KI ABSICHERUNG

Safe AI for Automated Driving

Abschlussbericht KI Absicherung

Gemeinsamer Abschlussberichts des Verbundprojektes KI Absicherung

(Öffentliche Version)

Version	1.0
Editoren	Dr. Stephan Scholz / PD Dr. Michael Mock
Projektkoordination	Volkswagen AG / Fraunhofer IAIS
Fälligkeit	31.12.2022
Erstellungsdatum	02.11.2022

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Dokumenteninformation

Autoren

Dr. Maram Akila / Fraunhofer IAIS

Dr. Andreas Albrecht / Robert Bosch GmbH

M.Sc. Karl Amende / Valeo Schalter und Sensoren GmbH

Fridolin Bauer / BMW Group

Dipl.-Ing. Frédérik Blank / Robert Bosch GmbH

Andreas Blattmann/ Universität Heidelberg

Nikolas Brasch / TUM

M.Sc. Michael Fürst / DFKI

Dr. Nicolas Gay / Luxoft

Sebastian Gerres / Merantix Momentum GmbH

Prof. Dr. Hanno Gottschalk / University of Wuppertal

Dr. Oliver Grau / Intel Corporation

Dr. Johannes Guenther / Intel Corporation

M.Sc. Korbinian Hagn / Intel Corporation

Dr. Philipp Heidenreich / Opel Automobile GmbH

Dr. Niels Heller / QualityMinds GmbH

Dr. Christian Hellert / Continental AG

M.Sc. Simon Heming / Robert Bosch GmbH

Dipl.-Ing. Martin Herrmann / Robert Bosch GmbH

M.Sc. Alexander Hirsch / Robert Bosch GmbH

Dr. Fabian Hüger / Volkswagen AG

Dr. Markus Huber / Mackevision Medien Design GmbH

M.Sc. Bastian Knerr / QualityMinds GmbH

PD Dr. Michael Mock / Fraunhofer IAIS

M.Sc. Svetlana Pavlitskaya / FZI

M.Sc Christian Pfister / ASTech GmbH

Dipl.-Ing. Andreas Rohatschek / Robert Bosch GmbH

M.Eng. Timo Sämann / Valeo Schalter und Sensoren GmbH

Dr. Martin Schels / Continental AG

M.Sc. Jonas Schneider / e:fs TechHub GmbH

M.Sc. Jan David Schneider / Volkswagen AG

Dr. Stephan Scholz / Volkswagen AG

Thomas Schulik / ZF Friedrichshafen AG

Dr. Loren Schwarz / BMW Group

Ulrich Seger / Robert Bosch GmbH

M.Sc. Bharat Shinde / Valeo Schalter und Sensoren GmbH

Martin Simon / Valeo Schalter und Sensoren GmbH

Dr. Thomas Stauner / BMW Group

Christian Witt / Valeo Schalter und Sensoren GmbH

Reviewer

Dr. Stephan Scholz / Volkswagen AG

PD Dr. Michael Mock / Fraunhofer IAIS

Fridolin Bauer / BMW Group

Dipl.-Ing. Frédérik Blank / Robert Bosch GmbH

Dr. Fabian Hüger / Volkswagen AG

Dr. Thomas Stauner / BMW Group

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 13.10.2022	Input auf Confluence	s.o.	s.o.
0.2	13.10.2022	Ausspielen Word und Layout	Bert Hildebrandt	EICT
0.3	13.-15.10.2022	Formelles Review und Layout	Dr. Nikos Papamichail	EICT
1.0	21.10.2022	Finalisierung	Dr. Nikos Papamichail	EICT

Inhaltsverzeichnis

Anhänge

Für die zahlreichen Einzelergebnisse wurde jeweils ein finaler Ergebnissteckbrief erstellt, die als Anhänge zum vorliegenden Bericht zur Verfügung stehen. Aufgrund der Größe der Dokumente sind sie aber nicht Teil dieses Berichts, sie wurden für jedes TP zusammengefasst und sind als zusätzliche 4 getrennte Dokumente verfügbar.

Anhang 1: KI Absicherung - Finale Ergebnissteckbriefe TP1 – KI Funktion

Anhang 2: KI Absicherung - Finale Ergebnissteckbriefe TP2 – Synthetische Daten

Anhang 3: KI Absicherung - Finale Ergebnissteckbriefe TP3 – Methoden und Maßnahmen

Anhang 4: KI Absicherung - Finale Ergebnissteckbriefe TP4 - Absicherungsstrategie

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung, Übersicht & Aufgabenstellung

Der Einsatz von KI ist ein Schlüssel für das hochautomatisierte Fahren. Im Projekt KI Absicherung haben KI- und Sicherheitsexperten aus Industrie und Wissenschaft eine Methodik für eine Sicherheitsargumentation, die Schwachstellen von KI-Funktionen systematisch identifiziert, messbar macht und entschärft entwickelt. Ziel des Projekts war es einen industriellen Konsens für einen methodischen Ansatz zur Absicherung von KI-Funktionen für den Anwendungsfall der Fußgängererkennung zu erreichen.

Die Absicherung von Funktionen, die KI-basierte Algorithmen nutzen, ist für die deutsche Automobilindustrie im internationalen Wettbewerb entscheidend. Im Projekt KI-Absicherung entwickelte ein Konsortium aus OEMs, Zulieferern, Technologieanbietern und wissenschaftlichen Einrichtungen einen "Industriekonsens" über eine Methodik, mit der inhärente Schwachstellen in KI-Funktionen identifiziert und systematisch entschärft werden können. Die Methodik beinhaltet auch einen systematischen Ansatz zur Ableitung einer stringenten evidenzbasierten Sicherheitsargumentation.

KI Absicherung ist ein Projekt der VDA-Leitinitiative autonomes und vernetztes Fahren aus der Projektfamilie Künstliche Intelligenz und maschinelles Lernen im automobilen Umfeld. Die KI Familie stellt eine einzigartige Kombination von Projekten dar, die für die deutsche Industrie- und Forschungslandschaft von herausragender Bedeutung sind. Domänenübergreifend legen alle vier Projekte und deren Zusammenspiel den Grundstein für die erfolgreiche Umsetzung von künstlicher Intelligenz für Fahrzeugkonzepte und -systeme der Zukunft. KI Absicherung zielt darauf ab, den abgesicherten Einsatz von KI im Fahrzeug zu ermöglichen; in KI Wissen wird bereits vorhandenes Wissen für KI nutzbar gemacht. KI Delta Learning steigert die Lernkompetenz der Netzwerke und das Projekt KI Data Tooling wird eine ganzheitliche Datenbasis sowie verschiedene Methoden und Werkzeuge für deren effiziente Nutzung im Rahmen des Trainings und der Validierung von KI-Funktionen im Fahrzeug bereitstellen.

Die KI Familie wird im Rahmen des Fachprogramms „Neue Fahrzeug- und Systemtechnologien“ (NFST) durch das Bundesministerium für Wirtschaft und Klimaschutz (BMWK) gefördert.

Die beteiligten Partner des Verbundprojekts KI Absicherung haben sich zu Projektstart das Ziel gesetzt Methoden und Maßnahmen für die Absicherung KI-basierter Funktionen bzw. Einzelmodule für das automatisierte Fahren zu entwickeln und deren Eignung hinsichtlich einer stringenten Sicherheitsargumentation zu untersuchen. Hierfür wurde der aktuelle Stand der Technik aufgegriffen und soweit vorangetrieben, dass erstmals ein gangbarer und im Expertenkreis anerkannter Weg hinsichtlich der prinzipiellen Herangehensweise zum Nachweis der Absicherbarkeit von KI-Modulen aufgezeigt werden konnte. Die hierzu erarbeitete konsensfähige Methodik wurde ebenso wie alle anderen Inhalte des Vorhabens an einem definierten abgrenzbaren Use Case erarbeitet

Hierfür wurde die Fußgängererkennung aufgrund der sicherheitstechnischen Relevanz von „schwächeren“ Verkehrsteilnehmern ausgewählt. Die Bestimmung bzw. Abschätzung der Fußgängerintention wurde hierbei bewusst ausgeklammert und sich auf die Objektdetektion, die semantische Zuordnung der Messpunkte und die Posenbestimmung von Fußgängern fokussiert.

Für die Ausplanung und Realisierung des Vorhabens wurde eine Projektstruktur mit einer Unterteilung in fünf Teilprojekte (TP) definiert. Die Hauptschwerpunkte und Zielsetzungen der einzelnen Teilprojekte lassen sich folgendermaßen zusammenfassen:

Teilprojekt 1 (TP1) mit dem Kurztitel „KI-Funktion“ lieferte den beispielhaften aber praxisrelevanten Untersuchungsgegenstand für die Erforschung einer Absicherungsmethodik für KI-Algorithmen. Hierzu wurden Algorithmen zur KI-basierten Fußgängererkennung entwickelt. Ziel war es den Stand der Technik repräsentativ abzudecken, wobei sich auf tiefe neuronale Netze, engl. Deep neural networks (DNNs) [55] [56] beschränkt wurde. Das erste Teilprojekt untergliederte sich hierbei in fünf Arbeitspakete:

- AP1.1 „Technische Plattform“ diente der Bereitstellung der technischen Plattform, der Projektinfrastruktur für alle weiteren APs und somit sämtliche für das Projekt notwendigen Tools und Arbeitsprozesse.
- AP1.2 „Anforderungen (Performance-KPIs)“ widmete sich der systematischen Erfassung von quantitativen und qualitativen Anforderungen an eine synthetisierte Datengenerierung an TP2. Zudem wurden in AP1.2 auch die funktionalen Anforderungen und Gütekriterien für die Entwicklung der KI-Funktionalität definiert.
- AP1.3 „Implementieren von Algorithmen zur Fußgängerkennung“ stellte die Basisalgorithmen zur Fußgängererkennung aus monoskopischen Bildsequenzen bereit. Diese Algorithmen basieren auf DNNs und wurden im Wesentlichen mithilfe der von TP2 bereitgestellten Daten generiert Damit stellen sie die funktionale Basis für TP 3 und 4 bereit.
- AP1.4 „Erweiterung der Fußgängererkennung um Tiefendaten“ widmete sich der Entwicklung von Algorithmen zur Prädiktion und Tracking von 3D Bounding Boxen von Fußgängern in Weltkoordinaten. Zu diesem Zwecke wurden die Bilddaten mit Tiefendaten aus der Simulation angereichert.
- AP1.5 „Erweiterung um Posenschätzung“ entwickelte Algorithmen zur Fußgänger Posenschätzung, die im Gegensatz zur vergleichsweise groben Lokalisierung von Fußgängern mittels Bounding Boxen auf die Bestimmung bzw. Schätzung der aktuellen Körperhaltung abheben.

Teilprojekt 2 (TP2) mit dem Kurztitel „Generieren von synthetischen Lern- und Testdaten“ stellte dem Konsortium einen synthetischen Datensatz zur Verfügung, der neben den einzelnen monoskopischen Bilddaten auch die korrespondierenden Metadaten enthält. Dies Metadaten waren hierbei von besonderer Bedeutung für die Bestimmung bzw. den Abgleich der Performance- und Qualitätsmetriken der KI Verfahren mit der sog. Groud Truth aus der Simulation. Das zweite Teilprojekt untergliederte sich hierbei in fünf Arbeitspakete:

- AP2.1 „Toolkette für synthetische Datenerzeugung“ erarbeitete eine technische Verarbeitungskette, die alle Verarbeitungsschritte umfasst, um von einer formalen Spezifikation einer Verkehrsszene zu den simulierten Sensordaten zu gelangen.
- AP2.2 „Corner Cases“ hatte zum Ziel Situationen bzw. Szenariene samt Kontext und dynamischen Objekten zu identifizieren, in der die KI-Funktionalität ein nicht erwartetes und funktional nicht hinlängliches Ergebnis liefert, obwohl ein korrektes Verhalten erwartbar gewesen wäre. Des Weiteren fungierte AP2.2 als zentraler „Gate-Keeper“ für die

Anforderungen an die Datengenerierung im gesamten Projekt. In dieser Rolle werden alle eingehenden Datenanforderungen konsolidiert und priorisiert an AP2.5 weitergeleitet.

- AP2.3 „Abstraktion von Sensorik“ beschäftigte sich mit der Frage der Übertragbarkeit von KI-Funktionen bezüglich ihrer Absicherbarkeit bei Änderung der Sensorik. Weiterhin wurde die Übertragbarkeit der KI-Funktionen bei Änderung der Sensorik durch Methoden des Transfer Learnings untersucht, sowie Konzepte zur Anpassung der Transfer Learning Methoden erarbeitet.
- AP2.4 „Bewertung synthetischer Daten“ erarbeitete eine Methode zur vergleichenden Bewertung von Datensätzen unterschiedlicher Qualität.
- AP2.5 „Datengenerierung und Noisy Data“ generierte synthetische Trainings- und Validierungsdaten unter Einbezug von bewussten Verunreinigungen bzw. Störungen von Sensordaten.

Teilprojekt 3 (TP3) mit dem Kurztitel „Methoden und Maßnahmen zur Absicherung von KI“ baute einen Werkzeugkasten aus bekannten und neuen Methoden und Maßnahmen zur Absicherung von KI auf. Hierzu wurden bestehende Ansätze angepasst, bewertet und mit dem Blick auf die Absicherbarkeit der zugrundeliegenden KI-Funktion erweitert.

- AP3.1 „Tracking State-of-Research“ übernahm die Aufgabe der Analyse und Überwachung der Forschungslandschaft sowie die Bewertung der entwickelten Mechanismen auf ihren Mehrwert hinsichtlich eines möglichen Beitrags zu einer stringenten Sicherheitsargumentation.
- AP3.2 „Wirksamkeits- und Sicherheits-KPIs für Machine Learning fokussierte auf die Identifikation und Belastbarkeit von Qualitätsmetriken.
- AP3.3 bis AP3.5 setzen sich explizit mit Methoden und Maßnahmen auseinander, die entweder kein (AP3.5 „Blackbox Methoden und Maßnahmen“), ein beschränktes oder aber über ein vollständiges (AP3.4 „White-/Greybox Methoden und Maßnahmen“) Wissen über die dahinterliegende KI Funktion haben, oder sogar funktionsverändernd wirken (AP3.3 „Funktional verändernde Methoden“).
- AP3.6 „Aggregierte Methoden und Maßnahmen“ setzte sich mit der Kombination von Methoden und Maßnahmen und deren Mehrwert hinsichtlich der Absicherbarkeit auseinander.

Teilprojekt 4 (TP4) mit dem Kurztitel „Gesamtheitliche KI-Absicherungsstrategie“ hatte zum Ziel die Definition und exemplarische Umsetzung eines systematischen Vorgehens zur Formulierung einer gesamtheitlichen Argumentation (Assurance Case) zur Absicherung einer KI-Funktion (Fußgängererkennung) vorzunehmen.

- AP4.1 „Strukturierung und Formalisierung des Eingaberaums“ wurde der Grundkontext für die Absicherung der Erkennung von Fußgängern im urbanen Kreuzungsbereich definiert und unter Verwendung einer geeigneten Beschreibungssprache beschrieben. Zudem war das Ziel eine Formalisierung und Strukturierung des gesamten Eingaberaums (Domänenanalyse) der KI-Funktion hinsichtlich funktionsrelevanter Kontextelemente (z.B. Verkehrsteilnehmer, Wetter, Objekte, Lichtverhältnisse) und Kontextdimensionen (Eigenschaft eines Kontextelementes oder eines Umwelteffektes) unter Nutzung der zu berücksichtigenden Variationsmöglichkeiten und Corner Cases aus AP2.2 zu erarbeiten und in eine Ontologie zu überführen. Außerdem

wurden physikalische Effekte und bekannte Zusammenhänge zwischen Einflussfaktoren formuliert („A-priori-Wissen“).

- AP4.2 „Gesamtstruktur der Argumentation“ hatte sich zum Ziel gesetzt, die Sicherheitsziele der eingesetzten Funktion (z.B. jeder relevante Fußgänger wird rechtzeitig erkannt, so dass ausgewichen oder gebremst werden kann), die geforderten Zielgrößen (zulässigen Wertebereiche), der KI-Gesamtfunktionskontext als auch die übergeordnete Systemarchitektur – soweit zum Nachweis der Absicherbarkeit benötigt – zu definieren.
- AP4.3 „Argumentation für eine abgesicherte KI-Funktion“ nahm die Arbeiten unter der Prämisse an, dass eine KI-Funktion aufgrund der inhärenten Komplexität und Datenabhängigkeit schwierig als alleinstehende Funktion ohne zusätzliche Maßnahmen als sicher argumentiert werden kann. Entsprechend war es das Ziel des Arbeitspaketes ein Vorgehen zur systematische Herleitung einer Sicherheitsargumentation auf Basis von Evidenzen basierend auf den Methoden und Maßnahmen aus TP3 und AP4.4 exemplarisch zu erarbeiten. Dies geschah ohne den Anspruch einen vollständigen Sicherheitsnachweis für eine konkrete Umsetzung einer KI basierte Fußgängererkennung erarbeiten zu können.
- AP4.4 „Testmethoden und Wirksamkeitsuntersuchung“ hat sich zum Ziel gesetzt Testmethoden zur Anwendung zu bringen, mit denen es möglich ist die in AP4.1 durchgeführte Domänenanalyse und die in AP4.2 definierten Sicherheitsziele zum einen eine Abdeckung des Eingaberaums und zum anderen die Wirksamkeit der Einzelmaßnahmen zu untersuchen
- AP4.5 „KI-Teststrategie und KI-Testplan“ widmete sich dem allgemeinen Vorgehen zum Testen einer KI-Funktion. Welche Teststrategie vor dem Hintergrund der Absicherbarkeit von KI-Funktionen am besten geeignet scheint, wurde hierbei unabhängig von der konkreten Funktionsausprägung betrachtet. Darauf aufbauend war das Ziel für die KI-Funktion Fußgängererkennung auf Basis des konkreten Assurance Case unter Verwendung der KI-Teststrategie auszuarbeiten, welcher KI-Testplan hierfür geeignet erscheint.

Teilprojekt 5 (TP5) mit dem Kurztitel „Projektmanagement und Dissemination“ hatte eine querschnittliche Rolle innerhalb des Projektes KI-Absicherung. Die Hauptaufgabe dieses TP bestand darin, sowohl organisatorische als auch technisch-inhaltliche Schnittstellen unter den einzelnen Partnern und Arbeitspaketen zu schaffen und zu pflegen.

- AP5.1 „Projektmanagement“ hatte die übergeordnete Aufgabe der technischen Gesamtkoordination mit der inhaltlichen und zeitlichen Organisation der technischen Arbeiten, sowie die Sicherstellung der Qualität der Ergebnisse. Hierzu wurden die Arbeiten der einzelnen Arbeitspakete inhaltlich abgestimmt und nachverfolgt sowie der Arbeitsfortschritt hinsichtlich der definierten Meilensteine bestimmt. Bei inhaltlichen Konfliktsituationen und Planabweichungen wurden Vorschläge für adäquate Gegenmaßnahmen erarbeitet.
- AP5.2 „Ergebnisverbreitung“ hatte die Aufgabe die Kommunikation des Vorhabens und dessen Ergebnissen nach außen voranzutreiben. Eine Hauptaufgabe bestand in der zielgruppengerechten Aufbereitung und Zur-Verfügung-Stellung von relevanten Projektinformationen durch geeignete Kommunikationsmittel. Zum anderen wurden Veranstaltungen (Halbzeit- und Abschlusspräsentation) konzipiert, vorbereitet und umgesetzt.

- AP5.3: "Kommunikation mit Normungsgremien" unterstützte das Vorhaben im Allgemeinen und TP4 im Besonderen bei der Kommunikation mit relevanten Normungsgremien und Zertifizierungsstellen. Hierfür galt es zunächst, relevante Gremien, Aktivitäten und Organisation zu identifizieren. Ausgangspunkt hierfür waren die zu Vorhabenbeginn bekannten, für das Vorhaben KI Absicherung relevanten Normen. In Anbetracht der immer greifbareren Markteinführung automatisierter Fahrfunktionen war davon auszugehen, dass sich über die Projektlaufzeit hinweg relevante Normierungs- und Standardisierungsaktivitäten – sowohl im Themenbereich der Absicherung als auch im Technologiefeld KI – ausweiten würden. Diese galt es zu beobachten. AP5.3 trug darüber hinaus die Verantwortung dafür, dass das „Zugehen“ und „Einbinden“ der relevanten Akteure projektweit abgestimmt stattfand.

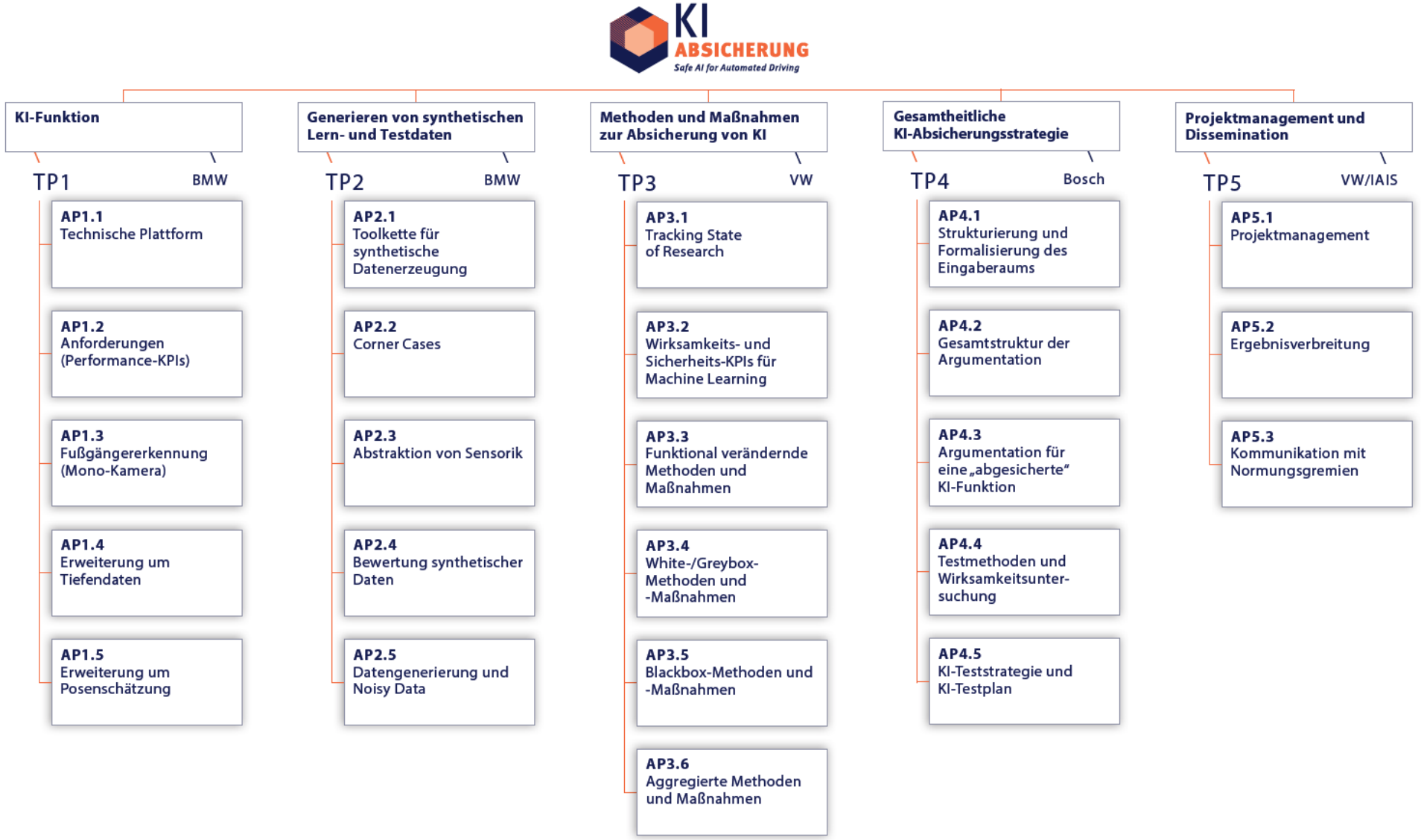


Abbildung 1.1: Projektstruktur

2 Gesamtansatz

Der in Abbildung 1 dargestellte Gesamtansatz zur Absicherung von KI-Funktionen im Fahrzeug, kann als das zentrale übergeordnete Ergebnis des Verbundprojekts "KI-Absicherung" aufgefasst werden, weil er die Grundmethodik darstellt, mit der es nach Überzeugung aller beteiligten Partner möglich ist, systematisch zu einer stringenten Sicherheitsargumentation für KI-basierte Funktionen zu gelangen.

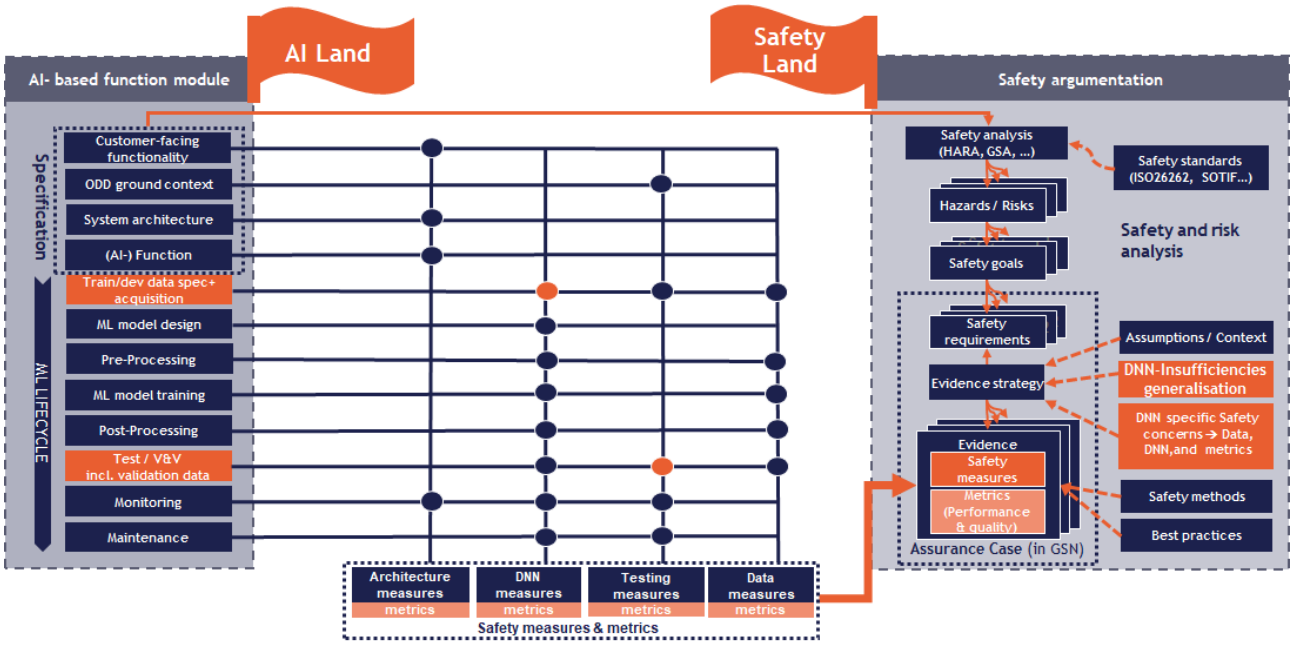


Abbildung 2.1: Gesamtansatz zur Absicherung von KI-Funktionen im Fahrzeug

Im linken Bereich der Darstellung sind die Spezifikation und die Entwicklungsschritte einer KI-basierten Funktion zusammengefasst, wohingegen auf der rechten Seite die Methodik zum Aufbau einer Evidenz-basierten Sicherheitsargumentation dargestellt ist. Diese stützt sich insbesondere auf die Sicherheitsmaßnahmen und Metriken, die bei der Entwicklung der KI-Funktion Anwendung finden, um eine hieraus eine stringente Sicherheitsargumentation ableiten zu können. Dabei sind die beiden Seiten nicht als separate, voneinander unabhängige Einheiten zu betrachten, sondern vielmehr in Prozessen miteinander verbunden, in denen die Zusammenarbeit zwischen KI-Entwicklern und Safety-Experten systematisch organisiert wird.

Die Spezifikation der KI-basierten Funktion ist naturgemäß der Ausgangspunkt, sowohl für den Aufbau der Sicherheitsargumentation, als auch für die Entwicklung der Funktion selbst. Neben der rein funktionalen Spezifikation der zu lösenden Aufgabe, wie im ausgewählten Anwendungsbeispiel der Fußgängererkennung, sind sowohl die Grundarchitektur und die Funktionalität zu spezifizieren, als auch der Einsatzbereich (Operational Design Domain, ODD), in welcher die Funktion angewendet werden soll.

Die Spezifikation der ODD ist jedoch gerade im sogenannten "Open World Context" eine besondere Herausforderung, da hier nicht alle Elemente und Eigenschaften der Umgebung, in der ein Fahrzeug zum Einsatz kommt, bereits zur Design-Zeit vollständig bekannt sind. Dennoch ist eine mögliche genaue Spezifikation der ODD in mehrfacher Hinsicht die Grundlage für eine abgesicherte KI-Funktion:

- zum einen bildet die ODD die Basis, um die für das Training der KI-Funktion benötigten Daten bereitzustellen,
- zum anderen definiert sie den Geltungsbereich der Sicherheitsargumentation und erzeugt hierbei insbesondere Anforderungen an die Test-Daten, um eine hinreichende Testabdeckung zu erzielen.

Dieser herausragenden Rolle der Spezifikation der ODD und der daraus abgeleiteten Spezifikation von Trainings- und Testdaten trägt das Projekt "KI-Absicherung" durch die detaillierte Spezifikation von Ontologien und Metadaten Rechnung. Die dazu entwickelten Beschreibungssprachen und Datenformate sind sowohl für Menschen verständlich, um eine nachvollziehbare Sicherheitsargumentation aufbauen zu können, als auch maschinenlesbar, um mit Tool-Unterstützung Daten-Analysen und Testauswertungen organisieren zu können. Weitere Einzelheiten und die Verbindung zu bestehenden Konzepten zur Beschreibung und Spezifikation von Daten, wie z.B. OpenDrive oder OpenScenario, werden in Abschnitt weiter untern näher erläutert.

Bei der Formulierung der Sicherheitsanforderungen ist zu berücksichtigen, dass eine KI-basierte Funktion potentiell fehlerhafte Ausgaben liefern kann. So ist es z.B. möglich, dass eine auf einem Plakat dargestellt Person von einem neuronalen Netz, das die Videobilder einer Fahrzeugkamera auswertet, fälschlicher Weise als eine "echte" Person klassifiziert wird. Hierbei wird die Möglichkeit, potentiell fehlerhafte Ausgaben zu liefen, von "KI-Entwicklern" üblicherweise als mangelnde Generalisierungsfähigkeit bezeichnet und stellt eine "Unzulänglichkeit" der Softwarefunktion im Sinne eines Safety-Experten dar.

Um dennoch in einer systematischen Art und Weise überprüfbare Sicherheitsanforderungen an eine KI-basierte Funktion formulieren zu können, hat das Projekt "KI-Absicherung" aus dem Expertenwissen der KI-Entwickler und beteiligten Safety-Experten zunächst sogenannte "DNN spezifische Sicherheitsbedenken" formuliert, welche diese Unzulänglichkeit systematisch einordnet und beschreibt. Diese spezifischen Sicherheitsbedenken unterscheiden sich vom Wesen her von den bekannten und bereits hinreichend untersuchten „klassischen Sicherheitsbedenken“, die in der Vergangenheit bereits bei den analytischen Verfahren der Umfeldwahrnehmung zu berücksichtigen waren. Diese sind weiterhin gültig. Hervorzuheben ist an dieser Stelle aber, dass sich das Projekt KI Absicherung rein auf die KI spezifischen Sicherheitsbedenken fokussiert, und die klassischen Aspekte der Sicherheitsargumentation ausklammert. Im Kontext des hochautomatisierten Fahrens sein hier auf die Schwesterprojekte der sog. Pegasus Familie verwiesen, mit denen das Förderprojekt KI Absicherung im engen Austausch stand.

3 TP1 KI Funktion

Wichtigste Ergebnisse und Ereignisse

Zentrale Aufgabe in Teilprojekt 1 war die Entwicklung und Implementierung von Basisalgorithmen zur Erkennung von Fußgängern. Dabei dienten Daten verschiedener Sensormodalitäten als Input der neuronalen Netze. Dieses Teilprojekt besteht aus auf fünf Clustern. In Cluster 1 wurde die Entwicklung einer technischen Plattform forciert. Diese liefert Tools, die die Arbeit der Entwickler unterstützt und Aufwände abnimmt, die ansonsten bei Vielen anfallen würden, und stellt sie an zentraler Stelle bereit. Diese Tools und Services wurden an zentraler Stelle bereitgestellt und mitunter durch das Feedback der Nutzer weiterentwickelt und kontinuierlich verbessert. Darunter fallen z.B. das Aufsetzen eines Ticketing-Systems für das gesamte Konsortium, Tooling und Management der Data Storage Plattform (DSP), eine Continuous Integration Pipeline für den Software Release Prozess (SRP), Tools für die Bereitstellung des KI Absicherungs-Datasets und eine Infrastruktur für das Hosting von Code Repositories. In Cluster 2 wurden Anforderungen an die KI-Funktion aus der Sicht des DNN-Entwicklers definiert und eine Abstimmung mit Schnittstellenpartnern im Projekt durchgeführt. Dafür wurde zum einen der zu verwendende synthetische Datensatz spezifiziert, inklusive Spezifikation der Sensormodalitäten Kamera und LiDAR, sowie den entsprechenden Annotationsformaten. Zum anderen wurden Anforderungen an die KI-Funktion definiert und geeignete Metriken für deren Evaluation festgelegt. Die Cluster 3-5 hatten die Aufgabe der eigentlichen Algorithmenentwicklung inne. Cluster 3 fokussierte sich dabei auf Ansätze zur Fußgängererkennung aus monoskopischen Videosequenzen (Einzelbildern einer einzigen Kamera). Cluster 4 hatte mit dem LiDAR ergänzend noch eine weitere Sensormodalität und damit Tiefendaten zur Verfügung. Im Cluster 5 wurden Algorithmen untersucht und entwickelt, die eine an ein Skelett erinnernde Pose der in den Daten auftretenden Personen schätzen.

Für die 2D-Bounding-Box Detektion auf Basis von Bilddaten wurde der Single-Shot Multibox Detector (SSD, https://link.springer.com/chapter/10.1007/978-3-319-46448-0_2) gewählt. Dieser wurde entsprechend den Anforderungen der Schnittstellenpartner adaptiert und auf den Projektdaten trainiert. Die Performanz wurde in einer ausführlichen Ablationsstudie untersucht und mit diversen Maßnahmen optimiert, darunter u.a. die Wahl von vor-trainierten Backbone Gewichten und eine Anpassung der Seitenverhältnisse und Anzahl der Anchor Boxen. Dieser Algorithmus war einer der prominentesten Algorithmen für Sicherheitsanalysen im Teilprojekt 4 und für die Entwicklung von Metriken und Evaluierungsmethoden im Teilprojekt 3.

Ein weiterer viel verwendeter Algorithmus war DeeplabV3+ bzw. DeeplabV3 (https://arxiv.org/pdf/1802.02611.pdf) für die semantische Segmentierung von Bilddaten - also die Zuordnung jedes einzelnen Pixels im Bild zu einer Klasse, wie Person, Auto, Gebäude, etc. Bei der Evaluation dieser Algorithmen wurde zunächst übereinstimmend mit der 2D-Bounding Box Detektion ein großer Domain Gap, d.h. visueller Unterschied zwischen Trainings- und Evaluationsdaten - durch die niedrigen Konfidenzwerte und hohe Falschdetektion der Prädiktoren festgestellt. Dieser Domain Gap besteht in den Trainingsdaten des synthetischen Datensatzes zwischen den Trainingsdaten (Tranchen 3 und 4) und den Evaluationsdaten (Tranche 5). Der Domain Gap ist auf die kontinuierliche Weiterentwicklung (bspw. durch neue Assets oder visueller Effekte) im synthetischen Datensatz zurück zu führen. Eine Änderung des Datensplits als Verbesserungsmaßnahme wurde vorgeschlagen und umgesetzt. Aus dieser resultierte die

Reduktion des Domain Gaps wodurch eine wesentliche Verbesserung der Model-Performanz erwirkt wurde.

Ansätze mit der zusätzlichen Sensormodalität LiDAR brachten die Herausforderung der Sensorfusion mit sich. Im Teilprojekt wurden Fusionsansätze auf Sensorebene, Merkmalsebene, Regressionsebene, "Late Fusion", sequentielle Fusion und "Temporal Fusion" untersucht. Für die Fusion auf Merkmalsebene wurde eine Kombination aus "AVOD- birds-eye-view" (LiDAR) in Verbindung mit der CenterNet-basierten Objekterkennung implementiert. Dafür wurden die Kameradaten im Merkmalsraum mit den Merkmalen aus dem AVOD-CenterNet mit dem "Orthographic Feature Transform-Netzwerk" (OFT) fusioniert. Der daraus resultierende Fusionsansatz schnitt bei dem Projektdatensatz gut ab und zeigte Ergebnisse, die mit den modernsten Objekterkennungsalgorithmen vergleichbar sind.

Für den Temporal Fusion Ansatz wurde als Basis der State-of-the-art-Algorithmus Faster-RCNN in Kombination mit Kalman- und Partikelfilter verwendet. Durch detaillierte Analyse der verschiedenen Architekturen konnten Komponenten der Filter so in die Faster-RCNN-Architektur integriert werden, dass Synergien genutzt wurden und eine fusionierte Architektur entstand. Für die Integration der Objektpermanenz (IOP) werden die Vorschläge von Partikelfilter und "Region-Proposal-Network" verkettet. Die zweite Stufe des Faster-RCNN und des Partikelfilters werden dann nacheinander ausgeführt.

Human Pose Estimation (dt. menschliche Posenschätzung) ist eine Aufgabe für Neuronale Netze bei der die Positionen der einzelnen Gelenke eines Menschen geschätzt werden. Die entsprechende Verbindung der erkannten benachbarten Gelenkspunkte ähnelt bzw. erinnert an das menschliche Skelett. Bisher hat diese Aufgabe noch wenig Einzug die Automotive Domäne gefunden, daher sind die Ansätze hier sehr forschungsorientiert. Neben überwachten Ansätzen (Daten enthalten Informationen des Zielzustands) wurden auch unüberwachte Ansätze untersucht. Diese nutzen geometrische Equivarianz und Invarianz unter Aussehensänderungen (z.B. Farbänderungen) aus. Veränderungen können bspw. durch Rotation und Farbveränderung erzwungen werden. Die Ansätze erzeugten sehr gute Ergebnisse auf öffentlichen Datensätzen, haben allerdings Probleme mit Verdeckungen, welche im Projektdatensatz einen großen Anteil darstellen.

Des Weiteren wurden spezielle Datenanalysen angestrebt, welche auch in Bezug zu aktuell existierenden öffentlichen Datensätzen gesetzt wurden. Die Untersuchungen ergaben, dass der im Projekt erzeugte synthetische Datensatz ein relativ hohes Aufkommen an unüblichen Posen wie z.B. liegende oder kriechende Personen hat. Dies ist im Sicherheits-Kontext von hoher Bedeutung, da Personen in solchen Posen ansonsten aufgrund des geringen Vorkommens in Realdatensätzen stark unterrepräsentiert sind, was einen negativen Effekt auf Erkennungsraten für Personen in solchen Posen hat.

3.1 AP1.1 Technische Plattform

Die entwickelte technische Plattform bestand aus Tools und Services, welche für TP1 Code Entwickler und deren Nutzer im Rahmen von den Arbeiten in AP1.1 entstanden sind. Diese Tools und Services wurden progressiv weiterentwickelt und mit Feedback der Nutzer allmählich verbessert.

Darunter zählen:

- Docker Images und Container für die Softwareentwicklung und Release im TP1
- Das Ticketing-System für das gesamte Konsortium
- Tooling und Management der Data Storage Plattform (DSP)
- Continuous Integration Pipeline für den Software Release Prozess (SRP)
- Gitlab / Bitbucket basierte Infrastruktur für das Hosting von Code Repositories
- Docker Image und Container für die automatische Berechnung von Metriken
- Docker Image und Container für die Auswertung semantischer Segmentierungs-Modelle
- Automatisierung von Tools für die Bereitstellung des KIA-Datasets (insb. Latest-and-Greatest, LaG)

Die Verbreitung von Software, sowohl intern als auch extern, erfordert die Einhaltung von Guidelines und Standards, damit die Software in verschiedenen Umgebungen bzw. Plattformen ausführbar ist. Darüber hinaus wwurde angestrebt, den User von der Komplexität der Installation und Software Abhängigkeiten mit dem darunterliegenden Betriebssystem zu abstrahieren. In AP1.1 wurde deshalb ein **Docker Image / Container für Softwareentwicklung bzw. für Software Release** (i.e. -verbreitung) erstellt, mit dem die Developer ihre eigene Software vor dem endgültigen Release testen konnten. Ein Docker Container ist im Grunde ein abgekapseltes Software-Paket, welches die zu verbreitende Software, alle ihre Dependencies (dazu notwendige Software Packages) und eine minimale Version eines Betriebssystems (Ubuntu Linux im Fall von KIA) beinhaltet. Darüber hinaus wurde der Docker Container während des Releases ebenfalls verwendet, um die in TP1 entwickelten Software automatisch zu testen. Die Bereitstellung des Docker Image bzw. Container und die passende Dokumentation sowie passende Tools für die Anpassung des Containers hat bereits in M2 der Projektlaufzeit stattgefunden.

Um die Softwareentwicklung, aber auch die Datenproduktion bzw. die Datenqualität systematisch zu betreiben, war es notwendig, ein geeignetes **Ticketing System** zu implementieren, welches anhand sog. Tickets einzelne Aufgaben bzw. Probleme und deren Lösung beschreibt. Das Ticketing-System wurde ebenfalls von AP1.1 für die Nutzung aller Konsortialpartner bereitgestellt. Die erste Implementierung geschah mit dem von der Firma Gitlab angebotenen Ticketing System, zumal hier andere Lösungen finanziell aufwändiger gewesen wären. Später im Projekt zeigten sich die Grenzen dieses Systems, denn die Anzahl von Tickets und die Abhängigkeiten zwischen diesen, insbesondere die Anforderungen an die synthetische Datenproduktion bzw. an die anschließende Qualitätsverfolgung, wuchs erheblich. Deshalb wurde im Rahmen von AP1.1 eine Migration des Ticketing-Systems auf das Jira System der Firma Atlassian durchgeführt, welches auf der Infrastruktur der Firma Luxoft lief.

Das **Hochladen und Herunterladen von Daten aus der Data Storage Plattform** (DSP, betrieben durch das Fraunhofer ITWM) wurde mithilfe von in AP1.1 entwickelten Tools durchgeführt, bis das KIA-Dataset Library angeboten wurde. Auch die logische Organisation der DSP, sowie das Management von Usern auf der Plattform wurden ebenfalls von AP1.1 koordiniert und organisiert.

Daneben war die Einhaltung von Qualitätsstandards bezüglich die Software-Verbreitung Aufgabe von AP1.1. Dafür wurde ein sog. **Continuous Integration (CI) Pipeline für das Software Release**

implementiert und während des gesamten Projektlaufzeit aufrechterhalten und verbessert. Die im Projekt entwickelte Software wurde einem Release Prozess unterzogen, der aus einer Reihe von Qualitäts-Tests auf einzelne Code Repositories und der darin enthaltenen Software besteht und in regelmäßigen Zeitabständen zu den sogenannten Release-Perioden durchgeführt wurde. Die CI Pipeline lief auf einem dedizierten Server mit ausreichenden Ressourcen: viel Storage, einem leistungsfähigen Prozessor und Hardware-Beschleunigung (GPUs) für die zu testenden AI-Funktionen. Eine Softwareinfrastruktur, zunächst das Open-Source Tool Jenkins und später eine reine Python-Entwicklung, lud die zu testende Software nach dem Erhalt eines vom Software-Entwickler getriggerten Event (Tagging auf dem Release Branch im Software Repository) automatisch und führte alle Tests der Reihe nach durch. Anschließend wurden die Test-Ergebnisse auf Confluence publiziert. Während des gesamten Prozesses wurde der Entwickler über den Zustand der Tests informiert, sodass er/sie passende Maßnahmen ergreifen konnte, falls ein Test fehlschlug. Es wurden fünf Tests im CI Pipeline implementiert: 1) Vollständigkeits-Test, um zu überprüfen, ob die Repository die abgemachte Verzeichnisstruktur hat bzw. alle angeforderte Dateien vorhanden sind; 2) Licensing-Test, um zu überprüfen, ob die Software und alle ihre Dependencies die im Projekt erlaubten Open-Source Software-Lizenzen erfüllen; 3) Software-Lauffähigkeitstest im Training-Fall, um zu überprüfen, ob die darin verwendeten KI-Modellen sich trainieren (anlernen) lassen; 4) Software-Lauffähigkeitstest im Inferenz-Fall, um zu überprüfen, ob die Software Klassifikation von Sensordaten durchführen kann; 5) Ausführung und Berechnung von Metriken auf vorberechneten Inferenzdaten. Neben TP1 konnte auch TP3 einen Teil der Funktionalität nach einer geeigneten Anpassung nutzen. Es wurden vier offizielle TP1 Software Releases (und zwei weitere nachgezogene Releases aufgrund der späteren Datenverfügbarkeit) während der Projektlaufzeit durchgeführt und für das gesamte Konsortium publiziert. AP1.1 bat alle beteiligten Partner um Unterstützung und führte spezielle Software-Anpassungen für die Beseitigung einzelner Schwierigkeiten in sämtlichen Releases durch.

Auch das **Hosting von Code Repositories und die Bereitstellung einer für die Projektzwecke passenden Repository-Struktur** fiel unter die Aufgaben von AP1.1. Zunächst auf der Gitlab-Plattform, später auf Atlassian Bitbucket migriert, wurde das Repository Hosting und dessen Struktur von AP1.1 koordiniert und gemanagt. Eine Migration des Plattformanbieters wurde aufgrund rechtlicher Fragen über die Speicherung der Daten sowie den Umgang mit Nutzerdaten (E-Mails, Namen, etc.) von verschiedenen Partnern angefordert. AP1.1 übernahm die Aufgabe der Migration von circa 150 Repositories und führte sie in kürzester Zeit durch. Die Daten lagen auf der Luxoft Infrastruktur auf Servern in Deutschland.

Die **Berechnung von Metriken aus den in TP1 entstandenen KI-Modellen** wurde explizit in der Vorhabensbeschreibung ausgelobt. Die Implementierung dazu kam in der zweiten Hälfte des Projekts, als die entwickelten Modelle eine gewisse Reife erlangt hatten. Darüber hinaus ersetzte die Berechnung von Metriken beide Lauffähigkeit-Tests in der CI-Pipeline, denn diese erforderten extrem viel Vorbereitungsarbeiten auf Seiten der Entwickler. Die Berechnung von Metriken brachte außerdem Mehrwert für andere Partner im Projekt, die Software aus dem TP1 Releases verwendeten, denn sie konnten Aussagen über die Performance-Verbesserung machen, bevor sie die Software einsetzen und deshalb wurde damit eine Beschleunigung aller im Projekt damit verbundenen Prozesse (insb. in TP3 und TP4) erreicht. Die Berechnung von Metriken, insb. für semantische Segmentierung, wurde mittels passender Docker Container implementiert und als Teil der CI Pipeline ausgeführt. Ebenfalls im Rahmen dieser Aufgabe wurde die

Containerisation einiger KI-Modelle bzw. TP1 Repositories, wie z.B. TP1 DeepLabV3, DeeplabV3+ und Detectron2 für das Training, sowie die Berechnung von Prädiktionen (Auswertung / Inferenz) auf Basis von KIA Daten implementiert.

Die **Bereitstellung des KI-Datasets** und zwar der korrigierten Version der Daten (sog. Latest-and-Greatest, LaG) mittels Fixes und weiteren Scripts aus dem KIA-Dataset-Library, entstand als Teil des Data Release Prozesses für TP2. Jedoch wurden die automatisierte Ausführung, das Herunterladen und Abarbeiten der Rohdaten aus der Datenproduktion, sowie die Anpassung der KIA-Dataset-Library für die Automatisierung der Aufgaben im Rahmen von AP1.1 gemacht. Sämtliche Rohdaten werden, sobald sie aus der Datenproduktion verfügbar sind, aus der Data Storage Plattform heruntergezogen, abgearbeitet (gefixt) und an die TPX Plattform für ihre weitere Nutzung in Projekten der KI Familie zur Verfügung gestellt.

3.2 AP1.2 Anforderungen an die KI-Funktion

Die Aufgabe von AP 1.2 war es die Anforderungen an die KI-Funktion aus der Sicht des DNN-Entwicklers zu definieren und innerhalb des Projekts abzustimmen. Im ersten Teil der Ergebnisse wird der synthetische Datensatz spezifiziert. Dies beinhaltet die Spezifikation der Sensormodalitäten Kamera und LiDAR, sowie des Annotationsformats. Weiterhin wird die Beschreibung von nominalen und extremen Basisszenarien aus Sicht des DNN-Entwicklers erarbeitet. Im zweiten Teil der Ergebnisse werden die funktionalen Anforderungen an die KI-Funktion definiert und geeignete Metriken für deren Evaluation festgelegt und abgestimmt. Dafür werden beispielhaft die kamerabasierte Fußgängerdetektion anhand von 2D und 3D Bounding Boxen, sowie die semantische Segmentierung verwendet. Ein wesentliches Ergebnis dabei ist unter anderem die Untersuchung der Performanz limitierenden Faktoren und die resultierenden Konsequenzen an die Entwicklung. Im Folgenden geben wir einen Überblick über die Hauptergebnisse von E1.2.1-E1.2.8.

In E1.2.1 wurden die Anforderungen an das Kameramodell abgestimmt. Als Ergebnis wurde ein idealisiertes Lochkameramodell mit einem horizontalen Öffnungswinkel von 60° und einer Auflösung von 1920x1280 Pixeln festgelegt, wobei die Kamera automobiltypisch hinter der Windschutzscheibe verbaut ist. Beim Rendering sollen optional typischen Fehlereffekte, wie unregelmäßige Verzeichnung, Vignettierung, Bewegungsunschärfe, oder Bildrauschen berücksichtigt werden.

Das Ergebnis für das Modell aus E1.2.2 ist ein idealisierter rotierender LiDAR, vergleichbar mit dem aus KITTI Datensatz bekannten Velodyne HDL-64E, der bei BIT TS bereits verfügbar war und in Tranchen 3-5 verwendet wurde. Ein realistischeres Modell eines gepulsten LiDAR wurde von Valeo zur Verfügung gestellt und im Projekt von BIT TS integriert. Im diesem Valeo Mobility Kit LiDAR wurden u.a. mehrere spezifizierte Abtastmuster und Öffnungswinkel realisiert, sowie eine Modellierung verschiedener Effekte wie mehrerer Echos und Rolling-Shutter. Das realistischere Modell war erst spät in der Projektzeit verfügbar und wurde in Tranche 7 verwendet.

In E1.2.3 war die große Herausforderung ein einheitliches Annotationsformat ähnlich wie in bekannten Datensätzen wie KITTI, nuScenes oder Cityscapes zu spezifizieren und gleichzeitig auf die Fähigkeiten und Ressourcen der Datenproduzenten BIT TS und Mackevision einzugehen. In dem so erarbeiteten Annotationsformat werden u.a. die Ordnerstruktur, Dateinamen und -formate der verschiedenen Sensoren und Annotationen wie 2D und 3D Bounding Boxen, semantische und instanziierte Segmentierung oder Fußgängerposen spezifiziert. Weiterhin sind

die Koordinatensysteme und entsprechende geometrische Transformationen spezifiziert. Dabei ist hervorzuheben, dass bei der synthetischen Datenproduktion die Annotationen vollständig automatisiert gewonnen werden und viele wertvolle Meta-Information zusätzlich verfügbar sind.

Eine Spezifikation des Datensatzes in Bezug auf die abgebildeten Szenen und Verkehrssituationen wurde in E1.2.4 und E1.2.5 erarbeitet. Die entsprechenden nominalen und extremen Basisszenarien wurden insb. aus Sicht der Algorithmen Entwickler spezifiziert und mit anderen Partnern aus AP 4.1 und P1 abgestimmt. Dabei wurde die in AP4.1 erarbeitete Ontologie für die Beschreibung der Operation Design Domain anhand von Zwicky Boxen verwendet. Diese beinhaltet u.a. eine Beschreibung des statischen Grundkontexts inklusive Straßen Infrastruktur, statische Objekte, Straßenoberfläche, Gebäude, dynamische Fußgänger- und Fahrzeugobjekte, Verkehrssituationen, Wetter und Beleuchtungsbedingungen. Die Variation der im Zentrum stehenden Fußgängerobjekte wurde dabei besonders ausführlich beschrieben. Für die Spezifikation der Verkehrsszenarien sind die NCAP-artigen Szenarien und die Design-of-Experiment Ansätze als Besonderheit zu nennen, die ein systematisches Vorgehen bei der Evaluation von DNN Algorithmen für sicherheitskritische Anwendungen ermöglichen. Aus Sicht der Algorithmen Entwicklung ist weiterhin eine Balance zwischen Einzelframes und Sequenzen notwendig, damit sowohl eine hohe Zahl von ausreichend unkorrelierten Einzelframes zur Verfügung steht als auch Sequenzen für die Entwicklung von Algorithmen für die Nachverfolgung. Als Beispiel für extreme Basisszenarien wurde die Variationen der Fußgängerpose verwendet. Weitere extreme Bedingungen können sich durch Variation der Kontrast- und Lichtbedingungen mit unterschiedlichen Sonnenständen abgeleitet werden.

