KI Absicherung

Forschungsprojekt
KI-Absicherung

Automatisierte **F**ahrfunktionen

Schlussbericht

Beitrag des Zuwendungsempfängers:

Valeo Schalter und Sensoren GmbH
Laiernstrasse 12
74321 Bietigheim-Bissingen

zu den Teilprojekten:

TP1 – KI-Funktion
TP2 – Generieren von synthetischen Lern- und Testdaten
TP4 – Methoden und Maßnahmen zur Absicherung von KI
TP4 – Gesamtheitliche KI-Absicherungs-strategie
TP5 – Projektmanagement und Dissemination

Laufzeit: **01.07.2019 – 30.06.2022**

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Wirtschaft und Klimaschutz unter dem Förderkennzeichen 19A19005H gefördert.

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

Ein Projekt entwickelt von der VDA Leitinitiative autonomes und vernetztes Fahren

Bietigheim-Bissingen, 12.12.2022

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

1. Kurzdarstellung

1.1. Aufgabenstellung

Eine der größten ungelösten Herausforderungen bei der Erforschung „Künstlicher Intelligenz“ (KI) und der limitierende Faktor ihrer intensiven Nutzung im Bereich des autonomen Fahrens stellt die Absicherung von KI-basierten Funktionen dar.

Ziel des Vorhabens war die Entwicklung und Untersuchung von Methoden und Maßnahmen für die Absicherung KI-basierter Funktionen für das automatisierte Fahren. Es wurde zudem eine Argumentation für eine abgesicherte KI-Funktion entwickelt.

Mit Hilfe der im Vorhaben gewonnenen Erkenntnisse soll durch Kommunikation mit normativen Gremien und Zertifizierungsstellen ein Industriekonsens bezüglich einer KI-Teststrategie unterstützt werden.

Für das Erreichen dieser Zielsetzung war es erforderlich, dass Experten aus bisher weitgehend unabhängig voneinander agierenden Fachrichtungen der KI-Algorithmik, der 3D-Visualisierung und Animation sowie der funktionalen Sicherheit einen Lösungsansatz erarbeiteten. Für das Vorhaben KI-Absicherung konnte ein schlagkräftiges Konsortium aus Industrie- und Wissenschaftspartner, welches die führenden Köpfe der benannten Fachrichtungen zusammenbringt, geschmiedet werden.

Der verantwortungsvolle Umgang mit „Künstlicher Intelligenz“ im Kontext von sicherheitsrelevanten Funktionen stellte eine besondere Herausforderung dar. Eine stringente Argumentationskette aufzubauen, die aus Expertensicht eine Absicherbarkeit von KI-Modulen hinreichend begründet, war ebenso erforderlich wie die Entwicklung von Methoden und Maßnahmen, die dazu geeignet sind, die nicht-funktionalen Eigenschaften der funktionalen Sicherheit über direkte oder indirekte Messmethoden zu bestimmen und zu bewerten.

Ziel war es, den derzeitigen Stand der Technik soweit voranzutreiben, dass erstmals ein gangbarer und im Expertenkreis anerkannter Weg (Code of practice) hinsichtlich der Absicherbarkeit von KI-Modulen aufgezeigt wird. Dabei sollte das nicht ausschließbare Restrisiko bei der Verwendung einer für den Menschen intransparent erscheinenden und damit häufig als „Blackbox“ bezeichneten Technologie besser bestimmbar bzw. abschätzbar werden. Dem Bild folgend besteht das zentrale Anliegen des Vorhabens KI-Absicherung darin, KI-Module transparenter zu machen bzw. aus einer „Blackbox“ eine immer hellere „Greybox“ werden zu lassen – eine Box, die sich von ihrem Verhalten her über verschiedenste Methoden und Maßnahmen gezielt beobachten, bewerten und beeinflussen lässt.

Analog zu biometrischen Messmethoden, die den Zustand eines Menschen bzw. eines Lebewesens bewerten oder überwachen, gilt es, KI-Messmethoden zu entwickeln und deren Bedeutung und Aussagekraft zu bestimmen. In der Medizin lassen sich ebenso wie bei der KI-Technologie die Methoden in verschiedene Klassen einteilen. So gibt es Diagnoseverfahren, die ausschließlich von außen Anzeichen für abnormales

Verhalten (Anomalie) beobachten. Andere wiederum brauchen mehr oder weniger einen direkten Kontakt, um z.B. die Körpertemperatur oder die Sauerstoffkonzentration zu messen oder bedienen sich vor der Messung injizierter Kontrastmittel, um die Messgenauigkeit zu erhöhen. Das Spektrum reicht bis hin zu invasiven Verfahren, die in den Organismus oder das System eingreifen. Die Aufteilung der im vorliegenden Projektvorhaben betrachteten Methoden und Maßnahmen folgte dieser Analogie, durch eine Unterteilung in modulverändernde, introspektive und externe Methoden und Maßnahmen. Diese werden ergänzt mit aggregierten Methoden und Maßnahmen, also der Kombinatorik der ersten drei genannten Klassen.

Mit diesem grundlegenden Ergebnis im Bereich der Methodik kann KI-Absicherung nicht nur die Sicherheit und Akzeptanz zukünftiger Fahrzeugtechnologien erhöhen, es schafft für die deutsche Automobil- und Zulieferindustrie auch einen wichtigen Baustein für ihre Wettbewerbsfähigkeit in der durch Digitalisierung getriebenen Ökonomie.

———

Inhaltliche Schwerpunkte des Partners Valeo

Valeo wird sich in allen vier inhaltlichen Teilprojekten einbringen. Schwerpunkte im ersten Teilprojekt setzt Valeo in der 3D Bounding Box Fußgängererkennung. Diese Erkennung wird im AP1.3 (Co-Lead-Rolle) auf Basis einer monokularen Kamera erzeugt und in AP1.4 (Lead-Rolle) auf Basis einer Fusion von Kamera und Lidar Sensoren. Im Teilprojekt 2 liegt der Fokus auf dem Einsatz von Transfer-Learning Methoden. In AP2.3 (Co-Lead-Rolle) werden diese Techniken eingesetzt, um die KI-Funktion auf ein geändertes Sensor-Setup anzupassen. In AP2.4 (Co-Lead-Rolle) untersucht Valeo die Anpassung der KI-Funktion bezüglich der Qualität der synthetischen Trainingsdaten. Im dritten Teilprojekt konzentriert sich Valeo auf die Entwicklung von introspektiven Methoden und Maßnahmen mittels Plausibilisierungstechniken (AP3.4, Lead-Rolle). Im Teilprojekt 4 liegt der inhaltliche Schwerpunkt von Valeo in der Strukturierung und Formalisierung des Eingaberaumes (AP4.1, Lead-Rolle).

Die konkreten Arbeitsaufgaben des Partners Valeo können zudem dem Kapitel 1.3 „Planung und Ablauf des Vorhabens“ entnommen werden.

1.2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Für das Vorhaben KI-Absicherung ergaben sich im Hinblick auf eine Reihe von Aspekten spezifische Herausforderungen, die es im Rahmen des Designs der Projektorganisation (Aufbau- und Ablauforganisation) entsprechend zu berücksichtigen galt. So, z.B. in Bezug auf die Relevanz der bearbeiteten Thematik.

Aufgrund Relevanz und Dringlichkeit des Themas für den Wirtschaftsstandort Deutschland strebte das Konsortium eine deutlich über das Einzelprojekt hinausgehende Wirkung an. Als Teil der VDA-Leitinitiative ist es Teil der AVF-Roadmap für eine kohärente und konvergente Technologie- und Methodenentwicklung [4].

Das Vorhaben KI-Absicherung spannt thematisch die folgenden beiden Wirkdimensionen auf:

• Die Potenziale der Schlüsseltechnologie KI sollen für die deutsche Automobilindustrie im Besonderen und den Wirtschafts- und Wissenschaftsstandort Deutschland im Allgemeinen verfügbar gemacht werden.

Im Selbstverständnis ist KI-Absicherung Teil der LI-Projektfamilie „KI-ML“[1] und damit eines Clusters von Projekten, in dem die wesentlichen und aufeinander bezogenen Bausteine für dieses Technologiefeld erarbeitet werden.

• Es werden neue, umfassende Ansätze und Lösungen für die Absicherung einer Fahrfunktion für automatisierte und vernetzte Fahrzeuge entstehen.

Im Vorhaben KI-Absicherung wird eine Nachweismethodik zur quantitativen und qualitativen funktionalen Absicherung einzelner KI-Funktionen auf der Ebene der Funktionsblöcke erarbeitet. Damit arbeitet es komplementär zu den Vorhaben der LI-Projektfamilie VVM[2]. Dort werden Validierungskriterien und -methoden für die gesamte Funktionskette in automatisiert fahrenden Fahrzeugen (maßgeblich beschränkt auf die Ebene der Systemarchitektur) definiert und damit Systemsicherheitsziele konkretisiert.

Aus der Zusammenführung der auf beiden Ebenen gewonnenen Erkenntnisse u.a. im Hinblick auf die Formulierung von Anforderungs-, Güte- und Qualitätskriterien (Safety Goals) wird ein deutlicher Erkenntnissprung für die Absicherung von AF-Systemen erwartet.

Um dieses Potential heben zu können stellen sich besondere Anforderungen an die Vernetzung und den Ergebnistransfer – sowohl im Hinblick auf die projektinterne Zusammenarbeit als auch im Hinblick auf die Kooperation nach außen. Über das projektinterne Management hinaus sollten deshalb weitergehende neue Kooperationsformate etabliert werden und ein Austausch über Projektgrenzen hinweg erfolgen.

Für das Management und die Projektorganisation des Vorhabens KI-Absicherung ergaben sich daraus neue zentrale Anforderungen:

[1] https://ki-familie.vdali.de/
[2] https://www.vvm-projekt.de/

(1) Eine vertragliche Grundlage, die eine erweiterte Nutzung der Ergebnisse möglich macht.

Hier wurden in KI-Absicherung unterschiedliche Stufen der Vertraulichkeit und Nutzungsmöglichkeiten für die geplanten Ergebnisse zugrunde und im Konsortialvertrag verankert. Angestrebt wurde eine neue „Musterregelung“, die es ermöglicht, sowohl den legitimen Interessen einer beschränkten Ergebnisnutzung der forschungstreibenden Partnern Rechnung zu tragen als auch die Öffnung für externe Dritte für die angestrebte Push-Wirkung auf den Kompetenzaufbau am Standort Deutschland.

(2) Eine systematische Herangehensweise auf einer übergeordneten Ebene für die gesamte Wirkkette der Projekte und ihrer Ergebnisverbreitung.

Neu ist hier, dass ein zielorientierter Ergebnisaustausch zwischen den entsprechenden Projekten ermöglicht werden sollte. Entsprechend muss die Steuerung der „verwandten“ Projekte in die Lage versetzt werden, den Projektfortschritt auch hinsichtlich der Einpassung in den Gesamtkontext zu überprüfen und gegebenenfalls Maßnahmen zu ergreifen, um auch den über das Einzelprojekt hinausgehenden Mehrwert einzulösen. Hier setzte KI-Absicherung auf eine übergeordnete gemeinsame Abstimmung von Fördermittelgeber und Industrie innerhalb einer Projektfamilie.

In Bezug auf die technische Koordination

Das Vorhaben KI-Absicherung zeichnet sich durch eine starke inhaltliche Vernetzung der Teilprojekte aus.

Die inhaltlichen Interdepenzen zwischen den inhaltlich-arbeitenden Teilprojekten TP1-4 und auf Ebene der definierten Ergebnisse wurden vorhabenübergreifend in einer Matrix zusammengetragen, welche Bestandteil der Vorhabenbeschreibung ist.

KI-Absicherung stand darüber hinaus vor der Herausforderung einer nahezu durchgehenden Parallelisierung der Arbeiten in einer vergleichsweise kurzen Bearbeitungszeit von 36 Monaten.

Im Hinblick auf die Ergebnisverbreitung

Aufgrund der hohen Bedeutung der Schlüsseltechnologie KI und der Größe des Vorhabens wird das Vorhaben sowohl auf Seiten des Fördermittelgebers BMWK als auch in der Industrie einen Leuchtturmcharakter haben.

Für die vielfältigen Zielgruppen gilt es geeignete Kommunikationsinhalte, -materialien und -kanäle zu schaffen. Aus der Verankerung innerhalb der LI-Projektfamilie heraus gab es hohe Abstimmungsbedarfe im Hinblick auf die Inhalte und das Timing der Außenkommunikation. Eine Bündelung von Aktivitäten in Form von gemeinsamen Events erscheint hier als Ziel führend.

Schließlich ergab sich mit Blick auf den Forschungsgegenstand und das Forschungsergebnis – u.a. die Implementierung von Deep Learning Modellen – die

Herausforderung, neue Demoformate für Sichtbarkeit und Visualisierung der Ergebnisse nach außen, jenseits von Fahrdemonstratoren, zu entwickeln.

Projektorganisation

Die Projektorganisation teilt sich in die strategische und operative Ebene und ist in Abbildung 1 dargestellt.

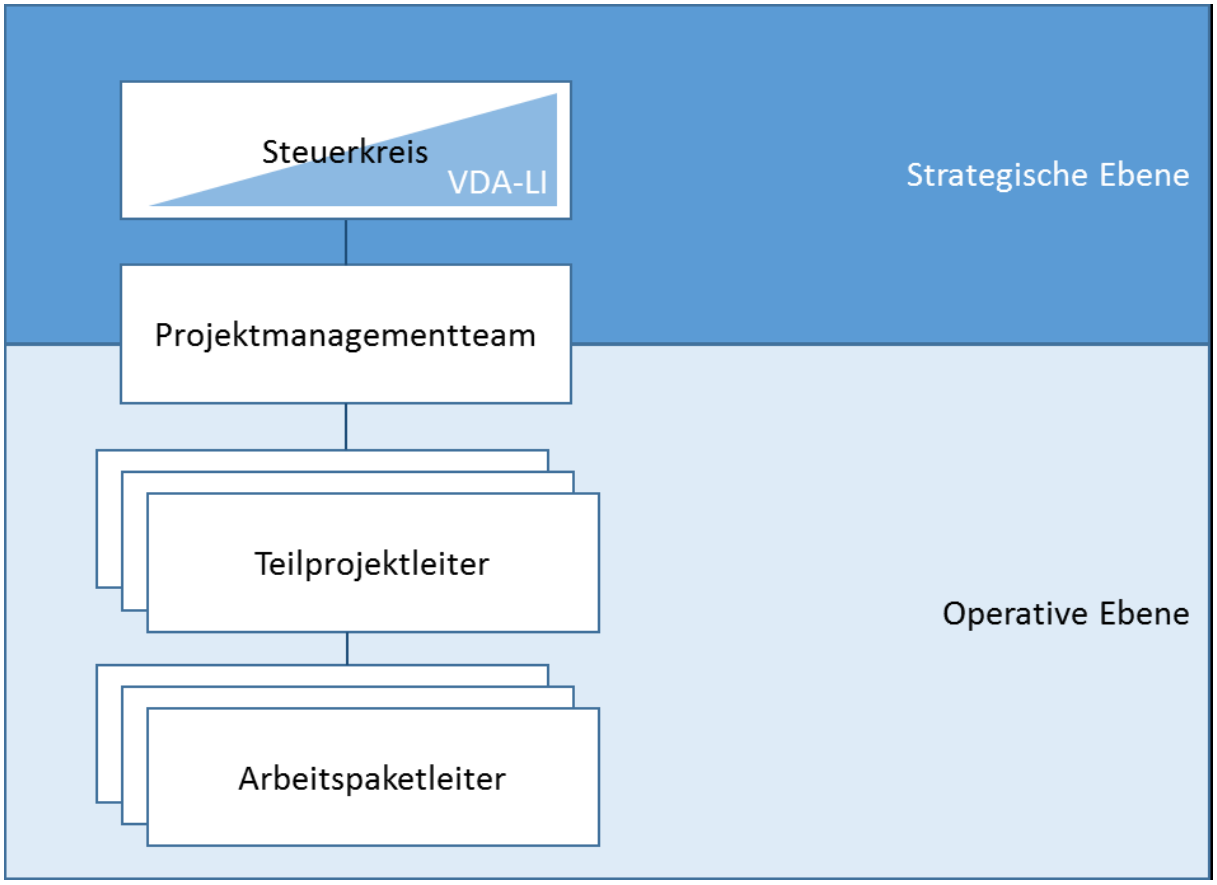


Abbildung 1: Projektorganisation

Neben der Projektorganisation gab es noch eine Reihe von übergreifenden Prozessen unter denen das Projektvorhaben durchgeführt wurde. Tabelle 1 zeigt diese mit Bezug zu den einzelnen Teilprojekten in der Übersicht.

