detrimental if the cosine similarity is too low. We outperformed stochastic weight averaging (SWA) on the BDD100K test data as well as under presence of a distribution shift on the ACDC data in performance and calibration. Furthermore, comparison with deep ensembles shows that our weight fusion has slightly better calibration and comparable performance to the deep ensemble with 3 members, albeit the latter requires three times the runtime. (Taken from conclusion: Saemann et al.: "Improving Predictive Performance and Calibration by Weight Fusion in Semantic Segmentation")
Potential Evidences for an Safety Assurance Case	I*t increases the generalization ability and confidence calibration.*

Title (Mechanism ID)	Improving Predictive Performance and Calibration by Weight Fusion (MECH-109816)
Link to papers	Saemann et al.: "Improving Predictive Performance and Calibration by Weight Fusion in Semantic Segmentation" (Not published yet. Planned for publication on ECCV 2022.)

5.8 E3.6.4 Final: nur projektintern für KI Absicherung verfügbar

5.9 E3.6.5 Final: nur projektintern für KI Absicherung verfügbar