Mit den zuvor definierten Bestandteilen ist es nun möglich die KI-Funktion in E1.2.6 funktional zu beschreiben. Dies beinhaltet die Spezifizierung des Eingangsformats, der Vorverarbeitung, der Hauptfunktion mit DNN Architekturen, der Nachverarbeitung und des Ausgangsformats. Im Falle der kamerabasierten Fußgängerdetektion anhand von 2D Bounding Boxen ist eine typische Vorverarbeitung die Skalierung und angepassten Normierung der Kamerabilder und eine typische Nachverarbeitung die Non-Maximum-Suppression und Skalierung der vom DNN bestimmten 2D Bounding Boxen. Zusätzlich wurden in E1.2.6 Performanz limitierende Faktoren identifiziert, welche die Performanz eines DNNs deutlich beeinflussen. Darunter sind die in E1.2.5 genannten extremen Bedingungen der Fußgängerpose oder extreme Kontrast- und Lichtbedingungen.

Für eine vergleichbare Auswertung der KI-Funktionen ist es außerdem notwendig einen gemeinsamen Satz an Metriken für die Performanz festzulegen, d.h. für den Vergleich zwischen dem Inferenzergebnis und der Ground-Truth. In E1.2.7 wurde dafür eine ausführliche Liste an Metriken für unterschiedliche DNN Anwendungen definiert. In Zusammenarbeit mit P1 wurden dabei auch sicherheitsrelevante Metriken berücksichtigt.

Abschließend wurde in E1.2.8 sichergestellt, dass die Ontologie ausreichend ist um die Performanz limitierenden Faktoren sowie die extremen Basisszenarien zu beschreiben. Die Ergebnisse wurden hier in einer Übersichtstabelle festgehalten.

3.3 AP1.3 Implementierung von Algorithmen zur Fußgängererkennung

Der Fokus in AP1.3 lag in der Implementierung von Algorithmen zur Fußgängererkennung aus monoskopischen Videosequenzen, d.h., mit Einzelbildern einer einzigen Kamera (vgl. Stereoskopie). Die Algorithmen können in fünf verschiedene Ansätze unterteilt werden: 2D-Bounding Box Detektion, Semantische-Segmentierung, Instanz-Segmentierung, 3D-Bounding Box

Detektion und 2D-Bounding Box Detektion unter Zuhilfenahme eines "Human Pose Estimators" (dt. Schätzer für menschliche Posen).

Mit Ausnahme der Posenschätzung sind die verbleibenden Detektionsansätze im automotive Kontext etabliert und im KI-Absicherung Projekt wurden diese, mit besonderem Fokus auf die 2D-Bounding Box Detektion und der Semantische-Segmentierung, als Grundlage der Arbeiten zur Absicherungsstrategie in TP3 und TP4 verwendet. Ein Großteil der Arbeiten an den Algorithmen in diesem Arbeitspaket entfiel deshalb auf die Implementierung, Evaluation und Anpassung einer Auswahl an bereits verfügbaren State-of-the-art (E1.31) Algorithmen an die Anforderungen der Methodenentwickler und weniger der Erforschung neuer Ansätze zur Personenerkennung. Durch die zentrale Rolle im Absicherungskontext des Projekts war ein reger Austausch mit Entwicklern in TP2 notwendig, um sicherzustellen, dass die synthetischen Daten qualitativ und quantitativ ausreichend sind um eine erwartbare Funktion der Algorithmen zu garantieren.

Zur 2D-Bounding Box Detektion, d.h., zur Lokalisation und Klassifizierung von Objekten in Bildern durch das Zeichnen eines Rechteckes pro gefundenem Objekt im Bild, wurde der Algorithmus Single-Shot MultiBox Detector (SSD, https://link.springer.com/chapter/10.1007/978-3-319-46448-0_2) ausgewählt und auf die synthetischen Projektdaten trainiert (E1.3.3a). Die Grundlage dieses Algorithmus basiert auf einem Objektdetektor für den Datensatz PasvalVOC07. Durch intensive Evaluation (E1.3.4) konnte festgestellt werden, dass die Standardkonfiguration des Detektors nicht geeignet war um für die Personenerkennung auf den sehr großen Bildern (1920x1280) eine ausreichende Erkennung zu generieren. Mit den Evaluationsergebnissen konnte dann eine signifikante Verbesserung erzielt werden, vornehmlich durch die Anpassung der Seitenverhältnisse der Anchor Boxes auf die Seitenverhältnisse der Personen in den Daten und der Verwendung eines auf ImageNet vor-trainierten ResNet50 Backbones.

Sowohl die Semantische-Segmentierung als auch die Instanz-Segmentierung versucht jeden einzelnen Pixel im Eingangsbild einer Semantischen Klasse zuzuordnen. Semantische Klassen im automotive Kontext sind beispielsweise Person, Auto, Straße etc. Die Instanz-Segmentierung geht einen Schritt weiter und ordnet nicht nur eine semantische Klasse zu (Klassifikation), sondern unterscheidet (Lokalisation) einzelne Objekte ähnlich der 2D-Bounding Box Detektion. Beide Ansätze wurden mit den Algorithmen DeeplabV3+ bzw. DeeplabV3 (beide Semantische-Segmentierung, E1.3.3b, https://arxiv.org/pdf/1802.02611.pdf) und Detectron2 (Instanz-Segmentierung, E1.3.3d, https://github.com/facebookresearch/detectron2) im Projekt umgesetzt. Auch diese Algorithmen wurden einer erweiterten Evaluation unterzogen (E1.3.4) um etwaige Schwachstellen in der Implementierung oder aber auch den Daten ausfindig zu machen. Übereinstimmend mit der 2D-Bounding Box Detektion Evaluation wurde ein großer Domain Gap, d.h. visueller Unterschied zwischen Trainings- und Evaluationsdaten, durch die niedrigen Konfidenzwerte und hohen Falschdeketionen der Prädiktoren festgestellt. Dieser Domain Gap bezieht sich auf die Trainingsdaten KI-A Tranche 3 und Tranche 4 und den Evaluationsdaten KI-A Tranche 5. Eine Änderung des Datensplits als Abschwächungsmaßnahme wurde vorgeschlagen und umgesetzt mit dem Resultat der Reduktion des Domain Gap.

Ein weiterer Ansatz zur Personenerkennung ist die 3D-Bounding Box Detektion (E1.3.3c). Hier wird versucht die Person im Bild nicht nur im 2D Bildraum zu lokalisieren, sondern im 3D Raum, d.h. im Kamera-Koordinatensystem. Die Schwierigkeit dieses Ansatzes ist vor allem die genaue Distanz zur Kamera aus einer Sequenz an Einzelbildern zu ermitteln ohne die Zuhilfenahme von

weiteren Sensordaten (vgl. AP1.4). Der Algorithmus wurde empfohlen als Verfeinerungsmethode eines Tiefendaten-gestützten 3D-Bounding Box Detektors, da die Erkennungsleistung ohne weitere Sensordaten stark hinter diesen zurückliegt.

Der letzte Ansatz zur Fußgängerdetektion ist ein 2D-Bounding Box Detektor der die geschätzte Pose einer Person (siehe AP1.5) verwendet, um eine verbesserte Prädiktion zu erzielen (E1.3.3e). Als Grundidee wird ein generatives Netzwerk (GAN) verwendet welches sowohl die Pose als auch das Erscheinungsbild enkodiert und danach wieder rekonstruiert. Mithilfe dieses Netzwerks kann nur durch eine Pose und eines Zufallvektors ein neues Bild einer Person mit genau dieser Pose erzeugt werden. Appliziert man diesen Ansatz auf Personen dessen Pose bekannt ist, aber stark okkludiert sind, z.B. verdeckt durch ein davorstehendes Auto, kann ein neues Bild erzeugt werden in welchen alle Personen un-okkludiert vor den jeweiligen Objekten zu sehen sind und welches- so die Hypothese - zu einer Verbesserung der darauffolgenden 2D-Bounding Box Prädiktion führt.

In dem AP1.3 wurden fünf verschiedene Ansätze zur Personendetektion aus monoskopischen Videosequenzen ausgewählt, (weiter-)entwickelt und sorgfältig evaluiert. Als Grundlage für die Zentrale Absicherungsstrategie im Projekt (Proof of Project Concept mit dem SSD) wurde ein besonderes Augenmerk auf die Evaluationsergebnisse dieser Algorithmen gelegt und hinsichtlich ihrer Prädiktionsperformanz auf den Projektdaten verbessert. Als Erst- und Hauptdatennutzer konnten viele wichtige Informationen zur Datenanforderung und Datenqualität an die Datenproduzenten zurückgespielt werden.

3.4 AP1.4 Erweiterung um Tiefendaten

Der Fokus von AP1.4 lag auf der Erweiterung der Fußgängererkennung mit LiDAR-Daten zur Erkennung von Fußgängern im 3D-space. Das Hauptaugenmerk in diesem Arbeitspackert lag auf der Verwendung von Daten aus zwei verschiedenen Sensormodalitäten - Kamera und LiDAR - und der Demonstration der Fusionsansätze auf verschiedenen Ebenen, um eine robuste Fußgängererkennung zu erreichen.

Für diese Aufgabe wurden insgesamt fünf Fusionsansätze und ein Ansatz mit einer Modalität ausgewählt. Zu den Fusionsansätzen gehörten die Fusion auf Sensorebene, Merkmalsebene, Regressionsebene oder "Late Fusion", sequentielle Fusion und "Temporal Fusion". Der Einzelmodalitätsalgorithmus verwendete nur einen LiDAR-Sensor für die Fußgängererkennung. Ein großer Teil der Aufgabe im Arbeitspaket bestand darin, die modernsten Algorithmen zu recherchieren, die ausgewählten Algorithmen zu implementieren und anhand der Projektdaten zu bewerten

Die Fusion auf Sensorebene (E1.4.1) zielte darauf ab, die Rohdaten von Kamera- und LiDAR-Sensoren zu fusionieren und die fusionierten Daten dem neuronalen Netz zur 3D-Fußgängererkennung zur Verfügung zu stellen. Als State-of-the-Art-Ansatz, der sich ausschließlich auf Lidar-Sensoren stützt, wurde PointPillars gewählt und durch das Anhängen von RGB-Informationen aus dem Kamerabild an die Punktwolke mit "Reverse Projection" erweitert. Die Auswertung des Projektdatensatzes zeigte, dass der Algorithmus nicht nur in nominalen Situationen, sondern auch in komplexen Situationen mit starker Verdeckung gut funktioniert.

Das nächste Ergebnis konzentrierte sich auf die Fusion auf Merkmalsebene (E1.4.2). Da die Objekterkennung nach dem Stand der Technik immer noch bildzentriert ist, wurde beschlossen,

für diesen frühen Fusionsansatz im bildähnlichen Merkmalsraum zu bleiben. Es wurde die Kombination aus "AVOD- birds-eye-view" (LiDAR) in Verbindung mit der CenterNet-basierten Objekterkennung gewählt. Die Kameradaten im Merkmalsraum wurden mit den Merkmalen aus der AVOD-CenterNet-Architektur mit dem "Orthographic Feature Transform-Netzwerk" (OFT) fusioniert. Der daraus resultierende Fusionsansatz schnitt bei dem Projektdatensatz gut ab und zeigte Ergebnisse, die mit den modernsten Objekterkennungsalgorithmen vergleichbar sind.

Die Fusion auf Regressionsebene (E1.4.3) - auch späte Fusion genannt - wurde anhand des "Frustum-ConvNet-Algorithmus" demonstriert. Der Algorithmus erzeugt Frustums für jede bereitgestellte 2D-Region, der der 2D Objektdetektor vorschlägt. Punktweise Merkmale werden hierbei mit Merkmalsvektoren auf Frustum-Ebene kombiniert. Diese konkatenierten Merkmalsvektoren bilden Merkmalskarten, die an ein vollständig gefaltetes neuronales Netz zum Ende-zu-Ende-Training in Kombination mit räumlichen Merkmalen von RGB-Daten weitergeleitet werden. Der Algorithmus wurde sowohl mit Projektdaten als auch mit öffentlichen Datensätzen evaluiert, und die Ergebnisse der Projektdaten zeigten gleichwertige Ergebnisse im Vergleich zu den trainierten Modellen der öffentlichen Datensätze.

Während die bisher erwähnten Ansätze eine Fusion von Kamera- und LiDAR-Daten verwendeten, stützt sich E1.4.4 nur auf eine Sensormodalität - LiDAR zur 3D-Fußgängererkennung. Hier wurden verschiedene zweistufige und einstufige LIDAR-Objektdetektoren eingesetzt und ein Punkt-Voxel-basierter Feature-Set-Abstraktionsansatz namens PVRCNN in das Projekt integriert. Die Methode verarbeitet zunächst die Punktwolke in Voxel und Sparse Convolutions, um Regionen von Interesse zu erzeugen. Dann werden Schlüsselpunkte und Voxel-Set-Abstraktionen verwendet, um die Merkmale zusammenzufassen und die Vorschläge zu verfeinern. Um die Leistung zu verbessern, wurde die LiDAR-Merkmalsextraktion mit Hilfe eines Graph Convolutional Neural Network und Transformatoren implementiert, um die Aufmerksamkeit auf die wichtigsten Merkmale zu lenken. Auswertungen von Projektdaten und öffentlich zugänglichen Metriken zeigten, dass zweistufige Ansätze genauer sind als einstufige Ansätze. Die Nutzung von Kontextinformationen unter Verwendung von Szenengraphen verbesserte die Erkennungsmetriken.

Der sequenzielle Fusionsansatz (E1.4.5) konzentrierte sich auf die sequentielle Verarbeitung der Daten von monokularen Kamerabildern und LiDAR-Punktwolken, um die 3D Bounding Box des Objekts zu erhalten. Dabei handelt es sich um einen zweistufigen Prozess, der in der ersten Stufe Kamerabilder und in der zweiten Stufe Punktwolken verwendet. Für diese Aufgabe wurde ein auf Frustum PointNet basierender Algorithmus integriert. Ein kamerabasierter 2D-Objektdetektor erkennt 2D-Regionsvorschläge für das Objekt. Diese Informationen werden an die zweite Stufe weitergegeben, um die entsprechenden Frustums aus der Punktwolke für jedes Objekt zu extrahieren. Die Punkte in den Frustums werden weiterverarbeitet, um zu den Parametern der 3D Bounding Box zu gelangen.

Der letzte Fusionsansatz (E1.4.6) zielte auf den zeitlichen Raum für die Fusion von Kamera- und LiDAR-Daten ab. Es wurde eine gründliche Recherche zum State-of-the-art durchgeführt, um die wichtigsten Einschränkungen der derzeitigen Algorithmen zu verstehen. Die gängigen Ansätze für die zeitliche Fusion basieren auf dem Tracking, und es wurde ein Ansatz entwickelt, um zeitliche Informationen in ein bestehendes zweistufiges Netz zu integrieren. Als Baseline wurde ein Faster-RCNN mit einem Kalman-Filter und einem Partikelfilter verwendet. Der Partikelfilter

wurde in das Faster-RCNN integriert, so dass die Vorschlags- und Verfeinerungsphasen zwischen den beiden Netzen geteilt wurden. Für die Integration der Objektpermanenz (IOP) werden die Vorschläge von Partikelfilter und "Region-Proposal-Network" verkettet und die zweite Stufe des Faster-RCNN und des Partikelfilters werden nacheinander ausgeführt.

Zusammenfassend wurden in diesem Arbeitspaket mehrere Ansätze zur Fusion von Kamera- und LiDAR-Daten auf verschiedenen Ebenen sowie ein Single-Modalitäts-Ansatz zur 3D-Fußgängererkennung erforscht, entwickelt, angewendet und evaluiert.

3.5 AP1.5: Human Pose Estimation

Human Pose Estimation (dt. menschliche Posenschätzung) ist eine Aufgabe für Neuronale Netze bei der die Position der einzelnen Gelenke eines Menschen geschätzt wird. Dazu wird ein Skelett definiert, welches vom biologischen Skelett inspiriert ist, aber nicht mit diesem übereinstimmt. So wird beispielsweise die Position von Handgelenk, Ellenbogen und Schulter, aber auch die Position der Augen und des höchsten Punkts des Kopfes geschätzt.

Die Posenschätzung ist bisher im automotive Kontext noch unüblich und wenig erforscht. Das AP 1.5 ist daher sehr forschungsorientiert und nicht im Fokus der Absicherung. Es wurde mit Ausblick auf eine zukünftige Absicherung an der Übertragung der Posenschätzung auf den automotive Kontext geforscht. Besondere Schwierigkeiten stellen die großen Abstände, massive Verdeckungen und die große Anzahl an Personen dar. Auf der anderen Seite ist die Varianz der zu erwartenden Posen deutlich geringer, da im Straßenverkehr bspw. keine Gymnastikposen zu erwarten sind.

Die Erforschung und Adaption der Posenschätzer für die automotive Domäne hatte dabei besonders drei Ansätze im Fokus: Überwachte Posenschätzung auf Kameradaten, Unüberwachte Posenschätzung auf Kameradaten und die Fusion von LiDAR und Kameradaten für die überwachte Posenschätzung.

Posenschätzer werden in zwei Kategorien unterteilt: Top-Down und Bottom-Up. Erstere detektieren zunächst jede Person im Bild und schätzen in einem zweiten Schritt die Pose jeder Person einzeln. Letztere lokalisieren zunächst alle Gelenkpositionen und verbinden diese Anschließend zu den Skeletten der Personen.

Bei der Erforschung der überwachten Posenschätzung (E1.5.2) haben sich die großen Distanzen und die hohe Auflösung der Bilder als eine Herausforderung für Bottom-Up Ansätze herausgestellt, während die gegenseitigen Verdeckungen für Top-Down Ansätze die Detektionsrate reduzieren. Daher wurde ein Hybrides Top-Down-Bottom-Up Verfahren entwickelt, welches zunächst Gruppen von Personen ausmacht und dann auf diesen Ausschnitten des Bilds in jeweils angemessener Auflösung Bottom-Up Ansätze anwendet.

Gegenseitige Verdeckungen stellen auch für die unüberwachte Posenschätzung (E1.5.4) eine Herausforderung dar. Die unüberwachte Schätzung nutzt geometrische Equivarianz und Invarianz unter Aussehensänderungen (z.B. Farbänderungen) aus. Bei dem in KI-Absicherung verfolgten Ansatz wird das Bild einer Person geometrisch augmentiert (z.B. rotiert) und einer Farbvariation unterzogen. Anschließend wird das Bild durch zwei Netze gegeben. Es wird eine Heatmap erstellt, welche bis auf die obige Rotation identisch sein muss (Ausnutzung der geometrischen Equivarianz). Anschließend wird mithilfe der Heatmap des Farbveränderten Bilds und der

Feature Repräsentation des rotierten Bilds das Originalbild wieder rekonstruiert. Es wird so zu sagen die geometrische Information des Farbveränderten Bilds und die Farbinformation des geometrisch veränderten Bilds kombiniert. Auf üblichen Datensätzen in der Posenschätzung funktioniert das Verfahren sehr gut, auf den KI-Absicherungsdaten stellt die große Menge an Verdeckungen eine Schwierigkeit dar.

Während die beiden obigen Ansätze nur Kameradaten benutzt haben, beschäftigt sich E1.5.3 mit der Fusion von Kamera und LiDAR. Ein Problem für lediglich Kamerabasierte Algorithmen ist die Tiefenwahrnehmung. Zwei Objekte, mit unterschiedlichen Größen können im Bild exakt gleich erscheinen, obwohl sie eine unterschiedliche Distanz zur Kamera haben. Verwendet man einen Laserscanner (LiDAR) so kann man die Distanz direkt messen und umgeht das Problem.

Der Fusionsansatz kombiniert LiDAR und Kamera in einem Featureencoder der aus der Objektdetektion (ähnlich zu AP1.4) stammt. In einem Top-Down Ansatz wird ein Vorschlag, wo das Objekt sich befinden könnte (3D Position), und ein Featurevektor der das Objekt beschreibt mit diesem Featureencoder erzeugt. Anschließend wird ein Posenschätzer angewendet, der basierend auf dem Featurevektor und der initialen 3D Position die Positionen der Gelenke schätzt und die 3D Position verfeinert. Da es im Projekt lediglich Posendaten mit Kameradaten gab, wurde der Ansatz nur auf öffentlichen Daten evaluiert; dort konnten überzeugende Ergebnisse erzielt werden. Eine Sensorfusion für die Posenschätzung macht im automotive Bereich durchaus Sinn.

Neben den Ansätzen zur Posenschätzung hat sich das AP1.5 auch mit den Anwendungen und dem Nutzen von Posen für die Sicherheit beschäftigt. So wurde in E1.5.6 evaluiert, ob Posen für die Verhaltensprädiktion von Fußgängern geeignet sind. Als Ergebnis der Untersuchung wurde festgestellt, dass es einige wichtige Features in der Pose gibt, die eine starke Korrelation zur Intention des Fußgängers haben.

Darüber hinaus wurde auch in einer Analyse des KI-Absicherungsdatensatzes analysiert, wie die Verteilung der Posen im Datensatz ist und ob es Defizite gibt. Die Beobachtungen wurden mit den anderen TPs geteilt und konnten dann für zukünftige Tranchen oder weitere Sicherheitsanalysen genutzt werden. Die Analysen haben gezeigt, dass die Varianz der Posen mit weiteren Datenlieferungen zugenommen hat und gelegentlich öffentliche Datensätze übertrifft. Besonders liegende oder kniende Personen konnten so nicht in dem Umfang in öffentlichen Daten identifiziert werden. Allerdings wurde auch bemerkt, dass öffentliche Daten eine größere Varianz in der Pose des Oberkörpers und den Armen aufweisen als die synthetischen Daten.

Zusammenfassend wurden mehrere Posenschätzer erforscht, entwickelt und analysiert. Darüber hinaus wurden auch Anwendungsfälle mit Blick auf die Sicherheitsrelevanz der Posenschätzung demonstriert. Wir gehen davon aus, dass die Posenschätzung in Zukunft eine zunehmend wichtigere Rolle in der sicheren KI für autonome Fahrzeuge spielen wird.

4 TP2 Generieren von synthetischen Lern- und Testdaten

Wichtigste Ergebnisse und Ereignisse

Synthetische Daten bieten wesentliche Eigenschaften, die sie für Sicherheitsanalysen empfehlen. So ist es möglich Szenen mit gezielten Eigenschaften zu erzeugen und darin einzelne Faktoren zu variieren. Insbesondere gilt das für kritische Verkehrssituationen. Präzise Ground Truth und umfangreiche Metainformation sind generierbar. Im Rahmen von TP2 wurden gemäß den Anforderungen des Projektes zwei Toolketten zur Erzeugung synthetischer Daten umgesetzt. Die auf real-time Rendering basierte Toolkette von Mackevision wurde dahingehend gestaltet, dass sie die Datengenerierung basierend auf einer eingeschränkten Datenspezifikationssprache unterstützt und sehr weitreichende Meta-Informationen liefert, die in Sicherheitsanalysen eingesetzt werden können, wie z.B. Informationen zur Verdeckung von Personen. Die Toolkette von BIT TS und Intel basiert auf physical-based Rendering mit Intel OSPRay und nutzt Sensor Plugins von Bosch und Valeo. Sie erlaubt die Nachbildung physikalischer Effekte wie insbes. Bewegungsunschärfe und den Effekt des zeitlichen Verlaufs eines LIDAR Scans. Durch die aus den Anforderungen des Gesamtprojektes bedingte Priorisierung von frühzeitig im Projekt gelieferten Daten mit hoher Variation konnten diese erweiterten Generierungsfeatures dem Projekt erst zu Projektende zur Verfügung gestellt werden. Der Großteil der Frames von BIT TS basiert auf einer Vorläuferversion dieses Toolings.

Insgesamt stellt TP2 dem Projekt ca. 370.000 Frames bereit. Die Frames zeichnen sich über den Projektverlauf durch eine zunehmende Varianz, zunehmende optische Effekte und zunehmende Ground Truth und Metainformation aus. Für alle Daten werden Ground Truth für Semantic Segmentation und Instance Segmentation, 2D und 3D Bounding Boxes und Tiefenkarten geliefert. Je nach Datentrache gibt es ferner LIDAR Sensordaten (ideal oder simuliert), Body Part Segmentation und Posen-Annotation. Die Metainformationen erlauben die Verknüpfung der Instanzen in den Frames mit den Assets im Asset-Katalog und deren Eigenschaften, wie Körpergröße oder Bekleidung von Fußgängern. Neben den von TP2 gelieferten Metainformationen erfolgt durch weitere Arbeiten im Projekt eine Anreicherung um weitere Informationen, etwa zur Relevanz der Objekte vor dem Hintergrund der betrachteten KI Funktion.

Wichtig für den Projektverlauf von TP2 war die Anforderung, die anderen Teilprojekte möglichst frühzeitig mit Daten zu beliefern. Dies führte insofern zu Mehraufwand als in vierteljährlichem Rhythmus die Produktion großer Datenmengen möglich sein musste. Eine längerfristige Entwicklung von Features war erschwert. Allerdings war es auch nur so möglich, Lernschleifen aus der Anwendung der Daten für Training und Test durchzuführen. Ferner war kritisch, dass die Anforderungen an die Daten sehr unklar waren. So entstand hoher Aufwand für die Verfeinerung und Bewertung von Anforderungen. Die letztendliche Priorisierung für die Umsetzung erfolgte im Rahmen des P1 Prozesses.

Basierend auf der im Projekt erkannten Wichtigkeit der Teilprojekt-übergreifenden Vernetzung bzgl. der Sicherheitsargumentation beteiligte sich TP2 mit Arbeiten im Evidence Workstream zu Performance Limiting Factors und leitete den Evidence Workstream zu Data Coverage. Beide sind wichtige Element der Sicherheitsargumentation. Bei dem Thema Datenqualität hat sich gezeigt, dass für die Anwendung die Datenqualität hinsichtlich der einfachen Nutzbarkeit der Daten zunächst mehr im Vordergrund steht als Qualität bzgl. spezifischer optischer Größen. Die Implementierung der „Latest-and-Greatest“ Daten Releases, die zum Release Zeitpunkt alle

bekannten Korrekturen und Verbesserungen enthalten, ist hieraus motiviert. Viele Analysen zur Datenqualität fließen über die Workstreams, insbes. jenen zu Performance Limiting Factors, in die Sicherheitsargumentation ein.

Neben der Toolentwicklung und Datengenerierung wurden in TP2 verschiedene Ansätze zur Suche nach und Konstruktion von Corner Cases entwickelt. Diese sind ebenfalls eine Grundlage für die Sicherheitsargumentation. Im Weiteren wurden in TP2 die Auswirkungen von Sensorparameteränderungen und Möglichkeiten zum Umgang damit untersucht.

Die Veröffentlichung der Daten aus dem Projekt ist geplant, um weitere Arbeiten zur Sicherheit von KI-basierten Perzeptionsfunktionen zu stimulieren. Wir sind überzeugt davon, dass die generierten Daten eine sehr gute Grundlage für viele weiterführende Forschungs- und Entwicklungsarbeiten bieten. Eine der Toolketten wird im Schwesterprojekt "KI Data Tooling" weiter eingesetzt. Erfahrungen zu Datenanfordeungen und zur Datenqualität wurden an dieses weitergegeben.

4.1 AP2.1 Toolketten für synthetische Datenerzeugung

Für die synthetische Datenerzeugung braucht es zum einen die 3D Eingangsdaten des zu simulierenden Szenarios und zum anderen die verarbeitende Software, welche daraus Bild-, Sensor- und Metadaten erzeugt. Die Herausforderung für AP2.1 in der Auswahl, Entwicklung und Verfeinerung dieser Software bestand in den teils widersprüchlichen, teils sich während des Projektes erst ergebenden oder ändernden Anforderungen: die Rendering Software soll realistische Ergebnisse liefern, schnell, flexibel und modular sein, verschiedene Sensorarten und Variationen unterstützen, sowie neue Eingabeformate lesen als auch zusätzlich umfangreiche Metadaten erzeugen. Schon vor Projektstart war klar, dass sich dies nicht mit einer einzigen Software erreichen lässt, sondern *zwei Toolketten* mit unterschiedlichem Schwerpunkt entwickelt und eingesetzt werden: die Game-Engine-basierte Toolkette von MackeVision mit Stärken in der Geschwindigkeit, Variationen und Effekten, sowie die physikalisch-basierte Toolkette von BIT-TS, Intel, Valeo und Bosch mit Fokus auf Modularität und der *realistischen Simulation* von verschiedenen Sensoren.

Die Wahl von zwei Toolketten führte zu der weiteren Herausforderung der Interoperabilität. Als Lösung geschieht der Austausch zwischen den Toolketten und die Anbindung verschiedener Module vorwiegend anhand spezifizierter Dateiformate. Vor allem das Khronos 3D Format glTF spielt eine zentrale Rolle, welches durch projektspezifische Erweiterungen ergänzt wurde: durch Einbinden von externen Szenenteilen und -objekten, Hintergrund und Beleuchtung mittels HDR Bildern oder eines prozeduralen Modells für Himmel und Sonne, sowie als wichtiges Ergebnis die Spezifikation realistischer Objektive, Bild- und LiDAR-Sensoren (siehe E2.1.2 und `EXT_cameras_sensor`). Die umfangreiche Unterstützung von glTF und UDIM Texturen und die Generierung von weitreichenden Groundtruth Metadaten wurde in Intel OSPRay Studio implementiert (weitere Details in E2.1.3).

Hauptaugenmerk bekam die Entwicklung der Module zur Sensorsimulation: das PSF Modul von Bosch, welches realistisch Kameras in Abhängigkeit der Entfernung, Sensorposition und Wellenlänge simuliert; das LiDAR Modul von Valeo mit Unterstützung von multiplen Echos, realistischen Abtastmustern und Varianten; und das Modul von Intel zum Einbringen von Unzulänglichkeiten von Bildsensoren wie Rauschen oder Linsenfehler (siehe E2.1.8 und E2.1.5). In Zuarbeit wurde der Renderkern OSPRay von Intel angepasst und erweitert, z.B. durch das

Ermöglichen einer flexiblen, natürlichen Beleuchtung mittels eines Himmels- und Sonnenmodells sowie mittels ausgemessenen Lichtquellen. Ein weiterer neuer Bereich betrifft die realistischen Sensoreindrücke durch die Berechnung von Bewegungsunschärfe und unterschiedliche Kameraverschlüsse (globaler und rolling shutter, weitere Details in E2.1.6).

Somit stehen nun mächtige, flexible und umfangreiche Werkzeuge für die synthetische Datenerzeugung zur Verfügung, welche eine sehr gute Basis für zukünftige Verbesserungen und weitere Sensormodule bilden.

4.2 AP2.2 Corner Cases

AP2.2. hatte zum Ziel eine inhaltlich breite Forschung für Corner Cases anzugehen. Hierzu wurde nach einer passenden Definition für „Corner Cases“ innerhalb der verschiedenen Unter-Arbeitspakete geforscht und als Basis der weiteren Arbeiten festgelegt: "*Ein Corner Case für eine KI-Funktionalität ist eine Situation (Szene samt Kontext und dynamischen Objekten), in der die KI-Funktionalität ein nicht erwartetes und funktional nicht hinlängliches Ergebnis berechnet, obwohl ein korrektes Verhalten erwartbar war.*" Hierbei war der Ansatz zum einen die Detektion von Corner Cases systematisch umsetzen zu können und weiterhin den Aufbau eines Corner Case Datensatzes zu untersuchen.

Daneben hatte AP2.2. einen integralen Bestandteil als „Gate-Keeper“ – das heißt als zentraler Ansprechpartner – für die Requirements in der Datengenerierung. Diese Requirements wurden nach Abstimmung und Priorisierung an die Datenproduzenten des AP2.5 weitergeleitet

Für die Corner Cases wurden Beschreibungssprachen für die Operational Design Domain vorgeschlagen. Initial wurde eine Corner Case Taxonomie erstellt, welche methodisch verschiedene Dimensionen bzw. Umstände erklärt, die zu Corner Cases führen können (z.B. Umweltstörungen oder Adversarial Attacks). Basierend auf der mit Experten im AP2.2 entwickelten Corner Case Taxonomie wurden dieser in diversen Arbeitspaketen weitere Dimensionen hinzugefügt (z.B. basierend auf Aufnahme der Umgebung (Wechselwirkung Szene-Sensor) oder der Verarbeitung der Sensordaten durch die KI-Funktionalität (Wechselwirkung Sensor-Intelligenz) bzw. flossen diese Ergebnisse in weitere Arbeitspakete ein. Um die Erstellung eines Corner Case Datensatzes zu beschleunigen wurde in diesem Rahmen ebenfalls ein Corner Case Detektor erstellt, welcher diese Grenzfälle automatisiert erkennen kann. Neben der Erstellung des synthetischen Datensatzes wurden parallel alternative Datenquellen für Corner Cases untersucht. Ziel war es, eine möglichst weit fortgeschrittene Automatisierung dieser Systeme zu entwickeln, um so eine hohe Anzahl an Corner Cases für das Projekt zu generieren. Dies hat eine höhere Qualität der Trainingsdaten zur Folge. Beispielhaft hierfür kann die Entwicklung eines Trajektorien-Datensatzes durch den Projektpartner DLR genannt werden. Hierbei werden verkehrsbedingte Grenzfälle und kritische Situationen im Datensatz detektiert. Dieser Datensatz wurde auf Basis einer Regelverletzung (Abbiegeverbot) vom Projektpartner Volkswaren AG mittels der Taxonomie aus dem E2.2.5 auf unterschiedliche Arten untersucht (z.B. mittels Klassifikatoren oder Pixelwerteigenschaften). Ergebnisse dieser Untersuchung waren Korrelationen zwischen den Kriterien (z.B. je größer eine Person, desto höher die Erkennungsleistung). Neben der automatisierten Erzeugung und Detektion von Corner Cases gab es durch QualityMinds und das DFKI weitere Untersuchungen bzgl. Augmentierungsmöglichkeiten für vorhandene Frames, z.B. durch synthetisch erzeugte Verdeckungen (Leaf Occlusion, Drop Occlusion) oder weitere Augmentierungen wie Nebel oder Rauschen. Diese Augmentierungen

wurden wiederum genutzt, um bei Finden einer geringen Detektionswahrscheinlichkeit einen Corner Case zu identifizieren. Hierbei konnte beobachtet werden, dass statistisches Rauschen die höchste Corner Case Rate hatte. Ein weiterer Teil des AP2.2 war die Erzeugung synthetischer Szenen für die Projektpartner, z.B. auf Basis der weiter oben genannten Verkehrssituation an der Leonberg Kreuzung. Valeo hat als Projektpartner hierbei die weitere Analyse der Identifikation von Corner Cases vorangetrieben. Die Identifikation von Corner Cases basiert hier auf einem Unterschied in den DNN-Ausgaben auf der Grundlage verschiedener Modalitäten (LiDAR und Kamera). Fortführend wurde durch QualityMinds ein iteratives Schema zur Identifikation von Corner Cases aufgesetzt. Durch Anwendung der Methode auf den KI-Absicherungs-Datensatz und eine semantische Segmentierungs-KI wurden insgesamt acht Merkmale gefunden, die in drei Gruppen unterteilt werden können: quantitative, wahrnehmungsbezogene und situative Merkmale. Um solche Merkmale zu finden, wurde ein iterativer Prozess definiert, bei dem jede Iteration aus vier Phasen besteht: Explorationsphase (Analyse des Datensatzes, z.B. durch Plotten der Daten), Hypothesenformulierungsphase (es wird eine bestimmte Leistungseinschränkung angenommen, z.B. Instanzgröße), Experimentierphase (Auswahl konkreter Merkmale und Bewertung, ob das Merkmal neu ist und einen signifikanten Einfluss hat) und die Ergebniskompilierungsphase (Anwendung der neu gefundenen Auswahlregel zusammen mit alten Regeln, um einen neuen Corner Case-Datensatz zu erstellen). Weitere Experimente im Bereich Occlusion / Zeitanalyse fanden ebenfalls statt.

Neben der Identifikation, Analyse und Erstellung von Corner Cases inkl. einem entsprechendem Datensatz wurde in diesem Arbeitspaket ferner die synthetische Datenerzeugung des Gesamtprojektes auf Requirements-Ebene adressiert. In der "Gate-Keeper" Rolle wurde die Qualität der Datenanforderungen in enger Abstimmung mit den Anforderern und Datenproduzenten sichergestellt, um so einen möglichst stark auf die Bedürfnisse der Anwender angepasste Datensatz zu erzeugen. Hierbei wurden auch Lessons Learned aus der Datengenerierung agil berücksichtigt und für die jeweils nächste Iteration der Datenanforderungen und Datengenerierung übernommen. Weiterhin wurde für die gewünschte Exaktheit von Datenanforderungen ein JSON Format entwickelt. Insgesamt wurden basierend auf den Datenanforderungen rund 360.000 Frames produziert.

Zusammenfassend konnten in diesem Arbeitspaket neben der Taxonomie zur Identifikation von Corner Cases auch praktische Ergebnisse bzgl. automatisierter Datensatz-Erzeugung von Corner Cases erreicht werden (inkl. Identifikation). Weiterhin gab es in Zusammenarbeit mit anderen Arbeitspaketen (AP2.5, AP2.1) und Prozess P1 die ständige Abstimmung und Weiterentwicklung bzgl. Anforderungsprozess und Datengenerierung, welche als Ergebnis den KI-Absicherungs-Datensatz hat.

4.3 AP2.3 Abstraktion von Sensorik

Das Training tiefer neuronaler Netze (DNNs) ist letztendlich ein Optimierungsproblem gegenüber den Trainingsdaten. Dies bringt mit sich, dass DNNs sehr sensitiv auf Änderungen in den Eingangsdaten (Domain Shift) sind. Eine relevante Klasse von Änderungen sind dabei Änderungen an Sensorparametern wir sie z.B. bei Verwendung unterschiedlicher Sensorvarianten und bei Integration in unterschiedliche Fahrzeuge auftreten. Kernziel von AP2.3 ist es die Auswirkungen solcher Parameteränderungen zu untersuchen, Verfahren zum Umgang mit Domain Shift zu untersuchen und weiterführende Maßnahmen hin zu einer Abstraktion von der konkreten Sensorik zu erforschen.

Dazu wurden relevante Klassen von Sensorparameteränderungen im Bereich des automatisierten Fahrens identifiziert und abgeleitet, welche Daten zur Untersuchung notwendig sind (E2.3.1 und E2.3.2). Diese wurden im Rahmen des Datenanforderungsprozesses im Projekt angefordert. Da nur AP2.3 diese spezielle Art von Daten benötigt, konnte deren Generierung erst spät im Priorisierungsprozess berücksichtigt werden. Bilddaten mit veränderten Kameraparametern lagen daher erst spät und Daten mir veränderten LIDAR Parametern lagen erst sehr spät im Projekt vor. Einige Untersuchungen wurden daher mit öffentlichen Datensätzen durchgeführt. Einige weitere Untersuchungen sind in ihrer Tiefe und ihrem Umfang durch die geringe dafür verbleibende Zeit eingeschränkt.

Die Auswirkung der Sensorparameteränderungen wurde für Semantic Segmentation, 2D und 3D Verfahren, sowie für Posenschätzung untersucht. In den Experimenten zeigt sich, dass Änderungen des Farbraums größere Auswirkungen haben als Kameraparameter wie Auflösung, Field-of-View und Kameraposition. Weitere Unterschiede ergeben sich durch die Art des Erkennungsverfahrens und den dafür genutzten Trainingsprozess. Im Training genutzte Datenaugmentierungen habe Einfluss darauf, wie sich Parameteränderungen auswirken. Analysen auf öffentlichen Datensätzen für die Wahrscheinlichkeitsverteilung von LIDAR Punkten zwischen einem Sensor auf Stoßstangenhöhe und einem auf Dachhöhe zeigen starke Unterschiede. Die Nutzung dieser Unterschiede zur Transformation zwischen / Abstraktion von diesem Sensorparameter war jedoch nocht erfolgreich.

Fine-Tuning Ansätze zur Optimierung auf geänderte Sensorparameter zeigen sich im Projekt erfolgreich. Auch ein geringer Umfang von Daten der Ziel-Domäne, im Experiment ca. 500 Frames, zeigen schon eine wesentliche Verbesserung. Die Experimente legen jedoch auch nahe, dass die Verbesserung auf der Ziel-Domäne mit Verschlechterung auf der Quell-Domäne einhergeht. Das legt den Einsatz spezialisierter Modelle nahe. Die untersuchten Optimierungsansätze basierend auf Domain Adaptation Verfahren zeigen unterschiedliche Ergebnisse. Wie auch in anderen Experimenten im Rahmen des Projektes zeigt sich, dass gelabelte Daten der Zieldomäne einem unsupervised Domain Adaptation Verfahren gegenüber zu bevorzugen sind. Ferner erwies sich die finden guter Hyperparameter für Domain Adaptation Verfahren als schwierig und war nicht immer erfolgreich.

Die Untersuchung von Domain Shift mit gezielt konstruierten synthetischen Daten, die die Manipulation einzelner Parameter zulassen hat sich bewährt. Wir gehen davon aus, dass mit dem Datensatz aus dem Projekt viele weiterführende Untersuchungen und Entwicklungen ermöglicht werden.

4.4 AP2.4 Bewertung Qualität und Relevanz synthetischer Daten

Die Verwendung synthetischer Daten für das Trainieren und Testen von Perzeptions-Algorithmen (z.B. DNNs zur semantischen Segmentierung), die final ihre Anwendung in der realen Welt finden, birgen Chancen, aber auch Risiken. Die Untersuchung von algorithmischen Anpassungen für diesen Fall, sowie die Untersuchung von Bewertungskriterien von synthetischen Daten sind Aufgabe des APs.

Zur gezielten Untersuchung von Datenkriterien auf die Deep Neural Network (DNN) Performance wurden Anforderungen von Parametervariationen in der Datengenerierung an die Datenproduzenten formuliert (E.2.4.1).

Zum Vergleich der Auswirkung unterschiedlicher Datensätze auf KI-Netzwerk-Performance (E.2.4.3) wurden weitergehende Experimente durchgeführt, die es vorsahen den Trainingsdatensatz bewusst in gewissen Dimensionen zu beschränken (z.B. Kontrast oder Posen), um die Auswirkung auf den gleichen Dimensionen in den Testdaten zu untersuchen. Hierzu wurde das SSD Modell mit der Tranche 5 von Mackevision verwendet. Für die Fußgänger-Posen kann eine Kausalität zwischen fehlender Posen in den Trainingsdaten und schlechte Performance auf den Testdaten festgestellt werden.

Ein besonderer Fokus wurde auf die Definition der sogenannten Design of Experiments (DOEs) gelegt. In diesen DOEs sollen unterschiedliche Eigenschaften in den Eingangsdaten auf die Performance der DNNs untersucht werden. Diese Eigenschaften werden im Projekt als Performance Limiting Factors (PLFs) bezeichnet und wurden in Abstimmung mit anderen APs / TPs ermittelt. Unter anderem wurden die Auswirkung von verschiedenen Kontrasten bei Fußgängern sowie deren Kontrast-Umgebung auf die DNNs Performance untersucht werden. Weitere Einflussgrößen sind die Licht- und Farbverteilung im Bild. Hierzu wurden Histogramme erstellt werden, die innerhalb und außerhalb der Fußgänger Bounding Box die Licht- und Farbverteilung im Bild beschreiben (z.B. Hauptfarbe, Farbvarianz).

Zudem wurde die Abhängigkeit zwischen der Shannon Entropy des Bildes/Fußgängers/Fußgängerumgebung und der Performance des DNNs untersucht werden. Zusätzlich sind die Posen der Fußgänger Teil der DOEs, bei der die Häufigkeit von gewissen Posen der Fußgänger innerhalb des Datensatzes ermittelt wurden. Die Posen können in z.B. 16 Cluster eingeordnet werden (abhängig von ihrer Drehung etc.) und ein Vergleich der Verteilung zwischen Trainings-, Validierungs und Testdaten gezogen werden.

Weiterhin wurden Arbeiten durchgeführt, die die statistische Signifikanz von Performance Limiting-Factor (PLF) Untersuchungen beleuchten. Hierzu wurde ein neues Testverfahren mit der Bezeichnung "Oracle Testing" vorgestellt. Die Frage nach der Übertragbarkeit der PLFs auf weitere Modelle, ist aus den durchgeführten Analysen unzureichend beantwortbar und bedarf weiterer Experimente.

Ein weiterer Teil des AP2.4 stellt die Domain Adaptation dar, bei der eine Minimierung der Domänenverschiebung vor allem zwischen synthetischen und realen Daten angestrebt wird. Ein Fokus liegt auf die Auswertung von kritischen Fälle, die naturgemäß in realen Daten nur unzureichend vorhanden sind. Zur optimierten Parametrierung mit einem Vergleich der Auswirkung unterschiedlicher Datensätze (E.2.4.4) wurde ein GAN Ansatz erarbeitet und Evaluierungen zur Distanzmessung zwischen realen und synthetischen Bildern durchgeführt. Weiterhin wurde die Domänenverschiebung zwischen öffentlich verfügbaren Datensätzen und dem KIA Datensatz messbar gemacht. Hierzu wurden Histogramme erstellt, die die DNN Performance auf den einzelnen Datensätze darstellen, und mittels der Wasserstein Distanz die Domänenverschiebung messbar gemacht. Dies kann ebenfalls ein Maß für die Ähnlichkeit der generierten synthetischen Daten zu den realen Daten darstellen. In einer weiteren Analyse wurden Datensätze anhand der mit ihnen erreichten Leistung unter Nutzung von Domain Adaptation Verfahren bewertet.

Für die Wirkkettenanalysen aus der Anwendung gezielt variierter Datensätze für grenzwertige Anwendungssituationen (E.2.4.5) wurde eine Datenmetrik basierend auf der Shannon Entropy zur Messung des Informationsgehaltes in den synthetischen Daten entwickelt. Weiterhin wurden systematisch Bildkontraste verändert und in Relation zu der Detektionsperformance gesetzt.

Anhand der durchgeführten Untersuchungen konnten Einblicke in die Funktionsweise der Mechanismen zur Domain Adaptation erlangt werden. Eine Bewertung von Qualität und Relevanz synthetischer Daten wurde auf experimenteller Basis durchgeführt und mündete in einer Liste von PLFs. Tiefergehende Analysen mit mehr Datensätzen und DNN Architekturen stellt den Ausblick dar.

4.5 AP2.5 Datengenerierung und Noisy Data

Für die unterschiedlichen Anforderungen des Trainings, der Validierung und des Testens von KI basierten Algorithmen und deren Auswertung, wurden in diesem Arbeitspaket synthetische Daten erzeugt. Diese Daten wurden in Tranchen im Dreimonatsrhythmus generiert und ausgeliefert, wobei jede Datenlieferung einem Ergebnis E2.5.1(x) entspricht. Für jede Tranche wurden die Inhalte und Merkmale gemeinsam mit AP2.2 und dem P1 Prozess koordiniert und definiert. Aufgrund zeitlicher Verschiebungen zum Projektstart und im Projektverlauf wurde E2.5.2j weggelassen. Die Gesamtanzahl der geplanten Frames wurde jedoch beibehalten in dem die entsprechenden Frames auf die vorherigen Lieferungen vorgezogen wurden.

Die Anforderungen an den synthetischen Datensatz im Projekt sind vielfältig. Durch automatisierte und systematische Datengenerierung soll der Datensatz viele Fußgänger in städtischen Szenarien mit hoher Varianz, unterschiedliche Licht- und Wetterbedingungen und umfangreiche Ground Truth- sowie Metadaten umfassen. Des Weiteren, soll er auch spezielle Szenarien und Corner Cases enthalten.

Bei der Datengenerierung wurden die beiden Toolketten von dem Projektpartner Mackevision und BIT Technology Solutions (BIT TS) / GUA8 entwickelt (siehe auch AP2.1). Die Toolkette von Mackevision basiert auf einer Game Engine und die Toolkette von BIT TS nutzt physikalisch basiertes Offline-Rendering, unter anderem mit der Render Engine OSPRay, die von Intel entwickelt wurde.

Die Daten wurden in Sequenzen erzeugt, die auf 3D Szenen und unterschiedlichen Szenarien aufbauen. Jede Tranche enthält eine Anzahl von Sequenzen. In der ersten Phase des Projekts (E2.5.1a – E2.5.1c) wurden existierende Tools und bestehendes Knowhow der Datenproduzenten genutzt, um möglichst früh Daten bereitstellen zu können. In den darauffolgenden Datenlieferungen (E2.5.1d – E2.5.1i) konnten durch die gezielte Entwicklung der Toolketten (AP2.1) neue Eigenschaften in den Frames und den zugehörigen Metadaten erlangt werden.

Grundkontext und Assets

Die Grundlage für die Datengenerierung sind zum einen 3D Elemente (Assets) für den Grundkontext, als auch Fußgänger und Objekte zur Erzeugung von Szenarien. Für die ersten drei Datenlieferungen wurde hier teilweise auf vorhandene Daten zurückgegriffen, jedoch wurden von Beginn an auch eine Vielzahl exklusiver Assets (Geometrie und Materialien) für das Projekt erzeugt. Insbesondere wurden über 40 animierbare Fußgängermodelle mit einer konsistenten Struktur für das Projekt produziert. Um die Anzahl und Vielfalt der Assets weiter zu vergrößern,

wurden zusätzlich Assets aus öffentlichen Bibliotheken bezogen, überarbeitet und in die Datengenerierung integriert.

Um einen hohen Grad an Realismus und Genauigkeit in der Datengenerierung abbilden zu können, wurden Messungen von Fußgängerbewegungen und Materialeigenschaften in den GUAs 6 und 7 durchgeführt (E2.5.2). Die aufgenommenen Skelettbewegungen von mehreren Personen und deren Interaktion wurde zur Datengenerierung auf die 3D Personenmodelle übertragen („retargeting“). Materialeigenschaften von gesammelten Materialproben, wie Putz, Textilien und Fahrbahnbeläge, wurden in GUA7 durch X-Rite mit einem TAC7 Scanner vermessen, prozessiert und in den Datenformation AxF und glTF bereitgestellt. Dadurch konnten die gesammelten Materialien in beiden Toolketten genutzt werden und dediziert Testdaten erzeugt werden.

Datensatz

Um die Anforderungen der Datenkonsumenten erfüllen zu können, wurden die Toolketten in AP2.1 kontinuierlich mit neuen Funktionen weiterentwickelt und zur Datengenerierung genutzt. Dadurch konnten mit jeder Datenlieferung neue Merkmale dem Datensatz hinzugefügt werden. Beispiele hierfür sind unterschiedliche Licht- und Wetterbedingungen, Sensormodelle, -parameter und -effekte, spezielle Szenarios, sowie umfangreiche Metadaten bzw. -annotationen.

Der gesamte Datensatz ist den Projektpartnern in Sequenzpaketen auf dem zentralen Datenspeicher zur Verfügung gestellt worden. Ein kleiner Auszug aus dem Datensatz ist in acht Abbildungen (Abbildung 4.1 bis Abbildung 4.8) dargestellt. Eine Übersicht der Daten pro Tranche und den Datentypen ist in Tabelle 4.1 und Tabelle 4.2 zu finden.

Tabelle 4.1: Überblick der Daten Tranchen.

Ergebnis (Tranche)	Frames BIT-TS	Frames Mackevision	Frames Summe	Delivery Notes
E2.5.1a (Tranche 1)	10.120	6.328	16.448	Data Delivery of Tranche 1: BIT-TS Data Delivery of Tranche1: Mackevision
E2.5.1b (Tranche 2)	20.006	10.144	30.150	Data Delivery of Tranche 2: BIT-TS Data Delivery of Tranche 2: Mackevision
E2.5.1c (Tranche 3)	36.533	---	36.533	Data Delivery of Tranche 3: BIT-TS
E2.5.1d (Tranche 4)	12.654	22.297	34.951	Data Delivery of Tranche 4: BIT-TS Data Delivery of Tranche 4: Mackevision
E2.5.1e (Tranche 5)	11.594	36.394	47.988	Data Delivery of Tranche 5: BIT-TS Data Delivery of Tranche 5: Mackevision
E2.5.1f (Tranche 6)	---	49.568	49.568	Data Delivery of Tranche 6: Mackevision
E2.5.1g (Tranche 7)	23.689	50.302	73.991	Data Delivery of Tranche 7: BIT-TS Data Delivery of Tranche 7: Mackevision
E2.5.1h (Tranche 8)	---	48.521	48.521	Data Delivery of Tranche 8: Mackevision
E2.5.1i	---	39.330	39.330	Data Delivery of Tranche 9: Mackevision

Tabelle 4.2: Überblick der verfügbaren Sensordaten, Ground Truth und Meta-Annotationen. Falls ein Datentyp nur aus einer Toolchain verfügbar ist, ist der Datenproduzent in eckigen Klammern genannt.