Tabelle 1: Übergreifende Prozesse in der Übersicht

	Beteiligte AP			
Prozess	**TP1**	**TP2**	**T3P**	**TP4**
Beschreibungssprachen- und Datenspezifikations-prozess	AP1.2	AP2.2 AP2.3 AP2.4 AP2.5	AP3.3 AP3.4 AP3.5	**AP4.1** AP4.4
Iterationsprozess Funktionen/Algorithmik	**AP1.1** AP1.2 AP1.3 AP1.4 AP1.5	AP2.1 AP2.5	AP3.1 AP3.2 AP3.3	AP4.1
Konsolidierungsprozess zum Kontext Gesamtfunktion und Systemarchitektur	AP1.2	AP2.1	AP3.2 AP3.5	AP4.1 **AP4.2** AP4.3 AP4.4
KPI-Konsolidierungs-prozess	AP1.3 AP1.4 AP1.5	AP2.2 AP2.4	**AP3.2** AP3.3 AP3.4 AP3.5 AP3.6	AP4.1 AP4.2 AP4.3
Datengenerierungs-prozess	AP1.1 AP1.2	**AP2.1** AP2.2 AP2.5	AP3.3	AP4.1

Je übergreifendem Prozess wurden Prozessverantwortliche definiert. Ihnen oblag es, im Vorhaben über die Projektlaufzeit dafür Sorge zu tragen, dass diese fachlichen TP-übergreifenden Abstimmungen fortlaufend und im Sinne des Gesamtprojektziels aktiv vorangetrieben wurden. Sie stellten sicher, dass der TP-übergreifende Informationsaustausch entsprechend des Prozesses stattfand und die Ergebnisse somit TP-übergreifend nahtlos ineinander übergehen konnten. Die Verantwortlichen sind auch integraler Bestandteil des Projektmanagementteams gem. Abbildung 1.

Projektbezogene Kompetenzen des Partners Valeo

Zum Abschluss sollen noch die projektbezogenen Kompetenzen des Partners VALEO kurz beschrieben werden.

Valeo bringt Kenntnisse aus Algorithmik für die Sensoren, Ultraschall, Radar, Lidar, Kameras und Laser-Scannern in das Vorhaben KI-Absicherung ein. Insbesondere in Bereichen künstlicher Intelligenz und bei Test und Absicherung ist Valeo sehr stark aktiv. Hier werden Forschungen für neue Sensortechnologien vor allem im Bereich der Laser Scanner und Kameras betrieben, die für Sensortechnologien nach 2020 zum Einsatz kommen werden.

Zuletzt war Valeo in verschiedenen Forschungsprojekten aktiv, unter anderem @City, VIDAS und Cloud-LSVA. In @City wird ein hochautomatisiertes Fahren im urbanen Bereich angestrebt. Die Interaktion mit schwächeren Verkehrsteilnehmern (wie z.B. Fußgänger) mit KI-Algorithmik stellt ein bedeutender Inhalt dar. VIDAS beschäftigt sich mit der Spezifikation von Use Cases und Testszenarien für automatisierte Fahrzeuge. Eine Entwicklung von KI-Algorithmik, sowie die Generierung von Daten für die KI-Funktion findet im Projekt VI-DAS statt. Die vorhandenen Kenntnisse aus diesen Projekten sollen auch in das Vorhaben KI-Absicherung eingebracht und weiterentwickelt werden.

1.3. Planung und Ablauf des Vorhabens

Das Vorhaben startete am 01.07.2019 und endete am 30.06.2022. Das Projekt war unterteilt in fünf Teilprojekte TP 1-5. Abbildung 2 zeigt die Projektstruktur in der Übersicht.

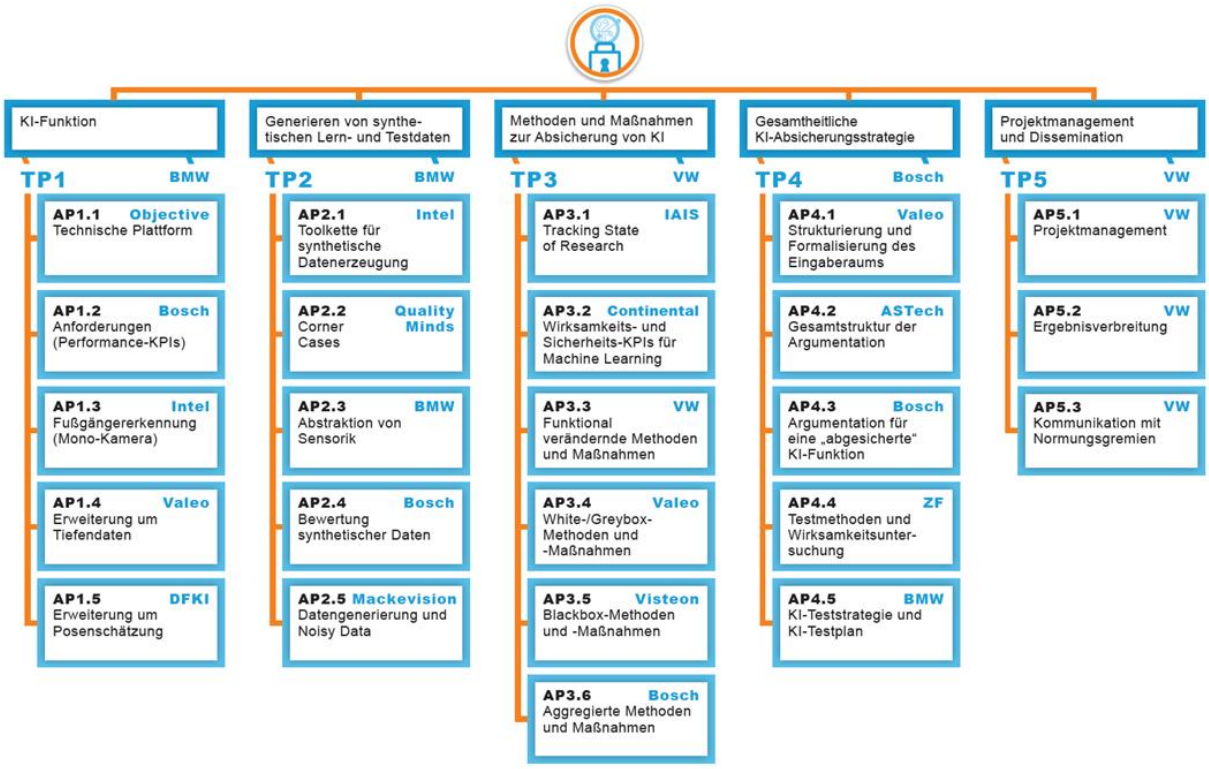


Abbildung 2: Projektstruktur KI-Absicherung

Für das Vorhaben KI-Absicherung wurden innerhalb der Projektlaufzeit fünf Projektmeilensteine (MS 1 – MS 5) definiert, anhand derer der Projektfortschritt und Projekterfolg erfasst wurde. MS 6 stellte den Projektabschluss dar. Vgl. hierzu Tabelle 2. In nachfolgender Tabelle sind auch die damit verknüpften Ergebnisse EX.X.X.x eingetragen.

Tabelle 2: Projektmeilensteine

ID	Titel	Fälligkeit	Relevante Ergebnisse
MS 1	Erste Version der funktionalen Algorithmen spezifiziert und implementiert. Erste Definition des Grundkontexts als sowie der Anforderungen an eine Argumentation liegen vor.	M6	E1.1.1a, E1.1.3c, E1.1.3d, E1.3.1, E1.3.3a-e, E1.3.5, E4.1.1; E4.2.1
MS 2	Rendering-Pipeline vollständig in Betrieb genommen	M9	E2.1.3-2.1.7, E2.2.6, E2.2.9, E2.3.2
MS 3	Erste Implementierungen von KI-Absicherungs-Methoden und Maßnahmen liegen vollständig für alle betrachten Klassen	M18	E3.3.1-5, E3.4.1-5, E3.5.1-6

ID	Titel	Fälligkeit	Relevante Ergebnisse
	von Mechanismen zur Bewertung und Steigerung der Absicherbarkeit von KI vor.		
MS 4	Es liegt eine erste Ausleitung der Argumentation der Sicherheitsziele auf die Sicherheitsanforderungen der KI-Funktion inklusive Safety Contract vor.	M24	E4.2.5, E4.2.7
MS 5	Alle Methoden und Maßnahmen sind spezifiziert, implementiert, bewertet und dokumentiert. Die modellhafte Beschreibung des strukturierten Eingaberaums inkl. einer Beschreibungssprache und eines maschinenlesbaren Formats ist definiert.	M32	E3.3.1-5, E3.4.1-5, E3.5.1-6, E3.6.4, E4.1.2a-c
MS 6	Alle Ergebnisse aus TP1, TP2, TP3 und TP4 liegen vor (Projektabschluss).	M36	alle

Die zeitliche Verortung der Projektmeilensteine im Projektzeitplan lässt sich auch der Darstellung in Abbildung 3 entnehmen.

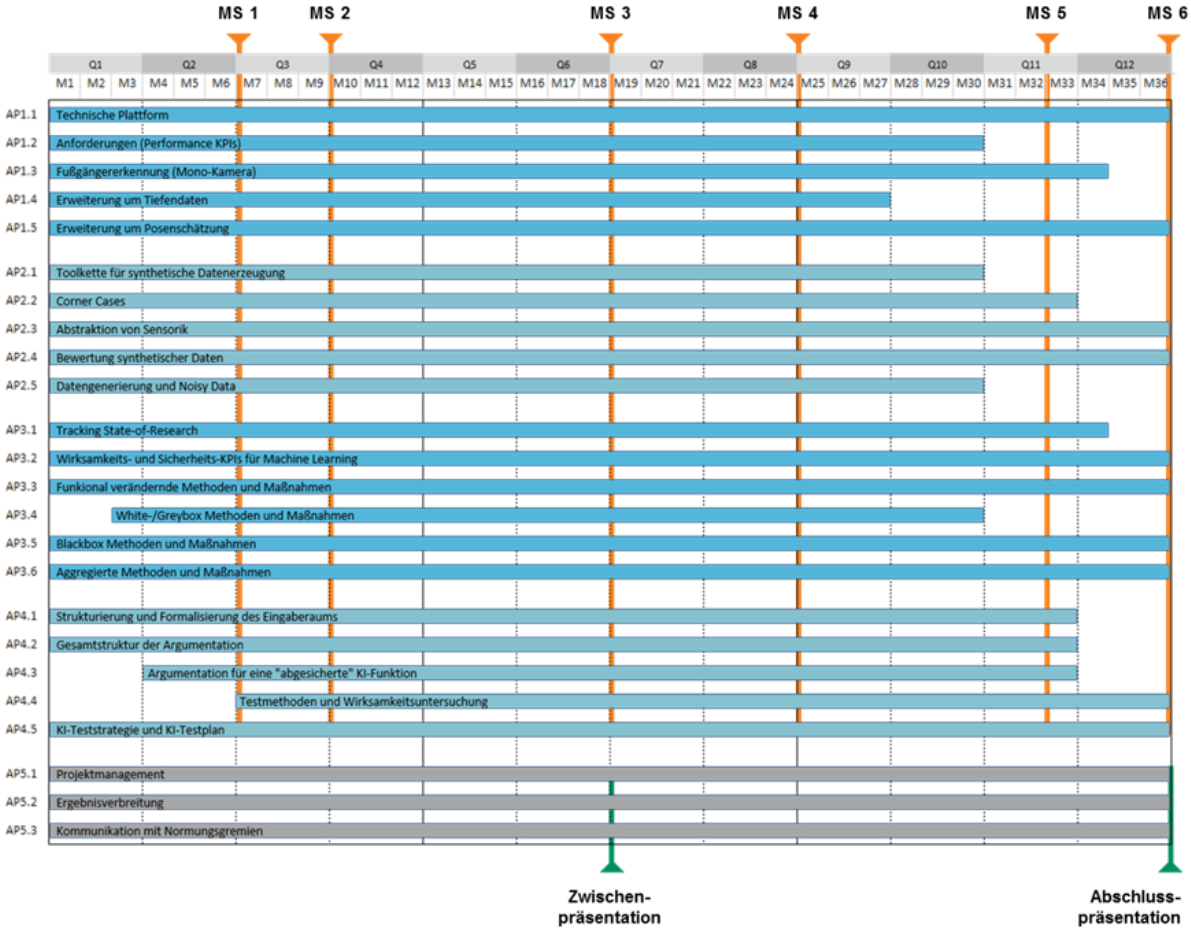


Abbildung 3: Projektzeitplan mit Meilensteinübersicht

Das Erreichen der Meilensteine bildete eine geeignete Basis für die Kommunikation des Projektfortschritts sowohl nach innen (innerhalb des Konsortiums) als auch nach außen (gegenüber dem Fördergeber und der Fachöffentlichkeit).

Vorhabensintern wurde das Erreichen der Meilensteine in Form von Meilensteinberichten (erstellt durch AP5.1 als E5.1.5) sowie nachrichtlich innerhalb des regulären Berichtswesens (erstellt durch AP5.1 als E5.1.6) dokumentiert.

Deliverables

Innerhalb des Vorhabens KI-Absicherung wurden auf Ebene der Arbeitspakete 147 inhaltliche Ergebnisse definiert. Auf Basis dieser Ergebnisse und deren Austauschs innerhalb des Konsortiums hat die Projektarbeit stattgefunden.

Das Konsortium des Vorhabens KI-Absicherung war mit dem Ziel angetreten, die nationale Automobilindustrie und Forschungslandschaft zu befähigen, bestehende Kompetenzen für die Schlüsseltechnologie KI im Bereich des automatisierten und vernetzten Fahrens auf- und auszubauen. Aus diesem Anspruch leitete sich die Notwendigkeit ab, Vorhabenergebnisse in der Breite auch in die Industrie sowie die Wissenschaft zu kommunizieren. Aus dem seitens der VDA-Leitinitiative formulierten Anspruch einer „hohe Anschlussfähigkeit und Verwertbarkeit der Projektergebnisse“

(VDA, 2017) ergab sich darüber hinaus die Notwendigkeit, auch, soweit möglich, Ergebnisse an Dritte bereitzustellen.

Daraus ergab sich die Notwendigkeit, Ergebnisse zusammenzufassen (zu Deliverables) und je Deliverable festzulegen, an wen diese verteilt werden.

Zusammenfassung von Ergebnissen zu Deliverables

Die 147 Ergebnisse aus TP1 bis TP4 lassen sich entlang der beiden Dimensionen „Inhalt“ und „Ergebnisform“ kategorisieren. Tabelle 3 fasst das Ergebnis zusammen. Hierbei zeigt sich das folgende Bild:

Während sich innerhalb der Dimension „Inhalt“ eine hohe Bandbreite an Kategorien zeigt, lassen sich die Ergebnisse entlang der Dimension „Ergebnisform“ auf wenige Kategorien verorten.

Im Vorhaben KI-Absicherung lassen sich folgende, mit Blick auf die definierten Ergebnisse, Ergebnisformen identifizieren:

- Dokumente, die z.B. Anforderungen, Methoden, Konzepte, Austauschformate, Beschreibungssprachen aufnehmen
- Softwareimplementierung (Object Code / Source Code)
- parametrierbare bzw. parametrierte Modelle (Sensormodelle, physikalische Modelle)
- DNN-Netze mit Struktur und Gewichten (einsehbar)
- generierte und Aufbereitete (Mess- oder Simulations)-Daten (inkl. Metainformationen / Labeln) + Motion Capture Erfassung / Playback (Content)
- Werkzeuge bzw. Plattformen, die im Vorhaben die notwendige technische Entwicklungsplattform bilden (IT-Infrastruktur)

Die identifizierten Inhaltskategorien lassen sich wiederum in größere Gruppen clustern. In Summe wurden sieben Gruppen (A-G) definiert, die ähnliche Inhaltskategorien clustern.

Tabelle 3: Kategorisierung der Ergebnisse nach Inhalt und Ergebnisform

		Dimension: Ergebnisform						
Gruppe	Dimension: Inhalt	Dokument	Software	DNN-Netz	Cotnent	IT-Infrastruktur	Σ	Σ
A	Methoden und Maßnahmen	1	10				**11**	18
	Maßnahmen		3				**3**	
	Methoden		4				**4**	
B	Deep-Learning-Modell (DL-Modell)		4	11			**15**	15
C	Anforderung	8					**8**	40
	Auswahl / Übersicht / Katalog / Zuordnung	9					**9**	
	Spezifikation	8			2		**10**	
	Definition	10					**10**	
	Taxonomie	1					**1**	
	Template	1	1				**2**	
D	Stand der Technik	6					**6**	33
	Evaluation	13					**13**	
	Guideline	14					**14**	
E	Methodik	4	3				**7**	13
	Modell	3					**3**	
	Konzept	3					**3**	
F	Tool		6			10	**16**	16
G	synthetische Daten				12		**12**	12
		81	31	11	**14**	19	**147**	

Weiterführende Informationen zu Planung, Inhalt und Umfang des Projektes können auch einer öffentlich zugänglichen Quelle[3] entnommen werden.

[3] https://www.ki-absicherung-projekt.de/fileadmin/user_upload/KI-Absicherung_Standardpraesentation_de_v1.5.pdf

Arbeitsschwerpunkte

VALEO selbst war in allen Teilprojekten TP 1 - 5 beteiligt. Die Arbeitsschwerpunkte seitens VALEO in den jeweiligen Teilprojekten werden in der folgenden Tabelle kurz umrissen. Die Detaillierung folgt in der eingehenden Darstellung in Kapitel 2.

Tabelle 4: Arbeitsschwerpunkte der Teilprojekte TP 1 bis 5 des Projektpartners VALEO

TP1	Seitens Valeo fand in TP1 eine aktive Beteiligung an allen Arbeitspaketen statt. Einen Schwerpunkt setzte Valeo bei der Algorithmen-Entwicklung zur Fußgängererkennung AP1.3 (Co-Lead) und AP1.4 (Lead). Im AP1.1 beschäftigte sich Valeo mit dem Testen des Docker-Images mit integrierter Entwicklungsumgebung. Im AP1.2 lag der Fokus auf einer Definition von Qualitäts-KPIs für 3D Bounding Box Fußgängererkennung und der Spezifikation des LiDAR Referenzsensors. Im AP1.3 fokussierte sich Valeo auf die Entwicklung einer 3D Bounding Box Fußgängererkennung basierend auf monokularen Kamerabildern. Weiterhin fand eine Rücktransformation in die 2D Bildebene statt, um diesen Ansatz mit 2D Bounding Box Fußgängererkennungs-Ansätzen zu vergleichen. Im AP1.4 hat Valeo neben der Arbeitspaketleitung eine sequentielle Fusion von Kamera und LiDAR Daten zur 3D Bounding Box Fußgängererkennung entwickelt. Dieser Ansatz baute auf eine 2D Bounding Box Fußgängererkennung aus AP1.3 auf und erweiterte diesen mit dem Einsatz von LiDAR Daten. Im AP1.5 lag der Schwerpunkt bei der Extraktion von der 3D Bounding Box Position von der 3D Skelett Posenschätzung, um eine Wechselwirkung von Posenschätzung und Fußgängererkennung zu untersuchen.
TP2	Valeo beteiligte sich mit einer aktiven Mitarbeit bei den AP2.1 bis 2.4. Der Schwerpunkt lag hier bei der Algorithmen-Entwicklung von teilüberwachten Domain Adaptation Ansätzen zur Einhaltung von Absicherungs-KPIs bei der Änderung von Sensor-Setup (AP2.3, Co-Lead) und Bildqualität (AP2.4, Co-Lead). Im AP2.1 lag der Fokus bei der Entwicklung eines LiDAR Sensormodells als Plugin für die Rendering Pipeline. Im AP2.2 untersuchte Valeo eine Korrelation zwischen Corner Case-Variationen mit Schwerpunkt auf Corner Cases für die KI-Funktion. Im AP2.3 fokussierte sich Valeo auf eine Annäherung der Absicherungs- und Qualitäts-KPIs bei Variation des Sensor-Setups durch Transfer Learning Techniken. Hierbei stand im Vordergrund die Anwendung vom klassischen Finetuning, sowie der Anwendung und Anpassung von Metriken und Netzwerkstrukturen, die eine relevante Ähnlichkeit von Merkmalsverteilungen innerhalb des DNNs zwischen den Domänen der unterschiedlichen Sensor-Setups erhalten. Ähnlich zum AP2.3 lag der Schwerpunkt bei AP2.4 in der Algorithmen-Entwicklung von Domain Adaptation Ansätzen. Die zu untersuchenden Ansätze bezogen sich hierbei auf GAN-Ansätze, um eine Einhaltung von Absicherungs- und Qualitäts-KPIs zwischen unterschiedlichen Qualitätsstufen der synthetischen Daten zu erreichen.

TP3	Valeo beteiligte sich in TP3 an allen Arbeitspaketen. Einen Schwerpunkt legte Valeo bei der Bearbeitung bei den Entwicklungs-Arbeitspaketen 3.3 bis 3.5. Für das AP3.4 übernahm Valeo sogar die Lead-Rolle. Im AP3.1 lag der Fokus auf einer kontinuierlichen State-of-the-art Analyse von introspektiven Methoden und Maßnahmen, der Publikation von Forschungsergebnissen und der Organisation von Workshops auf internationalen Konferenzen. Die Aktivität von Valeo in AP3.2 umfasste die Erstellung einer strukturierten Übersicht der verfügbaren KPIs und der Kommunikation von neu entwickelten KPIs aus AP3.4. Im AP3.3 fokussierte sich Valeo auf die Robustifizierung von KI-Funktionen durch eine Netzwerkspezialisierung und Ausnutzung zeitlicher Konsistenz in Videodaten. Die Netzwerkspezialisierung beinhaltete eine Erweiterung zu einem Multi Task Learning Netzwerk. Eine zeitliche Konsistenz in Videodaten erfolgte durch eine Merkmalskarten Fusion innerhalb des DNNs. Zusätzlich zur Leitung von AP3.4 lag einer der Arbeitsschwerpunkte auf der Erhöhung der Plausibilisierung der inneren Funktionsweise des Netzes. Dies beinhaltete die Erhöhung der Interpretierbarkeit von Merkmalskarten innerhalb des DNNs und der Kombination von geeigneten Heatmap Verfahren. Weiterhin dienten diese Ergebnisse der Entwicklung von Maßnahmen, die zu einer Absicherung der KI-Funktion beitrugen. Im AP3.5 beschäftigte sich Valeo mit der Evaluierung der Netzwerk-Vorhersage bei gezielter Manipulation der Eingangsdaten. Zudem wurden Corner Cases mit der DeepExplore Methode detektiert. Die in AP3.6 geplanten Beiträge beinhalteten die AP-Übergreifende Kombination von Heatmap Verfahren, sowie eine darauf aufbauende Metaklassifikation zur Steigerung der Absicherung der KI-Funktion.
TP4	Seitens Valeo fand in TP4 ein aktives Mitwirken in allen Arbeitspaketen statt. Der Schwerpunkt lag bei der Herausarbeitung von relevanten Grundkontexten, sowie der Spezifikation der Datenformate einzelner Dimensionen im AP4.1 (Lead). Im AP4.2 fokussierte sich Valeo auf die Argumentation der Sicherheitsziele basierend auf Eigenschaften verwendeter Sensorik und Netzwerkarchitekturen der KI-Funktionen. Im AP4.3 beschäftigte sich Valeo mit einer Nachweisstrategie für eine hinreichende Datenbasis als Grundlage für die trainierte KI-Funktion. Im AP4.4 lag der Fokus auf der Überprüfung der Abdeckbarkeit des Eingaberaums basierend auf Schätzungen von Dichteverteilungen von Datensätzen. Im AP4.5 behandelte Valeo die Erstellung eines Testplans bzgl. der Tiefensensoren-relevanten Garantien des Assurance Cases.
TP5	Valeo zeichnete verantwortlich für den „Iterationsprozess Funktionen/Algorithmik“ und hat im Projektmanagementteam die in diesem übergreifenden Prozess erarbeiteten Ergebnisse vorgestellt, sowie Anforderungen die sich aus der technischen Gesamtkoordination an diesen Prozess ergaben, in den Kreis der beteiligten AP kommuniziert und entsprechende Aufgaben definiert.

1.4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Der Stand von Wissenschaft und Technik an dem angeknüpft wurde, wird im Folgenden entlang der thematischen Gliederung der wissenschaftlichen und technischen Arbeitsziele aufgespannt.

- Algorithmen zur Fußgängererkennung implementierten
- Daten für Training und Test erzeugen
- Absicherbarkeit von KI-Modulen bewerten
- Unsicherheit erkennen und falsche Vorhersagen detektieren
- KI-Module absichern

Algorithmen zur Fußgängerkennung implementieren

Die Fußgängererkennung spielt im Automobilbereich seit vielen Jahrzehnten eine zentrale Rolle. Bereits in den Neunzigern gab es erste Prototypen der Automobilindustrie. Begründet ist diese Entwicklung u.a. durch den hohen Anteil der Fußgänger an den Verkehrstoten. Neben Zweiradfahrern sind sie mit Abstand die größte Gruppe von im Straßenverkehr getöteten Personen. Hier können Systeme zur automatischen Fußgängererkennung einen wichtigen Beitrag zur Vermeidung von Kollisionen leisten. Dies kann einerseits in Form von sogenannten Fahrerassistenzsystemen umgesetzt werden, die dem Fahrer des Fahrzeugs z.B. ein akustisches Warnsignal geben, wenn ein Fußgänger in einer potentiell gefährlichen Position registriert wird. Diese Systeme sind heutzutage bereits in Serie. Bei autonom agierenden Fahrzeugen sind die Anforderungen an die Fußgängererkennung nochmals wesentlich höher.

Die ersten Entwicklungen zur Fußgängererkennung waren rein bildbasiert. Dabei wurden die Personen mit sogenannten HOG (Histogram of Gaussians) Beschreibungen erkannt. Diese Algorithmen haben akzeptable Ergebnisse und auch ihren Weg in die Serie gefunden. In den letzten Jahren ist ein starker Trend hin zu Algorithmen, die auf tiefen neuronalen Netzen beruhen, zu erkennen. Diese KI-basierten Algorithmen haben die Genauigkeit nochmal auf ein wesentlich höheres Level gehoben.

Bei der Entwicklung von KI-basierten Algorithmen stehen meist Kriterien der Leistungsfähigkeit im Vordergrund, z.B. der Prozentsatz detektierter Fußgänger oder die Genauigkeit der Segmentierung im Bild [3]. Im Hinblick auf die kritischen Sicherheitsanforderungen darf beim autonomen Fahren jedoch die Optimierung auf solche Kriterien nicht das einzige Entwicklungsziel sein. Einerseits beantworten Performanzkriterien naturgemäß nicht die Frage, wie mit dem (oft niedrigen) Anteil von Fehlerkennungen umgegangen werden soll (z.B. 2% nicht erkannter Fußgänger). Andererseits müssen zur Gewährleistung der funktionalen und Gebrauchssicherheit auch andere Kriterien in den Fokus der KI -Algorithmen-Entwicklung rücken, die weniger intuitiv und v.a. kaum etabliert sind.

Für Anwendungen im Automotive-Bereich ist neben der Detektion von Fußgängern (= „reine“ Erkennung“) die Analyse ihres Verhaltens von essentieller Bedeutung, um eine bestmögliche Reaktion des Fahrzeugs zu ermöglichen. Grundlage hierfür ist nicht nur

die Lokalisation der Fußgänger (mittels so genannter Bounding Boxen), sondern auch eine präzise Erfassung ihrer Pose (Haltung einer Person) und die Modellierung der zeitlichen Veränderung derselben. Im Gegensatz zu einer vergleichsweise groben Verortung mittels Bounding Boxen kann so die genaue Körperhaltung und das zeitliche Verhalten erfasst werden. Diese Beschreibung bildet damit die Grundlage für die Klassifikation von aktuellen und für eine mögliche Prädiktion von zukünftigem Fußgängerverhalten.

Daten für Training und Test erzeugen

Trainings-, Validierungs- und Testdaten sind für die Absicherung von KI-Systemen essentiell. Dabei ist der Einsatz von realen, also mit Hilfe von Testfahrzeugen und entsprechenden Sensoren (Kameras, LiDAR, Radar, usw.) unabdingbar. Die Erfassung und Aufbereitung von realen Daten ist jedoch sehr aufwendig und teuer. Eine zentrale Herausforderung ist, dass neben den eigentlichen Sensordaten des Fahrzeugs auch Referenzdaten, sogenannte Ground Truth oder Labels, erfasst werden müssen. Für viele Aufgaben, z.B. Fußgängererkennung, können diese nur manuell generiert werden und können nicht automatisiert oder durch zusätzliche Sensoren erfasst werden. Dementsprechend gibt es kaum größere Datensätze, die zu Forschungs- und Entwicklungszwecken verwendet werden können. Dazu kommen rechtliche Einschränkungen, die aufgrund des Schutzes der Privatsphäre des Individuums eine Aufnahme im öffentlichen Raum derzeit in Europa nur stark eingeschränkt möglich machen.

Demgegenüber stehen Verfahren der synthetischen Datenerzeugung durch Simulation, die eine Reihe von Vorteilen bieten:

- Die Nutzungsrechte sind nicht grundsätzlich eingeschränkt. Vor allem können Parameter wie z.B. Fahrzeuggeschwindigkeiten oder Szenenkomplexität beliebig variiert werden.
- Im realen Raum riskante Aktionen können beliebig oft ohne Einschränkung nachgestellt werden.
- Synthetische Verfahren könne auch sehr einfach die für Training und Validierung benötigten ‚Ground truth-Daten' liefern.

Die Möglichkeit in der Simulation, Fahrsituationen beliebig variieren zu können, macht die synthetische Datenerzeugung besonders geeignet, um systematische Trainings- und Validierungsdaten zu erzeugen.

Nur mit Daten einer relevanten Abdeckung kann man die Absicherbarkeit der KI-Module gewährleisten. Heute existiert keine geeignete Datenbank an Validierungsdaten, auf deren Basis sich eine systematische Bewertung oder gar statistische Absicherung von KI-Modulen realisieren ließe. Deshalb gehen einige Studien von einem Brute-Force-Verfahren aus: die geeignete Varianz der Daten soll entsprechend über die Menge gewährleistet werden.

Eine Übersicht über derzeit verfügbare Datenbanken mit Trainings- und Validierungsdaten ist in Tabelle 3 dargestellt.

Tabelle 5: Übersicht etablierte Datenbanken für synthetische Daten (farbig hinterlegt) und Realdaten (weiß hinterlegt) im Automobilkontext

Datenbank	Beschreibung	Ground Truth	Jahr	Größe	Verwendete Sensorik	Relevanz für das Vorhaben	Quelle
ApolloScape	Große Datenbank, die laufend aktualisiert wird. Ziel sind 1M Frames. Datenbank ist frei verfügbar.	Semantische Segmentierung, Instanz Segmentierung, Straßenmarkierung, Tiefeninformation, Kamera Position	2018	>140t Bilder	3384x2710px Farbbild Tiefenbild aus LiDAR	Personen-Labels aus semantischer Segmentierung könnten verwendet werden.	[8]
BDD100K	Umfassende Datenbank mit amerikanischen Städten, zu verschiedenen Wetterbedingungen, Tageszeiten und Straßentypen.	Bounding Boxen, Befahrbare Areale, Straßenmarkierung, Instanz Segmentierung, Odometrie	2018	100t Sequenzen à 1200 Bilder	1280×720px Farbbild	Personen-Labels aus semantischer Segmentierung könnten verwendet werden.	[9]
Caltech Pedestrian Detection	Datenbank explizit für Fußgänger Erkennung. Nur Fußgänger sind annotiert.	Bounding Boxen, Instanz IDs, zeitliche Korrespondenzen, Verdeckungen	2012	250t Bilder	640x480px Farbbilder	Bounding Boxen könnten verwendet werden.	[10]
Cityscapes	Bildsequenzen aus verschiedenen deutschen Städten (~ 50 ➲ Sequenzen)	Semantische Segmentierung, Instanz Segmentierung, Bounding Boxen	2016	5000 Bilder (fein), 20t Bilder (grob)	2040x1060px , Farbbilder Tiefenbilder aus Stereo	Personen-Labels aus semantischer Segmentierung könnten verwendet werden.	[11]
Daimler Pedestrian Segmentation Benchmark	Datenbank für Fußgänger Segmentierung. Geringe, variierende Auflösung.	Form der Fußgänger	2013	785 Bilder	34x11px bis 468x267px Farbbilder	Formen könnten verwendet werden.	[12]