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
E2.5.1a (Tranche 1)	• RGB camera (exr, png)	• semantic group segmentation (png) • instance segmentation (png) [Mackevision] • instance segmentation (exr) [BIT TS] • depth (exr) [Mackevision] • depth (csv) [BIT TS] • depth (png) [BIT TS]	
E2.5.1b (Tranche 2)	• RGB camera (exr, png)	• semantic group segmentation (png) • instance segmentation (png) [Mackevision] • instance segmentation (exr) [BIT TS] • 2d bounding boxes (json) • 3d bounding boxes (json)	• class IDs for vehicles and pedestrians (json) [Mackevision] • sensor model (json) [BIT TS] • meta annotations: light source information (json) [Mackevision]

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
		• depth (exr) [Mackevision] • depth (csv) [BIT TS] • depth (png) [BIT TS] • camera matrices (csv) [BIT TS] • camera matrices (json) [Mackevision]	
E2.5.1c (Tranche 3)	• RGB camera (exr, png) • LIDAR, simple model (pcd) [BIT TS]	• semantic group segmentation (png) • instance segmentation (png) • instance segmentation (exr) [BIT TS] • 2d bounding boxes (json) • 3d bounding boxes (json) • depth (exr) • camera matrices (csv) [BIT TS]	• sensor model (json) [BIT TS]
E2.5.1d (Tranche 4)	• RGB camera (exr, png)	• semantic segmentation (png)	• meta annotations: light source information (json) [Mackevision]

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
	• LIDAR, simple model (pcd) [BIT TS]	• instance segmentation (png) • instance segmentation (exr) [BIT TS] • 3d bounding boxes (json) • 2d bounding boxes (json) • depth (exr) [Mackevision] • depth (csv) [BIT TS] • depth (png) [BIT TS] • camera matrices (csv) [BIT TS] • camera matrices (json) [Mackevision] • sensor model (json) • 3d skeleton data (json) [Mackevision] • 2d skeleton data (json) [Mackevision] • body part segmentation (png) [BIT TS]	• meta annotations: road surface mean color (json) [Mackevision] • meta annotations: road material id (json) [BIT TS] • meta annotations: entities array with pedestrians and corresponding instance id (json) • meta annotations: reference to asset catalogue base asset uuid for pedestrians (json) • meta annotations: pedestrian cloth color (json) [Mackevision] • meta annotations: HDRI reference (json) [BIT TS]

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
E2.5.1e (Tranche 5)	• RGB camera (exr, png) • RGB camera: processed with Intel error generator (AP2.1) -- sensor effects and advanced tone mapping (png) • LIDAR, simple model (pcd) [BIT TS]	• semantic segmentation (png) • instance segmentation (png) • instance segmentation (exr) [BIT TS] • 3d bounding boxes (json) • 2d bounding boxes (json) • depth (exr) • camera matrices (csv) [BIT TS] • camera matrices (json) [Mackevision] • sensor model (json) • 3d skeleton data (json) [Mackevision] • 2d skeleton data (json) [Mackevision] • body part segmentation (png) [BIT TS]	• meta annotations: light source information (json) [Mackevision] • meta annotations: road surface mean color (json) [Mackevision] • meta annotations: road material id (json) [BIT TS] • meta annotations: extended entities array with pedestrians and corresponding instance id (json) • meta annotations: reference to asset catalogue base asset uuid for pedestrians (json) • meta annotations: pedestrian cloth color (json) [Mackevision] • meta annotations: HDRI reference (json) [BIT TS] • meta annotations: light source sun position (json) [Mackevision] • meta annotations: more object information for objects (json) • meta annotations: occlusion information for pedestrians -- occlusion rate, occluding object, amount of occluded pixels (json) [Mackevision]

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
E2.5.1f (Tranche 6)	• RGB camera (exr, png)	• semantic segmentation (png) • instance segmentation (png) • 3d bounding boxes (json) • 2d bounding boxes (json) • depth (exr) • camera matrices (json) [Mackevision] • sensor model (json) • 3d skeleton data extended (json) [Mackevision] • 2d skeleton data extended (json) [Mackevision]	• meta annotations: light source information (json) [Mackevision] • meta annotations: road surface mean color (json) [Mackevision] • meta annotations: road material id (json) [BIT TS] • meta annotations: extended entities array with pedestrians and corresponding instance id (json) • meta annotations: reference to asset catalogue base asset uuid for pedestrians (json) • meta annotations: pedestrian cloth color (json) [Mackevision] • meta annotations: HDRI reference (json) [BIT TS] • meta annotations: light source sun position (json) [Mackevision] • meta annotations: more object information for objects (json) • meta annotations: occlusion information for pedestrians -- occlusion rate, occluding object, amount of occluded pixels (json) [Mackevision]

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
			• meta annotations: surface wetness condition by zwicky box [Mackevision] • meta annotations: light power applied to scene – sunlight [Mackevision] • meta annotations: sensor effect "lens flares" – toolchain specific parameters [Mackevision]

E2.5.1g (Tranche 7)	• RGB camera (exr, png)
	  • sensor module from Bosch [BIT TS]
	  • render engine OSPRay from Intel [BIT TS]
	• RGB camera: processed with Intel error generator (AP2.1) – sensor effects and advanced tone mapping (png) [BIT TS]
	• LIDAR (pcd) [BIT TS]
	  • sensor module from Valeo [BIT TS]
		• semantic segmentation (png)
		• instance segmentation (png)
		• 3d bounding boxes (json)
		• 2d bounding boxes (json)
		• depth (exr)
		• camera matrices (csv) [BIT TS]
		• camera matrices (json) [Mackevision]
		• sensor model (json)
		• 3d skeleton data extended (json) [Mackevision]
			• meta annotations: light source information (json) [Mackevision]
			• meta annotations: road surface mean color (json) [Mackevision]
			• meta annotations: extended entities array with pedestrians and corresponding instance id (json)
			• meta annotations: reference to asset catalogue base asset uuid for pedestrians (json)
			• meta annotations: pedestrian cloth color (json) [Mackevision]
			• meta annotations: light source sun position (json) [Mackevision]
			• meta annotations: more object information for objects (json)

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
		• 2d skeleton data extended (json) [Mackevision]	• meta annotations: occlusion information for pedestrians -- occlusion rate, occluding object, amount of occluded pixels (json) [Mackevision] • meta annotations: surface wetness condition by zwicky box [Mackevision] • meta annotations: light power applied to scene – sunlight [Mackevision] • meta annotations: sensor effect "lens flares" – toolchain specific parameters [Mackevision] • meta annotations: fog – toolchain specific parameters [Mackevision]
E2.5.1h (Tranche 8, WIP)	• RGB camera (exr, png)	• semantic segmentation (png) • instance segmentation (png) • 3d bounding boxes (json) • 2d bounding boxes (json) • depth (exr) • camera matrices (json) [Mackevision] • sensor model (json)	• meta annotations: light source information (json) [Mackevision] • meta annotations: road surface mean color (json) [Mackevision] • meta annotations: extended entities array with pedestrians and corresponding instance id (json) • meta annotations: reference to asset catalogue base asset uuid for pedestrians (json)

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
		• 3d skeleton data extended (json) [Mackevision] • 2d skeleton data extended (json) [Mackevision]	• meta annotations: pedestrian cloth color (json) [Mackevision] • meta annotations: light source sun position (json) [Mackevision] • meta annotations: more object information for objects (json) • meta annotations: occlusion information for pedestrians -- occlusion rate, occluding object, amount of occluded pixels (json) [Mackevision] • meta annotations: surface wetness condition by zwicky box [Mackevision] • meta annotations: light power applied to scene – sunlight [Mackevision] • meta annotations: sensor effect "lens flares" – toolchain specific parameters [Mackevision] • meta annotations: fog – toolchain specific parameters [Mackevision]
E2.5.1i (Tranche 9, WIP)	• RGB camera (exr, png)	• semantic segmentation (png) • instance segmentation (png) • 3d bounding boxes (json)	• meta annotations: light source information (json) [Mackevision]

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
		• 2d bounding boxes (json) • depth (exr) • camera matrices (json) [Mackevision] • sensor model (json) • 3d skeleton data extended (json) [Mackevision] • 2d skeleton data extended (json) [Mackevision]	• meta annotations: road surface mean color (json) [Mackevision] • meta annotations: extended entities array with pedestrians and corresponding instance id (json) • meta annotations: reference to asset catalogue base asset uuid for pedestrians (json) • meta annotations: pedestrian cloth color (json) [Mackevision] • meta annotations: light source sun position (json) [Mackevision] • meta annotations: more object information for objects (json) • meta annotations: occlusion information for pedestrians -- occlusion rate, occluding object, amount of occluded pixels (json) [Mackevision] • meta annotations: surface wetness condition by zwicky box [Mackevision] • meta annotations: light power applied to scene – sunlight [Mackevision] • meta annotations: sensor effect "lens flares" – toolchain specific parameters [Mackevision]

Ergebnis (Tranche)	Sensor Daten (Dateiformat) [Datenproduzent]	Ground Truth (Dateiformat) [Datenproduzent]	Meta-Annotationen (Dateiformat) [Datenproduzent]
			• meta annotations: fog – toolchain specific parameters [Mackevision]

Abbildung 4.1: Beispiele für Kamerasensorbilder aus Tranche 1 (Mackevision).

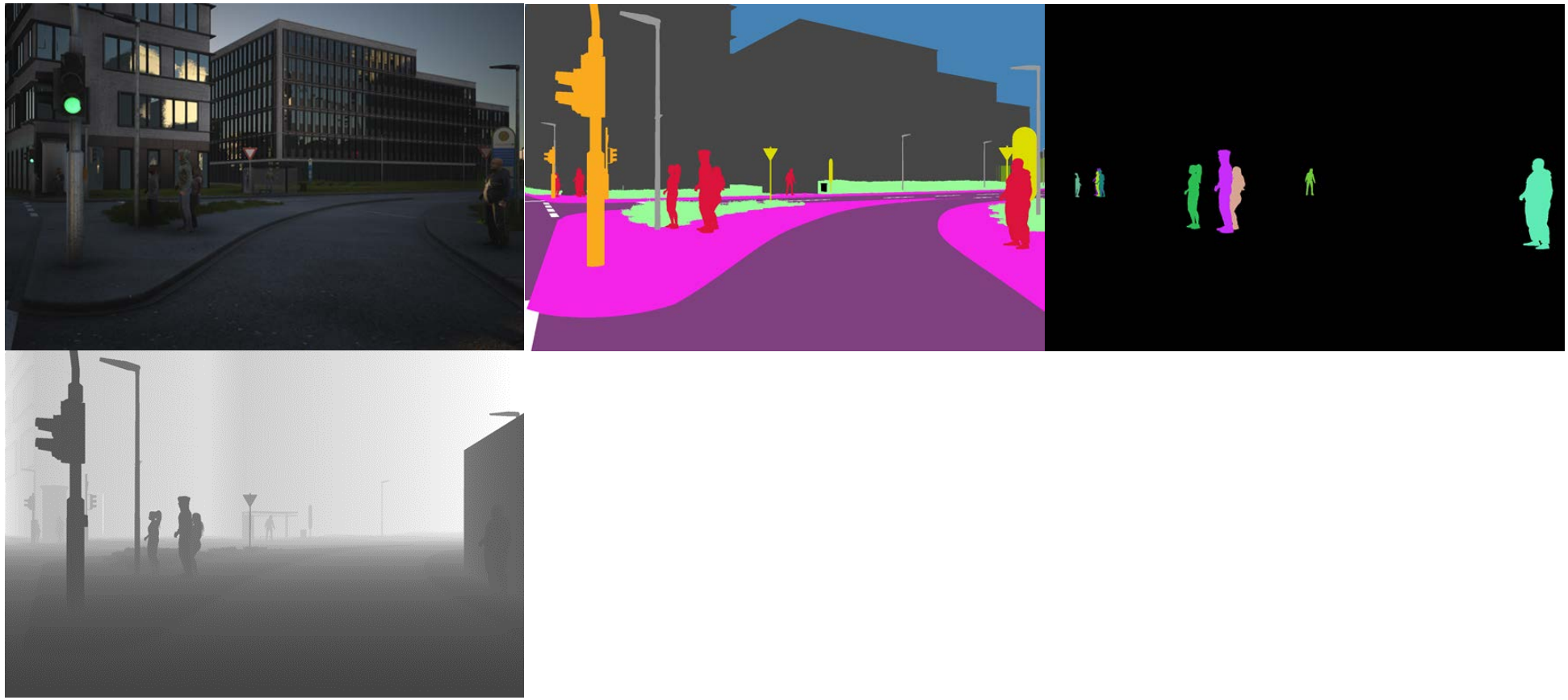


Abbildung 4.2: Ein Kamerasensorbild mit semantischer Gruppensegmentierung, Instanzsegmentierung von Fußgängern und Tiefeninformation aus Tranche 3 (Mackevision).

Abbildung 4.3: Ein Kamerasensorbild mit semantischer Gruppensegmentierung und vollständiger Instanzsegmentierung aus Tranche 4 (BIT TS).

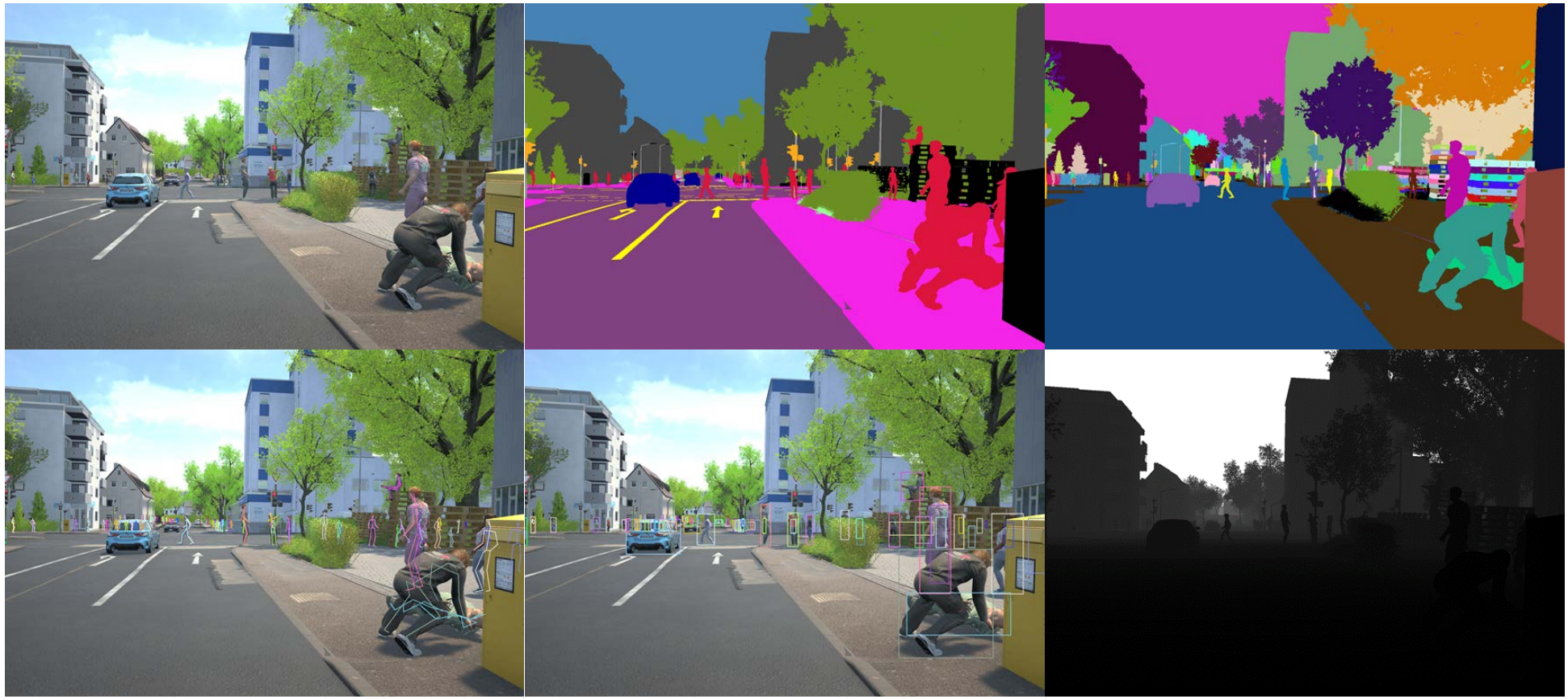


Abbildung 4.4: Ein Kamerasensorbild mit semantischer Gruppensegmentierung, vollständiger Instanzsegmentierung, Poseninformation, 2d Bounding Boxen und Tiefenkarte aus Tranche 5 (Mackevision).

Abbildung 4.5: Ausgewählte Kamerasensorbilder aus Tranche 6 (Mackevision) mit Nässe und Lens Flares (Linsenreflektionen).

Abbildung 4.6: Ausgewählte Kamerasensorbilder aus Tranche 7 (Mackevision) mit Vignettierung, Nebel und out-of-distribution Assets (Baustelle).

Abbildung 4.7: Ausgewählte Kamerasensorbilder aus Tranche 7 (BIT TS), bei der der Intel OSPRay Renderer zum Einsatz kam.

Abbildung 4.8: Ausgewählte Kamerasensorbilder aus Tranche 8 (Mackevision) mit Nachtszenarien.

5 TP3 Methoden und Maßnahmen zur Absicherung von KI

Wichtigste Ergebnisse und Ereignisse

Problemexposition

Um eine Absicherbarkeit von künstlicher Intelligenz für das automatische Fahren sicherzustellen, reicht reines Testen der abzusichernden Funktion aus zahlreichen Gründen nicht aus. Dazu gehören unter anderem der Aufwand für das Testen durch die große Vielfalt an möglichen Testfällen, die Sensibilität tiefer neuronaler Netze auf schädliche Bildstörungen (engl. Adverserial pertubations) durch natürlichen Schwankungen oder absichtlich herbeigeführte Manipulationen, das Nicht-Wissen über die Sensibilität trainierte neuronaler Netze bezüglich bestimmter semantischer Veränderungen wie Straßenbeschaffenheit oder Lichtverhältnisse und auch die Unklarheit auf welche Merkmale der Eingangsdaten sich das neuronale Netz sensibilisiert hat. Dies sorgt für die Notwendigkeit von Mechanismen zur Bewertung und Steigerung der Absicherbarkeit von KI. In anderen Worten: eine Argumentation für die Sicherheit einer KI-Funktion, also für das verlässliche (statistisch hinreichend sichere) Einhalten von Sicherheitszielen, benötigt die Ableitung von hochwertigen Informationen über die KI, um die Relevanz und Aussagekraft von Testfällen bewerten und belegen zu können.

Zusammenfassende Darstellung ausgewählter Ergebnisse

Den Kern der Ergebnisse von Teilprojekt 3 bildet eine systematische Bewertung der Effektivität von Mechanismen und Metriken zur Bewertung und Steigerung der Absicherbarkeit von KI. Grundlage bildet eine systematische Erfassung vom Stand der Technik in Form eines umfassenden Berichtes aus AP3.1 mit über 400 Referenzen und fortlaufendem Tracking vom Stand der Technik. Als zentrales Strukturierungsmerkmal wurde in AP 3.2 ein Konsens über DNN-spezifische Sicherheitsbedenken erarbeitet. Entlang dieser Strukturierung wurden in den APs 3.3-3.6 eine Vielzahl von Mechanismen und Metriken entwickelt, angewendet und bewertet. Hierbei wurde zwischen denjenigen Mechanismen unterschieden, die die Funktion an sich verändern („Funktional verändernde Methoden und Maßnahmen“, AP3.3), nicht verändern, aber die innere Struktur und alle Parameter des neuronalen Netzes kennen („White- und Greybox-Methoden und -maßnahmen“, AP3.4) und als Blackbox behandeln („Blackbox-Methoden und -Maßnahmen“, AP3.5). In AP3.6 wurden unter anderem Mechanismen entwickelt, die mehrere Mechanismen oder Modelle miteinander kombinieren.

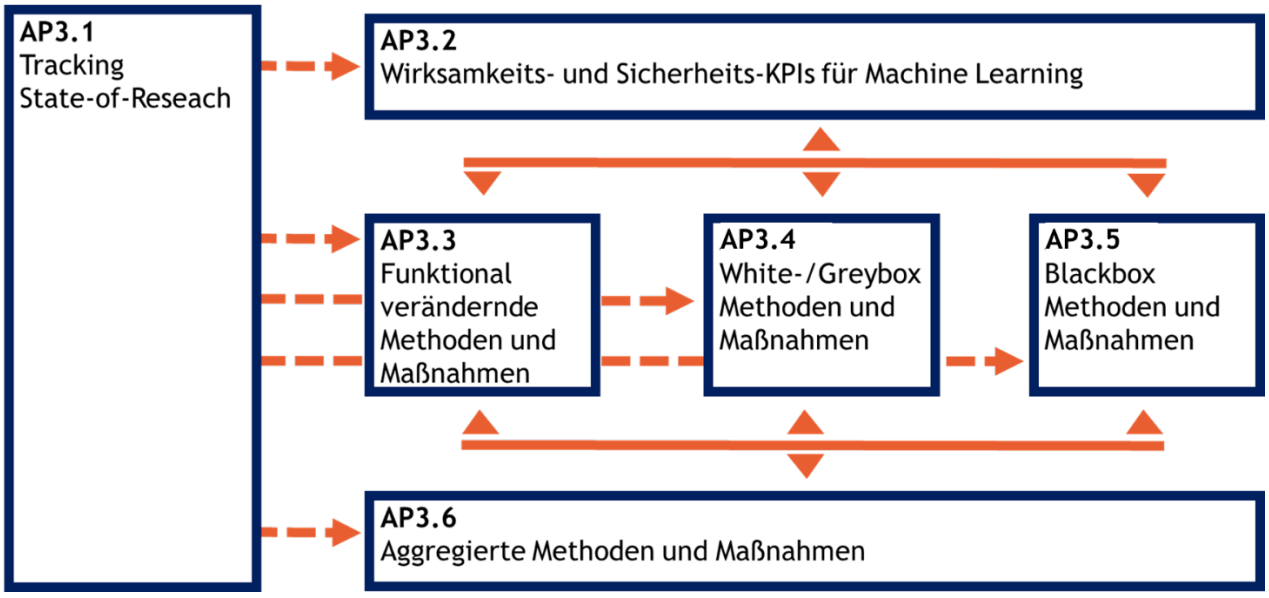


Abbildung 5.1: Struktur und Beziehungen TP3

Als zentrales Ergebnis liegt nun ein Baukasten von Mechanismen mit einer konsolidierten Dokumentation der einzelnen Mechanismen vor. Darin sind unter anderem die jeweils adressierten Sicherheitsbedenken zugeordnet, die Ergebnisse der Experimente aufgeführt, die Effektivität zur Baseline dargestellt und mögliche Evidenzen für die Sicherheitsargumentation zusammengefasst. Zu der Dokumentation gehören einheitliche Mechanismenbeschreibungen, ein Mechanismenkatalog und ein Metrikkatalog. Wesentliche Informationen zu den einzelnen Mechanismen befinden sich im Anhang (E3.3.x-E3.6.x).

Für einzelne Mechanismen wurden in Evidenzworkstreams gemeinsam mit TP4 dedizierte Argumentationsfragmente entwickelt.

Weitere Ergebnisse umfassen unter anderem folgende Punkte:

- die Publikationen einer Vielzahl von Mechanismen auf wesentlichen Konferenzen, sowie im einem Fachbuch des Springer Vrlags,
- die Ausrichtung von drei Ausgaben des Workshops SAIAD (Safe AI for Autononous Driving),
- sowie ein einheitliches Benchmark, ein Outputformat und ein Metriktool zum Vergleich von Mechanismen bzw. Modellvarianten.

Fazit & Ausblick

Es konnte gezeigt werden, dass die in TP3 betrachteten Safety Concerns grundsätzlich mit entsprechenden Mechanismen adressiert werden konnten. Die Ergebnisse - insbesondere die konsolidierte Liste von DNN-spezifischen Sicherheitsbedenken und die konsolidierte Dokumentation der Mechanismen u.a. in Hinblick auf Effektivität, Aufwand und Reife - bilden eine wichtige Entscheidungsgrundlage für eine mögliche Industrialisierung der Mechanismen. Die Mechanismen weisen dabei unterschiedliche Reifegrade auf. Bei der Verwertung in Serienprodukten müssen jeweils gezielt nach Stand der Technik und passend zu den Anforderungen der Funktion geeignete Mechanismen ausgewählt, implementiert und spezifisch bewertet werden.

5.1 AP3.1 Tracking State of Research

Ziel dieses Arbeitspaketes ist es, den wissenschaftlichen State-of-the-art der Methoden und Verfahren zur Absicherung von KI strukturiert zu erfassen und als Grundlage der weiteren Forschungsarbeiten in TP3 dem Konsortium zur Verfügung zu stellen. Dazu wurde zunächst eine Clustering - Struktur zur Erfassung des State-of-the-art erarbeitet und mit den beteiligten Partnern abgestimmt (E3.1.3). Diese Clustering Struktur bildet die Grundlage des gemeinsam erstellten State-of-the-art Berichts, der auf 93 Seiten über 400 Referenzen strukturiert beschreibt und zusammenfasst. Zu diesem Bereich haben Partner aus allen Teilprojekten von KI-Absicherung beigetragen, auch wenn sie nicht direkt in AP 3.1 beteiligt waren. Der Bericht mit dem Titel "Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety" wurde unter https://arxiv.org/abs/2104.14235 veröffentlicht und bildet das erste Kapitel des weiter unten beschriebenen Buches, das bei Springer veröffentlich werden wird.

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*

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

E3.1.2: Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety, https://arxiv.org/abs/2104.14235

Zudem wurde in AP 3.1 eine interaktive Literatur-Datenbank erstellt, in welche die Partner aus KI-Absicherung ihre Reviews aktueller wissenschaftlicher Publikationen einpflegen konnten (E3.1.3). Diese "Literatur-Repository" übernimmt die erarbeitete Clustering-Struktur, und bietet zusätzlich die Möglichkeit, Artikel zu verschlagworten. Ein öffentlicher Abzug (E3.1.4) wird im Web-Auftritt des Fraunhofer IAIS (AP-Leitung AP3.1) unter: https://jira.iais.fraunhofer.de/wiki/display/LiteratureRepositoryKIAbsicherung/Literature+Repository zur Verfügung gestellt.

Aus dem Projekt KI-Absicherung sind mehr als 50 wissenschaftliche Publikationen hervorgegangen (E3.1.5 - Austausch mit der wissenschaftlichen Community), darunter Veröffentlichungen auf hochrangigen Konferenzen wie ICCV, NeurIPS und CVPR. Auszeichnungen für die besten Beiträge wurden für die Workshops SAIAD (Safe AI for Autononous Driving) und WAISE (Workshop on AI Safety Engineering) verliehen, die jeweils auf den Konferenzen CVPR und Safecomp gehalten wurden. Der SAIAD-Workshop wurde von Mitgliedern des KI-Absicherungs-

Konsortiums vorgeschlagen, organisiert und fortlaufend durchgeführt (3 Ausgaben während der KI-Absicherung). Projektbeteiligte aus Wissenschaft und Industrie haben insgesamt 15 Kapitel für das Buchprojekt "Deep Neural Networks and Data for Automated Driving - Robustness, Uncertainty Quantification and Insights Towards Safety" mit 450 Seiten beigesteuert. Das Buch wird vom Springer Verlag publiziert ISBN 978-3031012327. Eine Liste ausgewählter Veröffentlichungen befindet sich unter https://www.ki-absicherung-projekt.de/en/publications.

5.2 AP3.2 Höherwertige Funktion KPIs für KI Funktionen

Um die Sicherheit von KI Funktionen nachweisen zu können, benötigt es belastbare quantitative Aussagen. Dazu sind Metriken notwendig, welche die relevanten Sicherheitsaspekte interpretierbar messen können. Dieses Arbeitspaket zeichnet sich vor allem durch eine hohe Kommunikation und Vernetzung mit anderen Arbeitspakten aus, da Metriken an verschiedenen Stellen im Projekt definiert und benötigt werden. Es wurden dazu relevante Metriken für dieses Projekt identifiziert und definiert. Dazu wurden verschiedene Schritt ausgeführt, die aus Abbildung 1 entnommen werden können.

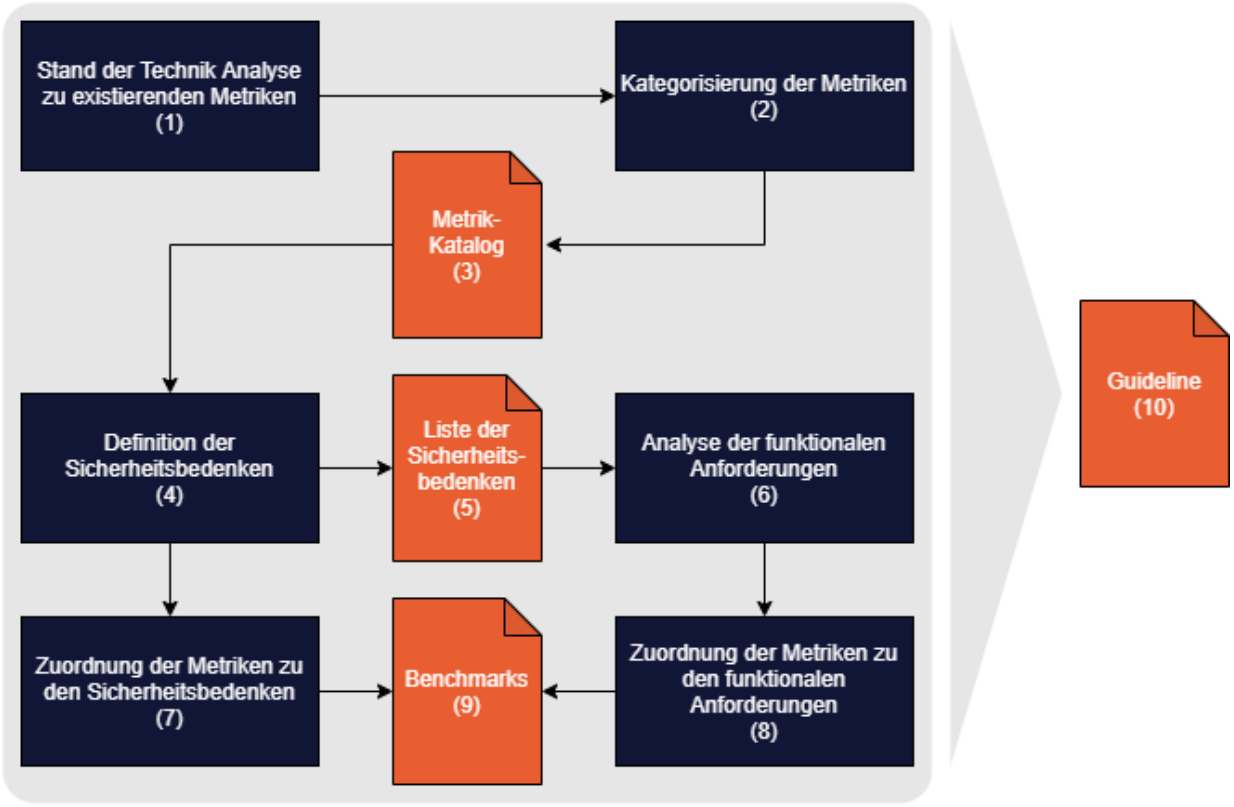


Abbildung 5.2: Übersicht zum Ablauf der Arbeiten in AP3.2

Nach einem Auftaktworkshop um das allgemeine Verständnis aufzubauen, wurde damit begonnen Metriken basierend auf dem Stand der Technik und den Mechanismen zu erheben (1). Anschließend wurde eine Kategorisierung der Metriken (2) vorgenommen, welche auf Basis eines allgemeinen Entwicklungsprozesses für KI Modelle basierte. Daraus resultierte ein Metrik-Katalog (3), welcher Informationen über die Metriken bzgl. deren Relevanz zur Sicherheitsargumentation enthält neben deren Beschreibung und Kategorisierung. Um schließlich relevante Metriken auswählen zu können, wurden zwei Ansätze gewählt. Zum einen wurden Sicherheitsbedenken (4) gegenüber tiefer neuronaler Netze mit dem Konsortium definiert, um die Eigenschaften der verwendeten Technologie darzustellen. Zum anderen wurden auf Basis der Systemanalyse die

funktionalen Anforderungen analysiert (6), um auch die Systemsicht zu berücksichtigen, wobei die Liste der Sicherheitsbedenken (5) berücksichtigt wurde, indem diese den funktionalen Anforderungen zugeordnet wurden. Anschließend konnten die Metriken zu den Sicherheitsbedenken und funktionalen Anforderungen zugeordnet (7 und 8) werden. Schließlich wurden sogenannte Benchmarks (9) definiert, mit dem Ziel eine Aussage darüber zu erhalten, ob die definierten Sicherheitsbedenken ausgeräumt sind. Die wichtigsten Erkenntnisse bei der Entwicklung von Sicherheitsmetriken wurden in einer Guideline (10) zusammengefasst.

Im speziellen wurden folgende Ergebnisse erarbeitet:

- Erstellung eines destillierten Metrikkataloges als auch einer Taxonomie (E3.2.2 und E3.2.5)
- Analyse aller funktionalen Anforderungen an die KI Funktion und Zuordnung von Metriken um deren Erfüllung messen zu können (E3.2.1)
- Definition der DNN-spezifischen Sicherheitsbedenken, welche Projektweite einen Konsens erreicht und auch in der Sicherheitsargumentation ihre Verwendung gefunden haben (E3.2.3). Siehe dazu auch die Abbildung 2.
- Analysen zur benötigten Datenbasis um belastbare Aussagen auf Basis der Metriken erhalten zu können (E3.2.4)
- Zusammenstellung der wichtigsten Erkenntnisse bei der Entwicklung von Metriken in einer Guideline (E3.2.5)

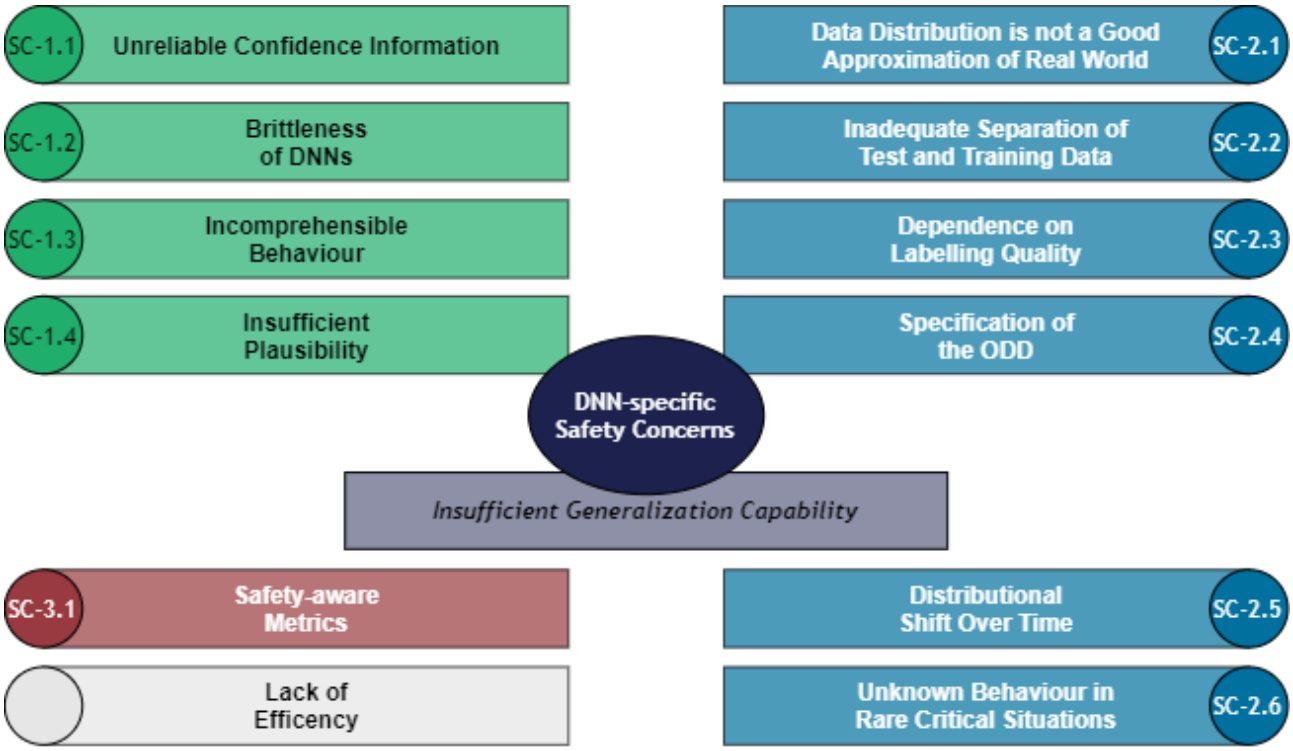


Abbildung 5.3: Definierte DNN-spezifischen Sicherheitsbedenken.

In diesem Arbeitspaket konnten ein gemeinsames Verständnis zum Thema Sicherheitsmetriken aufgebaut und exemplarische Sicherheitsmetriken definiert werden. Dabei wurde zunächst davon ausgegangen, dass die Standardmetriken wie z.B. Average Precision nicht ausreichen werden. Es hat sich aber schließlich herausgestellt, dass Standardmetriken eingesetzt werden

können, wenn Anforderungen aus der systemischen Betrachtung berücksichtig werden, wie z.B. die Definition eines relevanten Fußgängers. Abschließend betrachtet, können aus den Ergebnissen drei essentielle Bausteine für Sicherheitsmetriken aufgezeigt werden: 1) Daten-Metriken um die Aussagekraft einer Metrik einordnen zu können bzw. auch den gültigen Bereich 2) Metriken um eine Aussage über die Modell-Performanz treffen zu können und 3) Beschränkungen bei der Evaluierung die aus systemischen Sicht abgeleitet werden (z.B. Betrachtung von Fußgängern die maximal X Meter weit entfernt sind).

Um schließlich einen Industriekonsens bzgl. Sicherheitsmetriken schaffen zu können, wird empfohlen dieses Thema in einem weiteren Förderprojekt zu bearbeiten. Ausgangspunkt dazu sollte die erarbeitete Guideline zur Erstellung neuer Metriken sein.

5.3 AP3.3 Funktional verändernde Methoden & Maßnahmen

In AP3.3 werden sogenannte funktional verändernde Mechanismen entwickelt. Funktional verändernd heißt hierbei, dass auf die Funktion, welche von dem tiefen neuronalen Netzwerk (DNN) während des Trainings gelernt wird, verändert wird. Genauer gesagt wird in verschiedenen Wirkungsarten auf die Parametrierung, d.h. die Einflussgrößen, welche letztlich die Funktion darstellen, eingegriffen und somit die resultierende Funktion verändert. Entsprechend sind die in AP3.3 entwickelten Mechanismen sogenannte Whitebox-Methoden, welche weitreichenden Zugriff auf das DNN bzw. auf dessen Interna benötigen.

In AP3.3 wurden insgesamt über 25 Mechanismen, wovon insgesamt 19 im Laufe des Projektes finalisiert wurden. Durch serielle Workshops wurde der Austausch und die Kommunikation zwischen den Entwicklern gefördert und der Entwicklungsfortschritt der einzelnen Methoden im Arbeitspaket durch die Entwickler dargestellt. Die Untersuchungen der Mechanismen wurden in vielfach wissenschaftlich veröffentlicht und trugen somit aktiv der Forschungslandschaft bei.

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

In E3.3.1 wurden verschiedene Mechanismen zur bestmöglichen Datensatzselektion evaluiert. Unter anderem wurden drei Mechanismen entwickelt, welche im Rahmen eines active learnings, d.h. einer Selektion von Trainingsdaten mittels einer active learning Strategie, gezielt Daten aus einem Datensatz auswählen und darauf basierend ein Training des DNNs durchführen. Zwei Mechanismen benutzen dabei Unsicherheitswerte, welche auf verschiedene Arten ermittelt werden können, um die Auswahl der Daten durchzuführen. Der erste Mechanismus benutzt aggregierte Werte zu einer sogenannten Metaklassifizierung, um Unsicherheiten zu generieren, welche zur Inferenzzeit auf der Ausgabe des DNNs basiert.

Darüber hinaus wurde ein Mechanismus zur Kalibrierung von Unsicherheiten entwickelt und sowohl für 1D als auch für 2D Regressionsunsicherheiten untersucht. Der Mechanismus ermöglicht eine Optimierung der Unsicherheiten, sodass die Unsicherheiten zum einen der gewünschten Verteilung besser folgen und zum anderen selbst bei Domänenänderungen.

E3.3.2 Algorithmische Implementierung und Dokumentation für gezielte Datensatz-Veränderung

In E3.3.2 wurden mehrere Mechanismen entwickelt, welche sich mit adversarialen Daten beschäftigen. Dabei werden adversariale Daten unter anderem genutzt, um die Trainingsdaten des DNNs zu erweitern und so die Robustheit des DNNs mittels des veränderten Datensatzes zu

erhöhen. Ein Mechanismus hat hierbei untersucht, wie Lidar-Daten adversarial verändert werden können, d.h. wie zugrundeliegende Daten so verändert werden können, dass die resultierenden Punktwolken vom DNN nicht mehr erkannt werden können. Dies wurde durch Deformierungen von Fahrzeugen in den Daten erreicht. Die so erzeugten Daten werden dann verwendet, um wiederum die Trainingsdaten des DNNs zu erweitern. Die resultierenden neu trainierten Modelle sind in-domain genauso akkurat wie zuvor und weisen eine verbesserte Robustheit in out-of-domain Daten auf.

Zudem wurden adversariale Attacken für temporale Fusionsnetzwerke untersucht, d.h. DNNs, welche mehrere konsekutive Eingangsdaten zur Prädiktion verwenden. Hierbei wurden sowohl die Auswirkungen von adversarialen Attacken auf temporale Netzwerke untersucht, als auch eine Evaluierung der Robustheit dieser Netzwerke nach Training mit adversarialen Daten durchgeführt.

Ferner wurde in E3.3.2 eine Pipeline zur Augmentierung von Daten entwickelt. Diese nutzt synthetisch generierte Objekte, um existierende Daten bzw. Datensätze zu verändern und somit Erweiterungen der Szenen zu erzeugen. In der Pipeline werden diese Objekte, u.a. Fußgänger, Motorräder oder andere Fahrzeuge in die existierenden Bilddaten eingefügt und im Rahmen eines Postprocessings eine Angleichung der resultierenden Daten angewandt. Unter anderem können in diesem Postprocessing Licht- und Farbeffekte von Kameras in die neu erzeugten Datenpunkte integriert werden, sodass eine Reduzierung der durch die Augmentierung resultierenden Domänendifferenzen erzeugt wird.

E3.3.3 Algorithmische Implementierung und Dokumentation zur Analyse der Auswirkung von Netzwerk-Optimierung

In E3.3.3 wurden verschiedene Mechanismen zur Optimierung der DNNs erzeugt. Ein Mechanismus bestraft temporale Inkonsistenzen, d.h. starke Veränderungen in aufeinanderfolgenden Bilddaten einer Bildsequenz, mit einer neuen Loss-Funktion, sodass die trainierten DNN verbesserte temporale Eigenschaften aufweisen. Ein weiterer Mechanismus erhöht die temporale Robustheit durch die konfidenzbasierte Kombination von feature maps.

Zudem ermöglicht es ein Mechanismus mittels Wiener Filter, adversariale Attacken im Frequenzspektrum zu filtern und so als Präprozessierungsmethode das System mittels eines datenbasiert adaptierten Filters zu robustifizieren.

Weiterhin ermöglicht ein Mechanismus die Optimierung von DNNs mittels Pruning, wobei die Auswirkungen des Mechanismus auf die Robustheitseigenschaften des DNNs evaluiert wurden. In diesem Feld betrachtet ein weiterer Mechanismus sowohl die Robustheit als auch die Kompression von DNNs und optimiert beide Eigenschaften zur selben Zeit.

E3.3.4 Algorithmische Implementierung und Dokumentation zur Funktions-Robustifizierung

In E3.3.4 wurde ein Mechanismus entwickelt, welcher die Unsicherheiten für out-of-distribution (OOD) Objekte maximiert und somit die Detektierung von OOD-Objekten ermöglicht. Hierbei wird im Training die Entropie mittels einer speziellen Loss-Funktion bei OOD-Objekten maximiert, sodass die hohe Unsicherheit des DNNs nach dem Training auch auf unbekannten Objekte, welche nicht im Trainingsdatensatz enthalten waren, generalisiert. Für auftretende falsch-positive Detektionen wird ein Metaklassifizierer basierend auf aggregierten Metriken der

einzelnen Objekte genutzt, mittels dessen diese Vorhersagen entfernt werden und somit die falsch-positiv Rate reduziert werden kann.

E3.3.5 Maßnahmen-Taxonomie

In E3.3.5 wurde eine Baumstruktur basierend zunächst auf einer initialen Mechanismen-Taxonomie erzeugt. Anschließend wurde die Baumstruktur (bzw. die resultierende Taxonomie) in mehreren Iterationen an Reviews mit den Experten aus KI-Absicherung angepasst und weiterentwickelt. Der Taxonomie-Baum gibt einen Überblick über die wichtigsten Dimensionen von Mechanismen zur Absicherung von DNNs und stellt deren Relationen dar. Mit Der Taxonomie wurde ein Überblick über die verschiedenen Dimensionen der Absicherungsmechanismen geschaffen, welcher eine weiterführende Kategorisierung der Mechanismen erlaubt und eine Verwandtschaftsbeziehung von Mechanismen identifizierbar macht.

Die Mechanismen-Taxonomie wurde ebenfalls in den Methodenkatalog eingepflegt, um eine erweiterte Kategorisierung der Mechanismen zu ermöglichen. Ein partieller Ausschnitt der Taxonomie ist wie folgt zu ersehen:

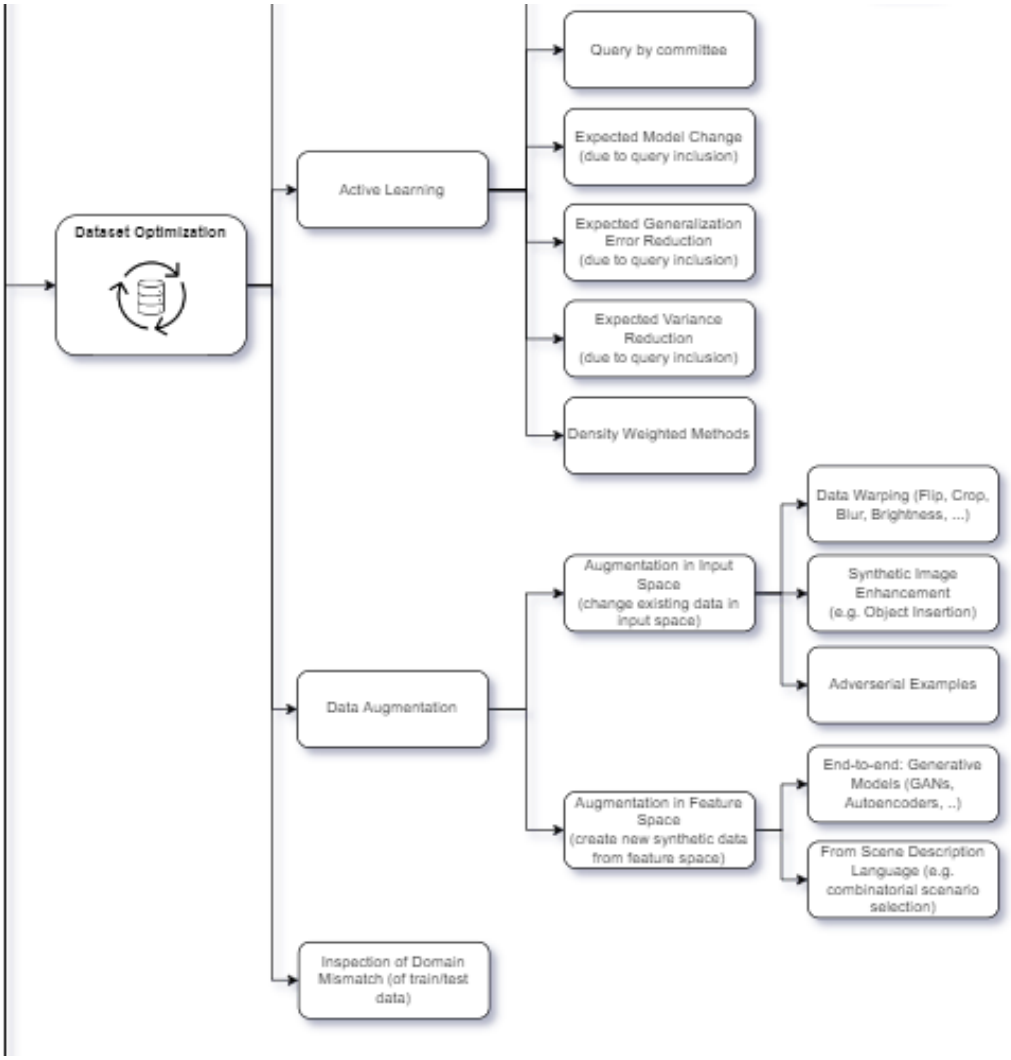


Abbildung 5.4: Partieller Ausschnitt der Taxonomie

5.4 AP3.4 White-/ Greybox-Methoden und -Maßnahmen

Das Arbeitspaket 3.4 beinhaltet die Entwicklung von introspektiven Methoden und Maßnahmen, die eine Mitigierung von Erklärbarkeitsdefiziten von Deep Neural Networks (DNN) anstreben. Introspektive Methoden und Maßnahmen verändern nicht die Eigenschaften des Modells, sondern analysieren die innere Struktur und haben daher einen evaluierenden und überwachenden Charakter. Zahlreiche Arbeiten die im Folgenden angerissen werden, stellen wertvollen Beiträge (Evidenzen) für die Sicherheitsargumentation dar und wurden auf internationalen Konferenzen und Workshops veröffentlicht.

Abbildung 5.5: Beispiel Heatmap.

Plausibilisierung der Funktionsweise des KI-Moduls (UAP3.4.1)

Die in diesem UAP betrachteten Methoden zielen darauf ab, die Funktionalität des KI-Moduls zu plausibilisieren. Zu diesem Zweck werden Heatmap-Methoden betrachtet (siehe Abb. 1), die im Bereich der erklärbaren KI bekannt sind. Heatmap-Methoden beziehen sich auf den Input des KI-Moduls und liefern Informationen über Bildregionen, die für das KI-Modul eine entscheidungsrelevante Rolle gespielt haben. Diese können visualisiert und somit vom Menschen interpretiert werden. Zu den implementierten Heatmap Methoden zählen unter anderem: LRP, SGLRP und DTD.

Zur Erkennung von false positives wurden wurden selbsterklärende Module zum DNN hinzugefügt (zur Generierung von attention heatmaps), sowie statistische Untersuchung der Co-Lokalität durchgeführt.

Weiterhin wurden generative Modelle entwickelt mit denen eine für den Menschen verständliche visuelle Darstellung der im DNN komprimierten Repräsentationen (latent space) durchgeführt werden kann. Für die generativen Modelle wurden Auto-Encoder und Generative Adversarial Networks (GAN) verwendet.

Ziel war hierbei, Auskunft über die Zuverlässigkeit der KI-Modul-Ausgabe für entsprechende Eingangsdaten zu geben (durch Unsicherheitsmodellierung) (UAP3.4.2)

Die in diesem UPA entwickelte Methoden zielen darauf ab, die Unsicherheit der vom neuronalen Netz gelieferten Ergebnisse zu modellieren. Mehrere Strategien für die Unsicherheitsmodellierungen wurden umgesetzt unter Berücksichtigung verschiedener Laufzeit

Kompromisse. Unter anderem wurde ein Multi-head Netzwerk verwendet, das auf dem gleichen Encoder operiert. Darüber hinaus wurde MC Dropout für Meta-Klassifikation zur Detektion von false positives umgesetzt, sowie mit der Intermediate Layer Variational Inference eine laufzeiteffizientere Variante von dieser implementiert. Zur robusten und echtzeitfähigen Unsicherheitsschätzung wurde ein Bayesian Neural Network mit einer Laplace Approximation durchgeführt.

Robustheitsprüfung durch Manipulation (E3.4.3)

Eine Robustheitsprüfung der Prediktion des DNNs unter gezielter Manipulation der Eingangsdaten wurde untersucht. Es wurde unter anderem ein Ansatz zur formalen Verifikation der Robustheit gegenüber verschiedene photometrische Transformationen vorgestellt. Damit ist es möglich, für einen gegebenen Input zu garantieren, dass das DNN gegenüber den untersuchten Transformationen robust ist. Durch Wegfall des Partners Visteon wurde der Umfang der durchgeführten Arbeiten in diesem UAP reduziert.

Online Anomalierkennung (E3.4.4)

Eine sicherheitsrelevante Voraussetzung für den Einsatz von Deep-Learning-Modellen in realen Anwendungen ist, dass sie in der Lage sind, kontinuierlich über ihre Limitierungen zu urteilen. Meta-Klassifizierung ist die Aufgabe, zu erkennen, ob die Vorhersage eines Modells richtig positiv oder falsch positiv ist, ohne Zugang zur Grundwahrheit zu haben. In diesem UAP wurde ein neues Verfahren zur Verbesserung der Meta Klassifikation durch Integration neuer Unsicherheitsmetriken vorgestellt. Hierzu wurden die Aufgaben der Objekterkennung sowie der semantischen Segmentierung betrachtet. In anderen Arbeiten wurde ein semi-supervised Klassifikator trainiert der das Ziel hat die Lokalität innerhalb des Merkmalsraums auszunutzen, um eine Klassifikation als normal bzw. anomal zu klassifizieren. In einer weiteren Arbeit kommt der Wasserstein k-means zum Einsatz. Die Ergebnisse zeigen eine Zunahme der relativen Häufigkeit von Falsch-Positiven mit zunehmendem Wasserstein-Abstand zum entsprechenden Clusterzentrum.