Datenbank	Beschreibung	Ground Truth	Jahr	Größe	Verwendete Sensorik	Relevanz für das Vorhaben	Quelle
Daimler Pedestrian Path Prediction	Datenbank über verschiedene Sequenzen von gestelltem Fußgängerverhalten	Bounding Boxen, Fußgänger Trajektorie, Odometrie	2013	68 Sequenzen	1176x640px Farbbilder	Bounding Boxen könnten verwendet werden.	[13]
GTA V	Datenbank generiert aus GTA V (Playing for Data & Playing for Benchmarks) zu verschiedenen Tageszeiten und Wetterbedingungen	Optischer Fluss, Semantische Segmentierung, Instanz Segmentierung, Odometrie	2017 / 2018	250t Bilder	1914x1052px Farbbilder	Personen-Labels aus semantischer Segmentierung könnten verwendet werden.	[5] [14]
HD1K Benchmark	Datenbank für Optischen Fluss. Wetter, Verkehrsszenarien und Lichtbedingungen wurden systematisch variiert.	Optischer Fluss	2017	>1000 Bilder	2560x1080px Farbbilder	Keine Relevanz	[15]
JAAD (Joint Attention for Autonomous Driving)	Datenbank über Fußgänger und Fahrerverhalten im Straßenverkehr. Die Aktionen der Fußgänger sind beschrieben. Aufnahmen wurden in Nord-Amerika und Ost-Europa durchgeführt.	Bounding Boxen, Fußgänger Verhaltensbeschreibung, Verkehrsbeschreibung	2016	346 Sequenzen à 5-10s	1920x1080px Farbbilder	Keine direkte Relevanz	[16]
KITTI Detection Dataset	Datenbank mit Aufnahmen aus Karlsruhe	2D/3D Bounding Boxen, Odometrie	2017	ca. 15t Bilder	1392x512px, Stereo Bilder	Bounding Boxen könnten verwendet werden.	[17]

Datenbank	Beschreibung	Ground Truth	Jahr	Größe	Verwendete Sensorik	Relevanz für das Vorhaben	Quelle
					LiDAR Information		
KITTI Flow & Semantic	Datenbank mit Aufnahmen aus Karlsruhe	Stereo Disparität, Optischer Fluss, Szenen Fluss, Semantische Segmentierung (1 Bild pro Szene)	2015	400 Sequenzen à 4 Bilder	1392x512px, Stereo Bilder LiDAR Information	Keine Relevanz	[17]
Mapillary Vistas	Datenbank mit Straßensituationen aller Kontinente, Wetterbedingungen, Tageszeiten. Verschiedene Kameras und Blickwinkel	Semantische Segmentierung, Instanz Segmentierung	2017	25t Bilder	Verschiedene Auflösungen	Personen-Labels aus semantischer Segmentierung könnten verwendet werden.	[18]
Synthia	Synthetische Datenbank verschiedener virtueller Städte & Wetterbedingungen & Tageszeiten. Dynamische und statische Objekte.	Semantische Segmentierung, Instanz Segmentierung, Odometrie	2016 - 2018	>200t Bilder	920x720px Farbbilder & Tiefenbilder	Personen-Labels aus semantischer Segmentierung könnten verwendet werden.	[19]
Virtual KITTI	Synthetischer Klon des KITTI Datensatzes	Semantische Segmentierung, Instanz Segmentierung, 2D / 3D Multi-Objekt Tracking	2016	50 Sequenzen (ca. 20k Bilder) aus 5 virtuellen Welten	1242x375px, Farbbilder & Tiefenbilder	Personen-Labels aus semantischer Segmentierung könnten verwendet werden.	[20]

Einige Datensätze können teilweise für das Projekt verwendet werden. Allerdings erfüllt keiner der Datensätze alle Anforderungen des Vorhabens KI-Absicherung.

Insbesondere für das automatisierte und autonome Fahren ist der Ansatz, KI-Funktionen allein in Feldstudien abzusichern, unrealistisch, da hierfür bei jeder Systemänderung geschätzt 240 Millionen Testkilometer gefahren und ausgewertet werden müssten [4]. Eine Vergleichbarkeit zwischen verschiedenen Sensor-Setups oder unterschiedlichen Fahrzeugtypen ist mit diesem Ansatz ebenfalls nicht gegeben, da reale Fahrten nicht exakt reproduzierbar sind. Bisherige Ansätze, möglichst viele Sensordaten aus dem täglichen Betrieb zu sammeln und das Verhalten des Fahrzeugs mit diesen zu trainieren, stoßen hier schnell an technische Grenzen, da insbesondere die kritischen Situationen sehr selten auftreten. Ferner lässt sich mit aufgenommenen Daten nicht das vollständige Fahrzeugverhalten testen (closed-loop), da in den aufgenommenen Daten die Fahrzeugtrajektorie des realen Fahrzeuges mit der Aufnahme fixiert ist und sich nicht verändern lässt (open-loop).

Synthetische Datensätze systematisch aus Simulationen zu generieren hat in diesem Zusammenhang ein großes Potential (synthetische Datensätze). So erreicht man eine hohe Abdeckung der Varianz und Komplexität sich verändernder Einflüsse wie Sensor-Setups, Umwelteinflüsse oder Fahrzeugtypen. Die entsprechende Trainings- und Validierungsdaten werden in einem ausreichenden Umfang zur Verfügung gestellt. Weiterhin können seltene kritische Ereignisse, sogenannte Corner Cases, verhältnismäßig leicht generiert werden. Kritische Szenarien, wie beispielsweise ein plötzlich auftauchendes Kind auf der Straße, sind mit realen Daten auf der Straße nur mit erheblichen Risiken und Kosten aufzunehmen. Bei synthetischen Straßen können solche Corner Cases definiert und simuliert werden. Zusätzlich ist eine Variation dieser kritischen Fälle verhältnismäßig einfach.

Bislang wurden Simulationssysteme für Fahrzeuganwendungen aus zwei Richtungen verfolgt: aus Sicht der physikalischen Fahrzeugsimulation sowie zur Fahrsimulation (z.B. zur Ausbildung).

Erstere Systeme, die aus diesen Ansätzen entstanden sind (z.B. von dSPACE oder IPG CarMaker) verfügen derzeit nicht über genügend realistische Darstellung und Syntheseverfahren. Ein hoher Grad an Realismus ist zwingend erforderlich um später aus den synthetischen Trainingsdaten Rückschlüsse auf reale Szenarien ziehen zu können. Trainings- und Validierungsdaten müssen somit in ausreichender Güte erzeugt werden. Ferner sind die existierenden physikalischen Fahrzeugsimulation komplexe proprietäre Systeme, die sich nicht zur systematischen Datenerzeugung eignen.

Erste Ansätze zur Verwendung von Fahrsimulation für Training- und Validierung haben mit Software aus dem Videospiele-Bereich experimentiert [5]. Aufbauend auf diesen Ergebnissen entstand der CARLA-Simulator für urbane Verkehrsszenen [6]. Diese Software basiert ebenfalls auf einem Simulationswerkzeug, das ursprünglich für Computerspiele ausgelegt wurde. Diese Technologie hat allerdings diverse Einschränkungen: Die generierten Bilder erscheinen auf den ersten Blick zwar

verhältnismäßig realistisch, werden aber durch eine Vielzahl von Approximationen generiert. Insbesondere wird die 3D-Szene oft durch einfache geometrische Formen modelliert und nur durch hochauflösende Texturen optisch ansprechend dargestellt. Durch den mangelnden geometrischen Detailgrad ist eine Simulation von aktiven Sensoren, wie LiDAR oder Radar, aber ausgeschlossen bzw. viel zu ungenau. Weiterhin können nur ideale Farbkamerasensoren simuliert werden. Aufnahmefehler, wie sie in realen Sensoren auftreten, können nicht ausreichend modelliert werden.

Im Unterschied zum Vorhaben PEGASUS (https://www.pegasusprojekt.de/de/) bzw. zur Projektfamilie V&V, die ebenfalls synthetische Simulationsdaten in die Absicherungskette einbauen, ist es für die Absicherung von KI-Funktionen notwendig Sensorrohdatensätze von allerhöchster Qualität zu erzeugen, um damit die KI-Wahrnehmungsmodule zu trainieren und zu testen. Der Blick liegt also nicht auf der Absicherung der Gesamtwirkkette (von der Wahrnehmung bis zur Aktorsteuerung), sondern auf der systematischen Absicherung von KI-basierten Teilfunktionen. Die Weiterentwicklung von Sensormodellen hat im diesem Zusammenhang eine hohe Bedeutung.

Die Sensorsimulationsmodelle schließen die Lücke zwischen der Umweltsimulation und den Funktionsalgorithmen (hier KI-Algorithmen). Damit wird die Einbeziehung der Umweltwahrnehmung durch die Sensoren auf die Güte der Gesamtfunktion in die Simulation ermöglicht. Die gängigen Simulationssysteme für Verkehrsszenen haben mittlerweile einfache Sensormodelle für Kamera, Radar aber auch LiDAR definiert und implementiert. Bei der Kamera geht man in der Regel von einem Lochkamera-Modell aus und vernachlässigt Effekte wie Verzeichnung, Unschärfe, Vignettierung aber auch Bildsensor-Rauschen. Physikalische Kameramodelle existieren, sind aber stark vereinfacht und damit unrealistisch. Auch realistische Modelle für Stereo-Kameras, Surround View oder Fisheye-Kameras sind nicht vorhanden. Mittlerweile ist es jedoch unumstritten das die Sensoreffekte einen erheblichen Einfluss auf die Performance von Algorithmen haben und daher mit in der Toolkette berücksichtigt werden müssen.

Die Qualität bei der Datensynthese von LiDAR-Systemen ist analog zur Kamera-Synthese stark von der Renderpipeline der Simulationsumgebung abhängig. Da LiDAR im Gegensatz zu Kameras ein aktiver Sensor ist, müssen mittels Ray-Casting aktiv Strahlen zurückverfolgt werden. Außerdem ist die Qualität stark von der Detailtreue sowie Oberflächenmodellierung der 3D Modelle abhängig. Für ADAS und autonomes Fahren liefert derzeit die Simulationsumgebung CARLA eine frei verfügbare Software an [6].

Bei einfachen Ray-Casting Simulationen kann jedoch ein erheblicher Unterschied zu echten Daten festgestellt werden [7], da unter anderem Gaußsches Rauschen sowie diverse Unschärfe-Prozesse innerhalb der Ray-Casting Pipeline sowie dem Sensormodell unberücksichtigt bleiben.

Absicherbarkeit von KI-Modulen bewerten

Um Absicherbarkeit von künstlicher Intelligenz für das automatische Fahren zu erreichen, wird ein reines Testen aus zahlreichen Gründen nicht ausreichen. Dazu gehören:

- Der hohe finanzielle, zeitliche und personelle Aufwand für das Testen von KI und das Bewerten der erzielten Performance (Versehen der Testdaten mit „Ground Truth“).
- Die Existenz von Adversarial Examples (leicht veränderte Eingabedaten, die das ansonsten korrekte Verhalten eines KI-Moduls hin zu einem Fehler verändern).
- Das Nicht-Wissen über die Stabilität und das Maß an Generalisierungsfähigkeit von trainierten KI-Funktionen.
- Maschinelles Lernen (insbesondere im Teilbereich des überwachten Lernens) ist ein automatisches Parametrieren eines Models anhand gegebener Beispieldaten (die Trainingsdaten), wobei die am Ende erzielten Parameter nicht mehr direkt interpretierbar sind. Dies sorgt dafür, dass bei neuronalen Netzen nach dem Training nicht klar ist, auf welche Merkmale der Eingangsdaten sich das neuronale Netz sensibilisiert hat. Dies erschwert ein „konzeptionell sauberes“ Testen.

Zusammengefasst besteht die Notwendigkeit der Entwicklung ganz neuer Mechanismen zur Bewertung von KI hinsichtlich der Absicherbarkeit. Solche Mechanismen können Methoden zur Überwachung von Training, Test und Inferenz, aber auch Maßnahmen zur Verbesserung der Laufzeiteigenschaften sein. Der Begriff Methoden bezeichnet hierbei Technologien und Software zur Analyse der Absicherbarkeit eines KI-Moduls (zum Zeitpunkt des Trainings, Testens oder der Ausführung). Maßnahmen sind sich hieraus ergebende Mechanismen zur Sicherstellung der Absicherbarkeit (aktive Interpretation der Ergebnisse, die durch Maßnahmen gefunden werden).

Solche Mechanismen (Methoden wie Maßnahmen) sind in Ansätzen aus der akademischen Welt bereits bekannt (siehe hierzu auch die umfassende Liste referenzierter Publikationen im Literaturverzeichnis). In Hinsicht auf die Absicherbarkeit von KI in sicherheitskritischen Anwendungen werden sie aber nicht oder nur sehr eingeschränkt hinsichtlich ihrer Aussagekraft bewertet: So werden (beispielsweise) aufgrund von nur eingeschränkt verfügbaren Trainings- und Testdatensätzen in akademischen Arbeiten tendenziell leichte KI-Aufgaben als Referenzfunktion für neue Technologien im Bereich der künstlichen Intelligenz genutzt. Dies betrifft sowohl die eigentliche Aufgabe (Klassifikation im Gegensatz zu komplexen Regressionen, dichten Klassifikationen wie semantischer Segmentierung oder Prädiktion) als auch den Input (Bild oder Text im Gegensatz zu multiplen Sensoren wie Radar, Lidar oder Video in hoher Auflösung).

Für eine Verwendung der Methoden im Rahmen der Absicherung oder einer Maßnahme zur funktionalen Verbesserung von KI zur Laufzeit muss eine Bewertung der Maßnahmen allerdings nachweisbar und anhand relevanter Funktionen erfolgen. Im bisherigen akademischen wie auch industriellen Anwendungsbereich von KI (vor allem in der Consumer Electronics-Industrie ohne sicherheitskritische Verwendung von KI oder mit einer regelbasierten bzw. menschlichen Kontroll- und Rückfallebene)

bestand eine derart strenge Notwendigkeit nach einem tiefen Verständnis von KI meist nicht. Dies sorgte dafür, dass der wesentliche wissenschaftliche und entwicklungstechnische Fokus auf eine Optimierung der Performance (z.B. Klassifikationsgenauigkeit, pixelweise Vorhersagegüte) von KI statt auf der Absicherbarkeit oder Transparenz dergleichen gelegt wurde.

Neben den beschriebenen technischen Herausforderungen, sind die Herausforderungen, die sich im Rahmen eines Einsatzes der Erkenntnisse und Ergebnisse der Grundlagenforschung in einem realen industriellen Umfeld – wie unteranderem dem automatisierten Fahren – ergeben, vielfältiger Natur:

- Die Bedingungen, unter denen die Methoden entwickelt und erprobt werden, sind sehr begrenzt. Die Methodenevaluation ähnelt stärker einem kontrollierten Experiment als einem umfassenden Test im Sinne der Ingenieurswissenschaften bis hin zu einer Zulassung, wie sie im industriellen Umfeld – mit dem Ziel einer späteren Verwertung in Produkten bzw. Dienstleistungen – notwendig wären.
- Die Erkenntnisse der Grundlagenforschung lassen sich insgesamt nur bedingt übertragen: Sie sind weit entfernt von einem systematischen Bewerten und damit auch von der darauf aufbauend möglichen Ableitung von Maßnahmen zur Absicherung von KI.
- Neben diesen Einschränkungen bei der Methodenentwicklung fehlen des Weiteren domänenspezifische Untersuchungen bzw. Vergleiche der Methoden, die einen differenzierten Blick auf die Ansätze erlauben.

Unsicherheit erkennen und falsche Vorhersagen detektieren

Seit 2015 beschäftigen sich Arbeiten im Bereich Deep Learning auch mit der Erkennung von Unsicherheit und mit der Detektion von falschen Vorhersagen zur Laufzeit.

Bei statistischen Modellen im Einsatz auf Daten unterscheidet man Unsicherheit in zwei Arten: Eingabeunsicherheit und Modellunsicherheit (vgl. [21]). Zusätzlich gibt es bei Klassifikationsaufgaben die Möglichkeit, die Klassifikationsunsicherheit zu messen, etwa in Form von Dispersion auf den A-posteriori-Wahrscheinlichkeiten.

- Die Berechnung der Modellunsicherheit in neuronalen Netzen galt lange Zeit als zu kostenaufwändig, weil man dafür nicht die Gewichte des Netzes, sondern an Stelle eines jeden Gewichts eine Verteilung lernen müsste. Mit [22] wurde jedoch eine Methode bereitgestellt, Monte-Carlo Dropout zur Inferenzzeit, mit der die Modellunsicherheit mit annehmbaren Rechenaufwand approximiert werden kann. Dieser Ansatz gilt seither als State-of-the-Art im Bereich Unsicherheitserkennung.
- Auch Eingabeunsicherheit wurde im Deep Learning erfolgreich zum Einsatz gebracht (vgl. [23]).