Offline Verifikation zur Netzwerkrobustheit (E3.4.5)

Bei der Offline Verifikation zur Netzwerkrobustheit (E3.4.5) finden unter anderem Untersuchungen der Verwendbarkeit von Heatmaps statt. Zum Testen der entwickelten Methoden wurden Testdaten die eine Nicht-Detektion zur Ursache haben erzeugt und auf ihre Rechenzeit optimiert. Die Korrelations-Suche zwischen Bild- und Heatmap-Domäne wurde mit Stand-der-Technik Objektdetektor-Architekturen wie Mask R-CNN und EfficientDet D1 durchgeführt. Der direkte quantitativ und qualitative Vergleich zeigt, dass aussagekräftige Einblicke in ihren Unterschieden geliefert werden können. Weitere Ansätze sind Methoden um Corner Cases und unterrepräsentierte Szenarien in den Trainingsdaten zu finden. Diese Ergebnisse können genutzt werden, um die Trainingsdaten mit zusätzlichen Daten anzureichern bzw. zu erweitern, um eine höhere Generalisierungsfähigkeit des Modells zu erreichen.

5.5 AP3.5 Blackbox-Methoden und -Maßnahmen

Das Arbeitspaket AP3.5 befasste sich mit Black-Box-Methoden zur Analyse und Verifikation neuronaler Netze. Die entwickelten Methoden und Mechanismen behandeln alle Teile der Verarbeitungskette des neuronalen Netzes - von den Eingabedaten über die Inferenz bis zu den Ausgabeergebnissen. Außerdem wurde im Rahmen von AP3.5 ein Framework für die Black-Box-Analyse neuronaler Netze unter gestörten Inputs entwickelt. Aufgrund dieser Diversität wurde

das Arbeitspaket in fünf Cluster gegliedert, die sich mit diesen unterschiedlichen Aspekten befassen.

Dieser Bericht zeigt die Ergebnisse aus dem Arbeitspaket 3.5 auf. Der Übersichtlichkeit halber werden nur die Mechanismen beschrieben, bei welchen schlüssige Ergebnisse erzielt wurden. Nicht in diesem Bericht enthalten sind Mechanismen, die beispielsweise im Laufe des Projekts aufgegeben wurden, weil sie sich während der Entwicklung nicht als Erfolg versprechend herausgestellt haben.

Cluster 3.5.1: Entwicklung von Methoden zur Bewertung der Abdeckung und Qualität der Eingabedaten

Cluster 3.5.1 befasst sich mit Black-Box-Methoden, die die Eingangsdaten neuronaler Netze betrachten. Die in diesem Cluster entwickelten Methoden bewerten deren Abdeckung des Eingaberaumes und Qualität. Umgebungseffekte tragen stark zur Gesamtleistung eines neuronalen Netzes bei Eingabedaten bei. Daher wurde ein Mechanismus entwickelt, der solche Effekte wie Nebel oder Schmutz erkennt und dem Gesamtsystem Wissen über diese Effekte zur Verfügung stellt. Anschließend können diese Informationen verwendet werden, um festzustellen, ob die KI möglicherweise Probleme mit den Eingaben hat und entsprechend zu reagieren, z. B. im Rahmen einer Strategie zum Umgang mit Unsicherheiten. Ein weiteres typisches Problem für KI-Systeme ist der Umgang mit Ausreißern in den Daten. Dieses Problem ist auch mit der Erkennung von Domain-Shifts verbunden. Gründe für solche Domain-Shifts können neue Arten von Verkehrsteilnehmern sein, z. B. Elektroroller, oder die Nutzung der Systeme in einem anderen Land, in dem sich die Straßeninfrastruktur oder das Straßenverhalten erheblich unterscheiden können. Es wurden zwei Mechanismen entwickelt, um bildweite Domänenverschiebungen zu erkennen: Ein Verfahren auf Basis von Variational Autoencodern und eines auf Basis von klassischen Autoencodern. Die Ergebnisse deuten darauf hin, dass die Autoencoder-basierte Methode für diese Aufgabe eine gute Performanz aufweist, während die VAE-basierte Methode nicht so gut abschnitt. Für lokale Domänenverschiebungen, d. h. auf Objektebene, wurde ein Mechanismus entwickelt, bei dem ebenfalls ein VAE-basierter Ansatz gewählt wurde. Für diese lokale Aufgabe erwies sich die VAE als besser geeignet und die Ergebnisse zeigen auf, dass Domänenverschiebungen auf Objektebene durch den Mechanismus erkannt werden können.

Cluster 3.5.2: Entwicklung von Methoden zur Unsicherheitsschätzung und Kalibrierung der Ausgabekonfidenzen

Um den Ergebnissen eines KI-Systems, im Falle von KI-Absicherung eines neuronalen Netzes zur Objektdetektion, vertrauen zu können, bedarf es zuverlässiger Konfidenz- und Unsicherheitsmaße. Cluster 3.5.2 enthält jene Mechanismen, welche bei der Bereitstellung genannter Maße helfen. Für Unsicherheitsmetriken wurden drei Mechanismen entwickelt, die sich auf verschiedene Arten von Unsicherheit konzentrieren. Die *Sampling-free Epistemic Uncertainty Estimation* wurde entwickelt, um einen Wrapper bereitzustellen, der das neuronale Netz kapselt und um eine Schätzung der epistemischen Unsicherheit, d. h. der Modellunsicherheit - erweitert. Der Mechanismus *Regression Calibration* hingegen bietet eine Möglichkeit, die Ausgaben für probabilistische Regressionsprobleme zu kalibrieren, was eine Schätzung der aleatorischen Unsicherheit ermöglicht, d. h. der Unsicherheit aufgrund verrauschter Eingabedaten. Auf der Seite der Kalibrierungsverfahren wurden Techniken für Bildklassifizierungs- und Objekterkennungsaufgaben entwickelt. Als Ergebnis wurde aufgezeigt,

dass solche Kalibrierungen von Objekteigenschaften wie Größe oder Position im Eingabebild abhängig sein müssen. Daher wurden Mechanismen für die multivariate Konfidenzkalibrierung, die Konfidenzkalibrierung für die Bildsegmentierung und die Nutzung von Unsicherheiten bei der Konfidenzkalibrierung entwickelt. Die entwickelten Mechanismen wurden von den beteiligten Projektpartnern als Open Source Projekt bereitgestellt, welche heruntergeladen und verwendet werden können.

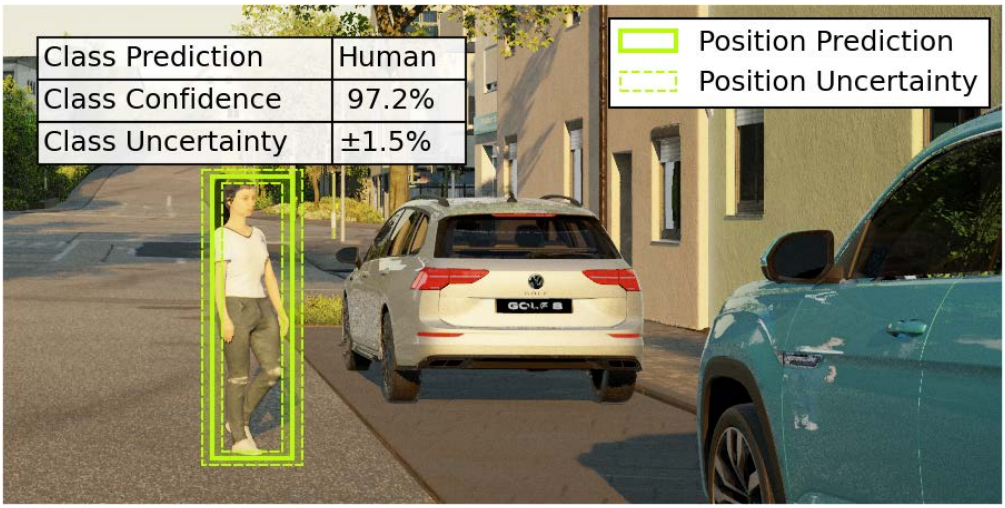


Abbildung 5.6: Schätzung der Unsicherheiten eines Neuronalen Netzes zur Fußgängerdetektion.

Cluster 3.5.3: Entwicklung von Methoden zu Adversarial Attacks und zur Robustifizierung der Netzwerke

Adversariale Beispiele, d. h. minimal gestörte Eingabedaten mit dem Ziel, ein KI-System zu einer falschen Ausgabe, wie einer Fehlklassifizierung zu provozieren, stellen ein Sicherheitsrisiko für sicherheitskritische KI-basierte Systeme dar. Daher wurden im Cluster 3.5.3 Mechanismen entwickelt, um solche Adversarial Attacks zu erkennen und zu bekämpfen. Zur Erkennung adversarialer Eingaben wurden zwei Methoden entwickelt: Die erste Methode verwendet interpretierbare Student-Netzwerke, um durch Wissensdestillation eine Selbstbeobachtung des zu untersuchenden Modells (Teacher-Modell) bereitzustellen. Auf diese Weise können sogenannte Image Shortcut Patches identifiziert werden, die das Netzwerk dazu veranlassen, eine Entscheidung zu treffen (Klassifizierung), während sie keine Informationen (Pixel) enthalten, die der jeweilig klassifizierten Klasse entsprechen. Diese Patches werden dann dahingehend bewertet, ob sie zu einer tatsächlichen falschen Vorhersage des zu testenden Netzwerks führen können. Somit hilft diese Methode aufzuzeigen, ob ein neuronales Netzwerk die richtigen Konzepte für die vorliegende Aufgabe gelernt hat und, wenn keine entsprechenden Patches gefunden werden, können Aussagen über die Robustheit des Netzwerks abgeleitet werden. Die zweite Methode verwendet ebenfalls einen Student-Teacher-Ansatz, konzentriert sich jedoch auf die Erkennung von Adversarial Attacks. Durch die Integration von zwei Student-Modellen in das zu testende Modell (Teacher) konnte gezeigt werden, dass Adversarial Attacks identifiziert und das Netzwerk robuster gegen gängige Angriffsmethoden gemacht werden kann. Ein weiterer Ansatz zur Robustifizierung neuronaler Netze, der entwickelt wurde, nutzt aufeinanderfolgende Bilder und Zeitreihendaten. Durch die Integration einer rekurrenten neuronalen Netzwerkstruktur (LSTM) in die Architektur und die Ausnutzung der Redundanz zwischen Zeitreihenbilddaten wurde die Robustheit gegenüber adversarialen Angriffen verbessert.

Cluster 3.5.4: Input-Augmentation-Techniken

Für Black-Box-Leistungsbewertungen neuronaler Netze muss die Leistung der Netze unter typischen Umgebungs- und Sensoreffekten analysiert werden. Cluster 3.5.4 stellt Methoden bereit, um Eingabedaten für neuronale Netze zu vermehren, indem Änderungen an den Eingabebildern vorgenommen werden. Die angewendeten Effekte werden dafür aus der ODD (Operation Design Domain) der Zielfunktion abgeleitet. Nach Anwendung dieser Effekte werden die Bounding Boxen, die das Netzwerk ausgibt, mit der Ausgabe bei ungestörten Bildern verglichen. Es wurde gezeigt, dass für die ursprünglichen Netzwerke ohne Robustifizierung das Verhalten der Netzwerke selbst bei leichten Eingangsstörungen nicht robust ist. Um diese Corner Cases zu identifizieren, wurde eine Methode entwickelt, die verschiedene Modalitäten von Sensoren (Kamera und Lidar) und Transformer-Architekturen verwendet. Unterschiede zwischen dem Output des Transformers und dem zweiten Sensortypen können dann darauf hindeuten, dass es sich bei einem bestimmten Datum um einen Corner Case handelt, der dann im Systemdesign entsprechend behandelt werden kann.

Cluster 3.5.5: Testing Framework

Cluster 3.5.5 stellt dem Projekt ein Testframework zur Verfügung, um neuronale Netze im Vorfeld auf ihre Robustheit hin zu analysieren. Dazu wurden neben Methoden zur Augmentierung und Störung von Eingangsbildern Metriken integriert, welche die Leistung des neuronalen Netzes unter diesen Transformationen messen. Auf diese Weise konnten die Mechanismen und Metriken anderer Projektpartner in das Framework integriert werden, um sie zu vergleichen und ihre Synergien zu nutzen. So wurden beispielsweise die Mechanismen aus Cluster 3.5.4 direkt in das Framework integriert. Das resultierende Test-Framework ist ein modulares Framework, das dem Benutzer Visualisierungen von adversarialen Angriffen und der Reaktion des Netzwerks bietet. Die Experimente können einfach mit Hydra konfiguriert werden. Experimente, die im Laufe des Projekts mit dem Framework durchgeführt wurden, halfen beispielsweise dabei, die Wirksamkeit von Robustifizierungsmethoden und eine umfassende Analyse der Projektnetzwerke unter üblichen Bildstörungen aufzuweisen.

5.6 AP3.6 Aggregierte Methoden und Maßnahmen

Das Arbeitspaket AP3.6 befasste sich mit der Entwicklung von Mechanismen im Bereich der funktionalen Redundanz, der Bewertungsredundanz und der aggregierten Mechanismen. Darüber hinaus hat dieses Arbeitspaket eine konsolidierte Dokumentation aller in TP3 entwickelten Mechanismen ermöglicht und zur Verfügung gestellt.

Dieser Bericht beschränkt sich zum Zwecke der Übersichlichkeit auf jene Arbeiten für welche schlüssige Ergebnisse erzielt wurden. Nicht in diesem Bericht enthalten sind somit jene Arbeiten, welche frühzeitig abgebrochen wurden, zum Beispiel, weil ein Partner das Konsortium verlassen hat, frühe Arbeitsergebnisse nicht aussichtsreich waren oder aufgrund von Kapazitätsproblem Aufwandsverschiebungen stattfanden.

E3.6.1 Auflösung funktionaler Redundanzen

Im Ergebnis E3.6.1 wurden Mechanismen zur Auflösung von funktionalen Redunanzen entwickelt. Das Ergebnis umfasst unter anderem die Ansätze Fused DNN und Mixture of Experts.

Im Rahmen des Fused DNN Ansatzes wurden verschiedene Methoden zur Fusion eines Netzes zur Objekt Detektion und eines Netzes zur Semantischen Segmentierung entwickelt und evaluiert.

Hierbei konnte erfolgreich gezeigt werden, dass die Fusion genutzt werden kann, um die Performanz für sicherheitsrelevante Fußgänger der Objekt Detektion zu erhöhen.

Beim Ansatz des Mixture of Experts werden mehrere Experten-Modelle mit Hilfe von semantisch unterschiedlichen Trainingsdatensätzen trainiert und ihre Prädiktionen mit Hilfe eines Gating Netzwerkes zu einer einzelnen Prädiktion fusioniert. Im Rahmen des Projektes konnte dieser Ansatz erfolgreich genutzt werden, um Bildregionen mit erhöhter Modellunsicherheit zu detektieren.

E3.6.2 Auflösung von Bewertungsredundanzen und Synergien

Im Ergebnis E3.6.1 wurden Mechanismen zur Auflösung von Bewertungsredunanzen entwickelt. Das Ergebnis umfasst den Ansatz der Robustness via Data Augmentations & Pre-Processors und die Untersuchung von Mixture of Experts Layers Embedded in CNNs.

Der Ansatz der Robustness via Data Augmentations & Pre-Processors zeigt auf, dass durch die Änderung des Trainingsprozesses mittels einer speziellen Daten Augmentierung (AugMix) und durch die Nutzung eines De-Noising Pre-Processing Moduls (VQVAE) die Robustheit eines Modells erhöht werden kann.

Beim Ansatz der Mixture of Experts Layers Embedded in CNNs wurde aufgezeigt, dass es möglich ist Experten-Modelle direkt in die Archtitektur eines CNNs zu integrieren. Weiterhin konnte gezeigt werden, dass sich die verschiedenen Experten, auch ohne die Nutzung verschiedener Datensätze, auf einzelne Teilbereiche des Eingaberaums spezialisieren können.

E3.6.3 Implementierung von aggregierten Methoden und Maßnahmen und Bewertung hinsichtlich KPIs

Im Ergebnis E3.6.3 wurden aggregierte Mechanismen entwickelt. Aufgrund mehrerer Change Requests hat sich hierbei der Fokus stark geändert. Das Ergebnis umfasst nun Arbeiten im Bereich Visual Analytics, ein real-time Ensemble via Weight Fusion und eine SOTIF Failure Mode and Effects (S-FMEA) Analyse.

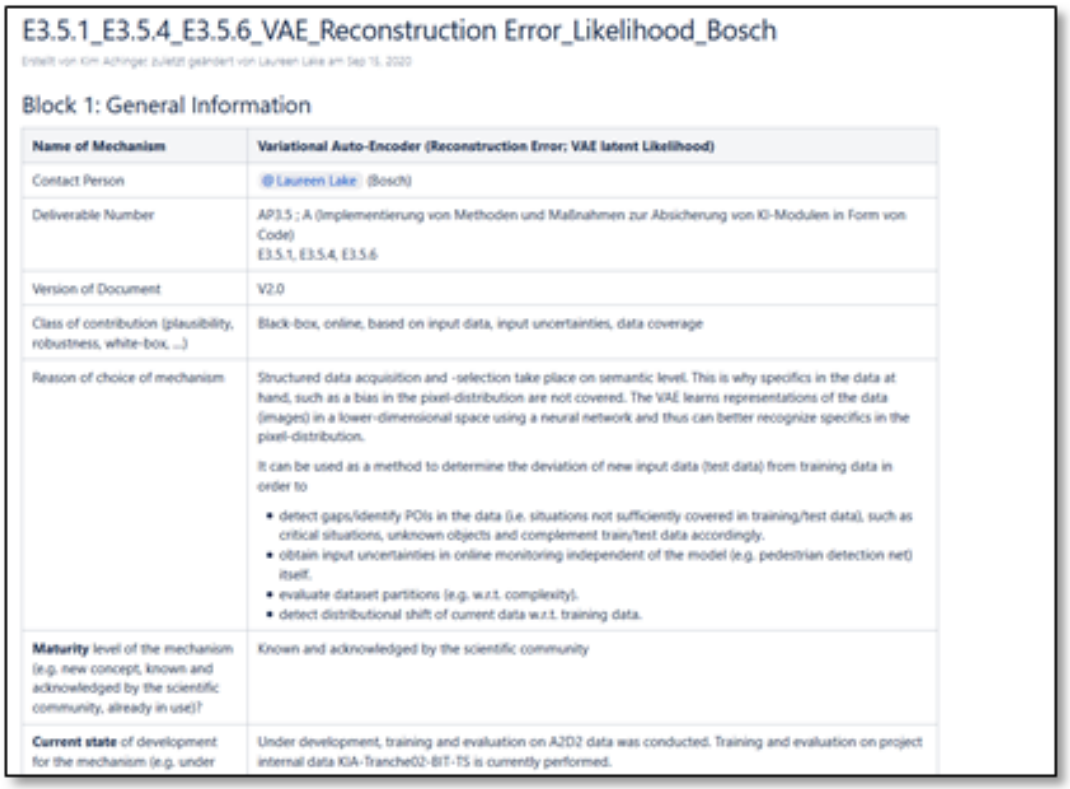
Visual Analytics ist hierbei ein Tool, welches durch iterative Anwendung in Verbindung mit menschlichem Expertenwissen systematische Schwachstellen in einem zu untersuchenden Modell finden kann. Im Rahmen von KI Absicherung, konnte die Funktionsweise erfolgreich demonstriert werden und mit Hilfe der im Projekt verfügbaren Metadaten Schwachstellen aufgedeckt werden.

Bei der Weight Fusion handelt es sich um ein Deep Ensemble welches durch die direkte numerische Fusion (gewichteter Durchschnitt) der einzelnen Modellgewichte eines Ensembles erzeugt wird. In der Folge ist dieses Ensembles analog zu einem einzelnen Modell echtzeitfähig. Dieser Mechanismus konnte erfolgreich angewandt werden und hat die Performanz gegenüber der Baseline steigern können.

Bei der S-FMEA Analyse wurden erfolgreich Triggering Conditions gemäß der SOTIF (ISO 21448) Definition gefunden. Dazu wurde das KI-Absicherung Benchmark Evaluation Tool genutzt und die Metainformationen der KI Absicherungsdaten herangezogen.

E3.6.4 Spezifikation, Implementierung, Bewertung hinsichtlich KPIs und konsolidierte Dokumentation aller Methoden und Maßnahmen aus TP3

Im Ergebnis E3.6.4 wurden alle TP3 Mechanismen in regelmäßigen Abständen mit Hilfe sogenannter Releases konsolidiert und an TP4 übergeben. Im Rahmen der Releases wurden die in E3.6.5 beschriebenen Templates verwendet und Release Requirements definiert, um die Nutzung dieser Templates und weitere Vorgaben zu kommunizieren und anschließend zu überprüfen.



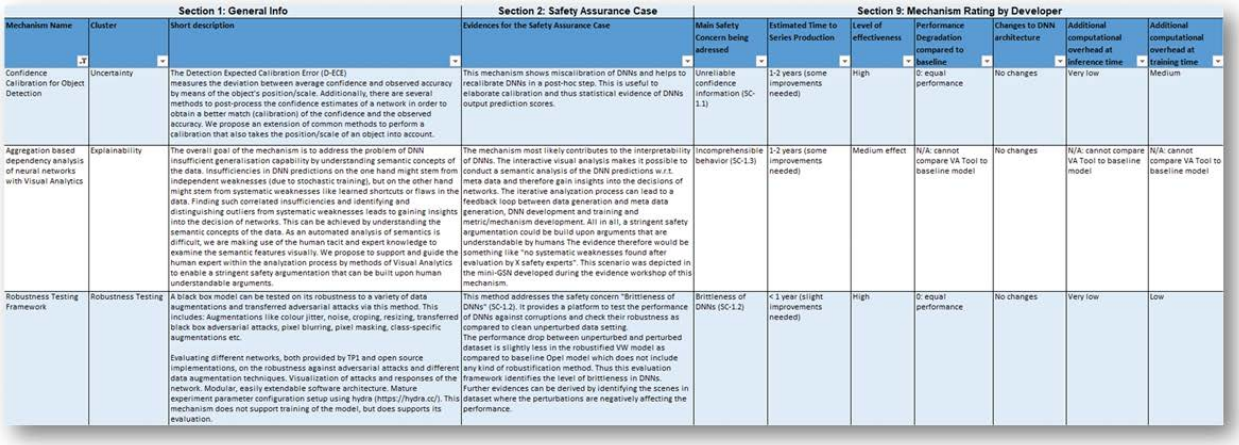
E3.5.1_E3.5.4_E3.5.6_VAE_Reconstruction Error_Likelihood_Bosch

Erstellt von Kim Achinger, zuletzt geändert von Laureen Lake am Sep 15, 2020

Block 1: General Information

Name of Mechanism	Variational Auto-Encoder (Reconstruction Error; VAE latent Likelihood)
Contact Person	@Laureen Lake (Bosch)
Deliverable Number	AP3.5 ; A (Implementierung von Methoden und Maßnahmen zur Absicherung von KI-Modulen in Form von Code) E3.5.1, E3.5.4, E3.5.6
Version of Document	V2.0
Class of contribution (plausibility, robustness, white-box, ...)	Black-box, online, based on input data, input uncertainties, data coverage
Reason of choice of mechanism	Structured data acquisition and -selection take place on semantic level. This is why specifics in the data at hand, such as a bias in the pixel-distribution are not covered. The VAE learns representations of the data (images) in a lower-dimensional space using a neural network and thus can better recognize specifics in the pixel-distribution. It can be used as a method to determine the deviation of new input data (test data) from training data in order to • detect gaps/identify POIs in the data (i.e. situations not sufficiently covered in training/test data), such as critical situations, unknown objects and complement train/test data accordingly. • obtain input uncertainties in online monitoring independent of the model (e.g. pedestrian detection net) itself. • evaluate dataset partitions (e.g. w.r.t. complexity). • detect distributional shift of current data w.r.t. training data.
Maturity level of the mechanism (e.g. new concept, known and acknowledged by the scientific community, already in use)?	Known and acknowledged by the scientific community
Current state of development for the mechanism (e.g. under	Under development, training and evaluation on A2D2 data was conducted. Training and evaluation on project internal data KIA-Tranche02-BIT-TS is currently performed.

Abbildung 5.7: Beispiel eines ausgefüllten Templates zur Mechanismenbeschreibung (gekürzter Screenshot, Mechanismenauswahl zufällig)



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

Abbildung 5.8: Beispiel des ausgefüllten Mechanismen Katalogs (gekürzter Screenshot, Mechanismenauswahl zufällig)

In Summe haben somit 6 Releases stattgefunden. Im letzten Release wurde der Fokus vor allem auf die Vollständigkeit und Ausführlichkeit der Dokumentation aller Mechanismen gelegt. Die

Durchführung eines jeden Releases ist mit Hilfe des entsprechenden Release Tables nachvollziehbar.

E3.6.5 Mechanismenbeschreibung für einheitliche Beschreibung und Bewertung der Methoden und Maßnahmen

Ergebnis E3.6.5 befasste sich mit der einheitlichen Dokumentation der TP3 Mechanismen. Zu diesem Zwecke wurden diverse Dokumentations-Templates erstellt:

Das Template zur Mechanismenbeschreibung, umfasst eine detaillierte Beschreibung eines jeden Mechanismus inkl. des experimentellen Aufbaus, der Ergebnisse, der Interpretation der Ergebnisse und die daraus abgeleiteten Evidenzen für eine Sicherheitsargumentation. Der Mechanismenkatalog fungiert als zusätzliche Informationsquelle. Hier werden ebenfalls alle Mechanismen aufgeführt, weiterführende Details gegeben und verschiedene Klassifikationen der Mechansimen aufgeführt. So gibt es hierin beispielsweise eine Einschätzung, wie weit ein Mechanismus von der Serienreife entfernt ist und wie hoch dessen Effektivität eingeschätzt wird. Für die Code Dokumentation wurden einheitliche Coding Guidelines und ein standardisiertes README Template für jedes Repo eingeführt. Abschließend wurde zusätzlich ein standardisiertes TP3 JSON Output Format definiert, welches die standardisierte Speicherung der Modeloutputs aller TP3 Mechanismen ermöglicht.

6 TP4 Gesamtheitliche KI-Absicherungs-Strategie

Wichtigste Ergebnisse und Ereignisse

Problemexposition

Eine Funktionsspezifikation wie sie in der klassischen Softwareentwicklung genutzt und anhand derer getestet wird, ist bei KI-basierten Funktionen nur noch bedingt formulierbar. Dies liegt einerseits darin begründet, dass eine KI-Funktion primär ihre "Funktion" bzw. funktionale Zusammenhänge aus den Trainingsdaten erlernt und nicht mehr durch anforderungsbasierte Funktionsbausteine beschrieben wird. Andererseits liegt es aber auch an der Größe und kombinatorischen Komplexität (Pixel & Pixelkombinationen) des Eingaberaums, der eine detaillierte Spezifikation unmöglich macht. Auch lässt sich die Funktionalität nicht wie gewohnt durch eindeutige Tests entsprechend definierter Anforderungen nachweisen. Hinzu kommt der teilweise Blackbox-Charakter von Deep Neural Networks (DNNs) und weitere Ursachen, welche zu Insuffizienzen von KI-Funktionen führen können. Deshalb bedarf es einer DNN-spezifischen Weiterentwicklung und Anpassung des Vorgehens bzgl. der Sicherheitsargumentation.

Zusammenfassende Darstellung ausgewählter Ergebnisse

Vor diesem Hintergrund wurden im Rahmen der Arbeiten in TP4:

- Eine Datenbeschreibungssprache, eine Ontologie und zahlreiche Daten-Metainformationen entwickelt, welche bei der Beschreibung des möglichen Datenraums, der genutzten Objekte in den Bildern als auch der dazugehörigen Datenabdeckung zum Einsatz kommen **(AP4.1).**
- Evaluiert, wie die Sicherheit eines DNNs zur Fußgängererkennung bis hin zu KI-Funktion argumentiert werden kann. Dabei wurden Teile wie das Restrisiko, die Definition der Operational Design Domain (ODD), die Ableitung von Sicherheitsanforderungen mit angemessenen Metriken, sowie die Definition und Einbindung von Safety Contracts für eine KI-Funktion behandelt **(AP4.2),**
- Eine neuartige Methodik für eine evidenzbasierte Sicherheitsargumentation unter Anwendung von funktionsspezifischen Absicherungsmethoden und -maßnahmen (TP3) entwickelt, welche sicherstellen soll, dass mögliche DNN-spezifische Sicherheitsbedenken ausreichend mitigiert worden sind, so dass die Sicherheitsanforderungen eingehalten werden können **(AP4.3)**.
- Neue Blackbox & GrayBox Testverfahren entwickelt und zur Anwendung gebracht, um die Performanz und Robustheit von DNNs sowie die Wirksamkeit von daten- und funktionsspezifischen Absicherungsmaßnahmen zu evaluieren **(AP4.4)**.
- Empfehlungen für verschiedene Klassen von Tests der KI-Funktion im Rahmen einer Teststrategie zusammengefasst und ein systematischer Daten-orientierter Entwicklungsprozess zur Beschreibung, Auswahl und Nutzung von Daten und Tests in Verbindung mit safety-kritischen DNNs und einer evidenzbasierten Sicherheitsargumentation weiterentwickelt und konsolidiert **(AP4.5)**.

Das Teilprojekt TP4 (gesamtheitliche Absicherungsstrategie) war charakterisiert durch seine "Klammerfunktion" am Anfang **(AP4.1)** und Ende **(AP4.2 bis AP4.5)** der Projekt-Entwicklungskette als auch von einer stark integrativen Arbeit im Hinblick auf die Nutzung, Integration und das Testen **(AP4.3 bis AP4.5)** von Ergebnissen aus vorgelagerten TPs.

Regelmäßige TP4-interne Diskussion- und Abstimmungsrunden, als auch die starke Mitwirkung aller TP4-Partner an den Evidence Workstreams (EWS) haben zur Entwicklung einer unter Projektpartnern konsolidierten Vorgehensweise hinsichtlich der Sicherheitsargumentation **(AP4.3)** und Teststrategie **(AP4.5)** für KI-basierte Funktionen geführt. Die beiden Prozesse **P1** (Beschreibungssprache & Datenanforderungsprozess) als auch **P3** (Konsolidierungsprozess Gesamtfunktion und Systemarchitektur) haben bei der Definition der benötigten und verfügbaren Daten, als auch zu der Spezifikation des Nutzungskontexts (ODD) und Funktionsspezifikation beigetragen.

Die gemeinsam entwickelte und aus der integrativen Arbeit mit anderen TPs resultierende evidenzbasierte Sicherheitsargumentation **(AP4.3)** stellt ein zentrales Ergebnis von TP4 dar. Hierbei wurde die Argumentation in Goal Structuring Notations als de-facto Standardvorgehen zur Strukturierung und Visualisierung der Sicherheitsargumentation etabliert. Die Ableitung von Evidenzen hat gezeigt, dass der Aufwand zur Entwicklung von Evidenzen nicht unterschätzt werden darf, insbesondere was die Kombination qualitativer und quantitativer Evidenzen angeht. Ordnungsmerkmal in der Argumentation waren die DNN-specific Safety Concerns. Das Vorgehen gemeinsam mit Feedbackgebern wie dem TÜV Süd (**GUA11**) reflektiert und Ausblicke für die partnerspezifische Weiterarbeit formuliert.

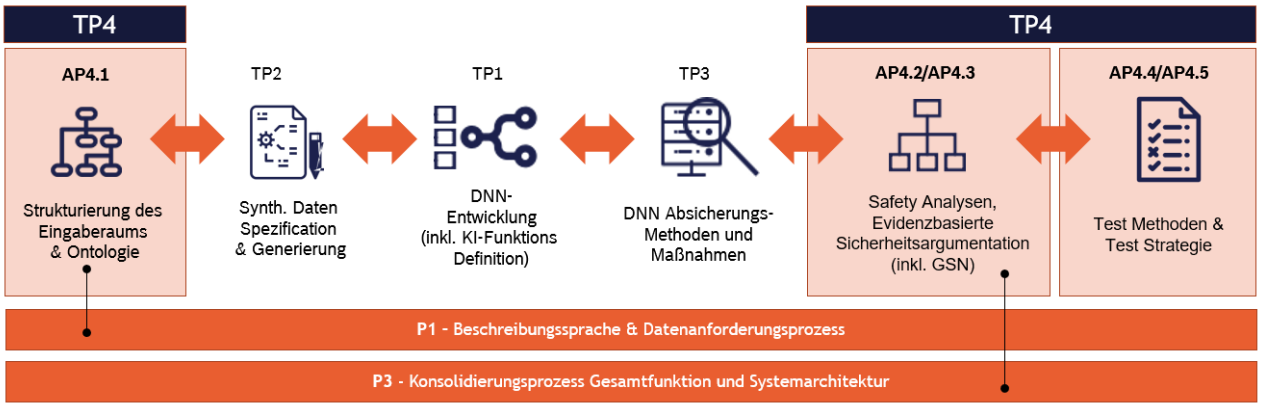


Abbildung 6.1: Positionierung der TP4-Arbeitspakete entlang Projekt-Entwicklungskette

Ausblick

Für anknüpfende Weiterarbeit werden folgende AP-spezifischen Arbeiten empfohlen:

- Auch wenn ein erster Abgleich der Ontologie aus KI Absicherung innerhalb der VDA Leitinitiative mit dem V&V Methoden Projekt und dem KI-Datentooling Projekt stattgefunden hat, so sollte in zukünftigen Projekten oder Standardisierungsaktivitäten ein Augenmerk auf die Zusammenführung der verschiedenen Ontologien und ODD-Beschreibung gelegt werden (z.B. im Rahmen der *OpenOntology* und *OpenODD* ASAM-Aktivitäten). Auch sollten die definierten Label- und Metainformationen, welche in der gemeinsamen Spezifikation **(E1.2.3 / E4.1.4b)** festgehalten wurden anderen Projekten oder der *OpenLabel* Initiative der ASAM als Input dienen **(AP4.1)**
- Die Benennung von konkreten Thresholds, sowie konkreten Zahlen als Eingangsgrößen für die Restrisikobetrachtung verbleiben für zukünftige Projekte (v.a. Serienprojekte). Zudem sollte die Wirkkettenanalyse auf Basis der (Gesamt-) Systembetrachtung erweitert werden **(AP4.2)**

- Hinsichtlich der Weiterentwicklung der evidenzbasierten Sicherheitsargumentation werden weitere partnerspezifische Verfeinerungen für konkrete Produkte / Wirkketten zur Ableitung quantitativer Evidenzen (z. B. Uncertainties), sowie Berücksichtigung des „open Context“ und der „unknown Unknowns“ empfohlen **(AP4.3)**
- Die Entwicklung eines Industriekonsenses zur Definition eines Vorgehens zur Analyse einer DNN Coverage unter Nutzung von geeigneten und definierten Metriken sollte weiter intensiviert und auch auf Realdaten ausgeweitet werden. Um die Absicherungsmethoden und Testmethoden noch stärker auf ihre Wirksamkeit und ihren Einsatzbereich hin überprüfen zu können, sollten die im Projekt KI Absicherung begonnenen vergleichende Untersuchungen zwischen den Methoden nochmals sehr systematisch untersucht werden. Insbesondere auch vor dem Hintergrund der Formulierung von möglichen Testende-Kriterien. Auch wenn im Projekt sehr gute Fortschritte bei der automatisierten Datenanforderung und Datenanalyse gemacht wurden, so waren hierbei häufig Einzellösungen oder -tools im Einsatz. Die Schaffung und Umsetzung eines gemeinsamen Frameworks zur automatisierten Auswertung der Daten basierend auf zentralen Metrik- und Datenvisualisierungsmöglichkeiten und vereinheitlichten Schnittstelle als auch zur automatischen Generierung von Datenanforderung sollte deshalb als wichtiger Startpunkt für Folgeprojekte definiert werden **(AP4.4)**.
- In Zukunft wird die weitere Sammlung von Erfahrungen (möglichst in praktischer Anwendung) und Weiterentwicklung/Industrialisierung der im Projekt entwickelten Testmethoden als Teil der Teststrategie notwendig sein. Die Weiterführung der proaktiven Kommunikation und Verankerung der Vielzahl von relevanten TP4-Ergebnissen aus KI-Absicherung in Richtung von normativen Gremien wie z.B. der ISO PAS 8800 wird auch nach Projektende eine wichtige Aufgabe bleiben **(AP4.5)**.

6.1 AP4.1 Strukturierung und Formalisierung des Eingaberaums

Neben der Funktionsspezifikation **(AP1.2)** bildet die Definition des Kontextes, in dem eine Funktion betrieben werden kann, die sogenannte Operational Design Domain (ODD) eine entscheidende Voraussetzung für die Entwicklung eines Sicherheitsnachweises **(P3.A)**. Denn nur wenn die Trainings- und Testdaten die ODD bzw. den möglichen Dateneingaberaum ausreichend abdecken, kann eine notwendige Bedingung für die Generalisierungsfähigkeit des DNNs erfüllt bzw. mit den entsprechenden Testdaten erst zuverlässig nachgewiesen werden. Grundsätzlich kann die ODD aus der Sicht von Felddaten und komplementär, aus der Sicht einer semantischen Strukturierung des Eingaberaumes betrachtet werden. Im Projekt KI Absicherung haben wir uns auf letzteres fokussiert und durch zahlreiche Experteninterviews, Literaturrecherchen, a-priori-Wissen und Experimente die Strukturierung und Formalisierung des (Daten)Eingaberaums durch Identifikation absicherungsrelevanter Kontextdimensionen (z.B. Performance Limitierende Dimensionen **(E1.2.3)**), Ausprägungen (Wertebereiche) und Kombinationen **(E4.1.2),** als auch physikalischer Zusammenhänge **(E4.1.3)** vorangetrieben und in Form einer Ontologie **(E4.1.4a)** maschinenlesbar zusammengefasst. Somit stellt eine detaillierte Beschreibung des Eingaberaums eine Grundvoraussetzung für eine Absicherung einer KI-basierten Wahrnehmungsfunktion dar. Darüber hinaus bildet die Fähigkeit der Eingaberaumbeschreibung auch eine Basis für das Design und die Analyse von Datensätzen.

Auch ist aus der Ontologie eine Beschreibungssprache **(E4.1.4a)** hervorgegangen, die es uns ermöglicht hat einerseits die mehr als 200 Assets aus dem TP2 Asset Katalog, welche im Rahmen

des Projektes bei der Datengenerierung genutzt wurden semantisch zu beschreiben und andererseits darauf basierend auch mehr als 70 relevante Metainformationen zu definieren und automatisiert bei der Datenproduktion zu erzeugen **(E4.1.4b)**. Beispiele sind der Höhenunterschiede zwischen Schulter & Fuß eines Fußgängers zur Bestimmung der Pose, der Drehungswinkel und Verdeckungsgrad eines Fußgängers relativ zur Ego-Kamera oder die Position der Sonne relativ zur Ego-Kamera. Zudem wurde die Beschreibungssprache auch dazu genutzt, um gezielte parametrierte Datenanforderungen in Form von maschinenlesbaren Datenanforderungs-JSONs an die Datenproduzenten zu richten. Für diesen Zweck wurden mehrere Softwaretools entwickelt, welche es uns ermöglicht haben mehr als sechs safety-kritische an NCAP angelehnte Verkehrs-Testszenarien mit mehr als 30.000 Bildern systematisch zu parametrieren und produzieren zu lassen **(E4.1.5)**.

Vor Beginn des Projekts war kein festgelegtes Verfahren zur detaillierten Entwicklung einer Beschreibung des Eingaberaums unter Berücksichtigung von Video-spezifischen Aspekten & Effekten vorhanden. In AP4.1 war es deshalb unser Ziel, diese Lücke zu schließen und ein solches exemplarisches Eingangsdomänenmodell für den Anwendungsfall der Fußgängererkennung in städtischen Umgebungen zu entwickeln **(E4.1.4b)**.

Die Grundlage für die im Projekt simulierten Szenarien stellen die entwickelten Grundkontexte dar **(E4.1.1)**. Hierfür haben wir mehrere repräsentative urbane Grundkontexte für die Fußgängererkennung im Kreuzungsbereich mit entsprechend typischer Infrastruktur, Wetter und Lichtanforderungen erstellt und als unsere Projekt-ODD definiert. Innerhalb dieser Grundkontexte werden die synthetischen Daten zum Training und Testen generiert. Sie umfassen statische Elemente, wie Kreuzungen, Straßen, Verkehrsinfrastruktur, Gebäude oder Vegetation. Ein Grundkontext liefert genügend Informationen, um die Erzeugung von Variationen eines bestimmten Szenarios zu ermöglichen, die eine weitere Bewertung der KI-Funktion bieten. Durch eine umfassende Analyse von Anforderungen wurden Grundkontexte mit synthetischen und realen Straßennetzwerken ausgewählt, für die Datenerstellung priorisiert und iterativ weiterentwickelt.

Abbildung 6.2: Beispiele für Grundkontexte (aus Tranche 5) als Bird's-eye view (Quelle: BIT-TS und Mackevision)

Die in **E4.1.2a** entwickelte Strategie zur Analyse des Eingaberaums setzt sich zusammen aus:

1. der Durchsicht vorhandener Datenquellen und existierenden Standards
2. Austausch in Expertenrunden für verschiedene Themenkomplexe

3. der initialen Einteilung von Elementen in Kategorien
4. der erneuten Revision durch die jeweiligen Experten
5. der Konsolidierung der Kontextelemente
6. der Beschreibung der Dimensionen durch physikalische bzw. messbare Einheiten, sofern dies möglich ist
7. der iterativen Verfeinerung basierend auf Datenanalysen, Absicherungsstrategien und Tests.

Als Ergebnis dieser Herangehensweise werden die absicherungsrelevanten Kontextdimensionen des Eingaberaums identifiziert und können in einer Ontologie repräsentiert werden.

Aufbauend auf den Grundkontexten wurde der Eingaberaum analysiert und relevante Kontextdimensionen für einen KI-Algorithmus im Bereich Fußgängererkennung im Kreuzungsbereich semantisch beschrieben. Wir haben uns entschieden, den Eingaberaum analog zum 6-Ebenen Modell aus dem Pegasus Projekt zu strukturieren. Basierend auf der entwickelten Strategie wurden die Kontextdimensionen erweitert, strukturiert und verfeinert. Die in **E4.1.2b** entstandene semantische Beschreibung des Eingaberaums wurde als kombinatorisches Modell basierend auf dem SCODE-Tool in **E4.1.2c** dargestellt. Dadurch werden kombinatorische Analysen des so aufgespannten Eingaberaums ermöglicht und Abhängigkeiten der einzelnen Kontextdimensionen untereinander aufgezeigt. Als Ergebnis liegt ein strukturierter Eingaberaum als SCODE Modell vor, welches ca. 250 absicherungsrelevante Kontextdimensionen semantisch beschreibt und in 22 Sub-Domänen (sogenannte Zwicky Boxen) klassifiziert. Jede Kontextdimension kann dabei verschiedene Werte bzw. Wertebereiche (sogenannte Alternativen) annehmen. In Summe enthält der Eingangsraum mehr als 1000 Alternativen.

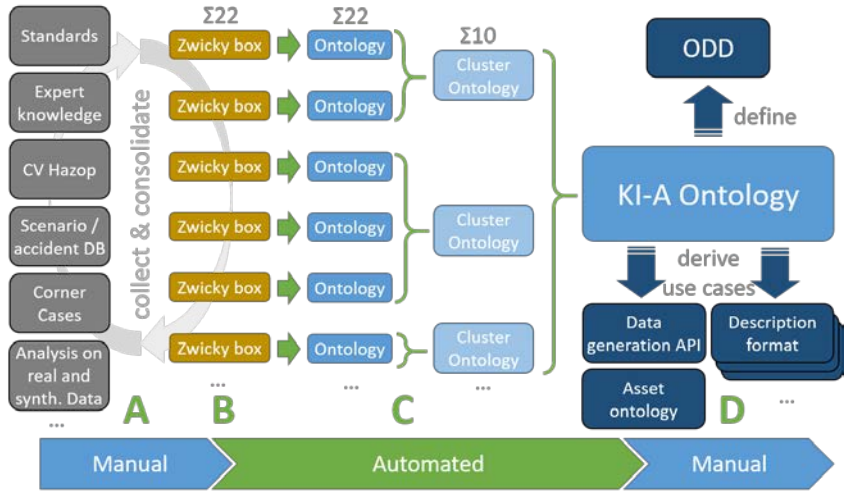


Abbildung 6.3: Übersicht der Strategie, um ein Modell des Eingaberaums zu entwickeln (A) und eine Ontologie abzuleiten (B-D), Quelle: Bosch

Das SCODE Modell wurde in **E4.1.4a** in eine Ontologie überführt und weiterentwickelt. Die in der Web Ontology Language (OWL) verfasste Beschreibung, erlaubt es weitere Verknüpfungen zwischen Kontextelementen maschinenlesbar zu formulieren. Die 22 Cluster sind für eine leichtere Handhabung entsprechend der in **E4.1.2b** herausgearbeiteten Relationen in 10 Sub-Ontologien zusammengefasst, die in einer Haupt-Ontologie zusammengeführt sind. Diese Sub-Ontologien sind thematisch nach Wetter, Lichtquellen, Sensor-Effekten, relativen

Positionsbewegungen, Fußgänger, Objekte, Straßenbedingungen, Farben Oberflächen und V&V Methoden geordnet. Die Teilontologie "V&V Methoden" bezieht sich hier auf Elemente welche im Schwesterprojekt "V&V Methoden" aktuell entstehen. Bei der Erarbeitung der Ontologie wurden insbesondere die verschiedenen Arten von Dimensionen berücksichtigt. So gibt es Dimensionen, denen sich:

a) Zahlenwerte oder Wertbereiche zuordnen lassen oder

b) durch eine Auswahl von abstrakten Alternativen beschreiben lassen

c) als Interaktionen zwischen verschiedenen Objektinstanzen darstellen lassen.

Mit Hilfe eines eigen entwickelten Tools (Assonto), konnten anhand der Ontologie alle Assets aus dem Projekt-Assetkatalog wie Fußgänger, Autos und Häuser effizient beschrieben werden.

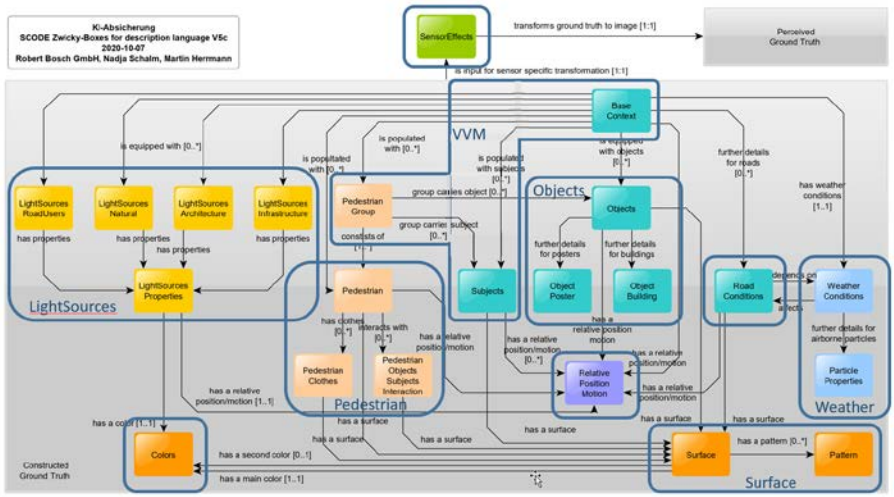


Abbildung 6.4: Übersicht der Cluster in SCODE und der Zusammenfassung zu Subontologien (blau umrahmt), Quelle: Bosch

In **E4.1.3** wurden verschiedene Ansätze daraufhin untersucht, inwiefern a-priori Wissen und physikalische Zusammenhänge genutzt werden können, um den Eingaberaum zu verkleinern bzw. neue Informationen abzuleiten. Das Wissen in der Ontologie kann dafür in der Semantic Web Rule Language (SWRL) formuliert werden. Darauf aufbauend können Wahrscheinlichkeitsketten aufgebaut werden, aus denen neues Wissen abgeleitet werden kann. Von der ursprünglichen Idee, mit diesem Vorgehen Kombinationen gänzlich auszuschließen zu können, wurde Abstand genommen, da sich für fast alle vermeintlich unmöglichen Kombinationen Gegenbeispiele finden ließen. Exemplarisch wurde eine Pipeline implementiert, die aus Inferenzergebnissen Wissen ableitet. Ein tiefes neuronales Netzwerk wird hier auf synthetische Daten angewendet und die Ergebnisse werden mit den Metadaten in eine Ontologie zusammengestellt. Regeln, die auf a priori-Kenntnissen basieren, werden verwendet, um neue Eigenschaften abzuleiten. Die angereicherten Daten werden angewendet, um Unzulänglichkeiten der Erkennungsergebnisse zu analysieren.

Um den verschiedenen Anwendungsfällen gerecht zu werden, wurden in **E4.1.4b** abhängig von den Anforderungen verschiedene Beschreibungsformate entwickelt. Um gezielt Datenanforderungen formulieren zu können, wurde das Data Specification and Description Format (DSDF) spezifiziert. Dieses ist ein JSON-basiertes Format, das auf den Elementen der

Ontologie aufbaut, um die Anforderung von Daten aus der Simulation zu formulieren. Es ist so konzipiert, dass es einen einzelnen Zeitpunkt darstellt und sich auf die Perspektive einer ausgewählten Kamera konzentriert. Die Action Description Language erweitert das DSDF um eine Beschreibung von Fußgänger- und Fahrzeugaktionen und erlaubt zeitliche Entwicklungen zu beschreiben. Die Ergebnisse und Erkenntnisse wurden in den OpenScenario Standardisierungsprozess eingebracht. Ein drittes Meta-Annotations-Format mit mehr als 50 Beschreibungseigenschaften dient der automatisierten Beschreibung der einzelnen Bilder und dort sichtbaren Szenen. Es wird für die Analyse und das Filtern von Datensätzen und Dateneigenschaften genutzt, aber auch um Korrelationen zwischen der Leistung der KI-Methoden und den Metadaten zu identifizieren und zu analysieren.

Abbildung 6.5: Höhenunterschied zwischen Schulter & Fuß

Abbildung 6.6: Drehungswinkel des Fußgängers zur Ego-Kamera

Abbildung 6.7: Fußgänger Verdeckungsgrad

Mehrere im Rahmen von **E4.1.5** entwickelte Datenanforderungstools haben es ermöglicht gezielte parametrierte Datenanforderungen in Form von maschinenlesbaren Datenanforderungs-JSONs an die Datenproduzenten zu richten. Mit ihrer Hilfe konnte der systematische Datenanforderungsprozess stark automatisiert bzw. konkretisiert werden. So konnten mit Hilfe der Tools Objekte aus dem Asset Katalog innerhalb einer Szene platziert und systematisch variiert werden. Insbesondere die konkrete, automatisierte und effiziente Anforderung von parametrierbaren safety-kritischen Szenarien hat es ermöglicht, dass mehr als sechs safety-kritische an NCAP angelehnte Verkehrs-Testszenarien mit mehr als 30.000 Bildern für das Testen der DNN Qualität zur Verfügung standen.

Auch wenn ein erster Abgleich der Ontologie innerhalb der VDA Leitinitiative mit dem V&V Methoden Projekt und dem KI-Datentooling Projekt stattgefunden hat und Anknüpfungspunkte vereinbart wurden, so sollte in Zukunft sichergestellt werden, dass diese Ontologie für eine Fußgängerdetektion im urbanen Kreuzungsbereich mit den Ontologien der anderen Projekte noch stärker zusammengeführt wird. Hierfür wäre es denkbar, dass die vorhandene Ontologie für eine Weiternutzung Open Source gestellt wird bzw. als exemplarisches Beispiel für die Kommunikation in Standardisierungsaktvitäten genutzt wird. Zudem sollte verstärkt eine Verlinkung der Ontologie und der ODD, wie sie im Rahmen des Projekts KI Absicherung

vorangetrieben wurde, stattfinden. Hierfür eignen sich ggf. die ASAM-Aktivitäten rund um *OpenOntology* und *OpenODD*. Des Weiteren wurden die definierten Label- und Metainformationen und das dazugehörige JSON-Format in einer gemeinsamen Spezifikation **(E1.2.3 & E4.1.4v)** festgehalten und mit dem Projekt KI Datentooling ausgetauscht, damit diese in Zukunft dort weitergenutzt und erweitert werden können. Auch hierfür ist ein Austausch der Spezifikation mit der ASAM *OpenLabel* sinnvoll.