Beide Konzepte können sowohl separat verwendet als auch miteinander verknüpft werden. Um die erkannte Unsicherheit zu quantifizieren, eignen sich verschiedene Dispersionsmaße und andere Methoden. Weit verbreitet sind die höchste Wahrscheinlichkeit der Ausgabe des neuronalen Netzes bzw. deren Abweichung von 100% oder die Entropie der Ausgabewahrscheinlichkeiten [22] [24]. Weiterhin gibt es auch ausgeklügeltere Ansätze, die sich weiterer Information aus dem Netz bedienen, etwa einer Outlier-Detektion auf dem vorletzten Level [25] oder Gradientenmaße über

beliebige Anzahlen von Lagen des Netzes [26]. Die genannten Konzepte geben auch Rückschlüsse auf Unsicherheiten im Klassifikationsvorgang ohne die Einbindung von Modell- oder Eingabeunsicherheit. Beim Einsatz von Modell- und/oder Eingabeunsicherheit können darüber hinaus weitere Größen wie die Varianz der Ausgaben über den Monte-Carlo Prozess betrachtet werden [22].

Zur Detektion von falsch klassifizierten Vorhersagen zur Laufzeit (ohne das Vorhandensein von Ground-Truth) wurde 2016 ein Baseline-Ansatz veröffentlicht. Dieser Ansatz stellt eine Maßnahme zur Klassifikation einer Vorhersage als vertrauenswürdig oder nicht vertrauenswürdig dar und kann kurz "Metaklassifikation" genannt werden. Die einfachste Form einer Metaklassifikation ist das Betrachten eines Unsicherheitsmaßes, etwa eines Dispersionsmaßes, und die Entscheidung ab welchem Schwellwert man der Vorhersage des Netzes vertraut oder eben nicht. Dabei gibt es zwei schwellwertabhängige Fehlerraten, die man gleichermaßen klein halten möchte: Die False-positive Rate (Übersehen eines kritischen Falls) und eins minus die True-positive Rate (Fehlalarm bei unkritischem Fall). Verschiedene Unsicherheitsmaße werden daher häufig schwellwertunabhängig hinsichtlich ihrer Performanz verglichen, z.B. mittels der Area Under Receiver Operatoring Characteristic Curve (AUROC) [27].

Die genannten Unsicherheitsmaße und Detektionsmechanismen finden u.a. auch im aktiven [28] und halbüberwachten Lernen [29], [30], [31], [32], [33] Anwendung. Beide Felder beschäftigen sich damit, mit möglichst wenigen Labeldaten performante neuronale Netze zu trainieren. Ausgangspunkt ist ein großer Pool an ungelabelten Daten und eine kleine Menge an initial gelabelten Daten. Auf den gelabelten Daten wird ein initiales neuronales Netz trainiert. Beim aktiven Lernen werden Vorhersagen des Netzes auf ungelabelten Daten bezüglich ihrer Unsicherheit geprüft und einige Daten mit hoher Unsicherheit nachgelabelt. Danach wird mit der nun etwas größeren Menge an gelabelten Daten trainiert. Dieser Prozess wird iteriert, bis die Modellgenauigkeit hinreichend oder stagniert ist. Beim halbüberwachten Lernen werden vorhersagen mit hoher Vorhersagesicherheit als korrekt vorhergesagt angenommen und in den Trainingsprozess eingespeist, der Rest funktioniert analog. Beide Konzepte können auch vereint werden, siehe etwa [34].

Bekanntermaßen lassen sich neuronale Netze austricksen. Es gibt einige Arbeiten, die zeigen, dass neuronale Netze durch leichte, fast unsichtbare Eingabedatenmanipulationen, sogenanntes Adversarial Noise, ihre Vorhersage verändern. Besonders auf Bilddaten wird dieses Konzept des Hackens von neuronalen Netzen angewendet, siehe etwa [35], [36], [37]. Es gibt Ansätze, diese Attacken auf neuronale Netze statistisch zu detektieren, siehe z.B. [38], [39].

KI-Module absichern

Die Anwendung der ISO 26262 "Funktionale Sicherheit" [41] hat bisher die notwendige Argumentation bezüglich Abwesenheit eines unangemessenen Risikos durch Versagen eines E/E-Systems unterstützt. Da dieser Standard jedoch maßgeblich Ansätze zur Vermeidung von systematischen Hardware- und Softwarefehlern und zufälligen Hardwarefehlern betrachtet, kann die alleinige Anwendung der ISO 26262 auf Entwicklungsprozesse beim automatisierten Fahren und insbesondere bei der Anwendung von KI-basierten Funktionen nicht die notwendige Systemsicherheit gewährleisten.

Ein erster Entwurf eines neuen Industriekonsens im Bereich Fahrerassistenzsysteme (ISO/PAS 21448 „Safety of the intended functionality“) soll das unangemessene Risiko, das durch jegliche Gefährdung, z.B. durch Systemgrenzen, erzeugt wird, minimieren. Jedoch sind die Analysen, die der SOTIF vorsieht, nur eingeschränkt für KI-basierte Funktionen verwendbar und weitere Maßnahmen werden notwendig. Weiterhin ist der SOTIF nur für Fahrerassistenzsysteme bis Automatisierungslevel SAE2 ausreichend. Für höhere Automatisierungslevel sind Anpassungen und Erweiterungen notwendig.

Auch ist der Ansatz, das System durch „Abfahren“ aller Situationen und Kombinationen zu validieren, aufgrund der großen Systemkomplexität beim automatisierten Fahren in Kombination mit der hohen Vielfalt von Situationen und äußeren Einflussfaktoren („offener Kontext“) technisch, zeitlich wie auch ökonomisch nicht realisierbar [4]. Insbesondere beim vielversprechenden Einsatz von Künstlicher Intelligenz für die maschinelle Wahrnehmung (Perception) eines automatisierten Fahrzeugs müssten alle möglichen Situationen, Betriebszustände, äußeren Einflussfaktoren und deren Kombinationen im Vorfeld mit repräsentativen Daten entsprechend trainiert und getestet werden, damit das Risiko einer unzureichenden Funktion minimiert und der sichere Betrieb der Funktion gewährleistet werden kann [42] [43] [44].

Ein vielversprechender Ansatz ist es deshalb, systemkritische Einflussfaktoren des Eingangsraums systematisch zu untersuchen und herauszuarbeiten. Damit können diese mit bekannten Methoden und Maßnahmen analysiert und reduziert werden. Hierfür werden zum Beispiel generische Methoden und Ansätze zur Beschreibung der Umgebung (Eingangs-, Ausgangs- und Testraums), in der ein automatisiertes Fahrzeug betrieben werden soll, sowie die Zusammenfassung des Eingangsraums in Äquivalenzklassen genutzt [45] [46] [47].

Die Anwendung auf Fahrerassistenzsysteme war bisher nur bedingt notwendig, da einfacher anzuwendende klassische Test-Methoden mit „Absicherungs-Datensätzen“ mit einer Bestimmung der True Positives / False Positives / False Negatives genutzt werden konnten. Diese Evaluationsmethode ist von einem Datensatz abhängig, von dem nicht erwartet werden kann, dass er für alle Situationen repräsentativ ist.

In der Literatur [45] [48] wird die wirksamkeitsorientierte Kombination von KI-spezifischen Methoden und Maßnahmen zur Sicherstellung der Absicherbarkeit von künstlicher Intelligenz als notwendiges Verfahren zum Nachweis der Absicherbarkeit

empfohlen. Im Vorhaben KI-Absicherung werden erste Ansätze von vielversprechende Methoden und Maßnahmen zur Sicherstellung der Absicherbarkeit (u.a. Explainable AI) erforscht und angewendet. Zusätzlich sind Strategien zur Absicherbarkeit einer spezifischen Funktion notwendig. Grundlegende Fragestellungen zur methodischen Entwicklung und Anwendbarkeit im KI Bereich von diesen Nachweisstrategien sind noch nicht ausreichend adressiert worden. Neueste Ansätze in diesem Bereich untersuchen, wie diese Fragestellungen zum systematischen und gesamtheitliches Vorgehen beantwortet werden können.

In [48] und [45] wird dargestellt, dass mittels Domänenanalyse der relevante Eingaberaum eingeschränkt werden kann und sich aus den Sicherheitszielen des Gesamtsystems Anforderungen an die Funktion ableiten lassen. Bertrand Meyer formulierte bereits 1992 einen Ansatz, das sogenannte Contract-based Design, in welchem Einschränkungen an Eingaben und Ausgaben genutzt werden, um die Zuverlässigkeit in softwareintensiven Systemen durch Anwendung von methodischen Richtlinien zu verbessern [49]. Hierbei formuliert eine Contract-based Design Spezifikation eine Garantie für das Ergebnis einer Funktion, unter der Annahme, dass die Eingabe ihre Spezifikation erfüllt. Analog hierzu kann der Ansatz des Contract-based Designs auch dafür genutzt werden, um zu spezifizieren, dass der Einsatz einer KI-Funktion in einer definierten Umgebung (hier: Assumptions) zu einem definierten Ergebnis (hier: Guarantee) führt.

Anhand von Sicherheitszielen können Guarantees abgeleitet werden, und anhand einer Domänenanalyse und Eigenschaften der Systemarchitektur können Assumptions abgeleitet werden. Diese bilden zusammen einen Safety Contract für eine KI Funktion (gegebenenfalls inklusive weiterer Funktionen). Für eine Funktion, die diesen Safety Contract erfüllt, kann nun validiert werden, dass sie auch das Sicherheitsziel erfüllt. Ebenso kann für eine konkrete Implementierung der Funktion verifiziert werden, dass sie den Safety Contract erfüllt. Die Validierung kann beispielsweise durch Expertenreviews geschehen, während die Verifikation beispielsweise durch strukturierte Testverfahren und mittels geeigneter Metriken (z.B. KPIs aus TP3) durchgeführt werden kann.

Angabe bekannter Konstruktionen, Verfahren und Schutzrechte, die für die Durchführung des Vorhabens benutzt wurden

Für die Durchführung des Vorhabens wurde maßgeblich auf den öffentlich verfügbaren Code zurückgegriffen, der keine Schutzrechte vorsieht. Dies umfasst sowohl das Deep Learning Framework Pytorch als auch Repositories zu DNN Entwicklungen. Daten zum Trainieren, Validieren sowie Testen der Entwicklungen wurden im Projekt im Teilprojekt 2 synthetisch erzeugt, sodass hier ebenfalls keine Schutzrechte verletzt oder Lizenzen erworben wurden. Lediglich für die Entwicklungsumgebung PyCharm wurden Lizenzen in Anspruch genommen, um die erweiterte Funktionalität dieses Frameworks für die Entwicklung der DNNs nutzen zu können.

Angabe der verwendeten Fachliteratur sowie der benutzten Informations- und Dokumentationsdienste

An Informations- und Dokumentationsdienste haben wir bei VALEO auf eigene Datenbanken im Intranet zurückgegriffen, sowie das Leibniz-Informationszentrum (Technik und Naturwissenschaften Universitätsbibliothek) genutzt. Eine Übersicht über weitere Quellen können der projekteigenen Webseite (https://www.ki-absicherung-projekt.de/veroeffentlichungen) entnommen werden. Weitere genutzte Quellen und Dienste sind dem Anhang 1 (Literaturverzeichnis) zu entnehmen.

1.5. Zusammenarbeit mit anderen Stellen

TÜV Rheinland Consulting GmbH
Der TÜV Rheinland war der Projektträger des Projektvorhabens KI-Absicherung und hat das Projekt im Auftrag des Ministeriums BMWK (ehemals BMWi) betreut.

EICT
Das Projektvorhaben wurde von einem Projektbüro begleitet, welches die Projektpartner neben der Unterstützung als Projektbüro auch bei der Ergebnisverbreitung (Webseitenerstellung, Projektfilm, Zwischen- und Abschlusspräsenation) begleitet und zum Projekterfolg beigetragen hat.

BIT Technology Solutions GmbH
BIT TS hat die Dienstleistung für den gemeinsamen Unterauftrag GUA 8 (Beitrag zur Schnittstellendefinition der Datengenerierungstoolkette) erbracht.

Neurocat GmbH
Neurocat hat die Dienstleistung für den gemeinsamen Unterauftrag GUA 10 mit folgenden Leistungen erbracht:

- Ausführbarer und dokumentierter Methodensatz zur Identifikation fehlender Testdaten
- Nachweis der Wirksamkeit der einzelnen Maßnahmen in Bezug auf die Sicherheitsziele
- Empfohlene Testmethoden für KI-Funktionen im Bereich Objektdetektion

2. Eingehende Darstellung

Die eingehende Darstellung beginnt zunächst mit einer allgemeinen Beschreibung der jeweiligen Teilprojekte und geht dann über in die partnerspezifische Arbeitsbeschreibung des Partners VALEO. Die Arbeitsfortschritte und Ergebnisse4 werden dann in zeitlicher Reihenfolge analog zu den während der Projektlaufzeit eingereichten Zwischenberichten dargestellt.

Teilprojekt 1: KI-Funktion

Das Teilprojekt 1 (Gesamt-Projektstruktur siehe Abbildung 2) ist in fünf Arbeitspakete gem. Abbildung 4 mit folgender thematischer Struktur unterteilt:

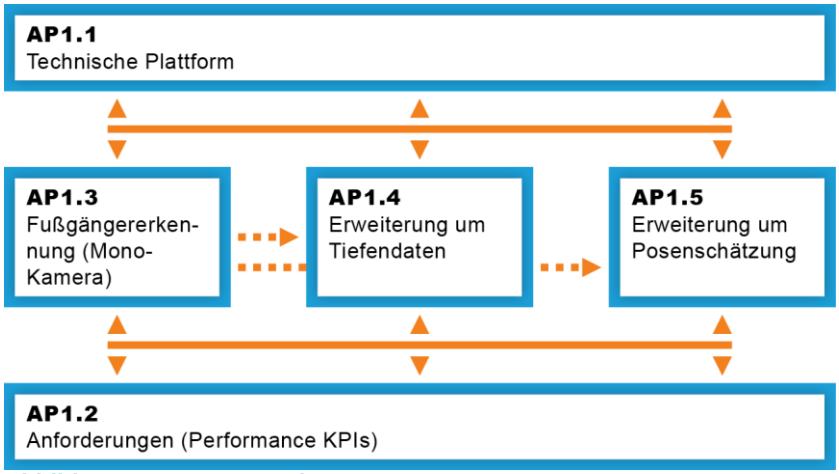


Abbildung 4: TP1-AP-Struktur

AP1.1 (Technische Plattform) stellt die technischen Grundlagen für die KI-Algorithmenentwicklung bereit und betreut diese im Projektverlauf. Hierzu zählen die Entwicklungsplattform (einheitliche Softwaretools und Bibliotheken zur Entwicklung von KI-Algorithmen und Code-Verwaltung), die Datenspeicherplattform (für alle Projektpartner zugreifbare Speicherlösung für Trainings- und Testdaten) sowie die Compute-Plattform (Rechenleistung zum Training von KI-Modellen). Zu den Umfängen von AP1.1 gehört auch das Release Management, also die Bereitstellung von Code-Paketen in 6-Monats-Zyklen.

AP1.2 (Anforderungen) koordiniert die Anforderungsspezifikation aus Sicht der Algorithmenentwicklung (Sicherheitsanforderungen werden in TP3 erarbeitet). Neben den funktionalen Anforderungen umfasst dies das Sensor-Setup, welches für die Generierung synthetischer Daten zugrunde gelegt wird und die Spezifikation der Ground-Truth-Labels (z.B. welche Klassen von Fußgängern sollen erkannt werden). Außerdem werden in AP1.2 Anforderungen an die Basisszenarien definiert sowie Gütemaße und Zielwerte aus Qualitäts- und Leistungsperspektive (z.B. Erkennungsraten, Verarbeitungsgeschwindigkeit).

[4] Ergebnisse werden in der Form Ex.x.x unter Nennung des jeweiligen Arbeitspaketes kenntlich gemacht.

AP1.3 (Fußgängererkennung – Bilddaten) konzentriert sich auf Basisalgorithmen für die Fußgängererkennung aus Bilddaten. Von den beteiligten Partnern werden mehrere Ansätze umgesetzt und anschließend verglichen. Neben Convolutional Neural Networks (CNNs) kommen Methoden der semantischen Bildsegmentierung und rekurrente Netze zum Einsatz.

AP1.4 (Fußgängererkennung – Bild- und Tiefendaten) befasst sich mit der effektiven Kombination von Bild- und Tiefendaten für die Fußgängererkennung. Es werden verschiedene Ansätze zur Fusion (z.B. Early und Late Fusion) gegenübergestellt. Jeder dieser Ansätze bringt Vor- und Nachteile aus funktionaler Sicht mit sich und kann sich unterschiedlich auf Sicherheitsziele auswirken.

Im AP1.5 (Fußgängererkennung – Posenschätzung) liegt der Schwerpunkt auf der Posenschätzung, d.h. auf der Ermittlung der exakten Körperhaltung der erkannten Fußgänger. Dies ist ein wichtiger Input für weiterführende Algorithmen, z.B. die Absichtserkennung.

Valeo war an den AP1.1-4 beteiligt.

AP1.1 Technische Plattform (2 PM[5])

Aufgaben Valeo:

Minimaler Docker-Container mit konfigurierter Entwicklungsumgebung (E1.1.1a)

• Testen des Docker-Images mit integriertem Framework basierend auf eines Standard Computer Vision Datensatzes wie z.B. MNIST oder Caltech Pedestrian.

• Testen des Trainings (GPU-Nutzung) und der Inferenz eines einfachen CNN's innerhalb des Docker-Images.

Stand der Arbeiten (31.12.2020):

Nachdem eine erste Version des Docker-Image von *Luxoft* bereitgestellt wurde, hat Valeo die Vorabversionen getestet und bei Fehlern entsprechende Verbesserungen mit eingebracht. Die CUDA Unterstützung wurde im Docker-Image mit integriert, so dass CNNs entwickelt und ausgeführt werden können. Der erste Release wurde seitens Valeo ausgiebig getestet.

Stand der Arbeiten (30.06.2020):

Die Arbeiten im Rahmen von AP1.1 wurden mit dem Release des Docker Containers abgeschlossen.

[5] Von Valeo ursprünglich eingeplanter Arbeitsumfang im Projektvorhaben

AP1.2 Quantitative und qualitative Anforderungen an die KI-Funktion (Performance-KPIs) (6 PM)

Aufgaben Valeo:

Definition der Referenz Lidarmodelle und ihrer Verbaupositionen (E1.2.2)

• Schwerpunkt auf der Definition von Lidarmodellen auf Basis der Erfahrungen in der Entwicklung von Lidarmodulen.

Spezifikation der Daten Annotation (E1.2.3)

• Schwerpunkt liegt auf der Annotation von Lidar Daten.

Definition funktionaler Anforderungen und Einteilung in Fähigkeitsstufen (E1.2.6)

• Schwerpunkt sind die funktionalen Anforderungen bezogen auf Lidar Daten.

Definition der Performanz- und Qualitäts-KPIs und ihrer Zielwerte (E1.2.7)

• Schwerpunkt liegt auf Performanz- und Qualitäts-KPIs bezogen KI-Algorithmen die Lidar Daten verwenden.

Stand der Arbeiten (31.12.2019):

E1.2.2: Der Referenz-LiDAR ist in Koordination mit E2.1 und E2.5 spezifiziert worden als Nah-Infrarot Laser Scanner mit rollendem Verschluss und einer Bildwiederholrate von 25 Hz. Das Scan-Pattern ist angelehnt an existierende Hardware. Zwei ausgezeichnete Verbaupositionen wurden ausgewählt: Angelehnt an den öffentlichen A2D2 Datensatz und an ein realistisches Serienfahrzeug.

Stand der Arbeiten (30.06.2020):

E1.2.2: Im ersten Halbjahr haben wir in enger Zusammenarbeit mit E2.1 ein Referenz-LiDAR-Scanmuster und zwei relevante Montagepositionen definiert. Die Einbaupositionen sind 1.) vordere Stoßstange, wie bei einem realistischen Serienauto, und 2.) Dach vorne Mitte, wie für das LiDAR in der vorderen Mitte im öffentlichen A2D2 (AEV)-Datensatz.

Stand der Arbeiten (31.12.2020):

E1.2.2: An der Spezifikation des synthetischen LiDAR Modells hat es im genannten Zeitraum keine wesentlichen Änderungen gegeben. Die Abstimmung mit den Projektpartnern in E2.1 wurde regelmäßig unter Rücksichtnahme auf die bestehende Spezifikation fortgeführt. Da die bisher spezifizierte LiDAR-Abtastrate von 25 Hz kein Vielfaches der Kamera-Abtastrate darstellt, wäre im Hinblick auf die Datenerzeugung eine Änderung auf beispielsweise 50 Hz evtl. sinnvoll und weiter innerhalb E2.1 zu

diskutieren.

Stand der Arbeiten (30.06.2021):

E1.2.2: Die Spezifikation des synthetischen LiDAR Modells wurde auf den neusten Stand gebracht. Es gab den Bedarf die Konfiguration flexibler zu gestalten. Hierzu wurde ein generischer LiDAR eingeführt bei dem jetzt beispielsweise die FoV-Parameter vom Nutzer gesetzt werden können. Der generische LiDAR richtet sich nach wie vor nach dem grundlegenden Scan Pattern mit der Aufteilung in Gruppen von Empfangseinheiten.

Es wurden folgende Modi eingeführt:

- Generic: hier kann die Konfiguration vom Nutzer gesetzt werden
- Standard: es wird die in E1.2.2 als Beispielkonfiguration genannte Konfiguration simuliert
- Realistic: es wird ein realistischer Mobility Kit LiDAR von Valeo simuliert

Weiterhin soll mit Tranche 6 der Rolling Shutter eingeführt werden.

Zudem wurde der LiDAR als Lichtquelle hinzugefügt, hierbei soll der LiDAR als Gaussian Beam modelliert werden. Es ist unklar, ob dies bis Tranche 6 eingeführt werden kann.

E1.2.3: Die Definition des LiDAR-Datenformats wurde vereinfacht. Es handelt sich jetzt um eine PCD Datei, die die Non Unique ID sowie die kartesischen Koordinaten und die Echo Pulsbreite der empfangenen Echos enthält.

Für die Instanzsegmentierung wird eine PCD Datei angelegt, die neben der Non Unique ID und der kartesischen Koordinaten die Instanz ID der Echos enthält. Hierbei wird die Zahl 0 für Punkte verwendet, denen keinen Instanz ID zugeordnet werden konnte.

Stand der Arbeiten (31.12.2021):

E1.2.2: Die Definition der verschiedenen Modi wurde verfeinert. Insbesondere das Scanpatterns des realistischen LiDAR wurde detailliert beschrieben. Der Rollingshutter wurde mithilfe von Intel implementiert und spezifiziert. Somit ist es jetzt auch möglich die Rotation des Spiegels korrekt zu beschreiben.

E1.2.3: Es wurde für die Ausgabe der Groundtruth eine zusätzliche “ideale” Punktwolkendatei angelegt. Diese Punktwolke unterliegt keinen Sensoreffekten wie dem Rollingshutter oder dem Sensitivitätsprofil. Es werden kartesische Koordinaten und Instanzsegmentierung ausgegeben.

Stand der Arbeiten (30.06.2022):

E1.2.3. Mit Tranche 7 wurde das von Valeo erstellte Sensormodell genutzt.

Ursprünglich hatten Kamera und LiDAR eine unterschiedliche Anzahl an Frames pro Sekunde, dies war dem Realismus geschuldet. Es wurde jetzt eingeführt, dass beide die gleichen Frequenzen implementieren, um auch leichter mit der Ground Truth arbeiten zu können.

In Zusammenarbeit mit Intel und Bit TS wurde das Verhältnis der Kamerabilder und der LiDAR Punktwolken zu den Bounding Boxen geprüft. Es stellte sich heraus, dass die Bounding Boxen zu den Punktwolken passen. Weiterhin ist es sinnvoll möglich die LiDAR Punktwolke auf die Kamerabilder zu projizieren. Somit ist **E1.2.3** abgeschlossen.

AP1.3 Implementieren von Algorithmen zur Fußgängererkennung (12 PM)

Aufgaben Valeo:

Implementierung der funktionalen Algorithmen: 3D Object-Information aus Video-Sequenzen (E1.3.3c)

• Mit Hinblick auf AP1.4 erweiterte Architektur des vielversprechendsten 2D Objekt Box Schätzer (Pixelkoordinaten) der Partner zum Zwecke eines neuen Netzwerkdesigns, welches in der Lage ist, 3D-Boxen für Fußgänger in Weltkoordinaten auf lediglich monokularen Kamerabildern zu schätzen.

• Training auf Daten mit festem Kamera-Viewpoint. Somit kann das Netz die extrinsische Kalibrierung indirekt selbst lernen.

• Erweiterter Detektor mittels 3D-Intersection Over Union basierend auf Benchmark Datensätzen. Die Ergebnisse auch werden mit den Entwicklungen der Objektpartner (2D-Boxen) mittels Rücktransformation in die Bildebene verglichen.

• Valeo untersucht in diesem AP die Effizienz bzw. Genauigkeit der von den Projektpartnern entwickelten Detektoren (Fußgänger) für 2D Objekt Bounding Boxen (z.B. YoloV3, Faster RCNN, DSSD, RetinaNet etc.) auf monokularen Kamera-Bildern mittels 2D Intersection Over Union basierend auf Benchmark-Datasätzen.

Stand der Arbeiten (31.12.2019):

Das Schlussfolgern von 3D Objektinformationen aus monokularen Kamerabildern erwies sich als herausfordernd. Ein Netz, welches dennoch eine gute Schätzung erlaubt, ist eine Sub-Version von *Frustum-PointNet*, welche ausschließlich auf dem Kamerabild arbeitet. Die prinzipielle Funktionsweise ist in folgender Abbildung veranschaulicht:

Abbildung 5: 3D Objektinformationen mit Hilfe einer Sub-Version von Frustum-PointNet

Als Eingabe erwartet das Netz einen 2D Bounding Box Vorschlag einer vorgeschalteten 2D Objekterkennung. Die Entscheidung ob und was für Objekte im Bild enthalten sind, unterliegt also ganz der Verantwortung des vorgeschalteten Schätzers. Erst anschließend schätzt das eigentlich betrachtete Netz die Eckpunkte einer potenziellen 3D Box im Kamerabild. Diese Aufgabe stellt sich als relativ einfach heraus, da keine Schlussfolgerung in die tatsächliche 3D Domäne erforderlich ist. Deutlich schwieriger ist die Distanz- und Ausdehnungsschätzung, die das Netz zusätzlich vornimmt - aufgrund der fehlenden Tiefeninformation des Bildes sind diese nur ein Richtwert und nicht hinreichend genau. Ein weiteres Optimierungsverfahren sorgt nun dafür, dass diese geschätzten Werte an die tatsächliche Realität angenähert werden: Ein Box-Regressor initialisiert eine 3D Bounding Box mit der geschätzten Ausdehnung und Position in der 3D Szene, projiziert diese mithilfe der intrinsischen Kamerakalibrierung in das Kamerabild und minimiert letztendlich die Distanz der eben projizierten Eckpunkte zu den vorher geschätzten Eckpunkten. Das Resultat sind, verglichen mit anderen monokularen Ansätzen, gut geschätzte 3D Informationen:

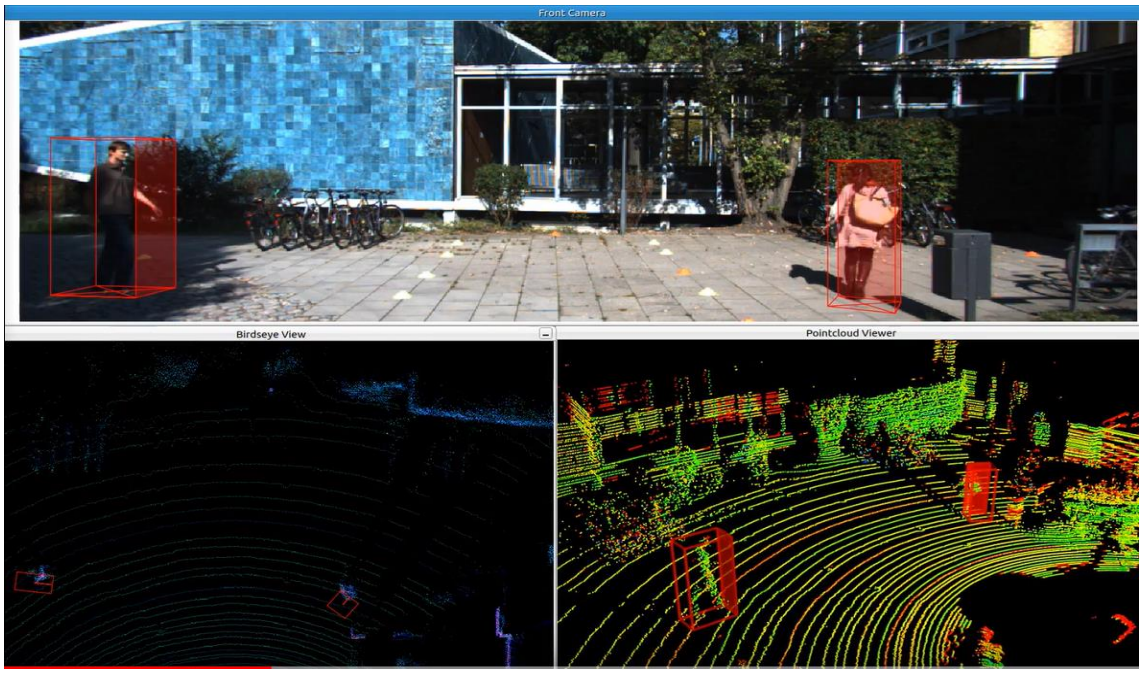


Abbildung 6: Fußgängererkennung mittels Image-Only Frustum-PointNet auf KITTI Testdaten

Die *PyTorch* Implementierung des Neuronalen Netzes ist öffentlich frei verfügbar. Sie wurde übernommen und angepasst, um das Netz auf Fußgänger im KITTI Datensatz zu trainieren. Obige Abbildung zeigt das Inferenzergebnis auf Testdaten. Das Netzwerk erhält als Input lediglich (a) das im oberen Bereich zu sehende Kamerabild

und (b) die 2D Bounding Box Vorschläge einer vorgeschalteten 2D Objekterkennung (hier: mittels Mask-RCNN). Die geschätzte 3D Box kann dann zu reinen Visualisierungszwecken in der zugehörigen Punktwolke veranschaulicht werden. Dort kann man auch eine leichte Verschiebung der Box um die eigentliche Position aus Sicht der Vogelperspektive erkennen. Dies zeigt, dass die 3D Position nicht optimal geschätzt werden kann - natürlich aufgrund der fehlenden Tiefeninformation im Bild. Die Schätzung der Objektausdehnung und dessen Orientierung funktioniert auf der anderen Seite mit diesem Ansatz sehr gut. Die generelle Genauigkeit hängt aber stark von der Genauigkeit der vorgeschalteten 2D Objekterkennung ab. Hier wurden ebenfalls verschiedene Ansätze untersucht, so zum Beispiel YoloV3 und Detectron2 (Mask-RCNN) von Facebook. Insbesondere wurden Probleme mit verdeckten Objekten beobachtet, die das Schätzungsergebnis des 3D Schätzers deutlich beeinträchtigen. Im weiteren Verlauf der Arbeit sollten diese Probleme noch adressiert werden.

Eine Bewertung des Ansatzes anhand einer Metrik konnte noch nicht durchgeführt werden. Problem ist die mangelnde Anzahl an Fußgänger Instanzen in bisher betrachteten Datensätzen. Um die Genauigkeit zu steigern wurden deshalb alle Instanzen für das Training verwendet, was momentan kein faires Benchmarking erlaubt (Hinweis: obige Abbildung zeigt KITTI Testdaten, auf die nicht trainiert wurde, jedoch sind für diese keine Annotierungen veröffentlicht, sodass auch hier keine Metrik berechnet werden kann).

Stand der Arbeiten (30.06.2020):

Der Ansatz wurde mit dem Ziel weiterverfolgt, die Inferenzfähigkeit des Netzes zu evaluieren. Öffentlich verfügbare Datensätze zur 3D Objekterkennung enthalten unzureichend viele Fußgängerinstanzen. Als Beispiel dient der weitverbreitetste Datensatz innerhalb der 3D Objekterkennung: KITTI. Dort gibt es lediglich 4,487 annotierte Fußgänger, während 28,742 PKWs vorhanden sind. Zusätzlich stehen zahlreiche unannotierte Bilder und gesamte Szenen-Sequenzen als Einzelframes zur Verfügung, um die Algorithmen qualitativ zu validieren und zu testen. Für ein qualitativ besseres Ergebnis wurden zunächst alle annotierten Fußgängerinstanzen für das Training herangezogen. Es wurden zwei verschiedene KITTI Videosequenzen durch das Netz gefüttert, wobei das Netz die Videos im Einzelframe-Stil verarbeitet. Im Folgenden sind Auszüge aus den Videos zu sehen, welche eine qualitative Einschätzung ermöglichen:

Abbildung 7: Visualisierung der 3D Bounding Boxes Prädiktion auf dem KITTI Datensatz

Für eine quantitative Beurteilung der Erkennungsfähigkeit können zahlreiche Metriken herangezogen werden, wobei alle auf Ground Truth Objekte angewiesen sind, um das Inferenz-Ergebnis mit dem ideal erwarteten Output zu vergleichen und dessen Abweichung zu quantifizieren. Dazu muss natürlich eine Aufteilung des Datensatzes in Trainings- und Validierungsdaten erfolgen, da ein Benchmarking auf Trainingsdaten keine gute Evaluierung darstellt. Leider enthält KITTI standardmäßig bereits unzureichend viele Fußgänger Instanzen, sodass wir die Trainingsdatenmenge nur ungern noch weiter aufteilen wollen. Stattdessen wurde zur Evaluierung mit separaten Validierungsdaten ein interner Datensatz herangezogen, welcher über hinreichend viele Instanzen verfügt. Nur so kann eine faire Evaluierung stattfinden.