6.2 AP4.2 Safety contracts, Restrisikobewertung & Gesamtstruktur der Argumentation

Im Rahmen von AP4.2 haben wir die Ableitung der Sicherheitsanforderungen für die KI-Funktion des Projekts KI-Absicherung entwickelt. Dabei wurde nicht nur der eigentliche Ansatz zur Erhebung dieser Anforderungen betrachtet, sondern auch die unterstützenden Aktivitäten zu deren Ableitung bearbeitet. Insbesondere wurden in der Anfangsphase des Projekts Anforderungen an die zu erarbeitende Sicherheits-Argumentation auf einer abstrakten Ebene definiert. Dabei standen Anforderungen hinsichtlich der Eigenschaften neuronaler Netze und des Maschinellen Lernens im Zentrum der Aktivitäten.

Ein weiteres frühes Ergebnis in AP4.2 ist eine Guideline für die Sicherheits-Argumentation in TP4. Diese wurde in Form eines exemplarischen GSN-Graphen erarbeitet. Dabei wurden beispielhaft Methoden und Maßnahmen für die Sicherheit einer KI-Funktion in den verschiedenen Ebenen des Entwicklungszyklus eingebaut. Die Grundideen dieser Guideline wurden in vielen Stellen des Projekts weitergeführt und an den konkreten Use-Case angepasst.

Das zentrale Werkzeug für die Erstellung der Argumentation ist der in AP4.2 entwickelte GSN-Editor auf Basis der eclipse IDE. Der Editor führt Strukturelemente anhand der Spezifikation der GSN-Notation ein und erlaubt die gestützte Erstellung von GSN-Graphen einheitlich über das gesamte Projekt hinweg. Der Editor erlaubt ebenfalls die Erstellung von Safety-Case-Patterns und somit standardisierte GSNs zu instanziieren. Dieser Editor wurde nicht nur in AP4.2, sondern auch in den sogenannten Evidence Workstreams und in AP4.3 genutzt.

Ein Problem bei der Ermittlung eines akzeptablen Restrisikos eines autonomen Systems, das mit einer KI-Funktion entwickelt wurde, ist, dass solange so ein System nicht in seine Umwelt entlassen worden ist, es schwierig ist statistisch verlässliche Daten hinsichtlich einer möglichen Gefährdung abzuleiten. Daher wurde in AP4.2 ein mehrstufiger Prozess zur qualitativen Erfassung des Restrisikos anhand der Methoden und Maßnahmen aus KI-Absicherung gewählt. Dazu wurden zuerst die ethischen Prinzipien im Kontext der Entscheidungsfindung mittels künstlicher Intelligenz aus externen Quellen gesammelt. Danach wurden die Methoden aus TP3 und AP4.4 diesen Prinzipien zugeordnet, soweit dies möglich war. Aus dieser Zuordnung wurde eine Abdeckung der KI-Absicherungsmethoden bzw. der Beitrag der einzelnen Methoden zu einem ethischen Prinzip bestimmt. Dieser Beitrag wurde anhand der GSN-Graphen propagiert und wir haben uns so einer quantitativen Einschätzung des Restrisikos unter Einbeziehung von Expertenwissen genähert.

Für die eigentliche Ableitung der Sicherheitsziele für die KI-Funktion in sogenannten „Machine Learning Safety Requirements“ (MLSR) wurde die inzwischen weitgehend akzeptierte STPA ("Systems Theoretic Process Analysis"-Gefährdungsanalysemethode) als Analysetechnik eingesetzt. Dazu wurden die Ergebnisse des P3-Prozesses als Eingangsgrößen verwendet.

Außerdem wurde auf die Einhaltung von ISO 26262 und ISO/PAS 21448 geachtet. Konkret wurden Hazards und deren zugehörige Sicherheitsziele auf Systemebene aufgestellt. Damit wurde eine funktionale Architektur als Control-Structure entwickelt. Als „Unsafe Control Action“ wurde der für uns wichtigste Aspekt untersucht, dass die Kollisionsvermeidungskomponente kein Bremssignal sendet, obwohl sich ein Fußgänger im Detektionsbereich befindet. Die zugehörigen „Loss Scenarios“ für die KI-Funktion wurden mittels Performance Limitationen, Safety Concerns und unter Berücksichtigung einer nachgeschalteten Perzeptionskomponente konstruiert. Diese Perzeptionskomponente erfüllte nachgelagerte Funktionen wie das Tracking von Fußgängern über die Zeit und die Fusion der 2-D Boundingbox aus dem SSD mit der „Distanzmessung“. Damit wurden insgesamt neun "Loss Scenarios" und sieben MLSRs aufgestellt. Insgesamt hat sich das Team stark an den gängigen Safety-Normen orientiert. Allerdings ist die Nennung von konkreten Akzeptanzkriterien in den MLSRs in diesem Rahmen nicht möglich gewesen (siehe auch weiter oben) und muss in anderen Projekten bearbeitet werden.

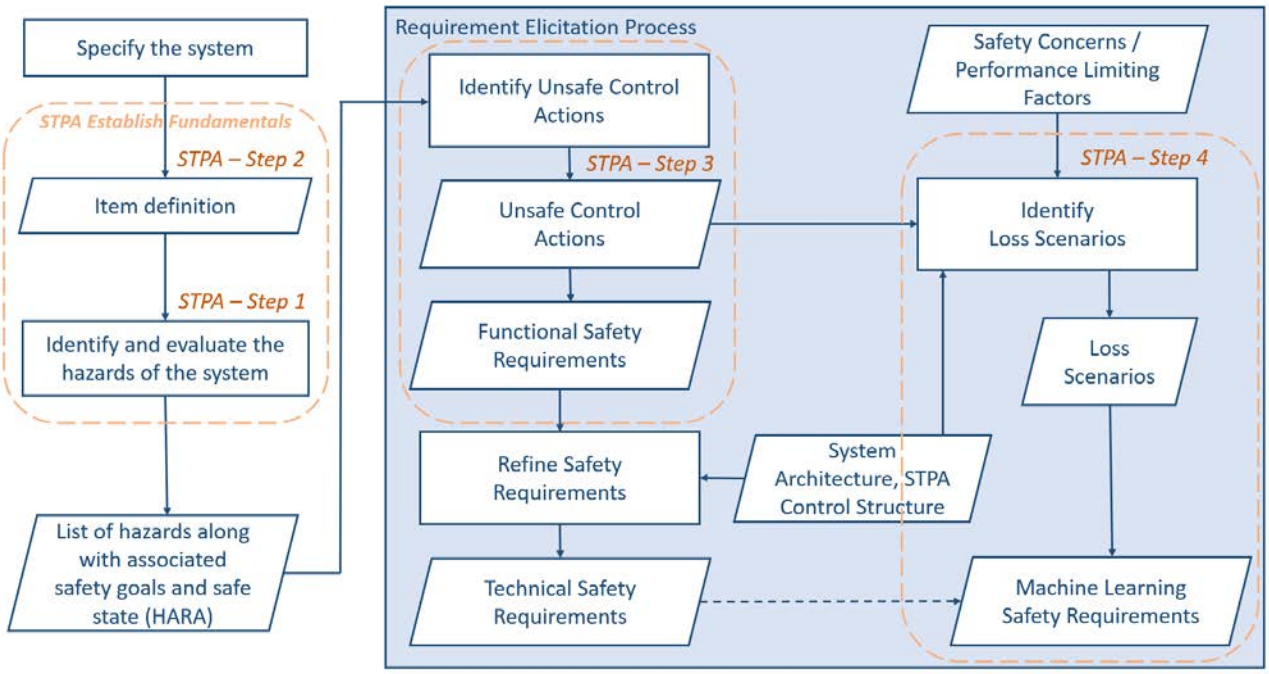


Abbildung 6.8: Vorgehen zur Ableitung der MLSR

Die Messbarkeit dieser MLSRs ist eine entscheidende Eigenschaft für die Sicherheitsargumentation. Allerdings ist in einem „Open World“-Kontext, in dem sich eine Fußgängererkennung bewegt, diese Messbarkeit mit starken Unsicherheiten behaftet. Um uns diesem Thema zu nähern haben wir die Metriken aus dem TP3 Metrik-Katalog hinsichtlich ihrer Eignung die MLSRs zu messen untersucht und das entsprechend festgehalten.

Die Argumentation für ein System mit einer Komponente, die als KI-Funktion realisiert ist, kann mittels „Safety Contracts“ modularisiert werden. Safety Contracts bestehen aus Annahmen und Garantien an das sie umgebende System. Allerdings bewegen wir uns in einem Open-World-Kontext mit sog. "long-tail" Verteilungen und können mit unserer KI-Funktion keine Garantien mit hundertprozentiger Sicherheit abgeben. Daher haben wir die Safety-Contracts mit einem subjektiven Unsicherheitswert ergänzt, der den "Belief" und die Erfüllung der Garantien repräsentieren soll. Zur Kombination von Unsicherheiten in einem GSN-Graphen wurden mit der Dempster-Shafer Evidenztheorie und den Bayesianischen Belief Netzen zwei Unsicherheitskalküle

anhand von MLSR12 (Forderung nach Robustheit gegen natürliche Pertubationen) auf ihre Eignung untersucht.

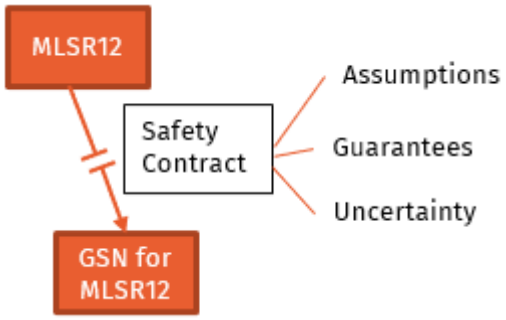


Abbildung 6.9: Safety contract

Schließlich wurde ein generelles Vorgehen zur Erhebung von Sicherheitsanforderungen für ein sicherheitskritisches System, welches eine auf maschinellem Lernen basierende Komponente zur Perzeption einsetzt, skizziert. Ein essenzielles Merkmal eines solchen Vorgehens ist die Notwendigkeit von vielen Iterationsschleifen in der Entwicklung. Dies ist nicht nur in der Entwicklung eines Neuronalen Netzes, sondern beispielsweise auch in der Auswertung von Metriken und der Applikation von KI-spezifischen Methoden und Maßnahmen erforderlich. Darüber hinaus ist die strukturierte Einbeziehung von Performance-Limiting-Faktoren, Safety-Concerns, einer Operational Design Domain (ODD) und seltenen Situationen (Corner-Cases) wichtig. Eine Sicherheits-Argumentation für eine KI-Funktion muss außerdem nach unserer Einschätzung zweistufig sein: Es muss nicht nur die unmittelbare Performance gemessen und argumentiert werden, sondern auch die Eignung der verwendeten Daten im Entwicklungsprozess aufgezeigt werden. Dabei unterstützt die Verwendung einer formalisierten Darstellung der subsymbolischen Eingaben in ein visuelles neuronales Netz die Argumentation. Die Verwendung von STPA zur Ableitung von Sicherheitsanforderungen hat sich im Projekt im Kontext einer KI-basierten Funktion bewährt.

6.3 AP4.3 Nachweisstrategie für eine „sichere“ KI-Funktion

Die zentrale Frage in KI-Absicherung ist, wie eine Lösung für die Absicherung der Perzeption von Fußgängern im Kreuzungsbereich aussehen kann. Bevor man sich mit möglichen Lösungen auseinandersetzt, muss man sich nochmal in Erinnerung rufen, um welche Problemstellung es sich bei der Absicherung einer KI-basierten Perzeption überhaupt handelt. Wir laufen bei der Absicherung von KI in verschiedene Probleme, weil es ein anderes Paradigma in der Entwicklung darstellt als bisher. Die Software wird nicht mehr explizit entwickelt, sondern das Netz wird trainiert und vieles der Verhaltensweisen steckt eher implizit in den Modellen. Die klassischen Wege zur Absicherung, die auf eine explizite Programmierung der Software abzielen, greifen hier nicht mehr. Wir können daher bei der Absicherung von KI nicht einfach auf bekannte Rezepte zurückgreifen, sondern müssen eine neue Herangehensweise finden.

Die folgende Abbildung zeigt den Ansatz, der in AP4.3 im Zusammenspiel mit anderen APs in KI Absicherung etabliert wurde. In diesem Ansatz werden Methoden und Maßnahmen so kombiniert, dass ein gesamtheitlicher Absicherungsansatz möglich ist. Diese Methoden und Maßnahmen haben das Ziel, sog. "DNN-specific Safety Concerns" zu mitigieren. Der Begriff "DNN-specific Safety Concerns" ist eine Symbiose aus den zu Beginn der Konsortialarbeit betrachteten Funktionalen Unzulänglichkeiten und den später aus Arbeiten in AP3.2 entstandenen Safety

Concerns. Zunächst muss herausgefunden werden, welche DNN-specific Safety Concerns oder Unzulänglichkeiten überhaupt mitigiert werden sollen. Dazu werden in AP4.3 die gesammelten funktionalen Unzulänglichkeiten und Safety Concerns aus AP3.2 verwendet, die nach einer Analyse der Merkmale neuronaler Netze entstanden ist. Diese Merkmale müssen messbar gemacht werden. Dazu wurden in AP3.2 Metriken eingeführt, die uns dies ermöglichen. Die einzelnen Methoden können dann, wenn die Unzulänglichkeiten und Metriken bekannt sind, in den Assurance Case einsortiert werden. Der Kern der Absicherungsargumentation besteht nun darin, den Erfolg der Bekämpfung der Unzulänglichkeiten durch den Nachweis der Wirksamkeit der einzelnen Methoden zu erbringen. Wenn also gezeigt werden kann, dass die schädliche Wirkung einer funktionalen Unzulänglichkeit durch eine oder verschiedene Methoden nachweisbar auf ein akzeptables Maß reduziert wird, dann kann man dies als Evidenz in der Argumentation formulieren. Und diese Evidenz wird dann in einer Goal Structuring Notation (GSN) visualisiert.

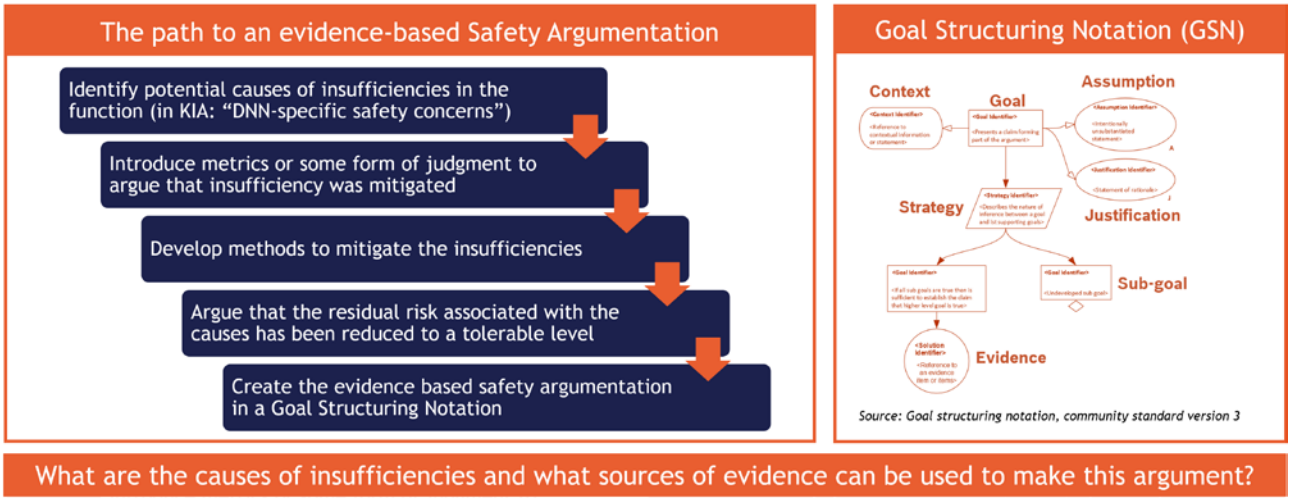


Abbildung 6.10: Vorgehen zur Entwicklung eines Assurance Case für die KI basierte Fußgängererkennung

Die Entwicklung des Assurance Case mit dem Kern der Sicherheitsargumentation beginnt mit der Festlegung der Sicherzeitsziele. Diese legen fest was gemacht werden muss, damit "Hazards" nicht auftreten. Dann werden schrittweise über die Analyse der Funktionalen Unzulänglichkeiten und Safety Concerns die Safety Requirements immer weiter verfeinert, um so zur Nachweisstrategie zu kommen Aus der Nachweisstrategie werden Evidenzen, also Beweismittel erarbeitet, die Grundbausteine für die Argumentation bilden.

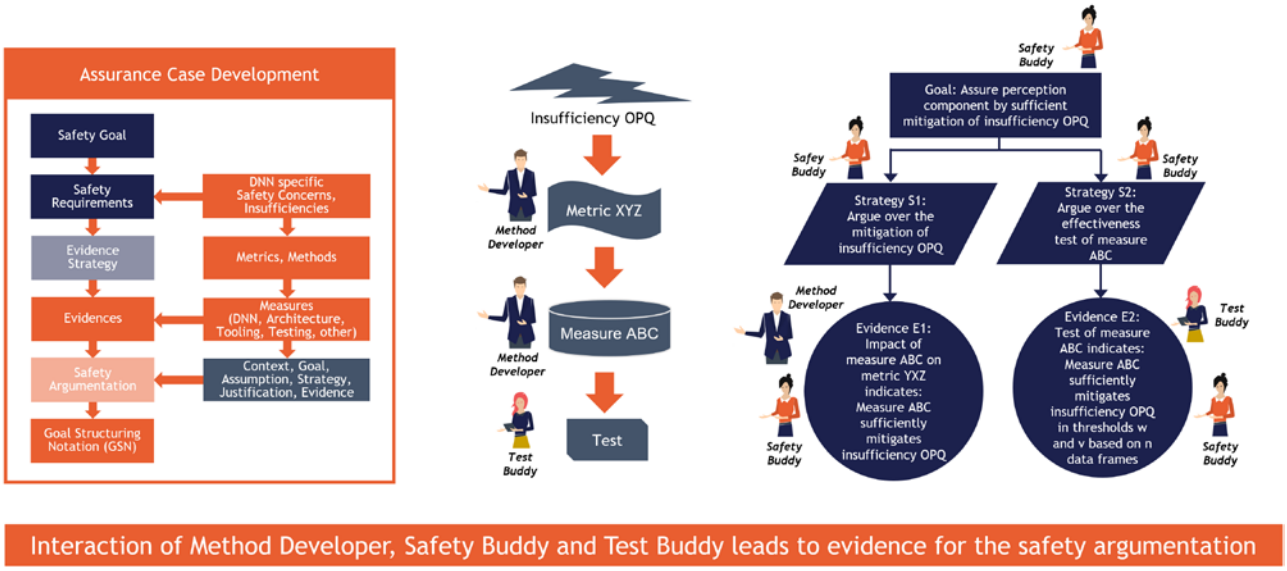


Abbildung 6.11: Evidenzbasierte Sicherheitsargumentation dargestellt in einer Goal Structuring Notation (GSN)

Um das System abzusichern, muss ein Gesamtabsicherungskonzept verfolgt werden. Dabei ist entscheidend, wie wirksam die Methoden und Maßnahmen im Hinblick auf die Sicherheitsargumentation sind und wie diese in das Gesamtkonzept einzahlen. Um dabei einen entscheidenden Schritt voranzukommen haben wir in Zusammenarbeit mit TP3 und dem Prozess P4 Evidence Workstreams eingeführt. In diesen Workstreams wird jeweils eine Methode genauer betrachtet, welche Unzulänglichkeiten und Safety Concerns dadurch mitigiert werden, und welche Evidenzen für die Sicherheitsargumentation daraus formuliert werden können. Dabei arbeiten Methodenentwickler aus TP3 mit sog. Safety Buddies aus AP4.3 zusammen. Darüber hinaus wurde auch aus AP4.4 mit sog. Test Buddies zu den Workshops beigetragen.

Die Arbeiten in AP4.3 wurden in drei Clustern organisiert. Die Aufteilung in die Cluster wurde so vorgenommen, dass in einem Cluster Teile für die Sicherheitsargumentation erstellt werden und im anderen Cluster diese Teile zum Ganzen zusammengesetzt werden. Ein weiteres Cluster ist für das Kompetenzmanagement vorgesehen. Im Cluster „Putting Things Together“ mit Lead von ASTech erfolgte die Konstruktion der gesamten Sicherheitsargumentation. In dem von ZF verantworteten Cluster "Creating Things" wurde die Konstruktion von Teilen der Sicherheitsargumentation durchgeführt. In dem von Bosch geleiteten Cluster „Expert Knowledge, Safety Buddies“ wurden Wissen und Kompetenzen für die Sicherheitsargumentation zusammengeführt und für die beiden anderen Cluster bereitgestellt.

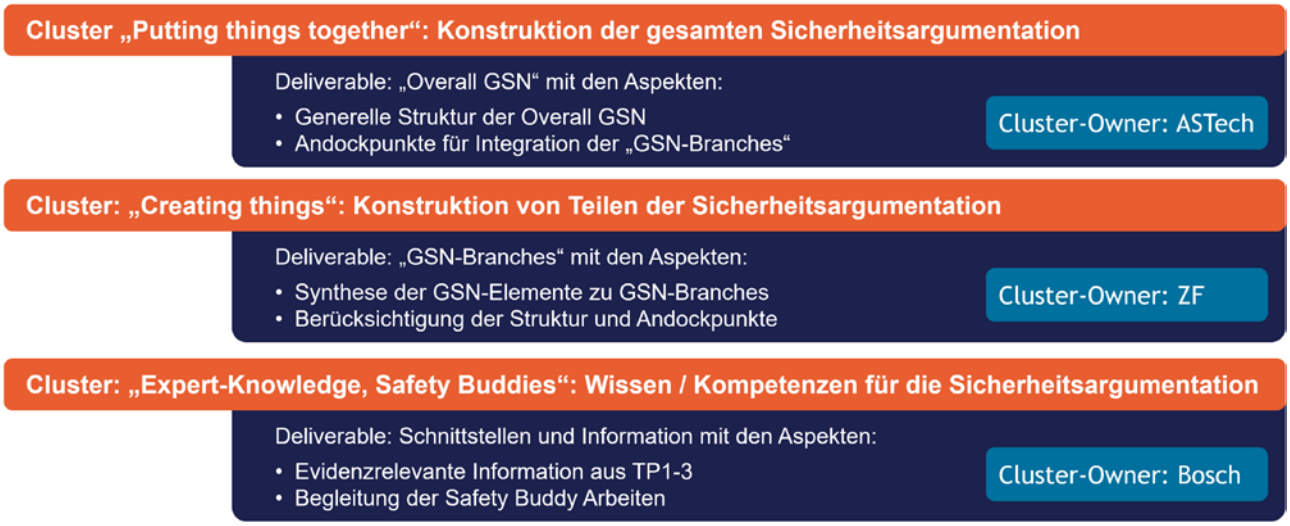


Abbildung 6.12: Umgesetzte Neu-Organisation von AP4.3 mit drei Clustern

6.4 AP4.4 Testmethoden & Bestätigung Wirksamkeit Projektergebnisse

Ein großer Vorteil in der Anwendung von tiefen neuronalen Netzen (DNN's) ist ihre Inferenzgeschwindigkeit und die Leistungsfähigkeit aus Daten automatisiert zu lernen. Leider birgt die Anwendung dieser Algorithmen jedoch Risiken, sodass diese auch eine Fehleranfälligkeit aufweisen. In diesem Arbeitspaket werden daher Testmethoden und Ansätze zur Prüfung der Wirksamkeit einzelner Absicherungsmethoden für DNN's aus TP3 erarbeitet. Hierbei stellen die mangelnde Interpretierbarkeit (sog. „black boxes“) und hohe Komplexität der DNN's, sowie die gleichermaßen hohe Dimensionalität und damit verbundene Variabilität des Eingaberaums Herausforderungen dar. Diesen wird mit verschiedenen Ansätzen begegnet, die zum einen Abdeckung, Leistungsvermögen und Extrapolationsfähigkeit von Methoden und Netzen im Datenraum untersuchen, aber auch strukturierte Ansätze, die auf Ebene der Neuronen formale Tests der Abdeckung durchführen. Auf technischer Seite wurden Konzepte entwickelt, um die Skalierbarkeit von Tests sowie deren Automatisierung zu gewährleisten.
Im nachfolgenden werden einzelne Ergebnisse dieser Stoßrichtungen exemplarisch vorgestellt.

E4.4.1a Hierbei sind die Testmethoden wie folgt in unterschiedliche Klassen einzuteilen. Hierzu zählen "White-Box" Testmethoden welche das Wissen über die eingesetzte und trainierte Netzwerkarchitektur benötigen und Black-Box Testmethoden, welche die eingesetzte KI-Funktion als geschlossenes nicht von innen erklärbares System betrachten. Während der Entwicklung der Testmethoden werden so genannte „Closed-Loop“ Ansätze verfolgt, indem die zu testende KI-Funktion direkt in den Prozess zur Testraumexploration eingebunden wird oder „Open-Loop“ Ansätze, in denen die Testergebnisse iterativ über eine zu definierende Samplegröße analysiert und weitere Testdaten angefordert werden können (Abbildung 6.13).

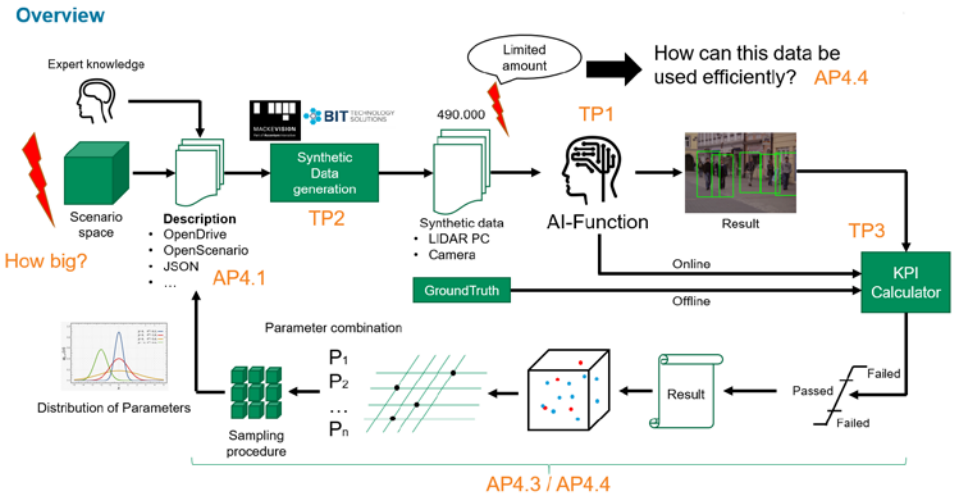


Abbildung 6.13: Beispielhaftes Konzept zur Testraumexploration

Hierbei ist zu beachten, dass die Testmethoden selbst wiederum eine lernende Komponente beim Sampling beinhalten können, um etwa die Parameter, welche die Performanz der KI-Funktion beeinflussen, zu identifizieren und die Effizienz zur Identifikation von fehlenden Testdaten zu steigern. Eine Teilmenge der Sampling Verfahren weisen anhand eines Experiments beispielhaft folgende Effizienz auf (siehe Abbildung 6.14) und es hat sich herausgestellt, dass sich mit dem Einsatz eines DNN's am effizientesten die kritischsten Eingabedaten, in diesem Fall synthetische Bilddaten, identifizieren lassen.

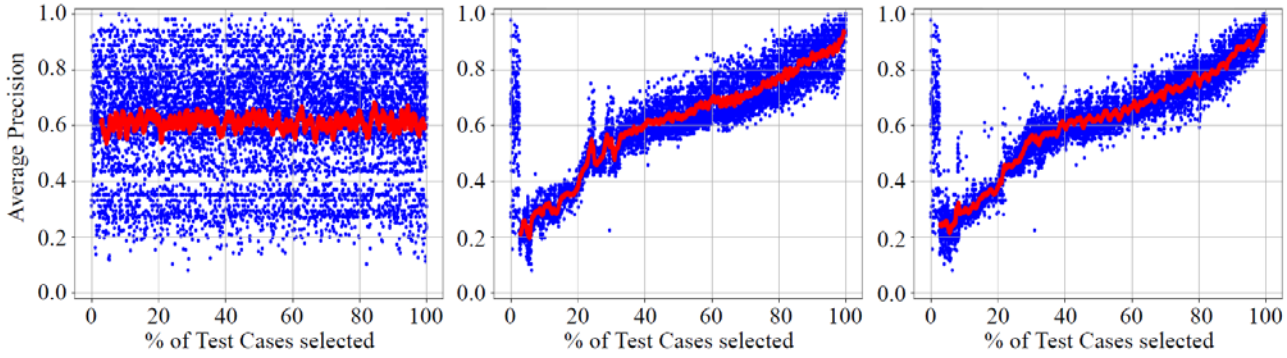


Abbildung 6.14: Effizienz diverser Samplingverfahren (random order, linear regression, neuronal network)

E4.4.3a Während obiger Ansatz als "Black-Box" Methode nur auf der Performance eines DNNs beruht, können auch innere Charakteristika durch "White-Box" Methoden wie Heatmaps oder Neuron Coverage Analysen herangezogen werden. Im Vergleich zu traditionell entwickelter Software, welche sich durch entsprechende dynamische Analysen hinsichtlich einer Anweisungs- und Entscheidungsabdeckung testen lassen, gestaltet sich der Aspekt „Code-Coverage“ bei einer KI-Funktion als schwierig. Weiter birgt die Forderung einer kompletten Coverage Analyse das Problem, dass bis heute noch keine einheitlichen Metriken definiert sind, welche es erlauben einheitlich eine Abdeckung des neuronalen Netzwerks zu untersuchen.

Um eine Abdeckung (Coverage) des neuronalen Netzwerkes untersuchen zu können, wurden Anforderungen wie folgt definiert:

- Es muss ein statisches (DNN) Modell eingesetzt werden, welches sich während der Ausführung nicht mehr ändert.
- Es muss ein Testendekriterium existieren, hier beispielsweise das Erreichen vollständiger Abdeckung.
- Die Metrik muss entsprechend eine vollständige Coverage zulassen, sodass das Testendkriterium auch prinzipiell erreichbar ist.
- Es muss bekannt sein, wie die Analyse während der Ausführung fortgeführt werden kann, z.B. der nächste Testschritt um die Coverage zu erhöhen, um nicht in einen Endpunkt zu verweilen, ohne einen vollständigen Test durchgeführt zu haben.

Weiter kann die Coverage in folgende Kategorien unterschieden werden:

- Semantische DNN-Modell Coverage, basierend auf der AP4.1 Ontologie
- Strukturelle “White-Box” Coverage

Im Rahmen der strukturellen „White-Box“ Abdeckung, wurde in diesem Zusammenhang nachgewiesen, dass sich z.B. mit wenigen Augmentierungen bereits eine hohe Abdeckung erzielen lässt. Die falsche Auswahl der Metrik birgt hierbei die Gefahr, irreführende Informationen zu liefern, welche nicht der eigentlichen Erwartungshaltung entsprechen, sodass ggf. eine hohe Abdeckung durch die ausgewählte Metrik nicht wirklich eine hinreichende Abdeckung sicherstellt.

E4.4.2 Die Definition und Auswahl geeigneter Metriken ist eine entscheidende Grundlage, um in der Wissenschaft vergleichbare Ergebnisse zu erzielen und es besteht die Notwendigkeit hierzu einen Standard zu erarbeiten. Im Grunde muss entweder eine vollständige Coverage Analyse oder eine erklärbare KI-Funktion vorliegen, um das "Safety Concern" "Incomprehensible behavior" zu mitigieren, sodass für jeden Lastfall (Eingabedaten) eine kritische Ausgabe in Bezug auf die Sicherheitsziele aus dem DNN ausgeschlossen werden kann. Weil aber beides nicht ohne weiteres nachgewiesen werden kann, sind Methoden mit systematischer Eingaberaumexplorations-fähigkeit. wie zu Beginn bei **E4.4.1a** beschrieben von entscheidender Bedeutung. Nach der Anwendung der entwickelten Testmethoden sind hierbei beispielhaft folgende Ergebnisse entstanden. Zwischen Performanz und einzelnen oder in Kombination variierten Eingabeparametern der Simulation, kann eine Korrelation nachgewiesen werden (Abbildung 6.15).

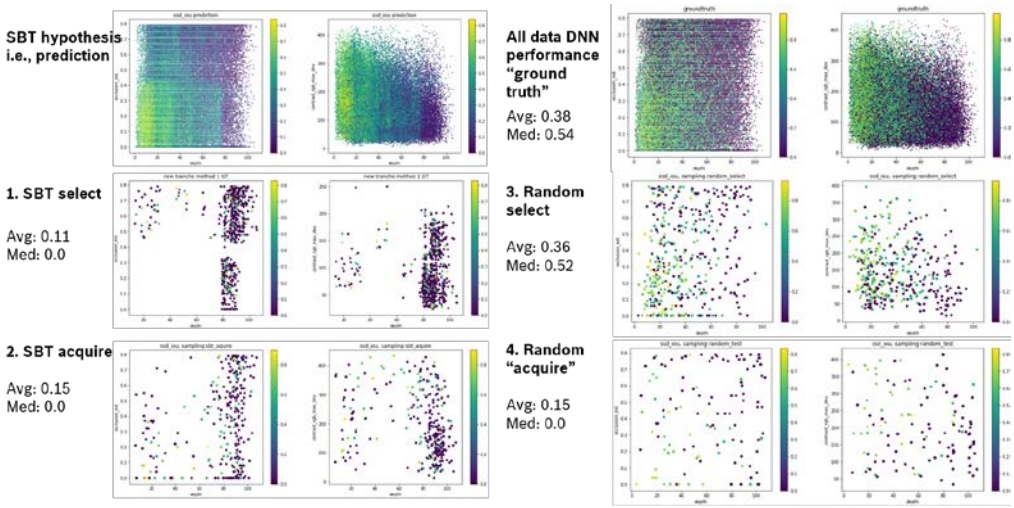


Abbildung 6.15: Exemplarische Auswertung aus der Testmethode "Search based testing"

Hierbei ist aber zu beachten, dass bei der vielfältigen Kombination von Eingabeparametern sich die Korrelation nicht immer auf nur eine Dimension zurückführen lässt. Weiter setzt dieser Korrelationsnachweis, eine gewisse Stabilität des DNN im Test voraus, welche mit gesonderten Testmethoden nachzuweisen ist.

Hierzu können Verfahren wie „Adverserial Attacks“ und Pertubationen auf den Eingabebildern angewendet. Einige Beispielbilder können aus der (Abbildung 6.16) entnommen werden.

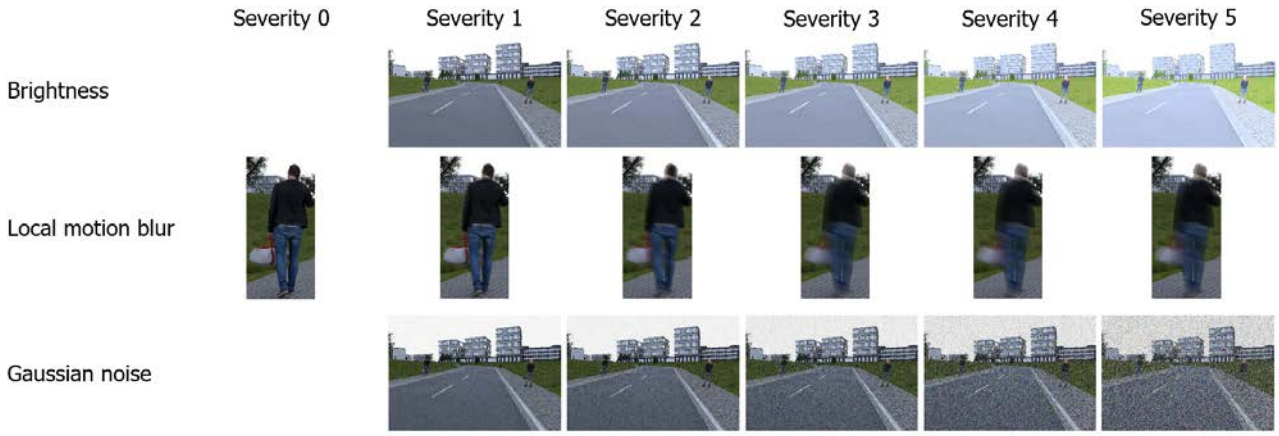
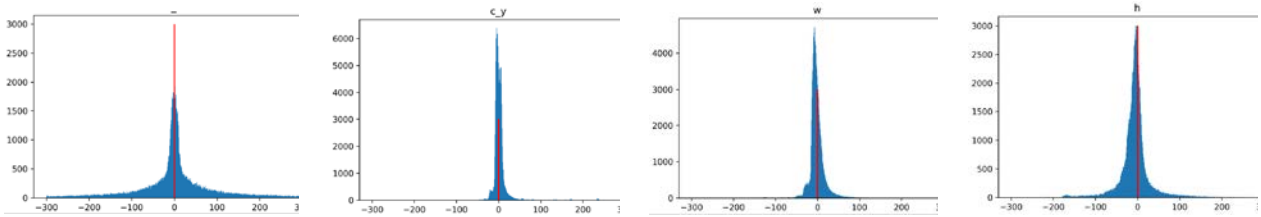


Abbildung 6.16: Beispiele für eine Pertubation auf Eingabebildern

Zusätzlich wird exemplarisch der Nachweis eines „**M**achine **L**earning **S**afety **R**equirement“ (MLSR) untersucht. Hierbei geht es darum, eine Fehlerverteilung der vom DNN prädizierten „Bounding Boxen“ zu untersuchen, welche zur weiteren Verarbeitung (bspw. mit Kalman-Filtern) einer Gauß-Verteilung genügen sollten. Dabei sind, wie in (Abbildung 6.17), folgende Ergebnisse entstanden:



a) Shape offset in x direction *b) Shape offset in y direction* *c) Width* *d) Height*

Abbildung 6.17: Fehlerdiagramm für Bounding Box Größen

E4.4.3b Eine weitere Fragestellung ergibt sich bezüglich der Extrapolierbarkeit von Tests und Ergebnissen. So ist der Eingaberaum des neuronalen Netzes, trotz ODD Spezifikation, nicht vollständig unter Kontrolle zu bringen. Dies kann unter anderem zu verschiedene Arten von Domain Shifts führen, die die Funktion des DNNs beeinträchtigen können und auf die in diesen Fällen Maßnahmen reagieren sollten. Als erste Untersuchungen wurden hierzu zunächst die Auswirkungen breit gefächerter Shifts untersucht, wie etwa der Wechsel von Tageszeiten (Tag zu Nacht) oder Wetterbedingungen (klar zu neblig). Derartige Betrachtungen können als semantische Ausführung der Augmentierungen in E4.4.2 verstanden werden und führen zu deutlich bemerkbaren Einbußen in der Performance. Neben diesen großflächigen Änderungen können aber auch lokale Variationen, wie zuvor nie gesehene "Arten" von VRUs (hierbei kann u.U. bereits ein neuer Bekleidungsstil ausreichen), eine Rolle spielen. Auch in diesen Fällen zeigten sich Reduktionen in der Leistungsfähigkeit des DNNs auf zuvor noch nicht gesehenen Fußgängermodellen.

In **E4.4.1b** wurde ein Konzept zur Skalierung der Testmethoden vorgestellt, welches exemplarisch bei den Methoden zur Robustheitsanalyse eingesetzt und getestet wurde. Durch die Skalierung ist es möglich die Methoden zur Robustheitsanalyse, sowie die Metrik-Berechnungen auf mehreren Instanzen auszuführen, sodass hierdurch eine große Zeitersparnis um den Faktor X (Anzahl der Instanzen) entsteht.

Neben der Testmethodenentwicklung ist in **E4.4.4b** ein Konzept zur Testautomatisierung vorhanden, worin die Testmethoden in unterschiedliche Automatisierungslevel unterschieden werden. Dabei bringen die Automatisierungslevel unterschiedliche Vor- und Nachteile mit sich, welche sich beispielhaft wie folgt äußern: Die in **E4.4.1a** entwickelten Testmethoden beinhalten "Closed-Loop" Ansätze, welche es erlauben den Eingaberaum voll automatisiert zu explorieren. Allerdings wird daraus eine Metrik zur Bewertung des Explorationsverhalten notwendig. Die "Open-Loop" Methoden hingegen zählen zu den semi-automatischen Methoden und erfordern hingegen nach jeder Iteration die identifizierten Testsamples zu analysieren.

Neben der Methodenentwicklung und dem Konzept zur Skalierung, ist in **E4.4.4a** ein Tooling entstanden, welches es ermöglicht die Metadaten aus dem KI-A Projekt über ein Datenbank-Backend Konzept zu verarbeiten. Hierbei können die Endanwender dieses Toolings eigene PYTHON Module implementieren und in das Computing Backend einbinden, sodass die zuvor entwickelten Methoden und Metrikberechnungen dort eingebunden werden können.

Zusammenfassend lässt sich festhalten, dass diverse Testmethoden entstanden sind, welche sich in unterschiedliche Klassen wie Black-Box, White-Box und Automatisierungsstufen einteilen lassen. Daraus sollte in Zukunft ein einheitliches Framework zum Testen der DNN basierten

Algorithmen entstehen, denn die Schnittstellen und definierte Metriken sind noch nicht so vereinheitlicht, dass eine problemlose Kopplung ermöglicht wird. Die exemplarischen Nachweise einer Wirksamkeit der einzelnen Methoden aus TP3 wurden durch die beteiligten Test-Buddies, Organisatoren und den anderen Beteiligten aus z.B. TP3 in den "Evidence Workstreams" vorangetrieben. Die integrative Arbeit zum effizienten Nachweis einer Mitigation der Safety-Concerns über alle Arbeitspakete hinweg stellt eine große Herausforderung dar und das Framework sollte in Zukunft die in einzelnen EWS betrachteten "Safety-Concerns" und die benötigten Methoden einbinden können. Neben der Entwicklung eines Frameworks, sind aber auch genauere Untersuchungen hinsichtlich der Neuron Coverage notwendig, um die diversifizierten Definitionen idealerweise zu vereinheitlichen. Darüber hinaus sollten an weiteren Methoden zur Untersuchung und Gewährleistung einer Extrapolierbarkeit geforscht werden.

6.5 AP4.5 KI-Teststrategie & KI-Testplan für Produktfreigabe

Ein Kernziel dieses Arbeitspakets ist es, aufzuzeigen wie eine Teststrategie für KI-basierte Wahrnehmungsfunktionen basierend auf den Projekterfahrungen gestaltet werden kann **(E4.5.1a+b)**. Zweck einer Teststrategie ist es, unter Berücksichtigung des Risikos des Entwicklungsgegenstands notwendige Tests, anwendbare Testmethoden und entsprechende Testende-Kriterien auszuweisen. Basierend auf einer Literaturrecherche des existierenden State-of-the-Art bzgl. genereller Anforderungen an eine Teststrategie und Anforderungen an eine Teststrategie des noch "jungen" Themenfelds von safety-relevanten daten-orientierten Verfahren **(E4.5.3a)** wurden eine Struktur für einen Teststrategie **(E4.5.1b)** entwickelt und die entsprechend eines Evaluationsprozesses ca. 15 relevantesten projektspezifischen Testmethoden & Testmaßnahmen evaluiert **(E4.5.1a)** und als Beispiele für die verschiedenen Tests bzw. Testschritte dort verortet.

Die folgende Tabelle zeigt die vier Testphasen für eine KI Funktion und die dazu nutzbaren Testaktivitäten. Für viele dieser Testaktivitäten konnten Methodenbeispiele aus dem KI-Absicherungsprojekt benannt werden.

Tabelle 6.1: Die vier Testphasen für eine KI Funktion und die dazu nutzbaren Testaktivitäten.

Datensatz-Analysen und Daten-Abdeckungsanalyse	Test der Komponenten des neuronalen Netzes	Überprüfung des Datenpools	ML-Integrations- und Qualifizierungstests
• Überprüfung der Unabhängigkeit von Test- und Trainingsdatensätzen • Analyse der Lücken in der Datenerfassung • Analysieren der Datentreue • Überprüfen der ODD-Abdeckung von Testsätzen	• Festlegung einer Strategie, die wiederholte Tests ermöglicht • Sicherstellen, dass die Ergebnisse der sicherheitsrelevanten Fälle ausreichend in den KPIs berücksichtigt werden	• Analyse der Labeling Qualität	• Analyse der statistischen Unabhängigkeit der ML Umfänge, abhängig davon normale Integrationstest oder Testaktivitäten analog zum Komponententest • Analyse von Ressourcenbeschränkungen

Datensatz-Analysen und Daten-Abdeckungsanalyse	Test der Komponenten des neuronalen Netzes	Überprüfung des Datenpools	ML-Integrations- und Qualifizierungstests
• Überprüfen, ob sicherheitsrelevante Fälle in den Testsätzen substantiell vertreten sind • Analyse der statistischen Relevanz von Testsätzen	• Durchführen von testsatzbasierten statistischen Tests • Durchführen von NN-Modell-Analysen/Überprüfungen • Durchführung von Tests auf der Grundlage von Eckfällen und Expertenwissen • Durchführen von suchbasierten Tests • Durchführung von abdeckungsgesteuerten Tests • Durchführen von Robustheitsanalysen		

Im Weiteren wurde ein ML Entwicklungs-Lebenszyklus aufgestellt, um die Testmaßnahmen besser verorten und in Beziehung setzen zu können. Dies ist ein Schlüsselelement zur Beherrschung der Qualität von ML in der Entwicklung. Der Lebenszyklus wurde im Projektverlauf mit einem übergreifenden, datenzentrierten Lebenszyklus auf Systemebene zusammengebracht und als standardisierbarer Prozess in kontinuierlicher Abstimmung durch mehrere Projektpartner systematisch formalisiert. Durch die unterschiedlichen Betrachtungsebenen beider Ansätze entsteht wesentlicher Mehrwert.

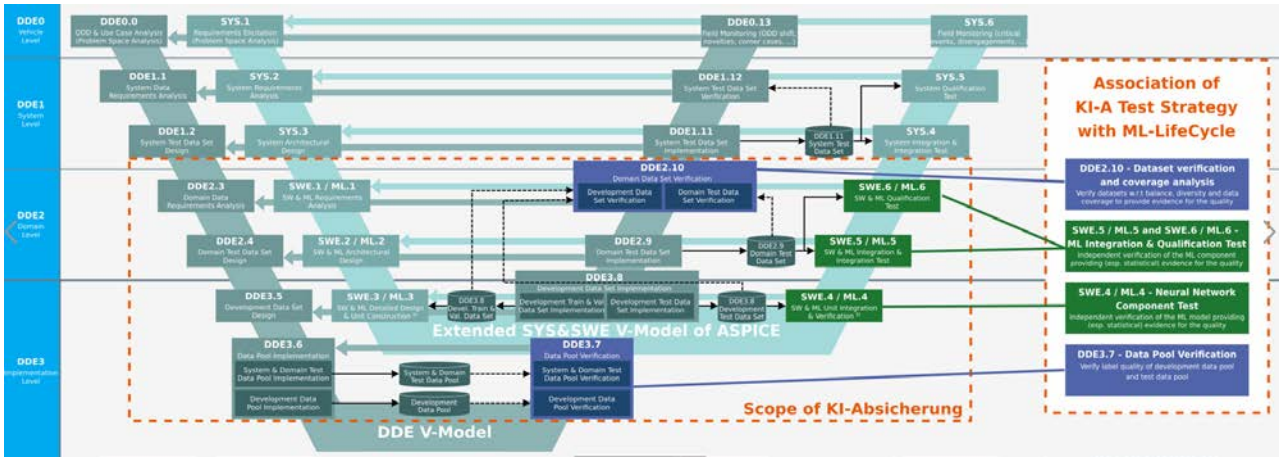


Abbildung 6.18: KI-Absicherung Teststrategie verknüpft mit KI-Entwicklungs-Lebenszyklus / datenbasiertes Entwicklungs-Prozess-Referenzmodell (Quelle: Bosch)

Um für die Teststrategie einen Katalog anwendbarer Methoden zu identifizieren, wurden die im Projekt entwickelten Testmethoden und Methoden aus der Literatur klassifiziert. Anders als zu Beginn des Projektes angenommen lassen die mit den Methoden gewonnenen Erfahrungen keine klaren Empfehlungen von Methoden zu. Dies ist ähnlich zur Situation bei Software Testmethoden allgemein. Die Teststrategie weist die Methoden daher aus und gibt einen Hinweis auf Erfahrungen aus dem Projekt. In jedem Fall ist die Kombination unterschiedlicher Methoden erforderlich. Auch wenn die Testmethoden selbst noch Gegenstand der Forschung sind, so zeigt der konsolidierte Katalog an Testverfahren dem Entwickler bzw. Tester eine gute Übersicht über den derzeitigen Stand der Technik inkl. wesentlicher Eigenschaften der Testmethoden je nach Testverfahren und gibt einen Hinweis auf Erfahrungen aus dem Projekt. Dies wiederum ist wichtig für die Auswahl von Verfahren für eine bestimmte Problemstellung in der Praxis. Das Risikomanagement erfolgt über den Assurance Case, in den die Testergebnisse als Evidenzen einfließen. Detailliertere Informationen zum datengetriebenen Lebenszyklus und der Teststrategie sind in **(E4.5.1)** angegeben.

Die Teststrategie wurde in Form eines Testplans für das in KI Absicherung genutzte Beispielnetz instanziiert und pilotiert. Der Schwerpunkt in der Testdurchführung lag dabei auf anforderungsbasierten und statistischen Tests. Die Tests weisen Defizite im Modell, den Daten und Anforderungen auf. Ein Beispiel ist, dass die Anforderung zur Erzeugung korrekter Bounding Boxen (MLSR03) Sonderfälle hoher Verdeckung nicht ausreichend berücksichtigt. Die Ergebnisse wurden in die Entwicklung zurück gespiegelt.

Abbildung 6.19: Fast vollständig verdeckter Fußgänger für den die Anforderung MLSR03 kaum erfüllbar ist, ohne dass damit notwendigerweise eine Gefährdung verbunden ist.

Einige Partner aus AP4.5 haben begonnen die Ergebnisse des Arbeitspaktes als auch wichtige Ergebnisse aus TP4 zur Sicherheitsargumentation an Standardisierungsgremien zu kommunizieren **(E4.5.3b)**. So wurde beispielsweise die Erstellung eines neuen Dokumentes zu Sicherheit von KI bei Straßenfahrzeugen unter dem Titel ISO PAS 8800 "Road Vehicles — Safety and artificial intelligence" in der ISO (International Organization for Standardization) beschlossen. Einige Partner waren an der Vorbereitung des Beschlusses beteiligt und wirken nun in der Erstellung mit.

Im Projekt "Zertifizierte KI" wurde ein Audit Katalog zur Entwicklung und Prüfung der Vertrauenswürdigkeit von KI veröffentlicht. Dieser Katalog wurde analysiert und in AP4.5

vorgestellt. Es wurde aufgezeigt, dass die Methodik von KI Absicherung dem Risiko-basierten Ansatz des Audit Katalogs genügt. Der Audit Katalog stellt eine Methodik dar, die dafür sorgt, dass KI-bedingte Risiken in sechs relevanten Dimensionen mitigiert werden. Die Methodik aus KI Absicherung kann insbes. bzgl. der Dimension Reliability als Konkretisierung des Audit Katalogs gesehen werden.

Die Arbeiten rund um den ML Entwicklungs Lebenszyklus / datenbasiertes Entwicklungs-Prozess-Referenzmodell haben inzwischen einen auch in der Praxis anwendbaren Stand erreicht. In Zukunft wird die weitere Sammlung von Erfahrungen, möglichst in praktischer Anwendung, und die weitere Industrialisierung der in der Teststrategie genannten Testmethoden notwendig sein. Die Weiterführung der proaktiven Kommunikation und Verankerung der Vielzahl von relevanten TP4-Ergebnissen aus KI-Absicherung in Richtung von Standardisierungsgremien, wie z.B. der ISO PAS 8800, wird auch nach Projektende eine wichtige Aufgabe sein.

7 Übergreifende Prozesse

7.1 P1 - Beschreibungssprachen- und Datenspezifikationsprozess

Eine Funktionsspezifikation, wie sie in der klassischen Softwareentwicklung genutzt wird, kann in Teilen zwar auch für eine KI-Funktion genutzt werden, doch im Gegensatz zur klassischen, anforderungsbasierten Top-Down Entwicklung von „per Hand“ programmierten Algorithmen erlernen ML-basierte Algorithmen, die teils nichtlinearen funktionalen Zusammenhänge direkt aus den Trainingsdaten. Auch lässt sich die Funktionalität nicht wie gewohnt durch eindeutige Tests entsprechend definierter Anforderungen nachweisen, sondern es wird anhand von repräsentativen Testdatensätzen evaluiert, ob die KI-Funktion die gewünschte statistische Performanz aufweist. Dies liegt einerseits in der hohen Anzahl an möglichen Eingabeparametern (Pixelwerte und Pixelkombinationen) und im nichtlinearen „Blackbox“ Charakter der tiefen neuronalen Netze begründet. Es kommt bei KI-basierten Funktionen dazu, dass ein wichtiger Teil der Funktionsspezifikation sich hin zu den Daten verlagert.