Konkret berechnet wurden (a) die durchschnittliche *Intersection over Union (IoU)* über alle 3D Erkennungen im Vergleich zur Ground Truth und (b) der *SRT-Score*. Da es sich um einen Ansatz handelt, der als Input 2D Box Vorschläge annimmt, treffen wir zunächst die Annahme einen idealen 2D Box Erkenner vorgeschaltet zu haben, der alle Ground Truth Fußgänger mit 100% IoU abdeckt. Damit können wir uns bei der Evaluierung vollkommen auf das 3D refinement im untersuchten Ansatz konzentrieren. Nachdem die Ground Truth 2D Box samt Bild durch das Netz gefüttert wurde, erhalten wir 3D Boxen, die wir nun mit der gleichen annotierten Ground Truth vergleichen. Man stellt fest, dass diese Boxen im 3D und im Durchschnitt nur noch mit ca. 15.8% IoU überlappen. Grund für diesen niedrigen Overlap ist der eigentliche Schritt der Tiefenauflösung und 3D Schätzung aus der monokularen Bildinformation, welcher das Genauigkeits-Bottleneck darstellt. Es ist wahrscheinlich, dass diese ermittelten 15% als upper-bound Genauigkeitsgrenze fungieren, d.h. ein System mit vorgeschalteten nicht-idealen 2D Detektor würde den finalen 3D Box overlap im Durchschnitt weiter drücken. Diese Annahme wurde durch Simulation von nicht-idealen 2D Boxen als Input bestätigt. Weiterhin wurde das System mit einem tatsächlichen 2D Detektor getestet (hier: YoloV3).

Weiterhin wurde der SRT-Score berechnet, da dieser perfekt die einzelnen Aufgaben der 3D Objekterkennung abbildet. Es ist eine Komposition aus einem Score, der

Rotations-, Translations- und Dimensionsschätzung getrennt beurteilt. Während eine perfekt geschätzte Box, aber um 180° gedrehte, auch 100% Überlappung im Vergleich zur Ground Truth zeigt, würde SRT die Fehlschätzung im Rotationsscore deutlich aufzeigen. Der gesamt SRT Score mit allen vereinigten Einzelscores für diesen Ansatz ist 0.466.

Stand der Arbeiten (31.12.2020):

Während dieses Entwicklungszyklus wurde ein Großteil der Arbeit an der Refraktionierung des Codes zur Verwendung des KIA-Datensatzes durchgeführt. Der Datensatz von Bit-TS wurde verwendet, da er die erforderlichen 3D-Ground Truth Daten enthält. Es wurden Datenlader implementiert, um den neuen Datensatz zu laden, damit Training und Inferenz auf diesem neuen Datensatz durchgeführt werden kann. Der Code wurde angepasst, um den KIA-Datensplit für Release 2 zu verwenden. Der neue Release-Management-Prozess wurde berücksichtigt und der entwickelte Code wurde an diesem Prozess angepasst. Der Code wurde in gitlab eingecheckt. Es wurde ein iterativer Prozess verfolgt, um die Probleme mit dem release process zu beheben und die automatisierten Trainings- und Inferenz-Smoketests mit Beispieldaten durchzuführen.

Da es noch keine Klarheit über die Lizenzierungsfrage in Bezug auf die vortrainierten Modelle gibt, wurde die Trainingsstrategie dahingehend geändert, dass die Parameter zufällig und nicht mit vortrainierten Gewichten initialisiert werden. Die ursprüngliche Version der Tranche3-Daten, die zur Verfügung gestellt wurde, wies mehrere Probleme auf, die die Leistung des Algorithmus behinderten und es konnten keine plausiblen Ergebnisse erzielt werden. Die Probleme im Zusammenhang mit dem Datensatz wurden denen, die die Daten generieren, kommuniziert. Wie in der vorherigen Berichtsphase erwähnt, verwendet dieser Algorithmus die Metrik der Scale Rotation Translation (SRT) score, um die Korrektheit der erkannten Bounding Boxes zu bewerten. Die fehlerhaften Bounding Boxes in der Ground Truth würden auch diese Auswertung behindern.

Darüber hinaus wurden Untersuchungen in der Richtung durchgeführt, den aktuellen two-stage monokularen Objektdetektor durch einen single-stage Detektor zu ersetzen, der die Notwendigkeit eines anfänglichen 2D-Detektors eliminieren und den Overhead reduzieren würde. Ergebnisorientierte Experimente wurden mit öffentlich verfügbaren Datensätzen und dem KIA-Datensatz durchgeführt, um den Kompromiss zwischen Leistung, Ressourcen und Komplexität des Algorithmus zu verstehen. Aufgrund von Zeitbeschränkungen nach der Bereitstellung der Hotfixes konnte dieser Ansatz jedoch nicht in diese Version aufgenommen werden. Ein regelmäßiger Austausch über developer sync und Feedback-Prozesse wurde auch mit Stakeholdern durchgeführt, um die während der Entwicklungsphase aufgetretenen Probleme zu besprechen. Zusätzlich zu den Entwicklungsaktivitäten, übernahm Valeo weiterhin die AP Co-Lead Rolle in diesem AP und beteiligte sich aktiv an den Projekttreffen, der Berichtsphase und der Bereitstellung von Feedback für die Datengenerierung.

Evaluation der untersuchten Deep-Learning-Modelle (E1.3.4)

Stand der Arbeiten (30.06.2021):

Während dieses Entwicklungs- und Berichtszeitraums wurde der Algorithmus für die 3D-Objekterkennung aus monokularen Kamerabildern von einem two-stage zu einem single-stage Algorithmus geändert. Der frühere Algorithmus verwendete 2D-Bounding Boxes als input für die Schätzung der 3D-Boxen. Dieser Ansatz war aufgrund der two-stage Berechnung und der Verwendung bestimmter geometrischer Optimierungstechniken für die Schätzung der 3D-Boxen langsam. Der Algorithmus war auch in Bezug auf die Genauigkeit nicht gut und konnte nur so gut sein wie der vorherige 2D-Detektor, der als vorheriges Modul benötigt wurde. Daher wird ein single-stage Algorithmus verwendet, der auf der keypoint estimation basiert.

Es wurden Dataloader implementiert, um die Projektdaten aus Bit-TS zu laden. In der Anfangsphase dieses Berichtszeitraums wurden im Rahmen des Projekts Hotfixes und Fix-Scripts für die Daten der Tranche 3 bereitgestellt. Es wurde versucht, die Modelle auf diesen korrigierten Daten zu trainieren, und weitere Probleme, die dabei auftraten, wurden den betroffenen Interessenvertretern mitgeteilt. Tranche 4 der Daten wurde ebenfalls in diesem Zeitraum zur Verfügung gestellt, aber das Problem mit den orientierten Boxen behinderte die Leistung des Algorithmus und führte zu Verzögerungen in der Projektentwicklung. Die für die Probleme bereitgestellten Hotfixes gelten nur für statische Objekte, und die dynamischen Objekte müssen vor der Vorbereitung des Datensatzes für das Training gefiltert werden. Der Code wurde angepasst, um dem Release-Management-Prozess im Projekt zu entsprechen. Die Auswertung der Ergebnisse für accuray ist für die nächste Release-Phase geplant, da verschiedene Kompatibilitätsprobleme beim Versuch, auf einer Kombination verschiedener Tranchen zu trainieren, aufgetreten sind.

Während des Dev-Sync wurden zwischen den Entwicklern regelmäßig Updates zum Stand der Projektentwicklung und zu Problemen mit der Datenfreigabe ausgetauscht. Valeo hat auch in diesem AP die Rolle des Co-Lead von AP1.3 weitergeführt und durch die Teilnahme an Projekttreffen sowohl auf inter- als auch auf intra-TP-Ebene einen Beitrag geleistet.

Stand der Arbeiten (31.12.2021):

Der Single-Stage Monocular 3D-Objekterkennungsalgorithmus wurde mit Bit-TS Tranche3 und Tranche4 trainiert. Der offizielle Datensplit des Projekts wird verwendet, um den Datensatz in Train-, Val- und Test-Datensätze aufzuteilen. Da der Ground Truth-Datensatz Box-Annotationen für alle Objekte in der Szene enthält, einschließlich verdeckter Fußgängerboxen, wurden diese verdeckten Objekte in einem Datenvorbereitungsschritt herausgefiltert. Es wurde ein occulusion-estimation Wert,

der von der Überlappung der 2D-Boxen abhängt und Objekte, die zu mehr als 40 % verdeckt sind, beim Training nicht berücksichtigt.

Der Code wurde als Teil von Release 3 veröffentlicht. Der Code wurde angepasst, um dem Release-Management-Prozess zu entsprechen. Beispielhafte Eingabedaten und qualitative Ergebnisse werden als Referenz zur Verfügung gestellt.

Abbildung 8: Visualisierung der 3D Bounding Box Prädiktion auf dem KIA Datensatz

Valeo übernahm auch weiterhin die Rolle des Co-Leiters von AP1.3 und beteiligte sich an der Planung und Kommunikation des Projekts. Regelmäßige Updates und Ideen wurden auch während der vierteljährlichen Dev-Sync-Sitzungen ausgetauscht.

Stand der Arbeiten (30.06.2022):

In diesem Berichtszeitraum wurde der monokulare 3D-Objekterkennungsalgorithmus auf dem kombinierten Datensatz von BIT-TS und Mackevision trainiert. Valeo beteiligte sich am Release 4 des Projekts und gab den Code als Teil des Freigabeprozesses frei. Für die Freigabe wurde der offizielle Projektdatensplit verwendet. Das trainierte Modell wurde anhand der Projektdaten evaluiert. Die qualitativen und quantitativen Ergebnisse zeigten, dass dieser einstufige Algorithmus, der auf der Schätzung von Schlüsselpunkten basiert, unseren zuvor verwendeten zweistufigen Ansatz übertrifft. Die Endergebnisse des Algorithmus werden hier als Referenz zur Verfügung gestellt. Darüber hinaus wurde der Abschlussbericht des Arbeitspakets erstellt und veröffentlicht.

Abbildung 9: Prädiktionen vom 3D Monocular Objekterkennungs Algorithmus auf dem KIA Datensatz

Valeo leistete weiterhin einen Beitrag zum AP in seiner Rolle als Co-Leiter. Ideen und Probleme wurden im Rahmen des Dev-Sync ausgetauscht. Valeo präsentierte seine Arbeit im Laufe des Projekts einschließlich der Endergebnisse in der Final Dev-Sync, zu der alle Partner der KI-Absicherung eingeladen waren.

AP1.4 Erweiterung der Fußgängererkennung um Tiefendaten (30 PM)

Aufgaben Valeo:

Valeo wird in diesem AP einen Algorithmus zur 3D Box Prediction designen, implementieren und evaluieren. Der Fokus von Valeo liegt dabei auf der sequentiellen Fusion

Sequentielle Fusion (E1.4.5)

• Vorverarbeitung von Training-/Test-/Validierungsdaten.

• Transformation der Punktwolke von LIDAR- in Kamerakoordinatensystem.

• Occlusion Correction

• Frustumerstellung (Sichtkegel) mittels Kamera 2D CNN aus AP1.3.

• Punktewolkenfilterung (Reduzierung der Punktwolke) mittels Camera CNN.

• Netzwerkarchitektur-Aufbau

• Auswahl einer geeigneten Basis Netzarchitektur zur Punktewolkenverarbeitung (z.B. PointNet, Frustum-Pointnet, Pointnet++, Voxelnet).

• Design der Gesamtarchitektur (Camera --> 2D CNN, Punk) für das Training

• Layer-Implementierung

• Training

• Hyperparameter-Auswahl (z.B. mit Gridsearch/evolutionäre Algorithmen basierend auf QualitätsKPIs aus AP1.2).

• Trainingsprozess-Monitoring (Regelmäßiges Validieren auf Validierungsdatensatz, Skripte zur Evaluation der Performance-KPIs coden, 3D Intersection Over Union).

• Nachverarbeitung

• Vereinen von Layern (z.B. Batch Normalization Layer in Convolutional Layer).

• Visualisieren von 3D Bounding Boxen in Kamera- und Weltkoordinaten.

• Evaluation der Methode auf Benchmark Datensätzen (aus AP2.5) mittels 3D IoU Metrik und Vergleich mit den anderen Fusionskonzepten aus E1.4.1 bis E1.4.4.

<u>Stand der Arbeiten (31.12.2019):</u>

Für die 3D Objektschätzung mittels Tiefen-/LiDAR-Daten wurde das Netzwerk Frustum-PointNet näher untersucht. Es kennzeichnet sich durch hoch akkurate Inferenzfähigkeit und öffentlich-freier Verfügbarkeit. Die prinzipielle Idee ist in folgender Abbildung visualisiert:

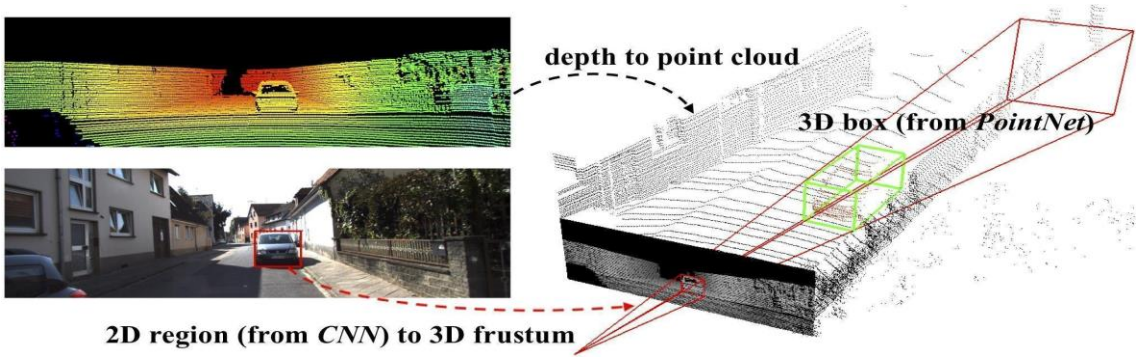


Abbildung 10: Illustrierung des Ansatzes zur 3D Objektschätzung mit dem Frustum-PointNet

Funktionsweise der Objekterkennung via Frustum-PointNet (Charles R. Qi et al: “Frustum PointNets for 3D Object Detection from RGB-D Data”)

Das Netz erwartet als Eingabe einen 2D Bounding Box Vorschlag eines bereits vorgeschalteten Netzes. Diese erste Stufe berücksichtigt ausschließlich das monokulare Kamerabild (Abhängigkeit zu E.1.3.x) und übernimmt die Verantwortung darüber, ob und welche Objekte erkannt werden. Anschließend schätzt Frustum-PointNet dann aus dem zugehörigen Ausschnitt der Punktwolke die 3D Geometrie des Objekts (wie z.B die tatsächliche Ausdehnung, Position und Orientierung des Objekts). Die gesamte Pipeline repräsentiert damit einen Ansatz zur sequentiellen Fusion von Kamerabild und LiDAR Punktwolke (E1.4.5).