Während deshalb bei der KI-Funktionsspezifikation eher die übergeordneten funktionalen Eigenschaften, welche die Funktion oder das System beherrschen muss, definiert werden (können), so muss durch eine (zusätzliche) Datenspezifikation gewährleistet sein, dass Anforderungen hinsichtlich der gewünschten Dateninhalte und Datenabdeckungen definiert werden. Dabei müssen die verwendeten Daten hinreichend die unterschiedlichsten Funktionsanforderrungen abbilden können, sodass die KI-Funktion die funktionalen Eigenschaften erfüllen kann. Hieraus entsteht der Bedarf die Daten für das Training als auch für das Testen beschreiben zu können, sodass die funktionale Umsetzbarkeit durch das Training sowie das Testen der Funktionalität mittels Daten sichergestellt ist. Grundsätzlich kann der möglich Dateneingaberaum aus der Sicht von Felddaten betrachtet werden, sowie komplementär dazu auch aus der Sicht einer semantischen Strukturierung des Eingaberaums, wobei beide zu berücksichtigen sind.

Im Rahmen des Projektes KI-Absicherung und des P1-Prozesses haben wir den Fokus auf eine semantische Beschreibung und einen Prozess zur Datenspezifikation gelegt. Beide wurde im Rahmen des Projektes und durch den P1-Prozess entsprechend vorangetrieben. Die im Projekt entwickelte und auf der Ontologie basierende Beschreibungssprache soll uns dabei unterstützen:

- Die genutzten Daten selbst als auch den Kontext und insbesondere die Randbereiche der Operational Design Domain (ODD), in dem die Funktion ausgeführt wird, zu beschreiben/zu spezifizieren **(E1.2.6, P3)**
- Die Hauptdimensionen des Eingaberaums und ihrer möglichen Variationen, die einen Einfluss auf die funktionale Leistung und Sicherheit einer DNN-basierten Funktion haben (Zwicky Boxes & Ontology) zu beschreiben **(E1.2.6, E4.1.4a)**
- Bei der Schätzung der Trainings- und Testdatenabdeckung unterstützen **(E4.4.1, E4.4.2)**
- Die Datendimensionen, die variiert werden sollen, zu beschreiben, um möglichst systematisch fehlende Kombinationen durch eine nachgelagerte synthetische Sensordatenproduktion erzeugen zu können **(E4.4.x, E2.5.x)**

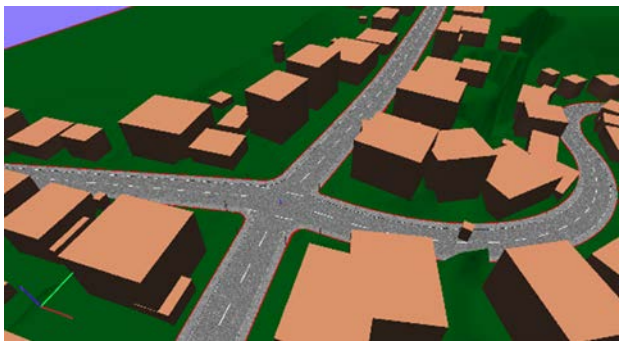
Der P1-Prozess angeleitet von Bosch und ZF hat bei der Erarbeitung dieser TP- und AP-übergreifenden Tätigkeitsfelder eine koordinierende Funktion eingenommen und Fachexperten

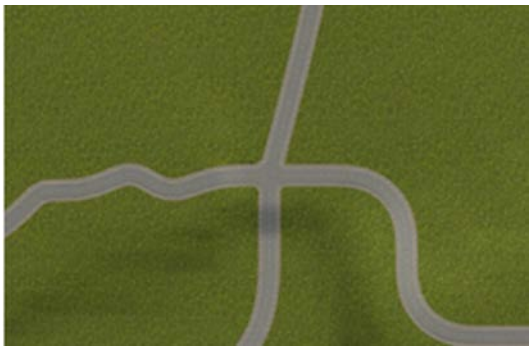
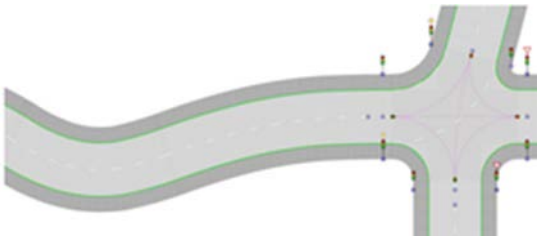
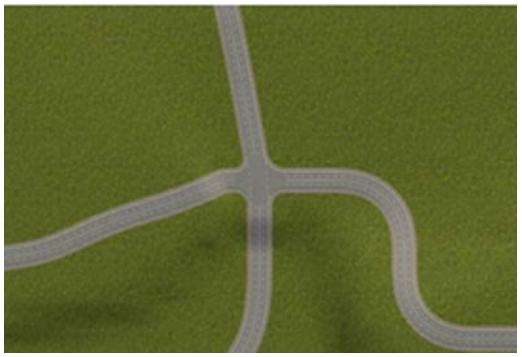
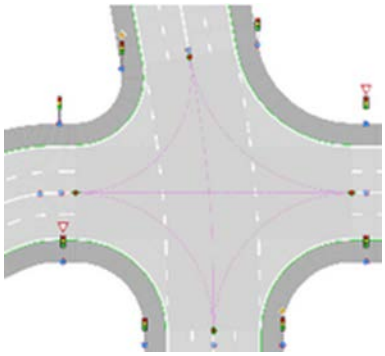
aus den Bereichen Grundkontextentwicklung **(E4.1.1)**, Ontologie **(E4.1.4a)**, Datenstrukturierung und Metadaten **(AP4.1.4b)**, KI-Funktionsentwicklung und KI-Funktionsspezifikation **(E1.2.6)**, Datenanforderungs-Management **(E2.2.6)**, Datenproduktion **(AP2.5)**, Absicherungsmaßnahmen-Entwicklung **(TP3),** als auch dem Testen von KI-Funktionen **(AP4.4, AP4.5)** zusammengebracht. Die Unternehmen BMW, Bosch, Mackevision, Opel, Qualityminds, Volkswagen, ZF (alphabetisch sortiert) haben sich als Vertreter ihrer APs am Prozess beteiligt und sich in den wöchentlichen Regelabstimmungen eingebracht.

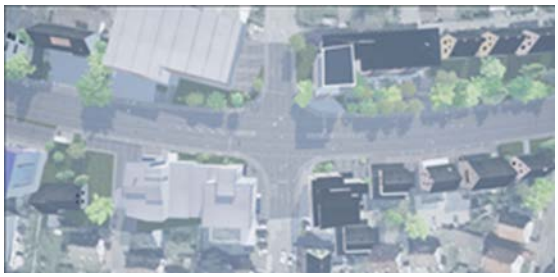
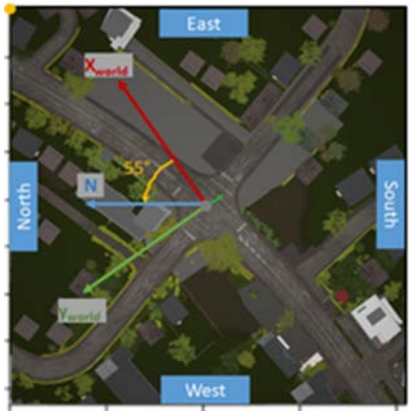
Im Rahmen der P1-Projektaktivitäten wurden maßgeblich folgende Themenfelder bearbeitet:

7.1.1 Grundkontext-Entwicklung

Der Grundkontext charakterisiert die Umgebung, in der die KI-Funktion in unserem Projekt trainiert, abgesichert und getestet werden soll. Die Entwicklung eines solchen Grundkontexts **(E4.1.1)** ist zeitaufwendig und muss als Grundlage für viele Datensequenzen dienen. Deshalb ist es wichtig, dass der Grundkontext möglichst viele Kombinationen an Dimensionen und Variationen ermöglicht. Auch soll er Corner Cases bzw. schwierige Situationen für ein DNN beinhalten. Entsprechend eines langen Kriterien-Katalogs, wurden Grundkontext Anforderungen an die Umgebung, in der eine Fußgängererkennung trainiert und abgesichert werden soll, in enger Kooperation mit **AP1.2** und **AP4.1** analysiert und entsprechend der Projektziele priorisiert. Die Eigenschaften des Grundkontextes sind in die Definition der Operational Design Domain (ODD = Gültigkeitsbereich einer KI-Funktion) geflossen bzw. wurden damit abgeglichen. Die Grundkontexte wurden von den Datenlieferanten als Grundlage für Umgebung, in der die Datensimulation stattfindet, genutzt.

Grundkontexte	Sequenzes:	
Synthetisch erzeugte urbane Kreuzung mit 2 Fahrspuren Die östliche Straße hat ein Höhenprofil (Gefälle). Die Ost- und Südstraße haben eine S-förmige Kurve (2 x 90°), die West- und Nordstraße werden mit einer einzigen Kurve generiert. Alle Straßen enden mit einer scharfen Kurve, die mit Gebäuden versehen ist, um den Himmelsbereich zu minimieren und möglichst vielfältige Hintergründe zu schaffen.	BIT-TS Sequenzen: 0070 - 0209	 

Grundkontexte	Sequenzes:	
Synthetisch erzeugte urbane Kreuzung mit 2 Fahrspuren Die Straßen haben ein komplexes Höhenprofil mit Steigungen bis zu 8 %. Die Ost- und Südstraße haben eine S-förmige Kurve (2 x 90°). Die Weststraße wird mit einem zufälligen Kurvenmuster generiert. In diesem Grundkontext enden alle Straßen mit einer scharfen Kurve, die mit Gebäuden ausgestattet ist, um die Himmelsfläche zu minimieren und möglichst vielfältige Hintergründe zu schaffen.	BIT-TS Sequenzen: 0147 - 0273	 
Synthetisch erzeugte urbane Kreuzung mit 4 Fahrspuren Die Straßen haben ein komplexes Höhenprofil mit Steigungen bis zu 8 %. Die Ost- und Südstraße haben eine s-förmige Kurve (2 x 90°). Die Weststraße wird mit einem zufälligen Kurvenmuster generiert. In diesem Grundkontext enden alle Straßen mit einer scharfen Kurve, die mit Gebäuden ausgestattet ist, um die Himmelsfläche zu minimieren und möglichst vielfältige Hintergründe zu schaffen.	BIT-TS, Sequenzen: 0301 - 0484	 

Grundkontexte	Sequenzes:	
<u>Urbane Kreuzung mit jeweils 2 und 4 Fahrspuren</u> Die Kreuzung befindet sich in Leonberg, Baden-Württemberg. Sie wurde photogrammetrisch digitalisiert. Sie wurde von Bosch als BackgroundIP dem Projekt bereitgestellt und von Mackevision entsprechend der Anforderungen im Projekt erweitert. Sie stellt die Grundlage für alle Datenlieferungen von Mackevision ab Tranche#4 dar.	Mackevision alle Sequenzen ab Tranche#4	 

7.1.2 Datenanforderungsmanagement

Im Gegensatz zur Verwendung von Realdaten, wo maßgeblich die "Natur bzw. reale Welt" vorgibt, welche Variationen in den Bildern enthalten sind, müssen bei synthetischen Daten die Funktionsentwickler und Tester beschreiben, welche Objekte, Varianz und Variationen in den Daten benötigt werden. Dies stellt für viele Beteiligte gewissermaßen eine zusätzliche Herausforderung dar, weil sie es nicht gewohnt sind normalerweise die Inhalte ihrer Daten zu beschreiben und auch weil die Datenlieferanten bzgl. der zu produzierenden Daten eine genaue Spezifikation unter Berücksichtigung der aktuellen "Machbarkeit" erwarten. Die Formulierung und Priorisierung von Anforderungen und Akzeptanzkriterien **(aller APs)** als Input sowohl für die Datenproduktion, als auch für ein Review der Requirements und der generierten Daten, wurden von den P1-Teilnehmern in wöchentlichen Abstimmungsrunden mit Vertretern aus allen Teilprojekten vorangetrieben. In enger Kooperation mit AP2.2 und TP2 wurde innerhalb von P1 ein Prozess etabliert, um die Datenanforderungen der verschiedenen Partner systematisch zu sammeln, zu konkretisieren und zu priorisieren.

Tabelle 7.1: Liste zum Vergleich von Anforderungen bzgl. Trainingsdaten und Absicherungsdaten (AP4.4)

Merkmal	Trainingsdaten	Test- und Absicherungsdaten
Variationen	• Maximum an möglichen Variationen und Kombinationen im gesamten möglichen Eingaberaum. • Frame2Frame Variation, soweit möglich • Beinhaltet repräsentative Corner-Cases / Edge Cases • Möglichst Sensorgenaue Nutzung entsprechender Sensormodelle in der Simulation • Fokus liegt auf Einzelbildern	Zusätzlich zu den Anforderungen an die Trainingsdaten: • bestimmte Corner Cases, die als "schwer" oder als "performanzlimitierend" eingeschätzt werden (insbesondere entlang der Ontologie-Dimensionen) • systematisch und gezielt variierbare und kombinierbare Dimensionen und Wertebereiche in einem Bild entsprechend bestimmter Testkriterien • Umsetzung konkreter Parametersätze ("kontrollierbare" bzw. “produzierbare” Variationen entlang definierter Parameter → Um mögliche Schwachstellen entlang von Testparametern zu identifizieren
Labeling / Metainformationen	Typische ground truth label Informationen (2D/3D Bounding Boxen, Semantische Segmentierung), die für das Training notwendig sind	Zusätzlich zu den Ground Truth Daten werden insbesondere Meta-Annotationen benötigt, die für tiefergehende Datenanalysen, Korrelations- und Sensitivitäts-Analysen genutzt werden können.
Bewertung der Qualität der genutzten Daten im Zusammenhang mit DNNs	Performanzmessung / Optimierung: Für relevante Performance Analysen und Vergleich von Algorithmen → Statistisch korrekte Verteilung von Objekten und Sicherheitskritischen Dimensionen	Performanzmessung auf: • bisher ungesehenen Objekten (Out-of-distribution) oder sich außerhalb der antrainierten / definierten ODD befinden (Out-of-ODD Datensätze) • Sollte die Spezifikation unterschiedlicher bzw. auch unrealistischer statistischer Verteilungen in Trainings-

Merkmal	Trainingsdaten	Test- und Absicherungsdaten
		/Validierungsdatensätzen zur Analyse der DNN-Robustheit ermöglichen • Statistisch vergleichbare Datensätze zu realen Daten vs. von der realen Welt statistisch abweichende Datensätze (z.B. Fußgängerverteilungen, Lichtverteilungen, ...) • Testdatensätze (e.g. NCAP-like), um die korrekt funktionierende KI-Funktion gegenüber in safety-kritischen Situationen aufzeigen zu können

Eine der größten technischen Herausforderungen bei der Umsetzung einer Vielzahl von Datenanforderungen der Nutzer lag bis zur Mitte des Projektes (Tranche#4/#5) an der Umsetzbarkeit von prozeduralen und parametrierbaren frame2frame Variationen, als auch benötigten Meta-Annotationen für tiefergehende Datenanalysen. In zahlreichen Workshops unter der Leitung des TP2-Leads wurde ein mögliches technisches Vorgehen zusammen mit den Datenproduzenten erarbeitet und die Anforderungen daran gespiegelt. Die Abstimmungen haben zu einer gemeinsam festgelegten TP2-Feature & Datenroadmap geführt, welche mit den P1-Teilnehmern abgestimmt, priorisiert und vereinbart wurde. Daran wurden die Anforderungen ausgerichtet.

Tabelle 7.2: Eigenschaften der verschiedenen Tranchen der Datenlieferung

Tranche	Beinhaltete Eigenschaften (Auszug)	Datanproduzent Mackevision	Datanproduzent BIT-TS
1, 2	• Vorbereitende Maßnahmen für die groß angelegte Datenproduktion (nicht Teil des veröffentlichten Datensatzes)	x	x
3	• Einführung von HDR-Bildern für unterschiedliche Beleuchtungssituationen (dunkel bis sehr hell) • Weiterer Hochlauf für die groß angelegte Produktion		x x
4	• Frame-zu-Frame-Variationen • Meta-Informationen über ASsetIDs verfügbar • GT-Segmentierung von Körperteilen • Einführung des prozeduralen Sonnenmodells	X X x	X X x
5	• Integration von Sensorrauschen (Post-processing) • Einführung von prozeduralen Wolken • Grund-Truth von Posendaten (Skelett) • Metainformationen zum Verdeckungsgrad	 X X x	x
6	• Umwelteinflüsse: Nässe, Sonnenlicht-Effekte • Out-of-Distribution Assets • Daten für verschiedene Kamerasensor-Parameter	X X x	

Tranche	Beinhaltete Eigenschaften (Auszug)	Datanproduzent Mackevision	Datanproduzent BIT-TS
7	• Datengenerierung mit Kamera- und LiDAR-Sensormodellen durch physikalisch basiertes Rendering mit OSPray		X
	• Daten für verschiedene LiDAR-Sensorparameter		x
	• Umgebungseffekte: Nebel, Vignettierung	x	
	• Meta-Informationen zu MoCap-Sequenzen	x	
8	Einführung von Nachtszenen, einschließlich Kunstlicht	x	
9	Einbeziehung spezifischer Nutzerwünsche bei der Datengenerierung (z.B. Kontrast, vermessene Materialien)	x	

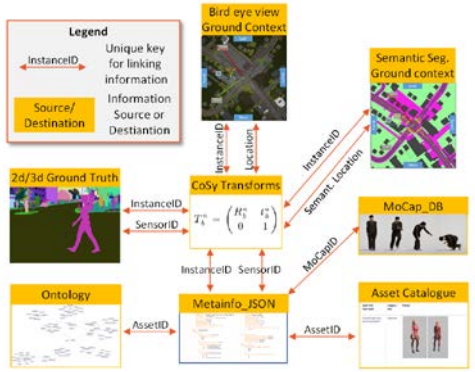
7.1.3 Enriched Metadaten (inkl. safety-relevanter Aspekte und Fileformat)

Metadaten spielen für Analyse- und Benchmarking Zwecke eine zentrale Rolle. Im Gegensatz zu konventionellen "ground-truth" oder "labeling" Daten, bieten diese eine Vielzahl an zusätzlichen Informationen, die die einzelnen Bilder und Objekte auf den Bildern charakterisieren. Dies sind z.B. die Sonnenposition und die Entfernung und Winkel der Fußgänger, die Hüft- und Kopforientierung der Fußgänger relativ zur Kamera, u.v.m.). Basierend auf der engen Zusammenarbeit von Datenproduzenten, Datennutzern und Algorithmen-Entwicklern aus AP1.2, TP2, AP4.1 & AP4.4 konnte mit Unterstützung von P1 ein Datenannotationskonzept entwickelt werden, welches in der Lage ist entsprechend von den Datenproduzenten zur Verfügung gestellte Informationen zusammen zu führen, um daraus relevante Metaannotationen zu produzieren **(E4.1.4b)**. Dieses Konzept basiert auf der Verknüpfung von Informationen basierend auf eindeutigen Datenschlüsseln wie dem Bilddateinamen, der SensorID, IntanzID, AssetID, MoCapID und der semantischen Lokation eines Fußgängers im gewählten Grundkontext **(E2.5.x)**. Hierbei werden diese Metainformationen:

- **Direkt während der Datenproduktion** gewonnen (z.B. Skelett-, Verdeckungs- und Lichtinformationen) **(E2.5.x)**
- Durch **Anwendung von intelligenten Bild-Filter- und Manipulationsalgorithmen** gewonnen **(E2.4.x)**
- Durch Nutzung von Post-processing-Algorithmen, welche unter Anwendung von Koordinatentransformationen und/oder Aggregationsmechanismen die vorhandenen Daten verknüpfen **(E4.1.4b)**.

In enger Zusammenarbeit mit den Datenproduzenten und P1-Teilnehmern wurde die Weiterentwicklung und Struktur des Meta-Annotations-Datenausgabeformat (**E4.1.4b)** (*general-globally-per-frame-analysis-enriched_JSON)* in einer Vielzahl von P1-Telcos TP-übergreifend vorangetrieben und abgestimmt. Daraus ist ein ausführliches Spezifikationsdokument **(E1.2.3 & E4.1.4b)** hervorgegangen, welches im Detail verfügbare Metriken definiert als auch beschreibt. Die verfügbaren Informationen wurden zudem mit der Ontologie bzw. mit zentralen Dimensionen aus der Ontologie verknüpft.

Datenannotationskonzept



RGB Kontrast einer Person zu ihrem Hintergrund

Höhenunterschiede von Schulter zu Fuß

Rotationswinkel der Fußgänger relativ zur Ego-Camera

Verdeckungsgrad der Fußgänger

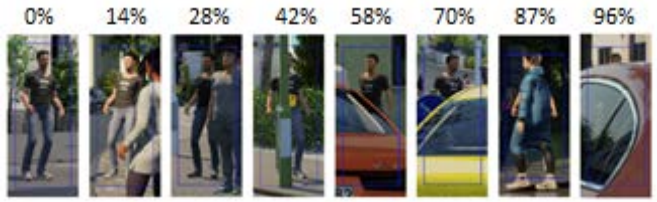
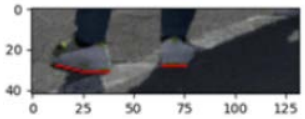
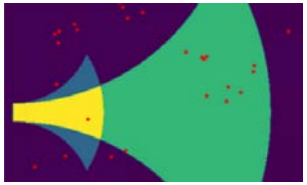
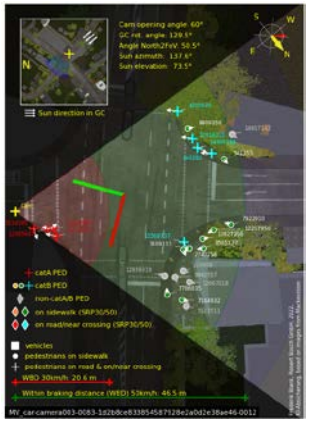


Abbildung 7.1: Enriched Metadaten

7.1.4 (Sicherheitsrelevante) Metadaten-Entwicklung

Eines der im Projekt definierten Safety Concerns sind die "Safety-aware metrics". Um dafür die notwendige Grundlage zu schaffen, wurde im Rahmen der P1 Aktivitäten in enger Abstimmung mit den Kollegen aus **AP1.2, AP4.2, dem P4-Prozess und AP3.2** mehrere metadaten-orientierte Definitionen und Implementierungen eines "Safety Relevant Pedestrian" entwickelt. Dabei werden Fußgänger nach bestimmten Kriterien verschiedenen Klassen zugeordnet (z.B. Bremsweg oder Lokation auf der Straße oder Gehweg), um Untersuchungen dann gezielt für diese unterschiedlichen Kategorien durchführen zu können.

Diese Fußgänger wurden unter Nutzung verschiedener Post-Processing & Visualisierungsskripte von Opel und Bosch entsprechend annotiert und visualisiert. Das folgende Beispiel zeigt anschaulich die entsprechende Entwicklungshistorie dar (alle Werteangaben sind nur exemplarisch).

Version #1	Version #2	Version #3	Version #4 (final)
Berücksichtigung aller Fußgänger, welche • im Field of View sind und • eine Entfernung <= 60 m von der Kamera haben • Weniger als 80% verdeckt sind	Berücksichtigung aller Fußgänger, welche • im Field of View sind und • eine Entfernung <= 60 m von der Ego-Kamera haben und • auf der Straße stehen	Berücksichtigung aller Fußgänger, welche • sich in einem möglichen Fahrschlauch bei einer Entfernung bis 20,6m (blau/gelb) bis und 46,5m (gelb/grün) befinden.	Berücksichtigung aller Fußgänger, welche • im Field of View sind und • sich in einem möglichen Fahrschlauch bei einer Entfernung bis 20,6m (blau/gelb) oder bis und 46,5m (gelb/grün) befinden. • sich auf der Straße oder dem Bürgersteig befinden • weniger als 80% verdeckt sind
 Quelle: Mackevision und Bosch	 Quelle: Bosch	 Quelle: Opel	 Quelle: Bosch

7.1.5 Formulierung von Datensatz-Anforderungen im Kontext von Safety Betrachtungen (inkl. NCAP-like Szenarien)

In Zusammenarbeit mit P1 unter Beteiligung von Bosch, ZF, IKA, Mackevision, QualityMinds und Opel wurde ein safety-kritisches Szenario mittels systematischer Variationen einzelner Parameter als Grundlage für mehrere Sequenzen der Macke Vision Datenlieferungen parametriert und produziert **(E4.1.5, E4.4.2, E1.2.4, E2.5.x)**. Das Ziel war es dabei in einer limitierten Anzahl von Bilder möglichst viele Kombinationen an Parametern zu produzieren und dennoch den parametrierbaren Eingaberaum möglich gut abzudecken. Diese wurden dann um weitere relevante safety kritische Szenarien seitens ZF und FKA erweitert. Als Basisszenarien wurden mehrere typische safety-kritische Szenen wie "Person taucht zwischen zwei geparkten Fahrzeugen auf und läuft auf die Straße" oder "Linksabbieger" gewählt. Als Variationsparameter wurden verschiedene Ontologie-Dimensionen herangezogen wie Sonnenposition, Wolken-Bedeckunggrad des Himmels, Nebeleffekte, Position des Ego-Fahrzeugs, Fußgänger-Position & Rotation, u.v.m.). Die untenstehende Grafik veranschaulicht anhand von einigen Beispielbildern das entwickelten *NCAP-ähnliche Szenario* und die dazu kontinuierlich entsprechend der Projektanforderungen entwickelte Tools. Während der durch die Zwicky-Boxen aufgespannte Datenraum für eine vollständige Simulation aller Kombinationen ca. 478 Milliarden Kombinationen benötigt hätte, wird durch die Anwendung der im Projekt weiterentwickelten Methode des Combinatorial Testings (hier 3-wise) die Anzahl an Kombination auf ca. 6700 Kombinationen für ein solches Szenario reduziert. Diese Datenanforderungen wurden in einem vereinbarten Datenanforderungsformat **(E4.1.4b)** als JSON gespeichert und von den Datenproduzenten in der Produktionspipeline entsprechend umgesetzt **(E2.5.x)**

Zusätzlich zu den sicherheitsrelevanten Szenarien wurde noch ein Posenszenario bestehend aus mehr als 10.000 Bilder in deiner Kooperation zwischen ZF/IKA, Bosch und Mackevision erstellt. Das Posen-Szenario wurde einerseits definiert **(E1.2.5)** um zu zeigen, wie mögliche Performanzlimitierende Faktoren (PLF) identifiziert und gemessen **(E1.2.6)** werden können, und andererseits, wie Experimente konzipiert werden können, um die Begrenzung des Netzwerks hinsichtlich solcher PLF gemessen **(E1.3.x)** und möglicherweise reduziert werden können **(E4.4.2)**.

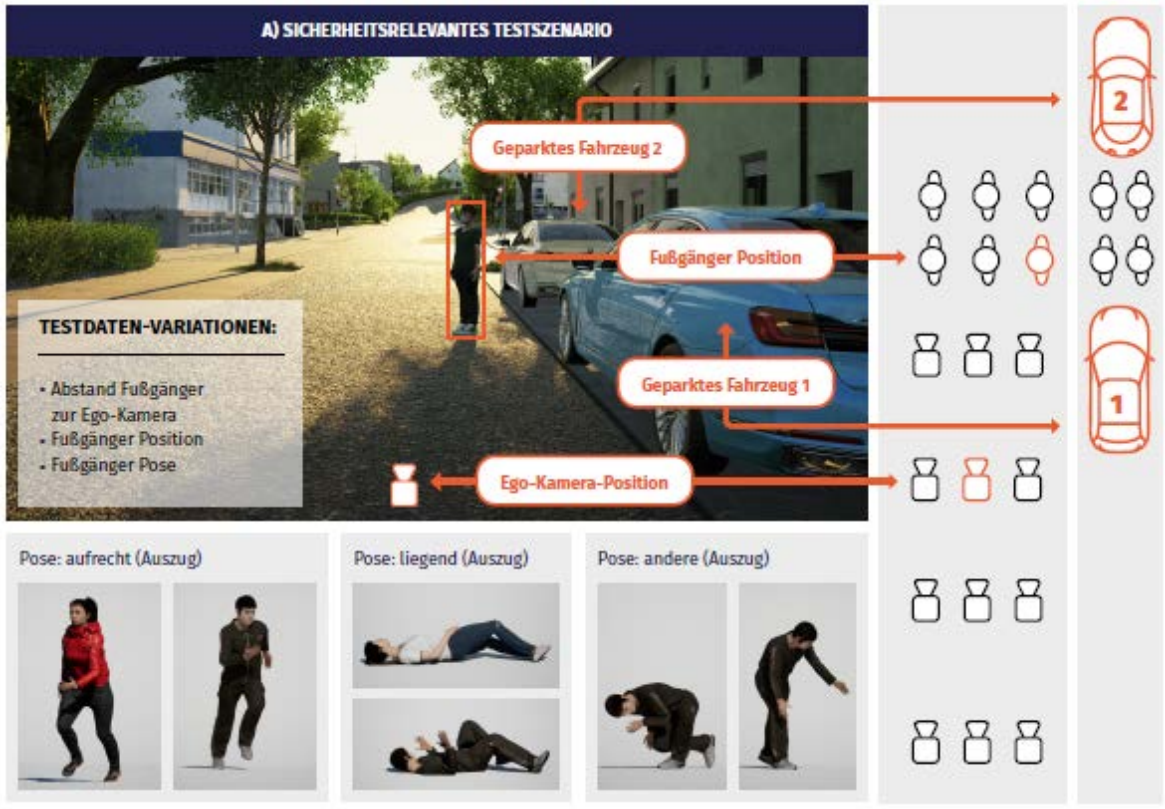


Abbildung 7.2: Parametrierte, sicherheitsrelevantes Test-Szenario (NCAP-like) aus der Vogelperspektive Quelle: Bosch, Mackevision

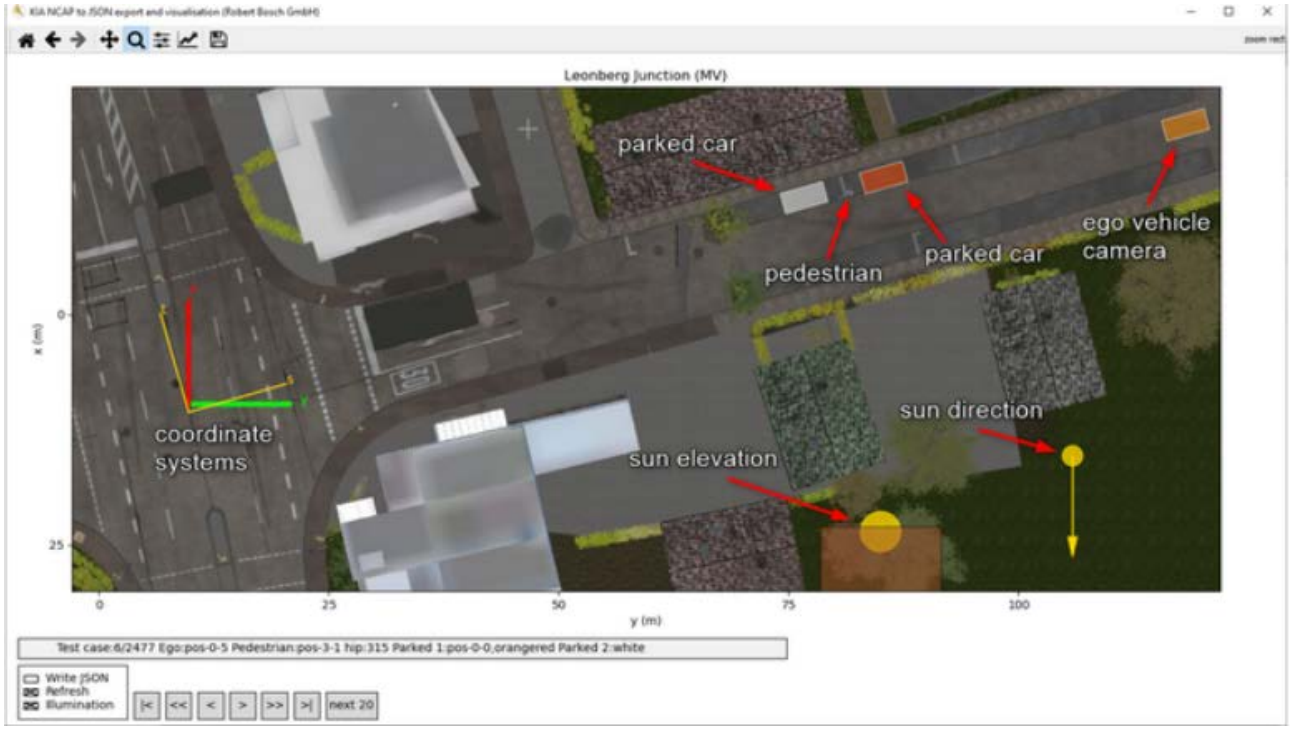


Abbildung 7.3: Tool zur Planung von parametrierten Test-Szenarien Quelle: Bosch

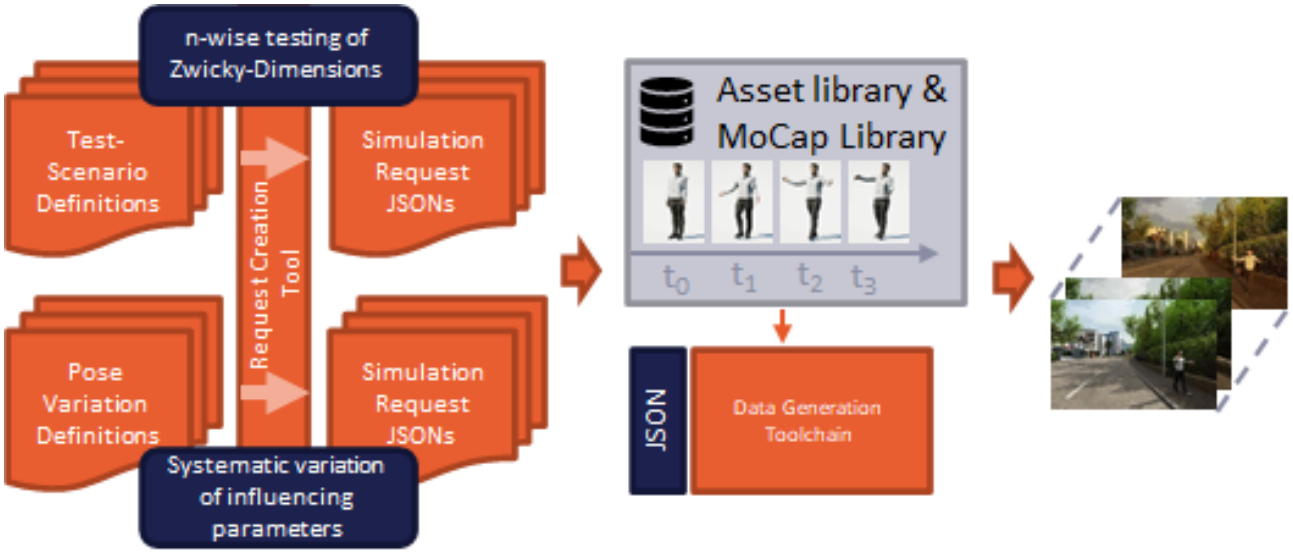


Abbildung 7.4: Visualisierung der Toolchain zur Verarbeitung der parametrierten Szenarien Anforderungen als JSON Quelle: ZF & Mackevision

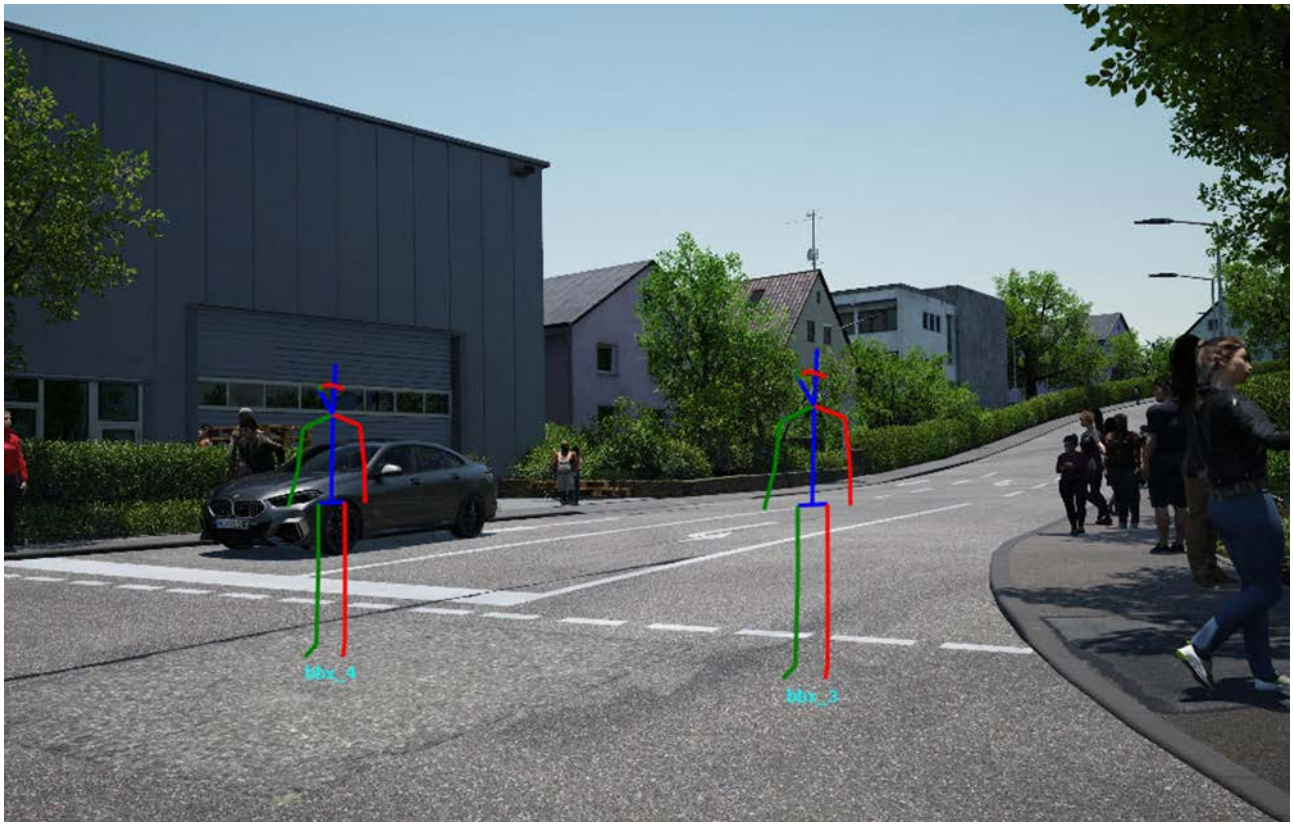


Abbildung 7.5: Visualisierung der Posenanforderungen für die Test-Szenarien Quelle: ZF/IKA, Mackevision

7.2 P2 - Iterationsprozess Funktionen/Algorithmik

Der Prozess *Entwicklung und Bereitstellung von DNN Modellen* beschreibt das Zusammenspiel aller Arbeitspakete bzw. Projektpartner, die an der KI-Algorithmenentwicklung beteiligt sind (v.a. AP1.3, 1.4, 1.5) oder dazu beitragen.

Die Ergebnisse der Algorithmenentwicklung werden wiederum iterativ über den Projektverlauf als Inputs in anderen Arbeitspaketen des Projekts benötigt (z.B. AP2.3, 2.4, 3.3, 3.4, 3.5, 4.5).

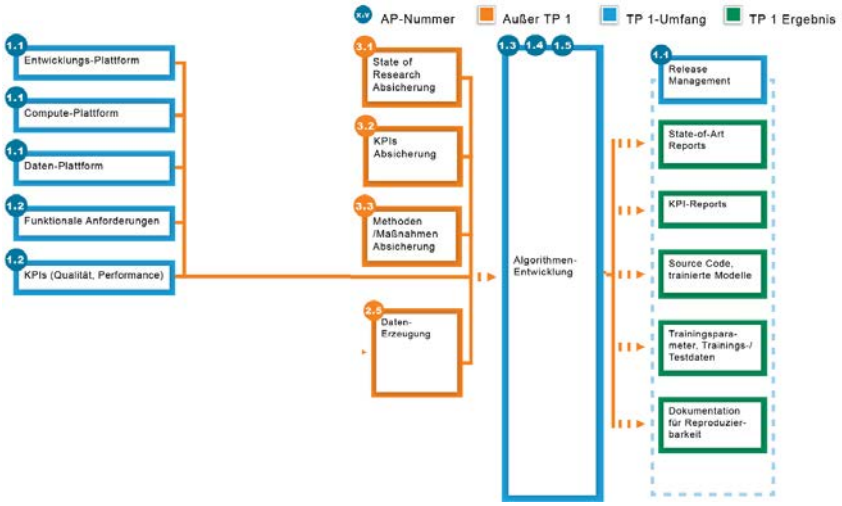


Abbildung 7.6: Zusammenspiel aller am Prozess Entwicklung und Bereitstellung von DNN Modellen beteiligten Arbeitspakete.

Im Mittelpunkt dieses Prozesses stand das Release Management (siehe blaue Box, oben rechts in der Abbildung). Das Release Management beschäftigt sich im Kern mit der Veröffentlichung (engl. release) von trainierten DNN Modellen für die Nutzung in TP2, TP3 und TP4. Inhalte der Veröffentlichungen sind stichpunktartig in den grünen Boxen dargestellt (siehe Abb.). Das Release Management ist im AP1.1 verankert und betrifft im Wesentlichen die DNN Entwicklungs-Arbeitspakete 1.3 bis 1.5. Zum Zeitpunkt des Schreibens wurden 5 Releases (Release 1, 1*, 2, 2*, 3) durchgeführt, die sich in den Anforderungen und verwendeten Daten unterscheiden. Die Release-Zyklen wurden den Zyklen der Datenveröffentlichung angepasst, sodass jeder neue Release auf eine neue Datenbasis beruht. Der Datensplit wurde dynamisch für jeden Release angepasst.

7.2.1 Release 1 / 1*

Das erste Release wurde zum M11 veröffentlicht und zur Übersicht in einer Tabelle dargestellt (siehe Release #1). Aufgrund der Verfügbarkeit von Trainingsdaten wurde teilweise auf öffentlich, verfügbare Datensätze zurückgegriffen. Soweit möglich wurde aber bereits der KIA Datensatz (Tranche 1) zum Training verwendet. Ein Vorschlag für die Teilung des KIA-Datensatzes in Trainings-, Validierungs- und Testdaten wurde durchgeführt.

7.2.1.1 Zur Vorbereitung des Releases wurde

1. eine Anleitung zum Arbeiten mit Gitlab, in dem der Code eingecheckt wird, verfasst (siehe 1.1.2c Ergebnisbericht (Ergebnis + erläuterndes Dokument).
2. eine Ausarbeitung der Ordner- und Code-Struktur in Gitlab erstellt sowie einen Überblick über den workflow in Gitlab gegeben (siehe 1.1.2c Ergebnisbericht (Ergebnis + erläuterndes Dokument).
3. eine Plattform zum Hochladen der trainierten Gewichte der DNN Modelle geschaffen und dokumentiert (siehe 1.1.2c Ergebnisbericht (Ergebnis + erläuterndes Dokument).
4. ein Tool zum Testen der Lizenzen, unter denen der Code steht, bereitgestellt (siehe 1.1.2c Ergebnisbericht (Ergebnis + erläuterndes Dokument).

5. ein Muster-README File erstellt, welches für jedes Repository ausgefüllt werden soll und als Dokumentation der DNN Modelle und seiner Auswertung dient. Mittels dieser Informationen soll eine Reproduzierbarkeit sichergestellt werden (siehe https://gitlab.com/kia2/code/templates/-/blob/master/README_repo_template.md).
6. ein Prozess für die Kennzeichnung von Repositories erstellt, welche für den Release vorgesehen sind (Stichwort: release und pre-release tag) (siehe 1.1.2c Ergebnisbericht (Ergebnis + erläuterndes Dokument).
7. eine Vielzahl von Meetings durchgeführt, um die genannten Punkte auszuarbeiten und zu diskutieren.

7.2.1.2 Der erste Release beinhaltet:

1. die Bereitstellung eines Dockerfiles, was die Basis für die Ausführung der DNN Modelle in Docker Containern darstellt.
2. die Bereitstellung von DNN Modellen inklusive Gewichte.
3. eine Dokumentation zur Installation und Ausführung der DNN Modellen, welche im gewissen Rahmen vereinheitlicht wurde (siehe jeweiliges README File).

Die Zeitpunkte der weiteren Releases ist im Anhang hier dargestellt: TP1 KI-Funktion

Aufgrund der Verfügbarkeit von weiteren Trainingsdaten aus dem TP2 (Tranche 2) wird ein weiterer Release angestrebt, welcher die gleichen DNN Modelle wie der Release 1 enthält, jedoch mit den Daten aus der Tranche 2 trainiert. Dieser wird mit Release 1* bezeichnet.

7.2.2 Release 2 / 2*

Für das Release 2 wurden die Daten aus der Tranche 3 zum verwendet. Hierzu wurden Erweiterungen bei der Automatisierung der Tests vorgegeben, um den Entwicklern die Arbeit zu erleichtern und eine hochwertige Qualität der Repositories sicherzustellen. Im Wesentlichen werden vier Level der Testautomatisierung angestrebt. Das erste Level testet die Vollständigkeit der Repositories gemäß der Vorgaben aus AP 1.1. Das zweite Level dient der Sicherstellung, dass im Code nur Lizenzen aus der KI-A Whitelist verwendet werden. Das zweite Level überprüft die Lauffähigkeit des Trainingsprozesses. Das dritte Level überprüft die Lauffähigkeit der Inferenz. Das vierte Level berechnet die Metriken, die gemäß AP1.2 für das jeweilige DNN relevant sind und erleichtert damit die Dokumentation und den Vergleich der DNN-Evaluierung.

Darüber hinaus wurde ein Release 2* durchgeführt, da mit der Veröffentlichung der Tranche 4 3D Bounding Box Labels geliefert wurden, sodass für das AP1.4, welches sich auf die Detektion von 3D Bounding Box fokussieren, ausreichend Daten für ein Training auf den KIA Datensatz zur Verfügung standen. Die Daten der neuen Tranche 4 wurden in einen Trainings- Validierung- und Testsplit eingeteilt, sodass dieser für die Entwicklung der DNN Modellen verwendet werden kann. Das Tool zur automatisierten Durchführung von Tests wurde von Gitlab zu Bitbucket übertragen und in seiner Funktionsweise verfeinert.

Aufgrund unterschiedlicher Datenkonsistenz und Kompatibilität der gelieferten Daten wurden Fix-Skripte implementiert und bereitgestellt. Es wird weiterhin daran gearbeitet, die Spezifikation des Annotationsformats bei der Datenproduktion vollumfänglich zu berücksichtigen, sodass eine Implementierung und Ausführung von Fix-Skripten nicht mehr

benötigen werden, um die DNN Entwickler zu entlasten. Zur Erhöhung der Vergleichbarkeit der TP1 DNNs wird die Einführung von Benchmarks angestrebt. Hierfür wurde ein Output Format für die verschiedenen Aufgaben wie semantische Segmentierung, 2D Bounding Box etc. für die DNN Ausgaben definiert und mit den Stakeholdern aus TP3 und TP1 abgestimmt.

7.2.3 Release 3

Die Trainingsdaten für den Release 3 sind auf die Tranche 3 und 4 von den beiden Datenproduzenten Bit TS und Mackevision beschränkt. Die Tranche 5 wurde erst kurz nach dem Release 3 veröffentlicht, sodass dieser erst Teil vom Release 4 sein wird. Die Kombination der Tranche 3 und 4 gilt sofern die Labels/Datenformate es zulassen. Der Zusatz "sofern die Labels/Datenformate es zulassen" bedeutet, dass nicht jeder TP1 Algorithmus alle Daten verwenden kann, z.B. sind semantische Segmentierungslabels nicht für Tranche 4 von Mackevision verwendbar. In diesem Fall werden die Daten von Tranche 4 von Mackevision ignoriert.

Bei der Durchführung des Release 3 wurde erneut die automatische Testtoolchain verwendet, die den Vollständigkeits-, Lizenz-, Inferenz- und Trainingstest durchführt. Trotz zahlreichen Diskussionen und Verbesserungen in Bezug auf die Toolchain und der Dokumentation für die Entwickler gab es Probleme bei den Tests. Dies führt dazu, dass viele Tests den Status "failed" haben, obwohl der Code von anderen Partner im Projekt ohne Probleme genutzt wurde. Weiterhin bestand laut Planung der Wunsch, die Toolchain weiter auszubauen, um zusätzlich eine automatische Metrik-Berechnung durchführen zu können.

Aus den genannten Gründen folgte eine Änderung der automatischen Testtoolchain. Im Gegensatz zu vorher, wurde die Inferenz- und Trainingstests eingeschränkt, sodass innerhalb der Test-Toolchain die Code Repositories nicht mehr ausgecheckt und installiert werden müssen. Dies stellt eine Erleichterung seitens der Toolchainentwicklung dar und beseitigt viele Fehlerquellen, die damit einhergingen. Um dem zusätzlichen Wunsch der Metrik-Berechnung gerecht werden zu können, werden die Predictions (von den Testdaten) der TP1 DNNs von den Entwicklern erzeugt und auf dem DSP hochgeladen. Die Toolchain liest automatisiert die Predictions sowie die zugehörigen Ground Truth Daten ein und berechnet die entsprechenden Metriken. Dieser Prozess beschränkt sich aufgrund mangelnder Nachfrage anderer DNNs auf SSD und DeeplabV3+.

7.3 P3 - Konsolidierungsprozess zum Kontext Gesamtfunktion & Systemarchitektur

7.3.1 Spezifikationen der Annahmen im Kontext der Gesamtfunktion

In diesem Teil des P3 Prozesses wurden funktionale und technische Herausforderungen durch Einholung von Feedback relevanter Stakeholder innerhalb des Projekts KI-Absicherung definiert. Die Identifikation dieser funktionalen Herausforderungen beinhalten u.a. die Detektion von sehr kleinen Fußgängern (Entfernung größer 30m), Fußgängern hinter Glaswänden und vielen weiteren. Zu den technischen Herausforderungen gehören v.a. Inkonsistenzen innerhalb der verschieden Daten Tranchen (z.B. unterschiedliche Definition der Labels) und das Fehlen von Komplexität in künstlich simulierten Daten.

Die "Operational Design Domain" (ODD) definiert den Kontext in welchem das System Anwendung findet. Diese ODD wurde im Rahmen des P3 Prozesses definiert und liegt in Form einer Tabelle

vor. Diese gibt ihr die Struktur und Übersicht, während die individuellen Bausteine in referenzierten "Zwicky Boxen" (aus SCODE Modell abgeleitet) ausführlich dargestellt werden.

Für die Gesamtfunktion (Fußgängererkennung) wurden in Kooperation mit E4.2.5 Annahmen abgeleitet:

- Allgemeine Aufgabe des Systems: Vermeidung einer Kollision mit Fußgängern durch Bremsen innerhalb spezifizierte Geschwindigkeiten
- Die Funktion "Fußgängererkennung" wird entweder durch das Fahrzeug oder den Kunden aktiviert
- Die Definition "Genug Zeit zum Bremsen" hängt von anderen (Umwelt-) Bedienungen ab, z.B. Wetter, Straßenverhältnisse (innerhalb der ODD)
- Allgemeine Voraussetzung: Deutschland, innerstädtisch, nur longitudinale Kontrolle, max 50 km/h
- Berücksichtigung der geschwindigkeitsabhängigen "Must Detect" Zone

7.3.2 System Architektur

Der P3 Prozess hat eine allgemeine Architektur für das in KI-Absicherung untersuchte System (Fußgängererkennung) definiert und mit den Projektpartnern abgestimmt. Diese Systemarchitektur ist in der folgenden Abbildung dargestellt. Sie beschreibt den groben Rahmen in welchem die KI-Funktion Anwendung findet.

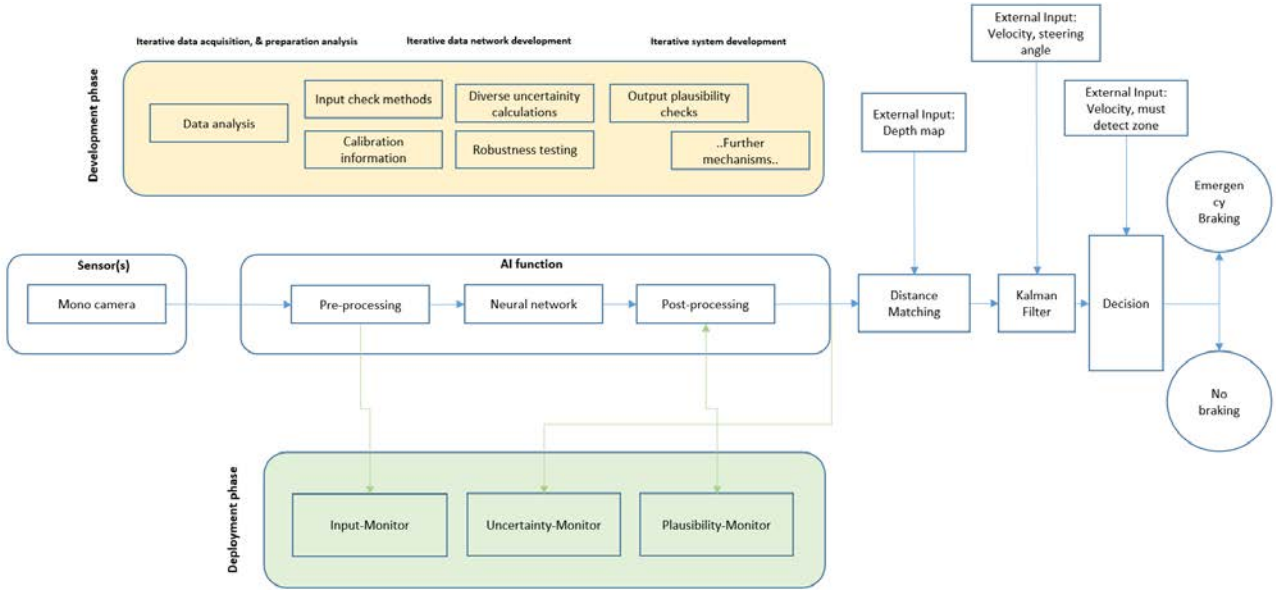


Abbildung 7.7: Systemarchitektur

7.3.3 Anforderungen für die KI-Funktion

Die abgeleiteten Anforderungen an die KI-Funktion lassen sich in funktionale, technische und Machine Learning Safety Requirements (MLSR) unterteilen.

Es wurde nur eine einzige funktionale Sicherheitsanforderung innerhalb des Projekts adressiert: *Das "Pedestrian Collision Avoidance" System soll alle Fußgänger innerhalb der*

geschwindigkeitsabhängigen "Must Detect" Zone erkennen und unter Berücksichtigung des nächsten Fußgängers in dieser Zone in allen möglichen Situationen in der ODD bremsen.

Diese funktionale Sicherheitsanforderung wurde anschließend zu technischen unter Berücksichtigung der Systemarchitektur weiterentwickelt. Ein Beispiel eines solchen ist: *Die Fußgängererkennung soll zuverlässig jeden Fußgänger in 4 aus 5 aufeinander folgenden Frames erkennen, solange sich das System in der ODD befindet.*

In einem finalen Schritt, wurden aus den technischen Anforderungen, Machine Learning relevante Anforderungen (MLSR) abgeleitet. Diese sind in sog. "atomic aspects" festgehalten, z.B. *Die Fußgängererkennung soll Fußgänger mit einem normal-verteilten Fehler in der Position der Bounding Boxen erkennen"*

Weitere Details zu dem Prozess der Ableitung von Sicherheitsanforderungen ist in E4.2.5 beschrieben

7.4 P4 - KPI-Konsolidierungsprozess

Um die Absicherungsstrategie für tiefe neuronale Netze zu ermöglichen und zu unterstützen wurden im Laufe des Projektes verschiedenste Metriken abgeleitet, definiert, berechnet und evaluiert. Solche Metriken beziehen sich auf:

- die funktionale Güte
- die Qualität der Datensätze
- absicherungsrelevante Eigenschaften der Netze an sich sowie
- absicherungsrelevante Eigenschaften der Mechanismen

Die Konsolidierung all dieser Metriken und deren Aggregation, damit diese in der Nachweisstrategie verwendet werden können, ist das Ziel dieses Prozesses. Das Ergebnis des Prozesses sind Metriken mit gleichartigen Beschreibung und Bewertungen, sowie die Aufbereitung dieser Metriken für die Integration in die Sicherheitsargumentation als Evidenzen. Zunächst wird der durchgeführte Metrik-Prozess vorgestellt, welcher im Projekt dafür gesorgt hat, dass die verwendeten Metriken beschrieben und die KI Funktion anhand dieser evaluiert wurde. Anschließend werden die sogenannten Evidenz Workshops und deren Nachfolger, die Evidenz Workstreams beschrieben, welche maßgeblich dazu beigetragen haben, Evidenzen für die Sicherheitsargumentation zu liefern.

7.4.1 Beschreibung und Schlussfolgerung des Metrik-Prozesses

Der Metrik-Prozess umfasst insgesamt vier Teilprozesse:

- Anforderung und Entwicklung von funktionale Metriken, welche die Performanz der KI Funktion evaluieren können.
- Anforderung und Entwicklung von Datenmetriken, welche zum einen die Datenqualität als auch die Abdeckung der Datenbasis bzgl. der Zieldomäne erfassen.

In allen nachfolgenden Prozessschaubildern markiert S den Startpunkt und E den Endpunkt. In Klammern sind die entsprechenden Arbeitspakete vermerkt, welche im jeweiligen Prozessschritt mitgewirkt haben.

Die Abbildung 7.8zeigt das Prozessschaubild für die funktionalen Metriken. Dieser Prozess besitzt zwei Startpunkte. Der erste beginnt bei der Definition von neuen Sicherheitsanforderungen (funktionalen Anforderungen) an die KI Funktion (1) aus systemischer Sicht. Nach der Definition werden in Abstimmung Anforderungen an benötigten funktionalen Metriken definiert (2) und abgestimmt (3). Der zweite Startpunkt verwendet bekannten Metriken sowie die vorher abgestimmten Anforderungen, um funktionale Metriken zu definieren und zu beschreiben (4). Anschließend werden diese Metriken konsolidiert (5) und schließlich in einen zentralen Metrikkatalog erfasst (6).

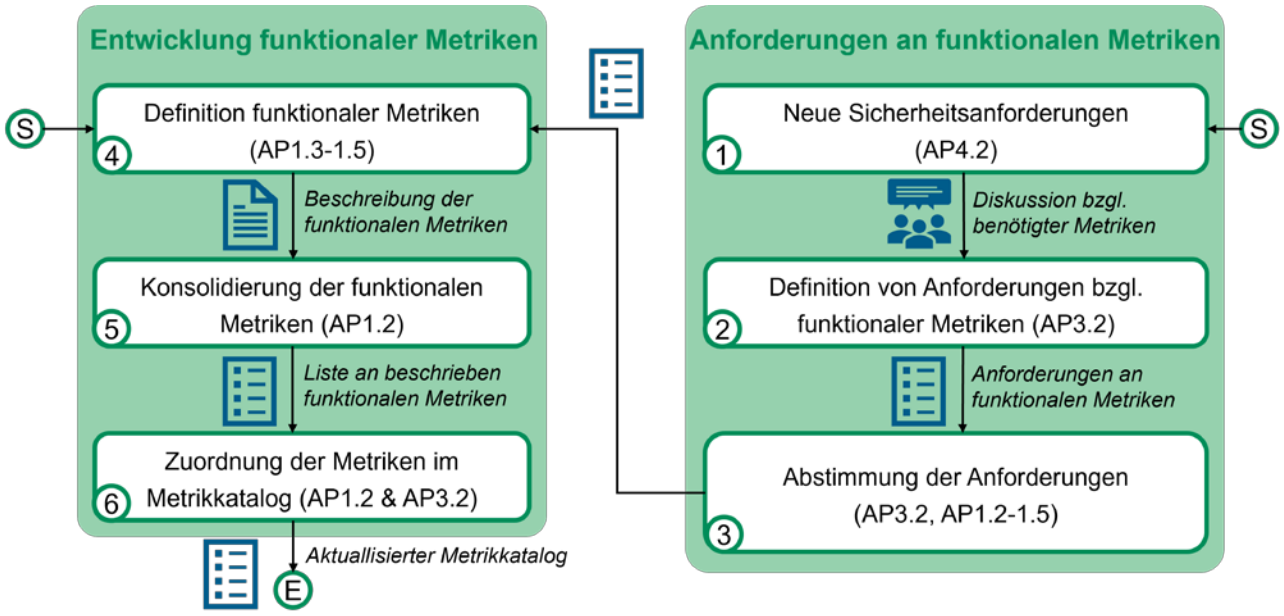


Abbildung 7.8: Prozessschaubild der funktionalen Metriken.

Da jede Metrikauswertung anhand von Daten durchgeführt wird, muss die Datenbasis eine gewisse Qualität aufweisen. Abbildung 7.9 zeigt dazu den Prozess, um geeignete Datenmetriken zu definieren. Dieser besitzt wie der Prozess für die funktionalen Metriken ebenfalls zwei Startpunkte. Der erste Startpunkt beginnt bei der Entwicklung einer Daten-Sicherheitsargumentation (1) aus welcher im nachfolgenden Schritt Anforderungen an die Datenbasis (2) bzw. an notwendige Datenmetriken abgeleitet werden. Schließlich werden die definierten Anforderungen abgestimmt (3). Der Zweite Startpunkt beginnt mit der Sammlung von verfügbaren Datenmetriken (4), wobei auch die Anforderungen berücksichtigt werden. Das Resultat ist eine Liste an Datenmetriken, welche schließlich verwendet wird, um Analysen auf der vorhanden Datenbasis durchzuführen (5). Anhand der Analysen können der Sicherheitsargumentation dann entsprechende Evidenzen (6) hinzugefügt werden. Abschließend werden relevante Datenmetriken im zentralen Metrikkatalog aktualisiert.

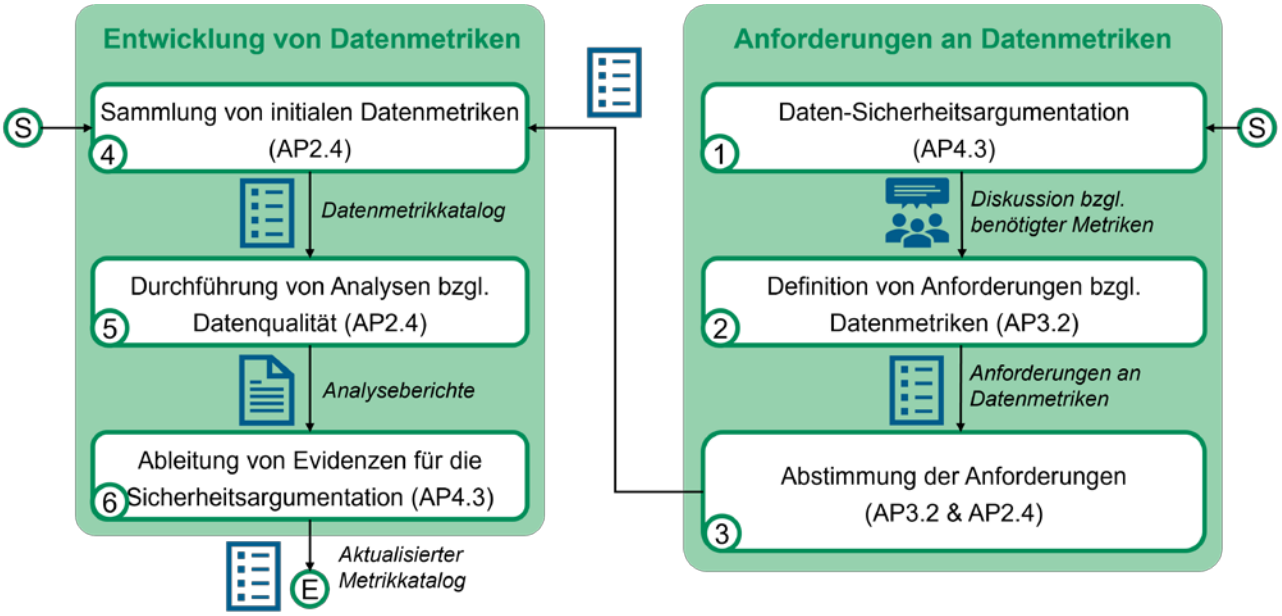


Abbildung 7.9: Prozessschaubild der Datenmetriken.

Die folgenden beiden Teilprozesse widmen sich der Definition und Validierung von Sicherheitsmetriken. Diese sind in Abbildung 7.10 und Abbildung 7.11 dargestellt. Dabei betrachtet der dargestellte Prozess in Abbildung 7.10 die Definition von Sicherheitsmetriken auf Basis der verwendeten Technologie und der Prozess in Abbildung 7.11 hat die systemische Sichtweise. In Abbildung 7.10 beginnt der Prozess bei der Identifikation von DNN-spezifischen Sicherheitsbedenken und Performanz limitierenden Faktoren (PLF) (1). Die Sicherheitsbedenken fassen die Eigenschaften von DNNs zusammen, welche potentiell zu Fehler führen können. Die PLFs beschäftigen sich mit (Umwelt-)faktoren welche die Leistung einer KI Funktion limitieren können. Anhand der DNN-spezifischen Sicherheitsbedenken werden Metriken basierend auf den verfügbaren Mechanismen definiert (2), welche zum einen messen sollen ob ein Sicherheitsbedenken ausgeräumt werden kann oder dazu beiträgt dieses zu mitigieren. Alle Metriken werden im Metrikkatalog dokumentiert und überprüft (3). Anschließend wird evaluiert, ob weitere Metriken benötigt werden (4) oder ob entsprechende Mechanismen entwickelt werden sollen (5). Nach der Mechanismenentwicklung werden diese evaluiert und einem Effektivitätstest unterzogen (6). Abschließend werden Testberichte erstellt, welche die Mitigation der DNN-spezifischen Sicherheitsbedenken oder PFLs nachweißen.

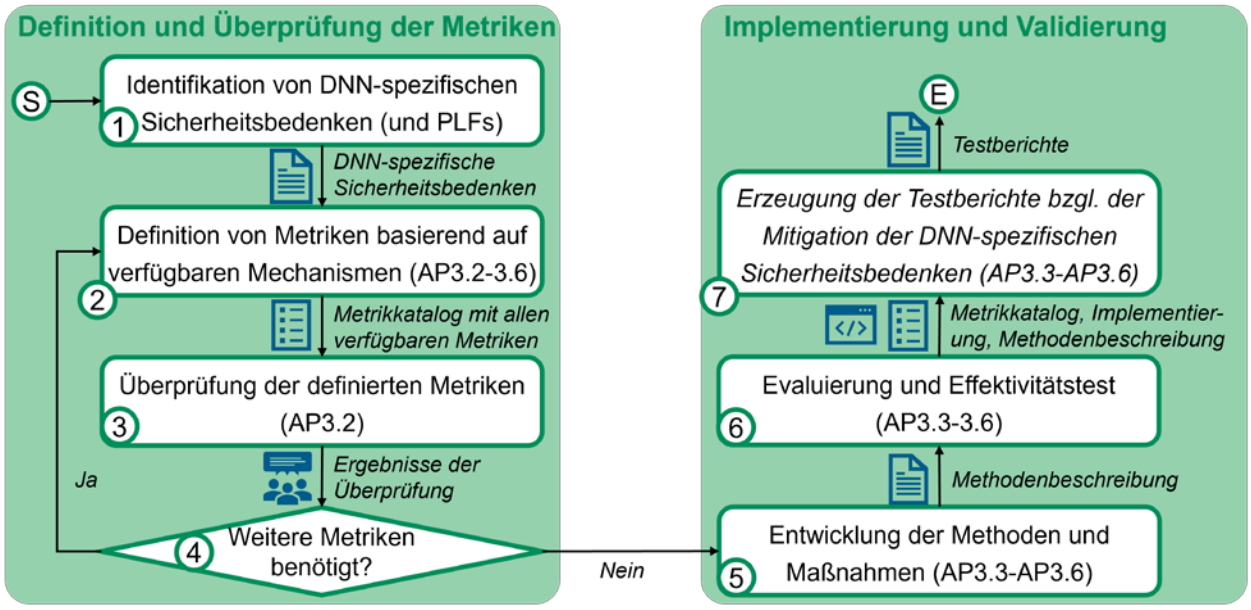


Abbildung 7.10: Prozessschaubild der Sicherheitsmetriken aus Technologiesicht.

Die Abbildung 7.11 beginnt bei der Definition von Sicherheitsanforderungen (1), welche prinzipiell funktionale Anforderungen darstellen an die entsprechende KI Funktion. Im Anschluss daran können Metriken definiert werden, welche einen Nachweis für die Erfüllung der Anforderungen erbringen können (2). Alle definierten Metriken werden überprüft (3) und in den Metrikkatalog eingetragen (5). Falls Metriken fehlen (4) müssen entsprechend neue entwickelt werden. Anschließend kann die Implementierung der Metriken und die Evaluierung stattfinden (6). Abschließend werden Testberichte erzeugt, welche zeigen ob eine Anforderung erfüllt werden konnte. Die Resultate fließend schließlich als Evidenzen in die Sicherheitsargumentation.

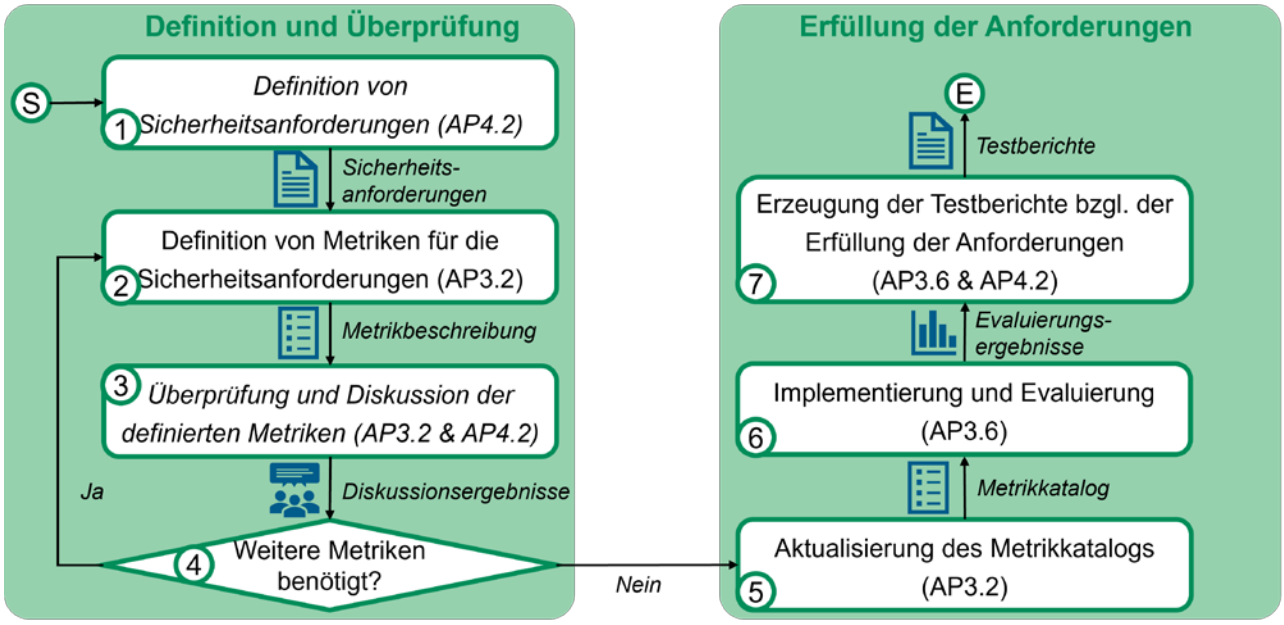


Abbildung 7.11: Prozessschaubild der Sicherheitsmetriken aus Systemsicht.

7.4.2 Evidenz Workshops and Workstreams

Während der Durchführung des Metrik-Prozesses wurde festgestellt, dass eine Integration der Ergebnisse anhand der Metriken in die Sicherheitsargumentation weiter Schritte benötigt. Daher

wurde eine Pilotphase organisiert: Die Evidenz Workshops. Diese Workshops hatten das Ziel für 7 ausgewählte Mechanismen eine Teil-Sicherheitsargumentation mit Evidenzen zu erstellen, für die entsprechenden Ergebnisse eines Mechanismus. Dabei folgt jeder Workshop einer definierten Agenda sowie zusätzlich einer Vorbereitung. Während der Vorbereitung wurde die Methode analysiert und eine Teil-Sicherheitsargumentation für diese erzeugt. Damit konnte herausgearbeitet werden, an welcher Stelle in der Sicherheitsargumentation ein Beitrag geliefert werden konnte. Weiterhin wurden geeignete Testmethoden identifiziert. Während des Workshops wurde schließlich folgende Punkte behandelt:

- Präsentation der Methode
- Vorstellung der entwickelten Teil-Sicherheitsargumentation
- Präsentation der Ergebnisse aus den Methoden-Experimenten
- Weiterentwicklung der Teil-Sicherheitsargumentation
- Vorstellung der Testmethoden

Nach einem Workshop wurden Kommentare und Feedback in die Sicherheitsargumentationen eingearbeitet und anschließend dem Projekt zu Verfügung gestellt. Die Teil-Sicherheitsargumentationen wurden anschließend innerhalb von TP4 in die Gesamtargumentation eingebettet. Eine Übersicht aller Workshops kann in der Tabelle 7.3 eingesehen werden.

Tabelle 7.3: Übersicht aller durchgeführten Evidence Workshops

Nr.	Methode	Datum	Kommentar
1	Variational Auto Encoder Reconstruction Error	10.07.2020	Neben dem Variational Autoencoder wurden auch ein GAN Autoencoder betrachtet. Im wesentlich tragen beide Methoden zur Out-of-Distribution Detektion bei, um bei Eingangsdaten welche deutlich außerhalb der Trainingsdatenverteilung liegen eine Fehldetektion der Fußgängererkennung zu vermeiden.
2	Local Uncertainty Realism via Ensemble Diversification	03.08.2020	Der Teil der Sicherheitsargumentation, welcher betrachtet wurde, umfasste wie vorherige Methode eine Out-of-Distribution Erkennung.
3	Hybrid Learning using Concept Enforcement	24.08.2020	Einbettung der Methode in einen generellere Sicherheitsargumentation zu den Themen Plausibilisierung sowie Verifikation und Validierung
4	Hybrid and Robustness-focussed Compression	29.09.2020	Trade-Off zwischen Robustheit und Effizienz (Inferenzzeit)

Nr.	Methode	Datum	Kommentar
5	Attention consistency validation	27.10.2020	Die Teil-Sicherheitsargumentation basiert auf dem Hauptziel, dass Fußgänger in einer Reichweite von bis zu 20 Metern mit genügender Präzision erkennt werden.
6	Meta LRP (Layerwise Relevance Propagation)	15.12.2020	Die Teil-Sicherheitsargumentation befasst sich damit, dass Fußgänger nicht erkennt werden sollen, wenn auch tatsächlich kein Fußgänger exisitert ("true negative").
7	Aggregation based dependency analysis of neural networks with Visual Analytics	15.01.2020	Die entwickelte Sicherheitsargumentation befasst sich vor allem mit der tool-gestützen Validierung durch mehrere Personen (Experten).

Nachdem sich die Pilotphase bewährt hatte, um Evidenzen zu bestimmen, wurde das Konzept anhand der DNN-spezifischen Sicherheitsbedenken auf das gesamte Projekt ausgerollt. Die Details dazu finden sich im Abschnitt Evidenz Workshops.

7.4.3 Schlussfolgerung

Im Projekt hat es sich bewährt sowohl von der technologischen als auch systemischen Seite an den Metriken zu arbeiten. Beide Prozesse können parallel ausgeführt und zentral über einen Metrikkatalog verwaltet werden. Des Weiteren hat sich gezeigt, dass eine gesonderte Betrachtung der funktionalen Metriken nicht unbedingt notwendig ist und auch von dem Prozess in Abbildung P4.3 erarbeitet werden kann. Es sei außerdem anzumerken, dass es bei den Datenmetriken den meisten Forschungsbedarf gibt, da im wesentlichen diese Aufschluss darüber geben können ob die durchgeführten Evaluierungen der Sicherheitsmetriken überhaupt relevant sind. Des Weiteren ist zu empfehlen, die Arbeiten an der KI Funktion und der Sicherheitsargumentation parallel durchzuführen und durch geeignete Prozess, wie Beispielsweise die Evidenz Workstreams, zusammenzuführen.

7.5 P5 - Datengenerierungsprozess

Der P5 Prozess hatte die Aufgabe die funktionalen Abläufe zur Datenerzeugung und dazu benötigte Infrastrukturen zu strukturieren und Schnittstellen AP-übergreifend zu definieren. Naturgemäß fallen zum Training und Validierung der KI-Funktionen große Mengen von Daten an. Diese werden durch den Spezifikationsprozess (vergleiche dort) gezielt angefordert (linker Block in Abbildung 50). Die Datenanforderung geschieht hier auf einem höheren Abstraktionsniveau, d.h. in einer unterspezifizierten Form als Beschreibung der logischen Kontextdimensionen. Das AP2.2 erstellt hieraus konkrete Szenarienbeschreibungen, die im Sinne einer medialen Umsetzung durch 3D-Simulationssysteme in konkrete Bild- und Sensorinformationen umsetzbar sind. Im Sinne von Computergrafiksystemen umschließt dieser Schritt, der operativ von AP2.5 umgesetzt wird den Aufbau von Umgebungsmodellen, Animationen und letztendlich den Prozess des ‚Rendering', also der Bild- und Sensordatenerzeugung.

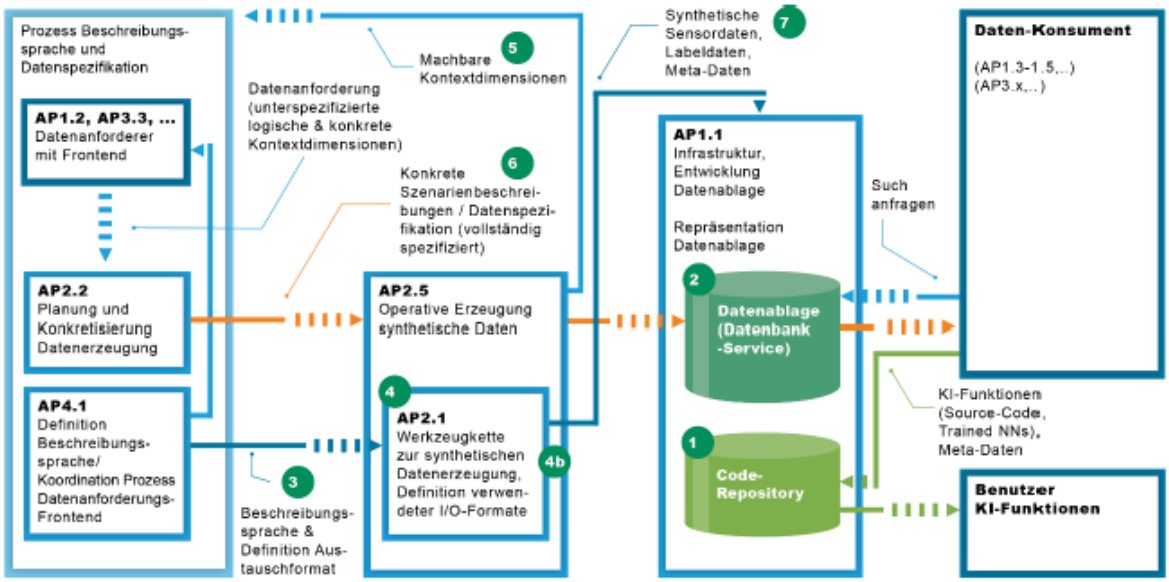


Abbildung 7.12: Logischer Ablauf von Datengenerierung und -fluss

Insbesondere hat sich P5 mit den folgenden Aufgaben beschäftigt, bzw. Ergebnisse erzielt:

- Zusammenstellung und Katalogisierung von Metadaten, die im Projekt entstanden sind
- Spezifikation von weiteren benötigten Metadaten, die nicht in anderen AP's behandelt wurden und insbesondere zum Austausch wichtiger Prozess-Metadaten benötigt wurden.
- Katalogisierung und Wartung einer Liste von 'datenproduzierenden' Funktionsmodulen, einschließlich deren Ein- und Ausgabedaten mit Referenzcharakter für das gesamte Projekt.
- Datenspeicherung, bzw. Spezifikation einer Datenbank
- Versionierung und Spezifikation eines Metadaten Headers, um Metadaten auch in Verarbeitungsketten nachvollziehen zu können

Der P5 Prozess hatte die Aufgabe den Aufbau einer durchgängigen Datenkette zu begleiten. Dies geschah in enger Zusammenarbeit mit Kollegen aus den Arbeitspaketen des TP1+TP2, aber auch im Weiteren den TPs 3+4.

8 Evidenz Workstreams

Motivation

Um die Sicherheit der KI Funktion zur Fußgängererkennung nachweisen zu können benötig es sogenannte Evidenzen für die Sicherheitsargumentation. Dabei können Evidenzen sowohl quantitativ als auch qualitativ sein. Im Projekt wurden sogenannte DNN-spezifischen Sicherheitsbedenken definiert, welche Eigenschaften bzw. Eigenheiten von tiefen neuronalen Netzwerken beschreiben, die eine Einfluss auf deren korrektes Verhalten besitzen. Um nachweisen zu können, dass diese Sicherheitsbedenken genügend ausgeräumt wurden benötigt es verschiedenste Metriken, Methoden und Maßnahmen. Die Evidenz Workstreams stellen dabei das zentrale Rahmenwerk da, um gezielt Evidenzen für die Sicherheitsargumentation zu erheben und entsprechende Teil-Sicherheitsargumentation für die Mitigation der Sicherheitsbedenken zu erstellen. Dabei wurden insgesamt 6 Evidenz Workstreams durchgeführt:

- Unreliable Confidence Information
- Brittleness of DNNs
- Incomprehensible Behavior & Insufficient Plausibility
- Performance Limiting Factors
- ODD Definition and Data Coverage
- NCAP Szenario Benchmark

Dabei wurden insgesamt 9 Ergebnisse definiert, welche von jedem Evidenz Workstream erarbeitet werden. Diese sind in folgender Abbildung dargestellt:

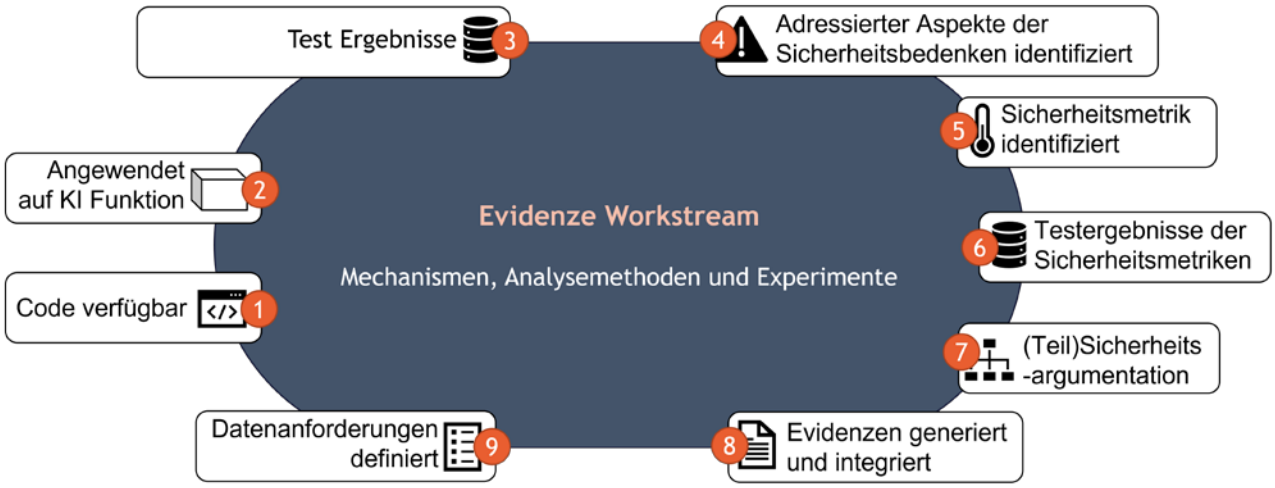


Abbildung 8.1: Neun Ergebnisse, die von jedem Evidenz Workstream erarbeitet werden

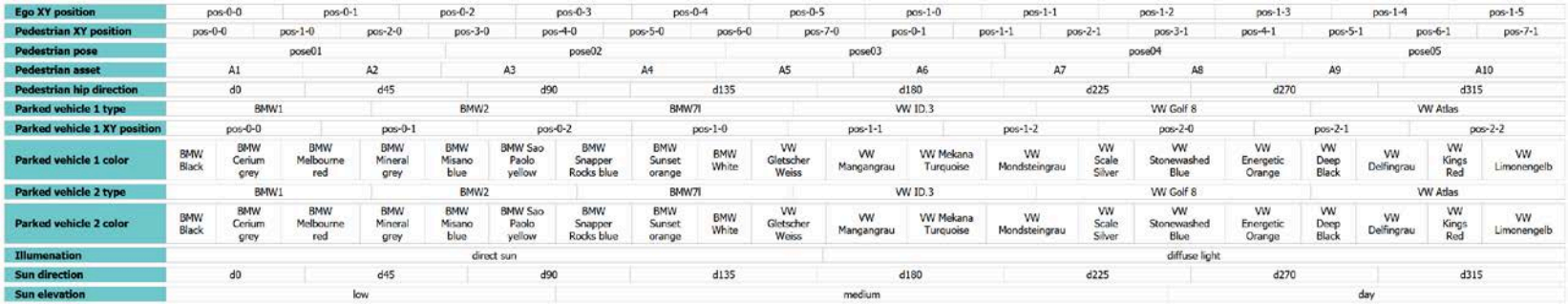
Zunächst muss der Programmcode des Mechanismus oder der Analysemethode vorliegen (1) und auf die Referenz KI Funktion angewendet (2) werden. Die ersten Testergebnisse (3) sollen ebenfalls dokumentiert werden, um die Effektivität des Mechanismus oder der Analysemethode bewerten zu können. Weiterhin beschreibt jeder Evidenz Workstream, welchen Aspekt eines DNN-spezifischen Sicherheitsbedenken adressiert (4) wird. Für die finalen Testergebnisse (6) werden Sicherheitsmetriken (5) definiert, um eine einheitliche Auswertung zu erreichen. Anhand

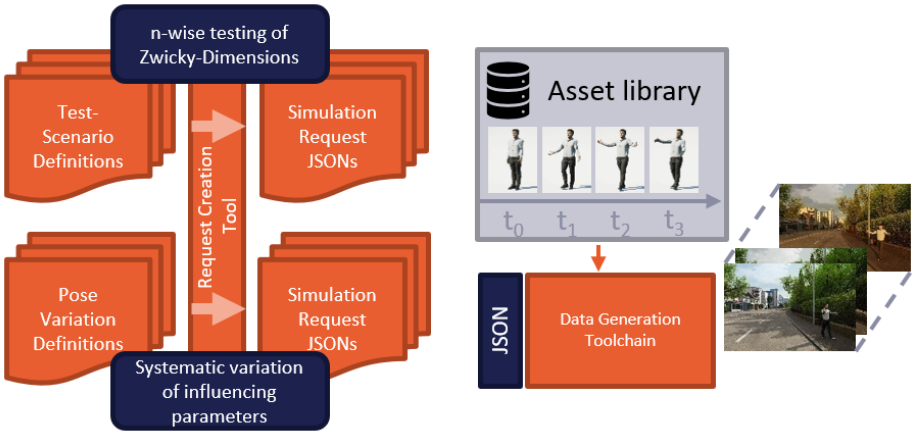
der Testergebnisse kann dann eine Teil-Sicherheitsargumentation erzeugt werden (7), welche dann mit den entsprechenden Evidenzen (8) befüllt wird. Die Teil-Sicherheitsargumentationen werden schließlich in die Gesamtsicherheitsargumentation integriert. Um diese Ergebnisse erreichen zu können ist die Mitarbeit von verschiedenen Experten gefragt:

- Methodenentwickler
- Test-Ingeniuer
- Sicherheits-Ingenieur

Die einzelnen Evidenz Workstreams sind im Nachfolgenden näher beschrieben.

8.1 Evidenz Workstream "Parametrized, safety relevant test scenarios for DNN assessment"

Titel	Parametrized, safety relevant test scenarios for DNN assessment
Beteiligte Partner	ZF Friedrichshafen AG (Organisator/Organisatorin) ZF Friedrichshafen AG, Bosch GmbH (Test Experte/Expertin) IKS Fraunhofer (Safety Experte/Expertin)
Motivation	In der Automobilindustrie werden Euro NCAP basierte Szenarien zur Bewertung der Sicherheit von Fahrzeugen genutzt. Diese Szenarien sind konkret und sehr detailliert beschrieben, um diese reproduzierbar wiederholen zu können und einen einheitlichen Benchmark zwischen unterschiedlichen Fahrzeugherstellern zu ermöglichen. Die Vereinheitlichung der Szenarien und dem damit verbundenen Ziel eine definierte Performanz eines Systems unter konkreten Umständen nachzuweisen, wird auch in diesem EWS verfolgt. Insbesondere sollen Schwachstellen der eingesetzten Objekterkennungsfunktion aufgedeckt und untersucht werden, sodass sich hierzu geeignete Methoden zu Mitigation anwenden lassen.
	 *Abbildung 8.2: "Zwicky Box" mit diskreten Parametern zur Erstellung eines kombinatorischen Testplans*

Titel	Parametrized, safety relevant test scenarios for DNN assessment
	 *Abbildung 8.3: Darstellung des Datenanforderungsprozesses*
Zusammen-fassung der Ergebnisse	Nach Auswertung der Inferenzergebnisse hat sich herausgestellt, dass die Hypothese einer Abhängigkeit von Pose und der damit verbundenen Variation des Seitenverhältnisses mit der Performanz bestätigt werden konnte. Aus den Ergebnissen war klar ersichtlich, dass liegende Fußgänger nicht erkannt wurden und diese Performanzlücke durch Hinzufügen von Bildern mit diesen Posen ausgeglichen werden konnte. Weiter wurde eine Argumentation zum "Safety Concern" SC2.4 - "Unknown behavior in rare critical situations" erstellt, indem die Themen "Unknown", "Rareness" und "Criticality" differenziert argumentiert wurden. Weiter wurde eine Realdatensatzanalyse zur Ermittung einer Auftretenswahrscheinlichkeit diskreter "Zwicky Box" Kombinationen durchgeführt und diese Methodik in die GSN integriert. Veröffentlichungen / Referenzen: 1. Christoph Gladisch, Christian Heinzemann, Martin Herrmann, Matthias Woehrle: *Leveraging combinatorial testing for safety-critical computer vision datasets.* In: Workshop on Safe Artificial Intelligence for Automated Driving (SAIAD) 2020, Seattle (USA), 14.06.2020

8.2 Evidenz Workstream "Analysis and Improvement of DNN Robustness"

Titel	Evidence Workstream: Analysis and Improvement of DNN Robustness
Beteiligte Partner	ZF Friedrichshafen AG (Organisator/Organisatorin) Merantix, FORTISS, Opel, Bosch, Neurocat (Test Experte/Expertin) Valeo, UnderstandAI (Safety Experte/Expertin) VW (Methodenentwickler/entwicklerin bzw. Beitragende)
Motivation	Für die zukünftige Nutzung von tiefen neuronalen Netzen (DNN) in sicherheitskritischen Anwendungen muss eine Robustheit gegenüber Störungen in den Eingabedaten während der Laufzeit nachgewiesen werden. Daher muss die Robustheit durch Anwendung von geeigneten Testdaten und Metriken (ISO/TR 4803 - Annex B) nachgewiesen werden. Weil DNN basierte Funktionen im Betrieb laufend Störungen ausgesetzt sind, ist die Nachweisführung zur Mitigation des Sicherheitsbedenken "Brittleness of DNNs" ein wichtiger Bestandteil der Sicherheitsargumentation.
Beschreibung der Beiträge	Im Rahmen von diesem EWS wurde eine Methode MECH-093532 zur Steigerung der Robustheit durch Augmentierungen untersucht, indem die Performance mittels der mAP Metrik auf unterschiedlich Augmentierten Datensätzen (siehe Abbildung 1) untersucht wurde. Die Augmentierung im Test umfassten dabei folgende Typen: • Sun / Brightness • Random Noise • Local Motion Blur Mit Hilfe dieser Augmentierungen (Abbildung 2) wurde der komplette KI-A Testdatensatz augmentiert und anschließend die Performance mittels der mAP Metrik ausgewertet (Abbildung 3).

Titel	Evidence Workstream: Analysis and Improvement of DNN Robustness

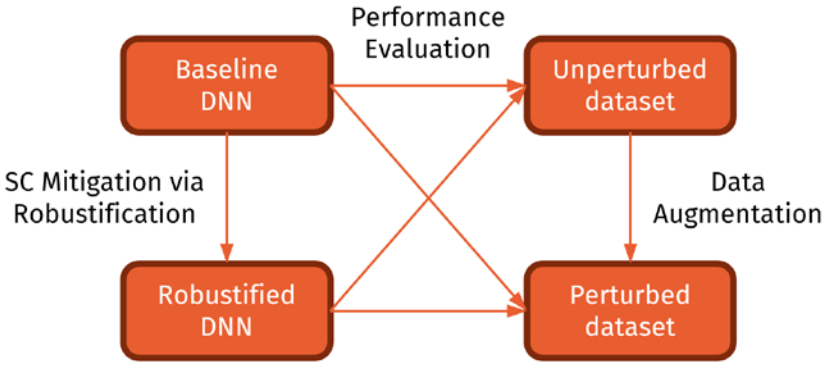
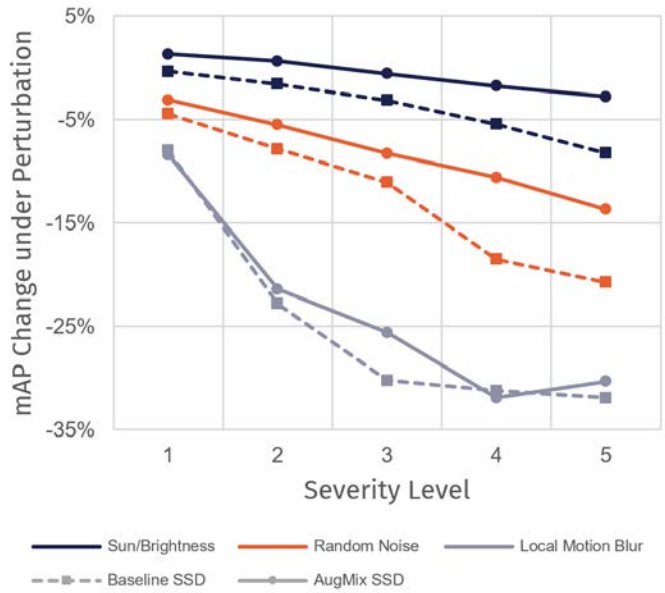


Abbildung 8.4: Auswertung der Performance auf unterschiedlichen Datensätzen

Abbildung 8.5: Beipiele der Augmentierungen in unterschiedlichen Stärken

Titel	Evidence Workstream: Analysis and Improvement of DNN Robustness
	 *Abbildung 8.6: Ergebnisse nach Auswertung mittels mAP*
	In Abbildung 3 ist ersichtlich, dass die mAP Metrik bei dem "robustifizierten" DNN für die beiden Augmentierungen Sun / Brightness und Random Noise eine Wirkung zeigt und wir von einer Steigerung der Robustheit ausgehen können. Die Ergebnisse für die Augmentierung "Local Motion Blur" hingegen zeigen Schwächen bei der Stärke 4 und es ist nicht erkennbar ob wirklich eine Robustheit für diese Art von Störung vorliegt. Dies kann zum einen an der Parametrierung der unterschiedlichen Stärken der Augmentierung liegen oder aber einer mangelnden Generalisierungsfähigkeit des DNN. Im Rahmen von diesem EWS wurde die Generalisierungsfähigkeit nicht abschließend untersucht, es ist aber davon auszugehen, dass durch Hinzufügen von weiteren Augmentierungen während des Trainings die Robustheit des DNNs weiter gesteigert werden kann.

Titel	Evidence Workstream: Analysis and Improvement of DNN Robustness
Zusammenfassung der Ergebnisse	Im Rahmen von diesem EWS wurden unterschiedliche GSNs für die Entwicklung und Absicherung von DNNs entwickelt. Hierbei beschreibt die GSN für die Entwicklung welche Methoden zur Steigerung der Robustheit angewendet werden können, und die GSN zur Absicherung die allgemeine Methodik zum Nachweis einer Robustheit mit Hilfe von "Safety Case Patterns". Die in dem Safety Case Pattern beschriebene Methodik, wurde später in einer Instanziierung genutzt, um konkrete Evidenzen mit Zahlenwerten zu generieren. Hierdurch entsteht ein Ansatz zum Nachweis einer Robustheit durch Anwendung einer Absicherungsmethode MECH-093532 und spezifischen Testmethoden über eine quantifizierbare Metrik, welche in einer formalen Beschreibung wie einer GSN festgehalten ist. Veröffentlichungen / Referenzen: 1. Nikhil Kapoor, Chun Yuan, Serin Varghese, Jonas Löhdefink, Roland Zimmermann, Serin Varghese, Fabian Hüger, Nico Schmidt, Peter Schlicht, Tim Fingscheidt: *A Self-Supervised Feature Map Augmentation (FMA) Loss and Combined Augmentations to Efficiently Improve the Robustness of CNNs.* In: 4. ACM Computer Science in Cars Symposium (CSCS 2020), Ingolstadt (online), 02.12.2020 2. D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, “AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty,” in *Proceedings - 8th International Conference on Learning Representations*, 2020: https://arxiv.org/pdf/1912.02781.pdf

8.3 Evidenz Workstream "Unreliable Confidence Information"

Titel	Evidence Workstream: Unreliable Confidence Information
Beteiligte Partner	BUW (Organisator/Organisatorin) Fraunhofer IAIS (Test Experte/Expertin) ASTech, BUW (Safety Experte/Expertin)

Titel	Evidence Workstream: Unreliable Confidence Information
	BUW, Fraunhofer IAIS, HS-Ruhrwest (Methodenentwickler/entwicklerin bzw. Beitragende)
Motivation	Die Ausgabe von tiefen neuronalen Netzen für Klassifikationsprobleme ist üblicherweise von probabilistischer Natur. Sie gibt Wahrscheinlichkeiten für mögliche Klassenzuordnungen an. Somit können die einzelnen Klassenwahrscheinlichkeiten als Konfidenz angesehen werden, dass das neuronale Netz eine korrekte Vorhersage bezüglich der jeweiligen Klasse trifft. In der Anwendung gibt es häufig Daten, für die das neuronale Netz keine präzise Vorhersage treffen kann. Dies kann eine Folge von unterschiedlichen Ursachen sein, wie zum Beispiel eine starke Variation von Objekten oder auch das Auftauchen von komplett unbekannten Objekten. In solchen Fällen sollte das neuronale Netz idealerweise eine Konfidenz angeben, welche zuverlässig angibt, ob der Vorhersage des Modells vertraut werden kann oder nicht. In der Praxis hat sich jedoch gezeigt, dass neuronale Netze häufig zu hohen Konfidenzen neigen, was häufig auch fehlerhafte Vorhersagen miteinschließt. In diesem Evidence Workstream tragen wir verschiedene Methoden zur Konfidenz Kalibrierung zusammen und erarbeiten ein Vorgehen für Evidenzen zum S unzuverlässigen Konfidenzen.
Beschreibung der Beiträge	Im Rahmen des EWS wurden verschiedene Dimensionen des Safety Concern "Unreliable Confidence Information" identifiziert, darunter besonders hervorzuheben sind Klassifikationsschärfe der Konfidenzschätzung ("Separation"), Kalibriertheit und Detektionsperformance ("MetaFusion-Performance"). Es wurden drei Methoden zur Mitigation des Safety Concern untersucht und evaluiert: • Gradient-Based Quantification of Epistemic Uncertainty for Deep Object Detectors MECH-173679 • Local Uncertainy Realism MECH-030262 • Multivariate Confidence Calibration MECH-043409 Zur Untersuchung der Effektivität der unterschiedlichen Mechanismen wurden zahlreiche Experimente auf Projektdaten, sowie externen Datensätzen durchgeführt, welche Einflüsse auf verschiedene Aspekte der Netzwerkvorhersage und deren Konfidenzschätzung messen. Die verwendeten Metriken wie beispielsweise AuROC ("area

Titel	Evidence Workstream: Unreliable Confidence Information
	under receiver operating characteristic" für Klassifikationsschärfe), ECE ("expected calibration error" für Kalibrierung) oder mAP ("mean average precision" für die resultierende Objektdetektion) spiegeln statistische Resultate wieder. Durch Vergleiche mit den unveränderten Objektdetektionsnetzen als Baseline konnten so bewertbare Tendenzen erkannt werden.

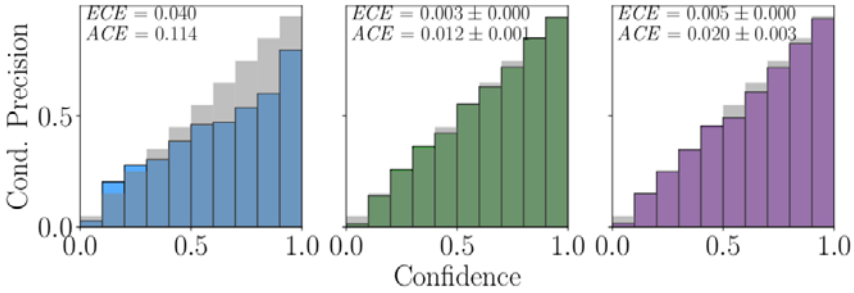


Abbildung 8.7: Kalibrierungsdiagramme der netzwerk-inhärenten Konfidenzschätzung (links), Monte-Carlo-Dropout (mitte) und Gradienten-basierter Konfidenz (rechts) auf externen Daten.

Titel	Evidence Workstream: Unreliable Confidence Information

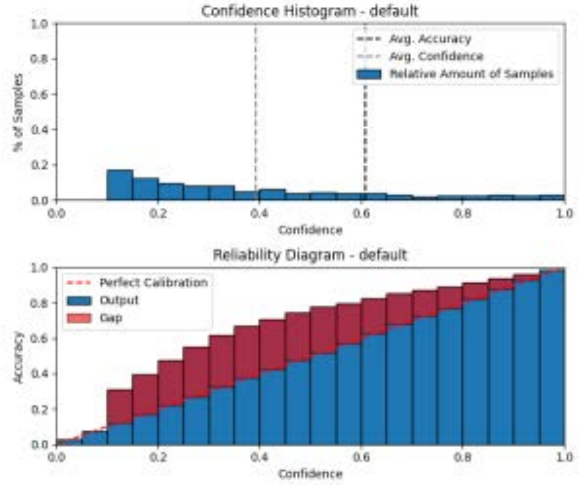


Abbildung 8.8: Bin-Besetzung und Kalibrierungsdiagramm für Projekt-interne Daten.

Die durchgeführten Experimente enthalten weitreichende Vergleiche mit Methoden aus der Literatur, in welchen Unterschiede und Gemeinsamkeiten herausgestellt werden konnten.

Mithilfe des in E4.4.1-3 entwickelten Test-Frameworks für Tests unter semantischer Differenzierung der Testdaten konnten die oben genannten Mechanismen weiterführender Tests unterzogen werden welche zusätzliche Tendenze n und Baseline-Vergleiche ermöglichten; darunter:

- Studien über die Sicherheits-Relevanz der zu detektierenden Fußgänger
- Threshold-Abhängigkeit der Vorhersagegüte
- Korrelation zwischen den geschätzten Konfidenzwerten und der berechneten IoU mit der Ground Truth.

Generierte Evidenzen wurden zu vier GSN-Strukturen verarbeitet.

Titel	Evidence Workstream: Unreliable Confidence Information
Zusammenfassung der Ergebnisse	Im Rahmen des EWS wurden insgesamt vier GSN-Strukturen erstellt, welche zentral HIER abgelegt sind. Zunächst wurde eine Struktur entwickelt, welche die Argumentation über alle Arten von Methoden und Maßnahmen zur Mitigation von unreliable confidence information darstellt. Diese übergreifende Struktur mündet in drei "Hauptstrategien". Für jede dieser drei Hauptstrategien wurde anschließend eine weitere GSN-Struktur entwickelt, welche diese Strategie weiter über Ziele (Goals) und Lösungen (Solutions) verfeinert. In jeder dieser GSN-Strukturen zu den drei Strategien sind die verfügbaren und zu verwendenden Safety Metrics für jede Solution aufgeführt. Die erste Strategie (Strategie 01) liefert Evidences, die durch die Bewertung von Metriken generiert wurden. Die dritte Strategie (Strategie 03) liefert Evidences durch die Anwendung von Testmethoden. Die zweite Methode wurde bisher noch nicht evaluiert (Stand: 09.05.2022). Von den Entwicklern der untersuchten Methoden wurden zudem mehrere Veröffentlichungen geschrieben: • A Novel Regression Loss for Non-Parametric Unvertainty Optimization - J. Sicking et al. LINK • Wasserstein Dropout - J. Sicking et al. LINK • Gradient-Based Quantification of Epistemic Uncertainty for Deep Object Detectors - T. Riedlinger et al. LINK • Multivariate Confidence Calibration for Object Detection - F. Küppers et al. LINK

8.4 Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"

Titel	Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"
Beteiligte Partner	Continental (Organisator/Organisatorin, Test) Fortiss (Safety Experte/Expertin) Continental, Fraunhofer IAIS, Fortiss (Methodenentwickler/entwicklerin bzw. Beitragende)

Titel	Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"
Motivation	Neuronale Netze zur Fußgängererkennung aus visuellen Daten haben die Eigenschaft, dass die Ausgabe für eine Eingabe nicht auf eine verständliche Weise nachvollzogen werden kann. Darüber hinaus werden Detektionen ohne Einbeziehung von etwaigen Plausibilitätserwägungen erstellt. Um diese sog. Safety Concerns im Rahmen von KI-Absicherung strukturiert zu bearbeiten haben die Partner Continental, Fraunhofer IAIS und Fortiss ihre Aktivitäten in TP3 und TP4 zusammen in diesem Workstream gebündelt. Dabei brachten die Partner ihre Expertise in der Sicherheitsargumentation, Hybriden lernen, Verfahren zur Erstellung von visuellen Heatmaps und der Verfahren der toolgestützten Einbringung von Expertenwissen mittel „Visual Analytics“. Alle Verfahren wurden auf verschiedenen Iterationen der KI-Absicherung KI-Funktion evaluiert. Aus den gesammelten Evidenzen wurden zwei GSN-Fragmente mit zugehörigen „Assurance Claim Points“ erstellt und an das TP4 übergeben.
Beschreibung der Beiträge	Insgesamt wurden 3 Ansätze zur Bearbeitung der ausgewählten Safety Conerns evaluiert. Bei der Verwendung von Heatmaps zur Ermittlung der Aufmerksamkeit wurden insbesondere folgende Experimente durchgeführt: • Auswertung der Attentioncluster unter Berücksichtigung der Detektionen • Verbesserung der Detektionen mittels einer Einbeziehung der Aufmerksamkeit des Netzes • Erstellung einer Liste von Texturmerkmalen, die positiv oder zerstörerisch zur Genauigkeit des Detektors beitragen • Untersuchungen zur Verteilung der Aufmerksamkeit in Eingabebildern

Titel	Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"

Abbildung 8.9: Messung der Aufmerksamkeit eines neuronalen Netzes via Merkmalswichtigkeit

Mittels des Visual Analytics Tools wurde eine große Anzahl von Experimenten gefahren. Wichtige Befunde auf dem KI-Absicherung Use-Case waren:

- Defekte in den Daten
- Untersuchung von ungewöhnlichen Fehldetektionen des neuronalen Netzes
- Evidenzen zur Wirksamkeit der gewählten Methode

Titel	Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"

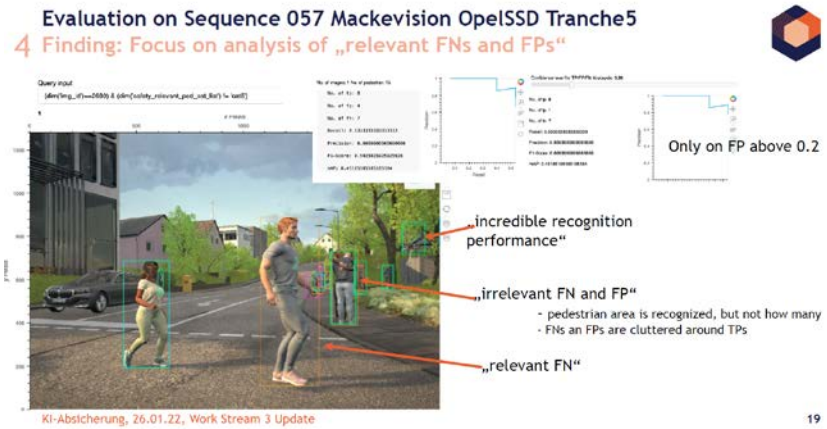


Abbildung 8.10: Assessment des KI-A Use-Cases mittels Visual Analytics

Mit der Methode zum hybriden Lernen und Messung der Stärke der Konzepteinbettung wurde das Label "Fußgänger" in Unterkonzepte anhand der Körperteile eines Menschen zerlegt. Die Erkennung dieser Konzepte wurde als Qualitätsmerkmal definiert. Dabei wurden folgende Experiment durchgeführt:

- Messung der Performanz des Konzeptmodells
- Beibringung von Evidenzen hinsichtlich der Wirksamkeit der Methode hinsichtlich
  - Performanz auf anderen, verwandten Use-Cases
  - State-of-the-Art Analysen

Titel	Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"

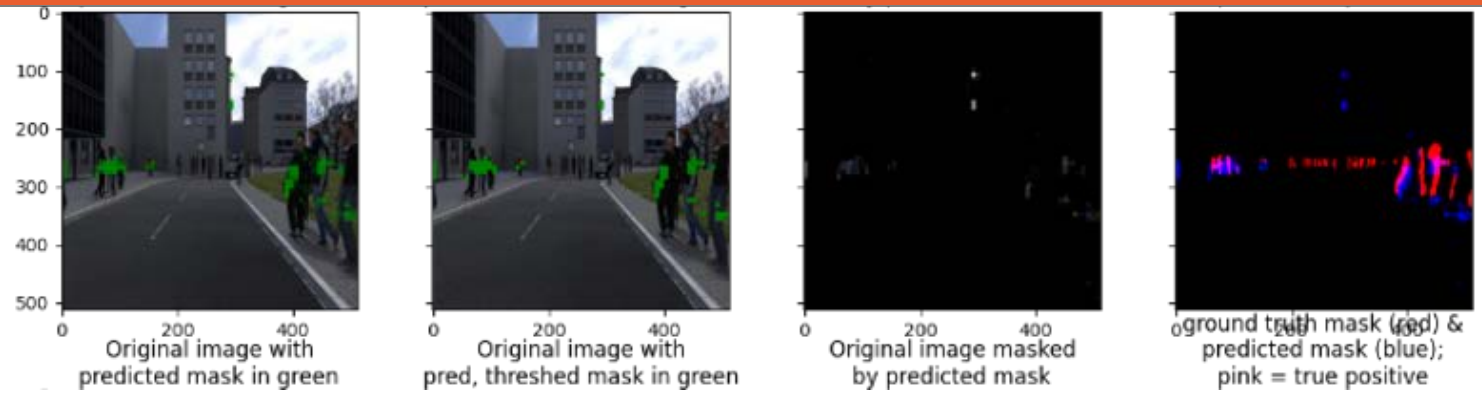


Abbildung 8.11: Resultate Körperteilerkennung zur Messung der Konzepttreue eines Netzes

Die generierten Evidenzen wurden in zwei GSN-Fragmenten für die beiden Safety Concerns verarbeitet. Dabei wurden die Anwendung des hybriden Lernens und des Heatmappings eher im Safety Concern "incomprehensible behavior" verortet, während die Resultate der Experimente aus Visual Analytics eher im Rahmen von "insufficient plausibility" bearbeitet wurden.

Titel	Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"

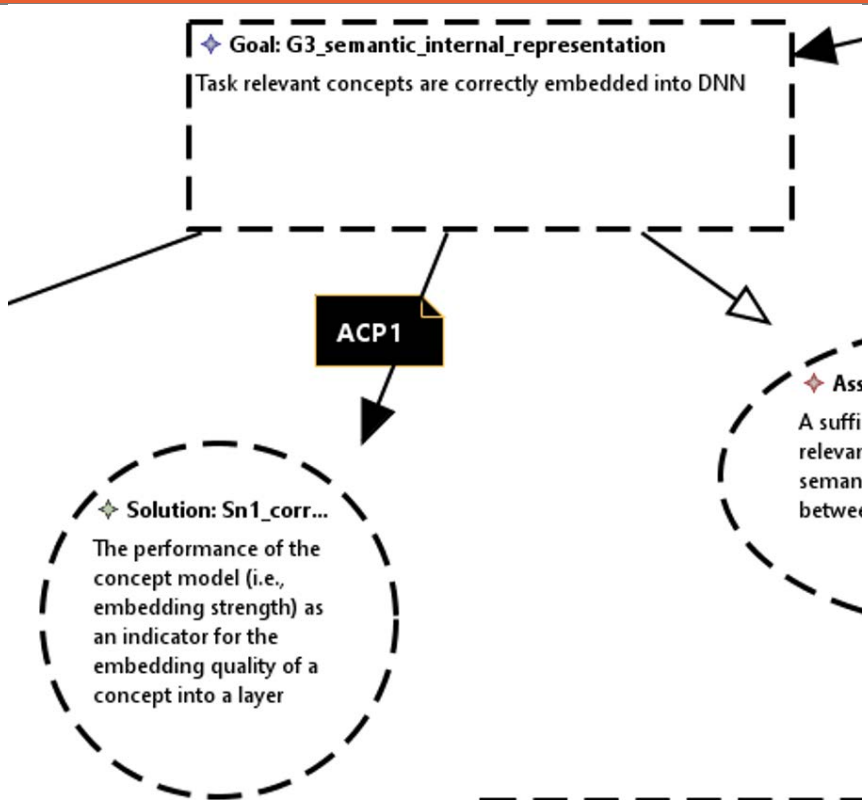
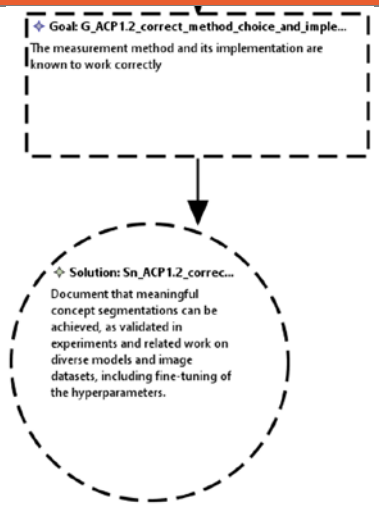


Abbildung 8.12: Verknüpfung Claim und Solution mittels ACP

Die GSN-Graphen verknüpfen die numerischen Resultate mit Claims aus dem Assurance Case. Dabei stellt sich natürlich die Frage in wie weit die genannten Evidenzen tatsächlich geeignet sind, die Claims in konstruktiver Weise zu belegen. Dazu haben wir als Strukturelement Assurance Claim Points (ACP), wo immer es uns im Rahmen dieses Projekts möglich war, eingeführt. In diesen ACPs wird die Unterstützung der behaupteten Verbindung in GSN aufgezeigt.

Titel	Evidence Workstream "Incomprehensible Behavior & Insufficient Plausibility"
	 *Abbildung 8.13: Ausschnitt ACP*
Zusammenfassung der Ergebnisse	Der Evidence Workstream war in der Lage einen ersten, erfolgreichen Brückenschlag zwischen auf der einen Seite den Methoden des Maschinenlernens bzw. der datengetriebenen Analyse und der Formulierung von Safety Concerns und Anforderungen in einem GSN herzustellen. Für die verwendeten Netze und Daten konnten allerdings nicht die gewünschten Eigenschaften nachgewiesen werden und lediglich eine Struktur des GSN der mit den dann verfügbaren Evidenzen konstruiert werden. Insgesamt zeigte sich mit aller Deutlichkeit die Schwierigkeit, in der numerische Resultate in einer sinnvollen Art und Weise in eine Argumentation aufgenommen werden kann.

8.5 Evidenz Workstream "Data Coverage and Data Distribution"

Titel	Data Coverage and Data Distribution
Beteiligte Partner	QualityMinds (Organisator) Bosch, Continental, Volkswagen (Sicherheitsexperten) BMW, Bosch, Fortiss (Methodenentwickler)
Motivation	Eine Voraussetzung für die sichere Verwendung von maschinellem Lernen als Grundlage für künstliche Intelligenz ist die Datenqualität. Schlechte Datenqualität lässt sich unter anderem durch zwei Eigenschaften charakterisieren: Erstens schlechte Abdeckung des Zielraums (einige Daten oder Datenbereiche fehlen im Datensatz) und zweitens Datenverzerrung (Die Verteilung der Daten entspricht nicht der Verteilung im Referenzraum). Diese Probleme sind für unterschiedliche Anwendungen unterschiedlich relevant: Eine Verzerrung mag akzeptabel sein, wenn man ein Datensatz bestimmte seltene Fälle zu test-zwecken überrepräsentiert. Für das Anlernen eine Künstlichen Intelligenz, kann ein verzerrter Trainingsdatensatz allerdings zur dazu führen, dass eine KI die Verzerrung reproduziert. Genauso kann eine schlechte Abdeckung in den Trainingsdaten mag akzeptabel sein, wenn die KI entsprechend gut über "die Lücken" interpolieren kann. Qualitätseigenschaften von Daten müssen also Kontextbezogen in eine Sicherheitsargumentation eingepflegt werden. Dieser Workstream auf das Sicherheitsbedenken "Data distribution is not a good approximation of real world (SC-2.1)".
Beschreibung der Beiträge	Drei Referenzräume wurden zur Gliederung der Evidenzen herangezogen: Physikalisch aufgenommene Daten, die Ontologie, und der latente Raum (eines neuronalen Netzes). Die Ergebnisse wurden in die Sicherheitsargumentationen für "Data Representativity" und "Data Fidelity" aufgenommen. **Referenzraum Physikalische Daten:** • Vergleich von Semantischen Segmentierungen auf KIAB-Daten und physikalisch aufgenommenen Daten mittels EMD (Earth Mover Distance)

Titel	Data Coverage and Data Distribution
	• Vergleiche von Kontrast, Luminanzwerten, Fußgängerverteilungen, und Posenverteilungen zwischen KIAB-Daten und physikalisch aufgenommenen Daten. **Referenzraum Ontologie:** • Abdeckung der Zwicky-boxen • Abdeckung der verschiedenen Ausrichtungen von Personen zur Kamera. • Betrachtung des gemeinsamen Auftretens von Merkmalen (kommen bestimmte *Kombinationen* häufiger oder seltener vor?) **Referenzraum Latente Neuonale Repräsentation:** Es wurde gezeigt, dass Schwerpunktpositionierung (Centroid Positioning) verwendet werden kann um zu überprüfen ob ein Neuronales Netz seinen Latenten Raum auf einem Datensatz "erschöpft".
Zusammenfassung der Ergebnisse	Im Workstream wurden erfolgreich sowohl bereits bestehende Evidenzen aus dem Projekt gesammelt als auch neue erarbeitet. Die gesammelten Evidenzen des Workstreams wurden zum Argumentation von "Data Fidelity" und "Data Representativity" verwendet. Die Einteilung der Ergebnisse nach einem Referenzraum stellte sich als hilfreich heraus. Auch wenn diese Einteilung nicht mehr Explizit in der Sicherheitsargumentation sichtbar ist, war sie sinnvoll für die Sammlung und Erstellung von Evidenzen.

8.6 Evidenz Workstream "Performance Limiting Factors"

Titel	Performance Limiting Factors
Beteiligte Partner	Intel (Organisator/Organisatorin) Bosch (Test Experte/Expertin)

Titel	Performance Limiting Factors
	Bosch, Fraunhofer IKS (Safety Experte/Expertin) VW, EFS, Stellantis, MackeVision, Fraunhofer IAIS (Methodenentwickler/entwicklerin bzw. Beitragende)
Motivation	Performanz-limitierende Faktoren (PLFs) können sich entweder durch einen direkt messbaren physikalischen Effekt oder indirekt durch ein Model eines solches Effektes ausprägen. Beispiele sind: Niedrige Kontrastverhältnisse zwischen einem zu erkennenden Objekt und dessen Hintergrund, Objektabstand zum Sensor oder der Verdeckungsgrad eines Objektes. Diese Faktoren haben vom Wesen her einen Einfluss auf die Performanz eines Perzeptionsalgorithmus. Zum Beispiel ist klar, dass mit steigendem Abstand zum Sensor oder aber bei zunehmenden Verdeckungsgrad die Erkennung eines Objektes ab einem bestimmten Grad nicht mehr möglich ist. Das Ziel dieses Work-Streams war es, die Signifikanz einzelner Faktoren nachzuweisen und Einfluss auf die Perzeptionsperformanz und Eignung zur Absicherung und Verbesserung von Maßnahmen zu untersuchen.
Beschreibung der Beiträge	Die Arbeit in dem WS gliederte sich in drei Teile: Zum einen wurde grundlegende Methoden und Algorithmen zur Messung von PLFs und geeignete Metriken für die Messung der Perzeptionsperformanz ausgewählt, als auch Datensequenzen ausgewählt. Ausgewählt wurde auch die Epistemic Uncertainty (MTRC-857563) zur Messung vor Unsicherheiten im Kontext der PLFs. Weiter wurden Nachweisstrategien zur Überprüfung der Signifikanz der PLFs entwickelt und angewendet. Dadurch konnte eine Reihe von PLFs als wichtig, bzw. signifikant bestätigt werden. Drittens wurde die Eignung von PLFs zur Verbesserung von TP3 Methoden, wie z.B. aktives Lernen untersucht. Ferner wurde eine Sicherheitsargumentation in Form eines Mini-GSN unter Einbeziehung aller Methoden und Erkenntnisse dieses WS entwickelt.
Zusammenfassung der Ergebnisse	Ausgehend von einer ursprünglichen Liste von ca. 16 Kandidaten für PLFs wurden die vielversprechendsten ausgesucht und mithilfe synthetischer und realer Daten (Cityperson) einer PCA Analyse unterworfen. Daraus ergab sich eine hohe

Titel	Performance Limiting Factors
	Signifikanz für Distanz (z.Sensor), sichbare Pixel, Verdeckungsgrad. Die untersuchten Kontrastwerte zeigten noch eine mittleren Einfluss, während die (2d) Position der Objekte kaum oder keinen Einfluss zeigte. Eine Untersuchung der Detektionsleistung über die Parameter Kontrast und Pose der menschlichen Objekte (gemessen wurde Schulter zu Fuß) belegte den starken Einfluss beider Faktoren. Abschließend wurden die entwickelten Methoden zum Zwecke einer Sicherheitsargumentation in einen Mini-GSN Baum überführt. Dieser strukturiert 6, als Entknoten ausgeführte Methoden: PCA, kombinatorische Tests, suchbasiertes Testen, re-training mit gefilterten Daten, DNN Modifikation und Alternativen auf Systemebene.

9 Ausblick

Die in KI-Absicherung erarbeitete Methodik zum Aufbau einer evidenz-basierten Sicherheitsargumentation für KI-basierte Funktionsmodule unter besonderer Berücksichtigung der DNN-spezifischen Sicherheitsbedenken resultiert aus der strukturierte Vernetzung zwischen den verschiedenen beteiligten Experten im Bereich KI, Safety, und Bildverarbeitung. Der damit erzielte fachliche Konsens wäre ohne das Projekt nicht möglich gewesen. Durch die breite Beteiligung von Industrie und Wissenschaft hat sich dieser Konsens als de-facto Standard in den Köpfen der Beteiligten und in den jeweiligen Unternehmen und Wissenschaftseinrichtungen etabliert. Darüber hinaus wird er aktiv in relevante Normungsgremien und Vorhaben eingebracht, sei es in die Definition der ISO PAS 8800 (Safety and Articifical Intelligence) oder in die Fortschreibung des Schwerpunktthemas "Mobilität" der DIN Normungsroadmap KI 2.0.

Die Ergebnisse aus KI-Absicherung wirken auch auf Forschungsprojekte, die über Automotive Anwendungen hinausgehen. In dem Forschungsprojekt "Zertifizierte KI" werden Methoden und Prüfverfahren für die vertrauensvolle Anwendung von KI in verschiedenen Anwendungsbereichen, wie etwa Cloud-Systemen oder in Dienstleistungssektor (z.B. Finanz- und Versicherungswesen) entwickelt. Das wie KI-Absicherung vom BMWK geförderte Projekt "Safetrain" untersucht Absicherungsmethoden und Sicherheitsnachweise für den Einsatz von KI-basierten Funktionsmodulen in fahrerlosen Regionalzügen.

Darüber hinaus werden Bausteine der in KI-Absicherung entwickelten Methodik als konkrete Artefakte sowohl in den Schwesterprojekten der KI-Familie in der VDA Leitinitiative "Vernetztes und autonomes Fahren als auch in der Entwicklung von abgesicherten KI-Funktionen in der Serienproduktion Eingang finden. Dazu zählen:

- die fein-granulare Ontology zur Definition der Operational Design Domain (ODD),
- der umfassende synthetische Datensatz inkl. Metadaten,
- der Baukasten an bewerteten KI-Funktionen und Absicherungsmethoden,
- und die in KI-Absicherung entwickelten Tools, Metriken und Testverfahren.

9.1 Verwertung in den Schwesterprojekten

Die KI-Absicherung ist Teil der KI-Familie (Abb.). Die KI-Familie stellt eine einzigartige Kombination von Projekten dar, die für die deutsche Industrie- und Forschungslandschaft von herausragender Bedeutung sind. Domänenübergreifend legen alle vier Projekte und deren Zusammenspiel den Grundstein für die erfolgreiche Umsetzung von künstlicher Intelligenz für Fahrzeugkonzepte und -systeme der Zukunft.

Kurz gesagt: KI Absicherung zielt darauf ab, den abgesicherten Einsatz von KI im Fahrzeug zu ermöglichen; in KI Wissen wird bereits vorhandenes Wissen für KI nutzbar gemacht. KI Delta Learning steigert die Lernkompetenz der Netze und das Projekt KI Data Tooling wird eine ganzheitliche Datenbasis sowie verschiedene Methoden und Werkzeuge für deren effiziente Nutzung im Rahmen des Trainings und der Validierung von KI-Funktionen im Fahrzeug bereitstellen.

Abbildung 9.1: KI Absicherung im Kontext der KI Familie

In der KI-Absicherung wurden zwei Toolchains mit unterschiedlichen Merkmalen entwickelt, insgesamt wurden 360.000 Frames produziert. Die Frames basieren auf den Anforderungen, die sich aus der engen Zusammenarbeit von Sicherheitsingenieuren, ML-Ingenieuren und Computergrafikexperten ergeben haben. Die Datengenerierung verwendet speziell entwickelte Assets und Motion Capturing. Die Daten wurden inklusive Annotationen und Dataloader den Schwesterprojekten zur Verfügung gestellt, sie sind auch nach Projektende auf der von KI Delta Learning und KI Data Tooling betriebenen Datenplattform zumindest bis zu deren Projektende verfügbar.

Die BIT TS/Intel/Bosch/Valeo-basierte Toolchain von KI Absicherung wird in KI Data Tooling weiter genutzt und verbessert. Weitere Transferthemen aus KI Absicherung sind:

- die Ontologie,
- das Glossar,
- Daten- und Toolchain-Anforderungen,
- Sensormodell-Schnittstellen,
- Labeling sowie Metainformationsspezifikation
- die synthetischen Daten inklusive Metadaten sind für die Schwesterprojekte der VDA L KI-Familie verfügbar. Sie werden auf einer von TPX/DLR beteriebenen Plattform bereitgestellt.

Neben den KI Schwesterprojekten existiert im Rahmen der VDA Leitinitiative noch die V&V-Familie. Zu dieser, auch als Pegasus-Familie bekannten Gruppe gehören die Projekte

- Pegasus: Autonomes Fahren Level 3, abgeschlossen Mai 2019
- VVMethoden: Autonomes Fahren oberhalb Level 3 in komplexen Fahrsituationen
- SetLevel4to5: Simulationsbasiertes Entwickeln und Testen von automatisierten Fahrzeugen als Teil des VVMethoden-Konzeptes, abgeschlossen August 2022

In diesen Projekten wurden bzw. werden Fragen der Validierung und Verifikation autonom fahrender Systeme behandelt und damit bestehen auch hier Anknüpfungspunkte zu KI

Absicherung. Während KI Absicherung die KI-Funktion und damit Einzelkomponenten betrachtet, nimmt die V&V-Familie die gesamte Systemebene in den Blick. Die Zusammenführung bzw. die Einbindung des KI Absicherungs-Konzeptes in die VVMethoden Systembetrachtung ist ein nächster Schritt, der in Folgeaktivitäten untersucht werden muss (s.u.).

Tabelle 9.1 listet alle mit den beiden Familien und den jeweiligen Projekten geteilten Ergebnisse von KI Absicherung auf.

Tabelle 9.1: Liste der mit den Schwesterprojekten geteilten Ergebnisse

Thema	Beispiele	Schwesterprojekte
Begriffsdefinitionen	• Glossare, Abkürzungen, Definitionen • Begriffsregister • Datenbankvariablen • Sensorparameterlisten • Functional Use Cases	• für alle Projekte der VDA-LI
Ontologie	• Alle (Teil-)Definitionen zur Ontologie, Szenarien-Beschreibungssprache & maschinenlesbare Formatdefinition • Zusammenführung der KI-A und VVM Ontologien in einem gemeinsamen Repository • Zwicky Boxen, ODD-Definition	• VVM • SetLevel4to5 • KI Data Tooling • KI DeltaLearning
Sicherheits-argumentation	• Argumentation, Assurance Case, Argumentationsstrategie inkl. Global Structuring Notation (GSN) • Sicherheitsziele	• VVM
Annotations-Spezifikation / Metadaten	• Definition der Bounding Box Formate • Definition des Skelettmodells für Posen • Definition von Metainformationen	• VVM • SetLevel4to5 • KI Data Tooling • KI Delta Learning
Sequenzdefinition / Corner Case Definition	• Anforderungen an Sequenzen • Beispielsequenzen	• VV- Methoden • SetLevel4To5 • KI Data Tooling • KI Delta Learning

Thema	Beispiele	Schwesterprojekte
Material-beschreibungen	• Liste der zu vermessende Materialien • Spezifikation bzw. Abgleich der Messdimensionen	• KI Data Tooling
Grundkontexte	• Austausch von einzelne vorhandenen Grundkontexten, um optimal voneinander zu profitieren Projektübergreifen kompatibel zu sein und es die Nutzungsrechte es zulassen • z.B. Beschreibungsdaten der benutzten Referenzkreuzungen. Konkretisierung mit Experten aus beiden Projekten	• KI Data Tooling
Sensormodell-Schnittstellen	• Kamera-Sensor Schnittstellendefinition • Lidar-Sensorschnittstellen Definition	• KI Data Tooling
Datenformat-/Datensatz-Spezifikationen	• Aufteilung von Datensätzen zB. in Trainings-, Test- und Validierungsdaten • Datensatzmodalitäten bzgl. Sensoreigenschaften (RGB-Bilder, Radar- und Lidar-Daten) • Generierung von selektierten Datensätzen (Bsp: Datensatz mit ausschließlich RBG-Bildern und Radar-Daten)	• KI Data Tooling • KI Delta Learning • VVMethoden
Anforderungen an Daten und Datengenerierung	• Exportierte Anforderungsliste oder „Highlights“ daraus.	• KI Data Tooling • KI Delta Learning • VVMethoden
Spezifikation Toolchain	• Wie sehen die Schnittstellen innerhalb der Toolchain aus • Spezifikation der zu importierenden Daten	• KI Data Tooling
Data Loader (Stk. 11.03.2021)	Inhalt des Repositories „KI-A-dataset“, das den Data Loader und weitere Scripte zur Benutzung der KI-A Daten enthält	• KI Data Tooling
Generierte Frames (Stk. 11.03.2021)	• Die in KI Absicherung generierten synthetischen Daten (Frames und Metainformationen)	• alle LI Projekte

Thema	Beispiele	Schwesterprojekte
	• Sofern das Schwesterprojekt selbst synthetische oder reale Daten erzeugt, ist die Bedingung ein entsprechender Beschluss des Schwesterprojektes, die dort generierten synthetischen und realen Daten ebenfalls KI-A zur Verfügung zu stellen.	

9.2 Nächste Schritte

Aufbauend auf den Ergebnissen von KI-Absicherung sollten weitere Schritte der vorwettbewerblichen Zusammenarbeit der Automobilindustrie und der Wissenschaft zur Sicherung des technologischen Fortschritts im Bereich des autonomen Fahrens angegangen werden. Dazu zählen u.a. die Bereiche:

- Tooling: Skalierbare Umsetzung von Absicherung- und Testwerkzeugen in einer Tool-Chain, Einbindung in MLOps und Entwicklungsprozesse, persistente Entwicklungsumgebung für automated driving, ...
- Realdaten: Absicherung im Zusammenspiel von synthetischen und Realdaten, Erzeugung von Metadaten für Realdaten, Einbeziehung von Live-Flottendaten, gemeinsame Corner Case Datenbank, ...
- Systemebene: Verbindung der Sicherheitsargumentation auf Ebene der Komponenten/KI-Ebene mit Ebene der Systemebene/Fahrzeug oder anderen Fahrzeugen (Car-to-X), Einbeziehung von Redundanzlösungen, ...
- Erweiterung der Sicherheitsargumentation unter Einbeziehung von Behörden und Zulassungsstellen: Probabilistische Sicherheitsargumentation, Bestimmung von Schwellwerten und Akzeptanzkriterien im gesamtgesellschaftlichen Kontext, ...

10 Publikationen, Präsentationen, Buchprojekt

Die Forschungs- und Entwicklungsarbeit hat in den drei Jahren Projektlaufzeit zu über 100 Veröffentlichungen und öffentlichen Präsentationen geführt. Der Großteil entfällt dabei auf Konferenzbeiträge und Papers in Fachzeitschriften.

10.1 Präsentationen bei Konferenzen und Fachtagungen

Beiträge mit Ergebnissen von KI Absicherung wurden bei den großen, etablierten Konferenzen in diesem Feld regelmäßig eingereicht und gehalten.

Tabelle 10.1 gibt eine chronologische Übersicht wieder. Besonders zu erwähnen ist sind sicherlich zwei Beiträge zur WAISE 2020:Das Paper “Revisiting Neuron Coverage and its Application to Test Generation” von den Projektpartnern IAIS und Bosch hat dort den "Best Paper Award" bekommen, dicht gefolgt auf dem zweiten Platz vom Paper “Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks", das ebenfalls in KI Absicherung vom Partner Bosch erstellt wurde.

Tabelle 10.1: Liste der Veröffentlichungen im Rahmen von Konferenzen und Fachtagungen

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
1	Detection of False Positive and False Negative Samples in Semantic Segmentation	Design, Automation and Test in Europe Conference (DATE) 2020	09 Mar 2020
2	Computational validation of perceptional functions	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Safe AI for automated Driving – Workshop (SAIAD) 2020 / Seattle, USA	14 Jun 2020
3	Self-Supervised Stability-Based Filter Pruning to Improve	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Safe AI for automated Driving – Workshop (SAIAD) 2020 / Seattle, USA	14 Jun 2020
4	Unsupervised Temporal Consistency Metric for Video Segmentation in Highly-Automated Driving	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Safe AI for automated Driving – Workshop (SAIAD) 2020 / Seattle, USA	14 Jun 2020
5	Robust Semantic Segmentation by Redundant Networks With a Layer-Specific Loss Contribution and Majority Vote	Workshop on Safe Artificial Intelligence for Automated Driving (SAIAD) 2020	14 Jun 2020
6	Leveraging combinatorial testing for safety-critical computer vision datasets	Workshop on Safe Artificial Intelligence for Automated Driving (SAIAD) 2020	14 Jun 2020
7	Self-Supervised Domain Mismatch Estimation for Autonomous Perception	Workshop on Safe Artificial Intelligence for Automated Driving (SAIAD) 2020	14 Jun 2020
8	Multivariate Confidence Calibration for Object Detection	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Safe AI for automated Driving – Workshop (SAIAD) 2020 / Seattle, USA	16 Jun 2020

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
9	Characteristics of Monte Carlo Dropout in Wide Neural Networks	International Conference on Machine Learning (ICML) - Workshop on Uncertainty & Robustness in Deep Learning 2020	17 Jul 2020
10	SemanticVoxels: Sequential Fusion for 3D Pedestrian Detection using LiDAR Point Cloud and Semantic Segmentation	IEEE International Conference on Multisensor Fusion and Integration (MFI) 2020 / Karlsruhe, Germany	14 Sep 2020
11	Structuring the Safety Argumentation for Deep Neural Networks	SafeComp - Workshop on Artificial Intelligence Safety Engineering (WAISE) / Lissabon, Portugal	15 Sep 2020
12	Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks	SafeComp - Workshop on Artificial Intelligence Safety Engineering (WAISE) / Lissabon, Portugal	15 Sep 2020
13	Revisiting Neuron Coverage and its Application to Test Generation	SafeComp - Workshop on Artificial Intelligence Safety Engineering (WAISE) / Lissabon, Portugal	15 Sep 2020
14	LRPD: Long Range 3D Pedestrian Detection Leveraging Specific Strengths of LiDAR and RGB	IEEE International Conference on Intelligent Transportation Systems (ITSC) 2020	20 Sep 2020
15	Runtime Optimization of a CNN Model for Environment Perception	IEEE Intelligent Vehicles Symposium (IV) 2020	21 Oct 2020
16	Towards Ontology-Based Corner Case Generation – Combining Ontology-Based Domain Modeling with Combinatorial Testing to Generate Semantic Corner Cases in Autonomous Driving	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2020 / Ingolstadt	02 Dec 2020

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
17	A Self-Supervised Feature Map Augmentation (FMA) Loss and Combined Augmentations Finetuning to Efficiently Improve the Robustness of CNNs	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2020 / Ingolstadt	02 Dec 2020
18	DNN Analysis through Synthetic Data Variation	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2020 / Ingolstadt	02 Dec 2020
19	Characterizing Data Sets for training and validation in automated driving	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2020 / Ingolstadt	02 Dec 2020
20	Increasing realism of synthetic datasets through additive sensor and lens artefacts	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2020 / Ingolstadt	02 Dec 2020
21	HPERL: 3D Human Pose Estimationfrom RGB and LiDAR	25th International Conference on Pattern Recognition	10 Jan 2021
22	Online Out-of-Domain Detection for Automated Driving	Dependable und Explainable Learning (DEEL) - Workshop on Machine Learning in Certified Systems 2021	14 Jan 2021
23	Second-Moment Loss: A Novel Regression Objective for Improved Uncertainties	International Conference on Learning Representations (ICLR) 2021	04 May 2021
24	Reevaluating the Safety Impact of Inherent Interpretability on Deep Neural Networks for Pedestrian Detection	Workshop on Safe Artifical Intelligence for Automated Driving (SAIAD) 2021	19 Jun 2021
25	Towards Black-Box Explainability with Gaussian Discriminant Knowledge Distillation	Workshop on Safe Artifical Intelligence for Automated Driving (SAIAD) 2021	19 Jun 2021
26	Exploration of Latent Spaces for Function Agnostic Domain Shift in Automated Driving	Workshop on Safe Artifical Intelligence for Automated Driving (SAIAD) 2021	19 Jun 2021

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
27	An Unsupervised Temporal Consistency (TC) Lossto Improve the Performance of Semantic Segmentation Networks	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Safe AI for automated Driving – Workshop (SAIAD) 2021 / Virtual	19 - 25 Jun 2021
28	Patch Shortcuts: Interpretable Proxy Models Efficiently Find Black-Box Vulnerabilities	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Safe AI for automated Driving – Workshop (SAIAD) 2021 / Virtual	19 - 25 Jun 2021
29	Behavior-Driven Synthesis of Human Dynamics	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) / Virtual	19 - 25 Jun 2021
30	Understanding Object Dynamics for Interactive Image-to-Video Synthesis	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) / Virtual	19 - 25 Jun 2021
31	Strategy to Increase the Safety of a DNN-based Perception for HAD Systems	IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 / Seattle, USA	19 - 25 Jun 2021
32	Verification of Size Invariance in DNN Activations using Concept Embeddings	International Conference on Artificial Intelligence Applications and Innovations (AIAI) 2021 / Kreta, Griechenland	25 - 27 Jun 2021
33	Effectiveness of Object Detection Calibration and Anomaly Detection Integration with respect to Safety-Related Metrics	IEEE International Conference on Dependable System and Networks (DSN) 2021 / Taipei, Taiwan	21 – 24 Jun 2021
34	Software architecture for human-centered reliability assessment for neural networks in autonomous driving	International Conference on Modelling in Industrial Maintenance and Reliability (MIMAR) 2021	29 Jun - 01 Jul 2021
35	PillarSegNet: Pillar-based Semantic Grid Map Estimation using Sparse LiDAR Data	IEEE Intelligent Vehicles Symposium (IV) 2021 / Nagoya, Japan	11 - 15 Jul 2021

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
36	Bayesian Confidence Calibration for Epistemic Uncertainty Modelling	IEEE Intelligent Vehicles Symposium (IV) 2021 / Nagoya, Japan	11 - 15 Jul 2021
37	From a Fourier-Domain Perspective on Adversarial Examples to a Wiener Filter Defense for Semantic Segmentation	International Joint Conference on Neural Networks (IJCNN) 2021	18 - 22 Jul 2021
38	Concept-based Pedestrian Detection	Projektübergreifender Workshop on DNN Interpretability #2 2021	06 Aug 2021
39	An Integrated Approach to a Safety Argumentation for AI-based Perception Functions in Automated Driving	Safecomp - Workshop on Artificial Intelligence Safety Engineering (WAISE) 2021 / York, Großbritannien	07 Sep 2021
40	ScrutinAI: An Iterative Workflow for the Semantic Analysis of DNN Predictions	International Workshop and Tutorial on eXplainable Knowledge Discovery in Data Mining (xKDD) 2021	13 Sep 2021
41	Keynote: Towards Safe AI for Automated Driving	European Dependable Computing Conference (EDCC) 2021	16 Sep 2021
42	Content Disentanglement for Semantically Consistent Synthetic-to-Real Domain Adaptation	IEEE International Conferenz on intelligent Robots and Systems (IROS) 2021 / Prag, Tschechische Republik	27 Sep 2021
43	Unsupervised Traffic Scene Generation with Synthetic 3D Scene Graphs	IEEE International Conferenz on intelligent Robots and Systems (IROS) 2021 / Prag, Tschechische Republik	27 Sep 2021
44	MEAL: Manifold Embedding-based Active Learning	IEEE International Conference on Computer Vision (ICCV) - Workshop on "Embedded and Real-World Computer Vision in Autonomous Driving" (ERCVAD) 2021 / Montreal, Kanada	11 Oct 2021

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
45	Deployment of Deep Neural Networks for Object Detection on Edge AI Devices with Runtime Optimization	IEEE International Conference on Computer Vision (ICCV) - Workshop on "Embedded and Real-World Computer Vision in Autonomous Driving" (ERCVAD) 2021 / Montreal, Kanada	11 Oct 2021
46	Semantic Concept Testing in Autonomous Driving by Extraction of Object-Level Annotations from CARLA	IEEE International Conference on Computer Vision (ICCV) - Workshop on "Embedded and Real-World Computer Vision in Autonomous Driving" (ERCVAD) 2021 / Montreal, Kanada	11 - 17 Oct 2021
47	Entropy Maximization and Meta Classfication for Out-of-Distribution in Semantic Segmentation	IEEE International Conference on Computer Vision (ICCV) 2021 / Montreal, Kanada	11 - 17 Oct 2021
48	iPOKE: Poking a Still Image for Controlled Stochastic Video Synthesis	IEEE International Conference on Computer Vision (ICCV) 2021 / Montreal, Kanada	11 - 17 Oct 2021
49	Adaptive test case selection for DNN-based perception functions	IEEE International Symposium on Systems Engineering (ISSE) 2021 / Wien, Österreich	13 Oct 2021
50	On the Necessity of Explicit Artifact Links in Safety Assurance Cases for Machine Learning	International Symposium on Software Reliability Engineering (ISSRE) 2021	24 Oct 2021
51	Improved Sensor Model for Realistic Synthetic Data Generation	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2021 / Ingolstadt	30 Nov 2021
52	Leveraging Interpretability: Concept-based Pedestrian Detection with Deep Neural Networks	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2021 / Ingolstadt	30 Nov 2021
53	Real-time Uncertainty Estimation Based On Intermediate Layer Variational Inference	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2021 / Ingolstadt	30 Nov 2021

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
54	Towards Safe AI for Automated Driving	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2021 / Ingolstadt	30 Nov 2021
55	Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings	European Conference on Computer Vision (ECCV) 2022 / Tel-Aviv, Israel	30 Nov 2021
56	Geometrically Realistic Adversarial Attacks for LiDAR-based Object Detection via Point Deformations	International Conference on 3D Vision (3DV) 2021 / London	01 Dec 2021
57	High-fidelity procedural data synthesis for validation and training of perception functions	Conference on Visual Media Production (CVMP) 2021 / London, Großbritannien	6 - 7 Dec 2021
58	SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation	Conference on Neural Information Processing Systems (NeurIPS) 2021	06 - 14 Dec 2021
59	Gradient-Based Quantification of Epistemic Uncertainty for Deep Object Detectors	Conference on Neural Information Processing Systems (NeurIPS) 2021	06 - 14 Dec 2021
60	High-fidelity procedural data synthesis for validation and training of perception functions	Conference on Visual Media Production (CVMP) 2021 / London, Großbritannien	6 - 7 Dec 2021
61	Laplace Approximation with Diagonalized Hessian for Over-parameterized Neural Networks	Conference on Neural Information Processing Systems (NeurIPS) 2021	06 - 14 Dec 2021
62	Using ontologies for data set engineering in automotive AI applications	Design, Automation and Test in Europe Conference (DATE) 2022 / Antwerpen, Belgien	14 - 23 Mar 2022
63	Synthetic Data Production Based on a Game Engine for Applications in Automated Driving	Eurographics 2022 / Reims, France	28 Apr 2022

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
64	Deep Variational Data Synthesis for AI Validation	DIN, Erster Workshop "Visuell-explorative Bewertung neuronaler Netze"	04 May 2022
65	Nutzung von Unsicherheiten von KI-Systemen als Teil eines systematisierten Entwicklungsprozesses	safe.tech 2022 / Munich	10 May 2022
66	CertainNet: Sampling-free Uncertainty Estimation for Object Detection	IEEE International Conference on Robotics and Automation (ICRA) 2022 / Philadelphia, USA.	23 May 2022
67	Data related considerations - from BMWK-funded project “KI Absicherung” in VDA LI KI	VDA NA052-00-32-14	30 May 2022
68	ScrutinAI: A Visual Analytics Approach for the Semantic Analysis of Deep Neural Network Predictions	International EuroVis workshop on Visual Analytics (EuroVA) 2022 / Rome, Italy	13 Jun 2022
69	Gradient-Based Quantification of Epistemic Uncertainty for Deep Object Detectors	IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022 / New Orleans, USA	19 - 24 Jun 2022
70	3D-VField: Learning to Adversarially Deform Point Clouds for Robust 3D Object Detection	IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022 / New Orleans, USA	19 - 24 Jun 2022
71	Performance Prediction for Semantic Segmentation by a Self-Supervised Image Reconstruction Decoder	IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022 / New Orleans, USA	20 Jun 2022
72	Is Neuron Coverage Needed to Make Person Detection More Robust?	IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022 / New Orleans, USA	20 Jun 2022
73	High-Resolution Image Synthesis with Latent Diffusion Models	IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022 / New Orleans, USA	23 Jun 2022
74	Workshop safe.ai	Wissenschaftkongress Ingolstadt 2022	01 Jul 2022

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
75	Highly Automated Corner Cases Extraction: Using Gradient Boost; Quantile Regression for AI Quality Assurance	International Conferenz on Data, Science, Technology and Applications (DATA) 2022 / Lissabon, Potugal	11 - 13 Jul 2022
76	Revisiting the Evaluation of Deep Neural Networks for Pedestrian Detection	International Joint Conference on Artificial Intelligence, European Conference on Artificial Intelligence (IJCAI-ECAI), AISafety Workshop 2022 / Wien	24 Jul 2022
77	Feasibility of Inconspicuous GAN-generated Adversarial Patches against Object Detection	International Joint Conference on Artificial Intelligence, European Conference on Artificial Intelligence (IJCAI-ECAI), AISafety Workshop 2022 / Wien	24 Jul 2022
78	Object Permanence in Object Detection; Leveraging Temporal Priors at Inference Time	international Conference on Pattern Recognition (ICPR) 2022	21 - 25 Aug 2022
79	Application of STPA for the Elicitation of Safety Requirements for a Machine Learning based Perception Component in Automotive	SafeComp 2022 / Munich	06 - 09 Sep 2022
80	Automating Safety Case Change Impact Analysis for Machine Learning Components	SafeComp 2022 / Munich	06 - 09 Sep 2022
81	Safety-Related Metrics: Assessing the Calibration of a Neural Network and Anomaly Detection	SafeComp 2022 / Munich	06 - 09 Sep 2022
82	Uncertainty Quantification and Resource-Demanding; Computer Vision Applications of Deep Learning	The John von Neumann Institute for Computing (NIC) Symposium 2022 / Jülich	29 - 30 Sep 2022
83	Laplace Approximation for Faster Uncertainty Estimation in Object Detection	IEEE International Conference on Intelligent Transportation Systems (ITSC) 2022 / Macau, China	08 Oct 2022

#	Titel der Publikation	Konferenz, Veranstaltung	Datum
84	Parametric and Multivariate Uncertainty Calibration for Regression and Object Detection	European Conference on Computer Vision (ECCV) 2022 / Tel-Aviv, Israel	23 - 27 Oct 2022
85	Improving Predictive Performance and Calibration by Weight Fusion in Semantic Segmentation	European Conference on Computer Vision (ECCV) 2022 / Tel-Aviv, Israel	23 - 27 Oct 2022
86	Balancing Expert Utilization in Mixture-of-Experts Layers Embedded in CNNs	European Conference on Computer Vision (ECCV) 2022 / Tel-Aviv, Israel	23 - 27 Oct 2022
87	Gradient-Based Quantification of Epistemic; Uncertainty for Deep Object Detectors	European Conference on Computer Vision (ECCV) 2022 / Tel-Aviv, Israel	23 - 27 Oct 2022
88	Towards Improved Intermediate Layer Variational Inference for Uncertainty Estimation	Workshop on Safe Artificial Intelligence for Automated Driving (SAIAD) 2022 / Tel Aviv	24 Oct 2022
89	Validation of pedestrian detectors by classification of visual impairing factors	Workshop on Safe Artificial Intelligence for Automated Driving (SAIAD) 2022 / Tel Aviv	24 Oct 2022
90	Adversarial Vulnerability of Temporal Feature Networks for Object Detection	Workshop on Safe Artificial Intelligence for Automated Driving (SAIAD) 2022 / Tel Aviv	24 Oct 2022
91	Suppress with a Patch: Revisiting Universal Adversarial Patch Attacks against Object Detection	IEEE International Conference on Electrical, Computer, Communications and Mechatronics Engineering 2022, Maldives	16 - 18 Nov 2022
92	SynPeDS – A Synthetic Dataset for Pedestrian; Detection in Urban Traffic Scenes	Association for Computer Machinery (ACM) - Symposium Computer Science in Cars (CSCS) 2022 / Ingolstadt	08 Dec 2022
93	Gradient-Based Quantification of Epistemic Uncertainty for Deep Object Detectors	IEEE CVF Winter Conference on Applications of Computer Vision (WACV) 2023 / Hawai	03 - 07 Jan 2023

10.2 Publikationen in Fachzeitschriften

Tabelle 10.2: Liste der Veröffentlichungen in Fachzeitschriften

	Titel der Publikation	Journal	Datum
1	Die Gefahren lauern vor allem hinter den Ecken - Corner Cases und ihre Tücken	German Testing Magazin	April 2020
2	Testing Deep Learning-based Visual Perception for Automated Driving	Journal ACM Transactions on Cyber-Physical Systems, Special Issue on Artificial Intelligence and Cyber-Physical Systems	Herbst 2021
3	MASS: Multi-Attentional Semantic Segmentation of LiDAR Data for Dense Top-View Understanding	IEEE Transactions on Intelligent Transportation Systems	2022
4	Concept Embedding Analysis: A Review	Artificial Intelligence Review	/
5	What Should AI See?	AI and Ethics, Arxiv	/

10.3 Weitere Ergebnisverbreitung

Darüber hinaus wurden für die Ergebnisverbreitung auch zusätzliche Kanäle genutzt um das Fachpublikum zu erreichen und die im Projekt erarbeiteten Konzepte auf breiter Front bekannt zu machen.

- Die Veröffentlichungen sind alle auf der externen Projektwebseite aufgelistet, die auch noch ein ige Zeit über das Projektende hinaus online bleibt: https://www.ki-absicherung-projekt.de/veroeffentlichungen
- Zu Beginn der Projektlaufzeit wurde im Rahmen von AP 3.1 "Tracking des State-of-Research in den Bereichen Bewertung, Plausibilisierung und Erklärung von KI-Methoden" eine umfangreicher Übersicht des aktuellen State of the Art erstellt und als Whitepaper veröffentlicht (https://arxiv.org/abs/2104.14235).
- Die hohe Qualität dieses White Papers führte zu der Idee mit der breiten Expertise im Konsortium ein Buchprojekt zu starten. Aufbauend auf der State of the Art Bericht des Whitepapers, dass als Anfangskapitel einging, wurden weitere Kapitel zu den verschieden Aspekten zusammengestellt. Das entstandenen Buch “Deep Neural Networks and Data for Automated Driving – Robustness, Uncertainty Quantification, and Insights Towards Safety” (Springer, https://link.springer.com/content/pdf/10.1007/978-3-031-01233-4.pdf, Editoren: Fingscheidt, Gottschalk, Houben) ist inline und als gedruckte Ausgabe verfügbar.
- In der Automobiltechnischen Zeitung (ATZ, Ausgabe 7-8 2022) wurde eine Artikelreihe zu Themen der Projekte der VDA Leitinitiative veröffentlicht. Der sechsseitiger Beitrag von KI Absicherung "Methodik zur Absicherung von KI im Fahrzeug" ist nun auf der Homepage

verfügbar und wurde beim Abschlussevent auch als Sonderdruck in Deutsch und Englisch für das Publikum bereitgestellt.

- Zu Projektabschluss wurde ein 4-minütiger Film zum Projekt und seinen Ergebnissen produziert, der auf dem Abschlussevent erstmals präsentiert wurde und seitdem unter https://www.youtube.com/watch?v=QdxmGtSb5_s öffentlich zugänglich ist.
- Auch die Präsentationen der Abschlussveranstaltung sind auf der Homepage verfügbar und ermöglichen einen guten Überblick über die Ergebnisse aus den Teilprojekten (https://www.ki-absicherung-projekt.de/final-event-results).

10.4 Präsentation bei externen Stakeholdern

Zusätzlich wurde das Projekt bzw. Ergebnisse oder Teilergebnisse des Projektes bei verschiedenen Veranstaltungen externer Stakeholder präsentiert, um das in KI Absicherung erarbeitete Konzept bekannt zu machen und den Einfluss auf das Fachgebiet zu erhöhen. Zum Teil wurden dazu auch Vertreter externer Stakeholder auch zu Projektveranstaltungen eingeladen. Dies waren z.B. andere Projekte zum Thema Absicherung:

Stakeholder oder Veranstaltung	Datum	Partner
Vd TÜV	21.02.2020	IAIS
BSI	24.03.2020	IAIS
ERCIM	26.05.2020	IAIS
TÜV Süd	23.06.2020	IAIS
XR Expo	26.06.2020	Mackevision
OPENLabel	23.07.2020	DFKI, Bosch
TÜV AI Conference	05.10.2020	IAIS
The connected Car and Autonomous Driving	26.10.2020	IAIS
Fraunhofer Solution Days	27.10.2020	IAIS
ML4 Virtual Vehicle Developement	25.02.2021	BMW
Heidelberger Bildverarbeitungsforum	02.03.2021	Volkswagen
BMWI/VDA, "Autogipfel": Durch Kooperation an die Spitze. Die Automobilindustrie gestaltet den digitalen Wandel	03.03.2021	Volkswagen
Projekt Confiance.AI	15.04.2021	IAIS
DIN, Workshop "Ansätze zur Prüfung von KI-Systemen"	23.06.2021	IAIS

Stakeholder oder Veranstaltung	Datum	Partner
DNN InterpretabilityWorkshop	26.08.2021	Opel
Fraunhofer Solution Days	07.08.2021	IAIS
WAISE Workshop	07.09.2021	IAIS
EDCC Keynote	16.09.2021	Volkswagen
Projektr Delta Learning - Halbzeitveranstaltung	07.10.2021	Bosch, ZF
ITS World Congress Hamburg	11.10.2021	Bosch
Projekt Safetrain	23.02.2022	IAIS
BSI	18.10.2022	IAIS

10.5 Presseberichte

Die Aktivitäten zur Ergebnisverbreitung haben auch dazu geführt, dass nicht nur Fachmedien zu KI Absicherung berichtet haben. Die meisten Artikel sind online verfügbar.

- https://www.die-stadtzeitung.de/index.php/2019/09/06/bergische-uni-beteiligt-sich-an-ki-absicherung/
- https://idw-online.de/de/news722994
- https://www.uni-heidelberg.de/de/newsroom/kuenstliche-intelligenz-fuer-automatisiertes-fahren
- https://digitales.nrw/de/aktuelles/nachrichten/bergische-uni-beteiligt-sich-forschungskonsortium-ki-absicherung
- https://www.iais.fraunhofer.de/de/presse/presseinformationen/presseinformationen-2020/presseinformation-200526.html
- https://idw-online.de/de/news747957
- https://www.it-daily.net/it-management/digitale-transformation/24352-ki-absicherung-wie-autonomes-fahren-sicherer-wird?highlight=autonomes%20fahren
- https://www.internationales-verkehrswesen.de/ki-absicherung-wie-autonomes-fahren-sicherer-wird/
- https://www.wissenschaftsregion-bonn.de/news-termine/news/news-details/pm7374-ki-absicherung-wie-autonomes-fahren-sicherer-wird/
- https://www.abitur-und-studium.de/Blogs/Universitaet-Wuppertal/Bergische-Uni-beteiligt-sich-an-Forschungskonsortium-KI-Absicherung
- https://www.elektroniknet.de/elektronik-automotive/assistenzsysteme/ki-funktionsmodule-machen-autonomes-fahren-sicherer-176834.html

- Elektronik automotive Ausgabe 7-8/20
- https://www.intelligent-mobility-xperience.com/xxx-a-949073/
- https://www.industry-of-things.de/autonomes-fahren-bis-2022-soll-sicherheit-garantiert-werden-a-954649/
- https://www.egovernment-computing.de/autonomes-fahren-bis-2022-soll-sicherheit-garantiert-werden-a-957316/
- https://www.elektronikpraxis.vogel.de/autonomes-fahren-bis-2022-soll-sicherheit-garantiert-werden-a-957360/