Die frei verfügbare *PyTorch* Implementierung wurde übernommen und gemäß den Ansprüchen des Projekts angepasst. Anschließend wurde das Netzwerk auf Fußgänger des KITTI Datensatzes (vgl. http://www.cvlibs.net/datasets/kitti/) neu trainiert. Bisherige Ergebnisse ermöglichen bereits eine visuelle Bewertung der grundlegenden Performance und Genauigkeit des Ansatzes:

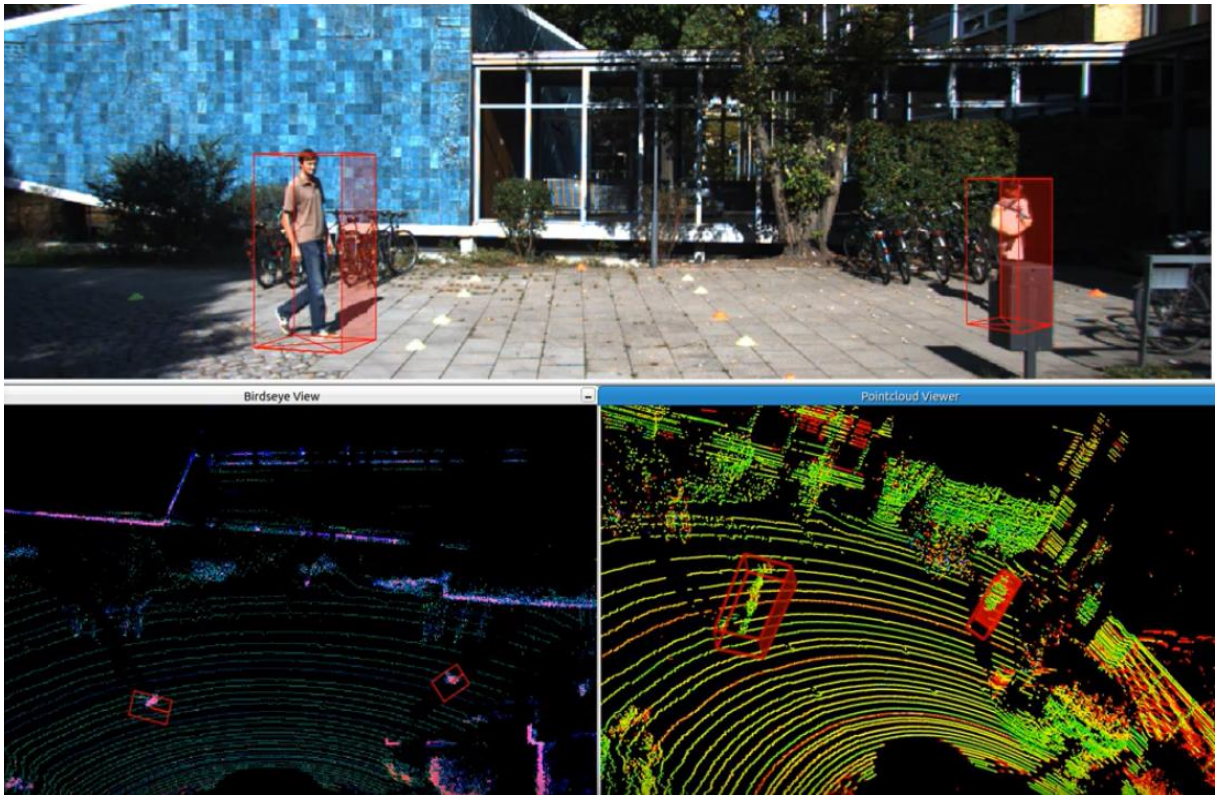


Abbildung 11: Fußgängererkennung mittels Frustum-PointNet auf KITTI Testdaten

Das Netz profitiert von den 3D Daten des LiDAR Sensors, über sie können 3D Eigenschaften robust geschlussfolgert werden, insbesondere die 3D Ausdehnung und Position. Hinsichtlich Orientierung ist auffällig, dass spärliche Punktwolken hier leicht zu inkorrekten Ergebnissen führen. Mit Verweis auf E.1.3.3c wäre es beispielsweise sinnvoll bereits in der ersten Stufe weitere Merkmale aus dem Kamerabild zu extrahieren, welche für die Orientierungsschätzung Mehrwert bringen. Eine weitere Erkenntnis ist, dass die Genauigkeit des Frustum-PointNets stark von der Genauigkeit des vorgeschalteten 2D Detektors abhängt. Insbesondere wurden Probleme mit verdeckten Objekten beobachtet. In naher Zukunft sollte deshalb auch die 2D Erkennung weiter verbessert werden. Auch sollte in finalen Systemen, Frustum-PointNet auf den vorgeschalteten Detektor abgestimmt werden, d.h. im Training sollte man die 2D Box Erkennungen eines tatsächlich verwendeten 2D Detektors berücksichtigen.

Aufgrund der mangelnden Anzahl an Fußgänger in bisher verwendeten Datensätzen wurden alle Objektinstanzen für das Training verwendet, so konnte die Genauigkeit der Schätzung maximiert werden. Im Gegenzug bedeutet dies, dass noch keine Benchmark-Metrik berechnet werden konnte, da ein Benchmark auf Trainingsdaten

keine faire Beurteilung erlaubt (Hinweis: obige Abbildung zeigt die Inferenz auf KITTI Testdaten, für die aber keine Ground Truth Daten veröffentlicht wurden, weshalb auch hier keine Metrik berechnet werden kann.).

Weiterhin wurde die Nutzung des *AUDI* Datensatzes untersucht. Auch dieser enthält - ähnlich wie KITTI - eine unzureichende Menge an Fußgänger Instanzen. Hinzu kommt das viele Objekte sogar unpräzise annotiert wurden, weshalb die Arbeit am Datensatz vorerst wiedereingestellt wurde. In vergangenen Wochen wurde auch der Datensatz *NuScenes* untersucht, welcher deutlich mehr und auf dem ersten Blick auch saubere Annotierungen zeigt. Ein Training des Netzes auf diesen Datensatz steht allerdings noch aus. Für alle Datensätze gilt: es wurde *Python* Code entwickelt, um die jeweiligen 3D (und 2D) Bounding Boxen im jeweiligen Format zu visualisieren.

Stand der Arbeiten (30.06.2020):

Wie in AP1.3 ausführlicher beschrieben, wurden auch hier die gleichen KITTI Videosequenzen durch das Netz gefüttert, wobei das Netz über Parameter verfügt, die auf allen Instanzen des Datensatzes optimiert wurden. Da dieser Ansatz hier mit der tatsächlichen Punktwolke arbeitet, wird die Erkennung in der LiDAR Punktwolke visualisiert. Im Folgenden sind Punktwolken-Auszüge mit 3D Erkennung zu sehen, welche eine qualitative Einschätzung ermöglichen:

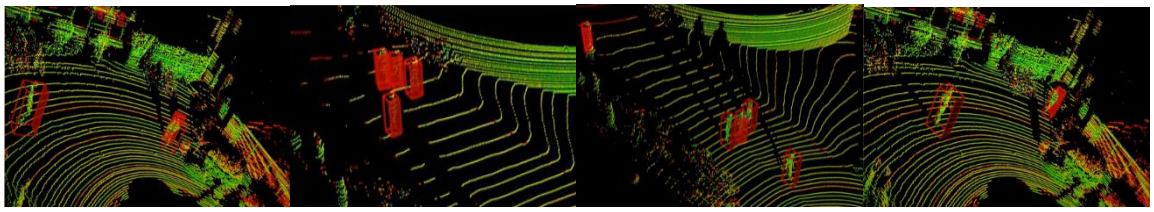


Abbildung 12: Visualisierung der 3D Bounding Box Erkennung anhand der LiDAR Punktwolke

Für eine quantitative Beurteilung der Erkennungsfähigkeit können zahlreiche Metriken herangezogen werden, wobei alle auf Ground Truth Objekte angewiesen sind, um das Inferenz-Ergebnis mit dem ideal erwarteten Output zu vergleichen und dessen Abweichung zu quantifizieren. Dazu muss natürlich eine Aufteilung des Datensatzes in Trainings- und Validierungsdaten erfolgen, da ein Benchmarking auf Trainingsdaten keine gute Evaluierung darstellt. Leider enthält KITTI standardmäßig bereits unzureichend viele Fußgänger Instanzen, sodass wir die Trainingsdaten Menge nur ungern noch weiter aufteilen wollen. Stattdessen wurde zur Evaluierung mit separaten Validierungsdaten ein interner Datensatz herangezogen, welcher über hinreichend viele Instanzen verfügt. Nur so kann eine faire Evaluierung stattfinden.

Konkret berechnet wurden (a) die durchschnittliche *Intersection over Union (IoU)* über alle 3D Erkennungen im Vergleich zur Ground-Truth und (b) der *SRT-Score*. Da es sich um einen Ansatz handelt, der als Input 2D Box Vorschläge annimmt, treffen wir zunächst die Annahme einen idealen 2D Box Erkenner vorgeschaltet zu haben, der

alle Ground Truth Fußgänger mit 100% IoU abdeckt. Damit können wir uns bei der Evaluierung vollkommen auf das 3D refinement im untersuchten Ansatz konzentrieren. Nachdem die Ground-Truth 2D Box samt Bild und LiDAR Punktwolke durch das Netz gefüttert wurde, erhalten wir 3D Boxen, die wir mit der gleichen annotierten Ground Truth vergleichen. Man stellt fest, dass diese Boxen im 3D und im Durchschnitt nur noch mit ca 54.7% IoU überlappen. Es ist wahrscheinlich, dass diese ermittelte durchschnittliche IoU als upper-bound Genauigkeitsgrenze fungieren, d.h. ein System mit vorgeschalteten nicht-idealen 2D Detektor würde den finalen 3D Box overlap im Durchschnitt weiter drücken. Diese Annahme wurde durch Simulation von nicht-idealen 2D Boxen als Input bestätigt.

Weiterhin wurde noch der SRT-Score berechnet, da dieser perfekt die einzelnen Aufgaben der 3D Objekterkennung abbildet. Es ist eine Komposition aus drei Scores, welche Rotations-, Translations- und Dimension Schätzung getrennt beurteilen. Während eine perfekt geschätzte Box, aber um 180° gedrehte, auch 100% Überlappung im Vergleich zur Ground Truth zeigt, würde SRT die Fehlschätzung im Rotations Score deutlich aufzeigen. Der gesamt SRT Score für diesen Ansatz ist 0.634.

Stand der Arbeiten (31.12.2020):

Während diesem Entwicklungszyklus wurde die erste Version des LIDAR-Datensatzes in der Tranche 3 zur Verfügung gestellt. Nur die Daten von Bit-TS wurden für die Entwicklung berücksichtigt. Nach einer ersten Evaluierungsphase dieser Daten, wurden jedoch einige Probleme mit den 3D-Annotationen festgestellt. Auch die Rotationen und die Ausrichtung der Objekt-Boundingboxen waren nicht korrekt. Gefundene Probleme und Unklarheiten wurden an die Datenerzeuger weitergeleitet.

In der Zwischenzeit wurden Datenlader implementiert, um den neuen Datensatz zu laden, um Training und Inferenz auf diesem neuen Datensatz durchzuführen. Der Code wurde angepasst, um den KIA-Datensplit für Release 2 zu verwenden. Der neue Release-Management-Prozess wurde berücksichtigt und der entwickelte Code wurde an diesem Prozess angepasst. Der Code wurde in gitlab eingecheckt. Es wurde ein iterativer Prozess verfolgt, um die Probleme mit dem release process zu beheben und die automatisierten Trainings- und Inferenz-Rauchtests mit Beispieldaten durchzuführen.

Aufgrund von Konflikten und Lizenzierungsproblemen mit den vortrainierten Modellen wurde die Verwendung von vortrainierten Gewichten für das Training vermieden. Es wurde versucht, das Modell von Grund auf zu trainieren, indem die Parameter zufällig initialisiert wurden. Die inkrementelle Verbesserung der KIA Daten verzögerte die Entwicklung in diesem AP und führte das Training nicht zu akzeptablen Ergebnissen. Wie in der vorherigen Berichtsphase erwähnt, verwendet dieser Algorithmus die Metrik der Scale Rotation Translation (SRT) score, um die Korrektheit der erkannten

Bounding Boxes zu bewerten. Die fehlerhaften Bounding Boxes in der Ground Truth würden auch diese Auswertung behindern.

Der Algorithmus verwendet eine 2D Bounding Box aus dem Kamerabild als Vorinformation, um 3D-Boxen aus den LIDAR-Daten zu erkennen. Derzeit werden beim Training 2D-Boundingboxen aus den Ground Truth Daten verwendet. Es wurden Untersuchungen in dieser Richtung durchgeführt, um einen 2D-Bounding-Box-Detektor zu integrieren, der die Objekte aus den Kameradaten erkennen würde. Der 3D-Detektor würde dann diese Boxen als Input für die weitere Erkennung verwenden. Da dies noch einer ausführlichen Evaluierung bedarf, wurde es noch nicht in das Repository aufgenommen. Es wurden Hotfixes für die Daten zur Verfügung gestellt, um einige der Probleme zu beheben, aber aufgrund der zeitlichen Begrenzung vor dem Release 2 war es schwierig, ein funktionierendes Modell zu erstellen.

Ein regelmäßiger Austausch über developer sync und Feedback-Prozesse wurde auch mit Stakeholdern durchgeführt, um die während der Entwicklungsphase aufgetretenen Probleme zu besprechen. Zusätzlich zu den Entwicklungsaktivitäten übernahm Valeo weiterhin die AP Lead Rolle in diesem AP und beteiligte sich aktiv an den Projekttreffen, der Berichtsphase und der Bereitstellung von Feedback für die Datenerzeuger.

Stand der Arbeiten (30.06.2021):

Während dieser Berichterstattung Phase wurden weitere Arbeiten zur Vorbereitung der Projektdaten und zum Training des Algorithmus auf den Projektdaten durchgeführt. In der Anfangsphase wurde der Code begutachtet, um die Probleme im Zusammenhang mit dem Release-Management-Prozess zu beheben, und es wurden Smoke-Tests durchgeführt, um die Smoke-Tests in der CI-Pipeline zu bestehen. Da die Tranche3-Daten aus dem Projekt immer noch Probleme mit der Ausrichtung der 3D-Bounding Boxes aufwiesen, war die Leistung des Algorithmus nicht optimal, was die Freigabe der trainierten Gewichte verzögerte. Die Hotfixes und die im Projekt bereitgestellten Fix-Skripte beheben einige Probleme mit den Daten, allerdings nur bei den statischen Assets; die dynamischen Assets müssen vor dem Training des Modells herausgefiltert werden.

Der Datenlader wurde geändert, um sich an die bereitgestellten Korrekturen anzupassen. Der aktualisierte Code wurde als Teil einer Zwischenversion, Release2*, veröffentlicht. Der Code entspricht dem Release-Management-Prozess und besteht den Completeness- und Trainingstest. Der Lizenztest schlägt weiterhin fehl, da nicht alle im Code verwendeten externen Bibliotheken in der Whitelist der Projektlizenz enthalten sind.

Ein regelmäßiger Austausch von Ideen und Entwicklungsfortschritten fand im Rahmen eines Dev-Sync statt. Rückmeldungen zu den Problemen mit den Projektdaten wurden

an die betroffenen Interessengruppen weitergeleitet. Valeo übernahm in diesem AP weiterhin die Rolle des AP-Lead, indem es aktiv an den Projekttreffen teilnahm und auch verschiedene Treffen auf TP-Ebene plante und organisierte.

Stand der Arbeiten (31.12.2021):

Der auf Frustum-Pointnet basierende Sqeuential-Fusionsalgorithmus wurde auf den kombinierten Daten von Bit-TS Tranche3 und Tranche4 trainiert. Das offizielle Datensplit des KI-A-Projekts wurde verwendet, um die Daten in einen Train-, Val- und Testdatensplit zu unterteilen. Ein Parameter zur Schätzung der Verdeckung wurde verwendet, um die verdeckten Pedestrian-Boxen im Trainingsdatensatz herauszufiltern. Valeo nahm an Release3 teil und der Code wurde als Teil des Releases veröffentlicht. Der Code wurde so angepasst, dass er mit dem Release-Management-Prozess übereinstimmt. Probleme im Zusammenhang mit der CI-Pipeline wurden den betroffenen Projektpartnern mitgeteilt.

Einige qualitative Ergebnisse sind hier als Referenz aufgeführt. Valeo hat seine Rolle als Leiter von AP1.4 aktiv fortgesetzt. Aspekte wie der Zeitplan für die Freigabe, Probleme mit der Datenfreigabe und Deadlines wurden rechtzeitig an alle Partner weitergegeben. Es wurden auch vierteljährliche Entwickler-Synchronisationstreffen organisiert, um Ideen auszutauschen und projektbezogene Fragen zwischen den Entwicklern zu klären.

Stand der Arbeiten (30.06.2022):

Während dieses Berichtszeitraums wurde der sequenzielle Fusionsansatz zur Erkennung der Schätzung einer 3D-Bounding-Box des Objekts auf den kombinierten Daten von BIT-TS Tranche 3, Tranche 4 und Tranche 5 trainiert. Es wurde das offizielle Projektdatensplit verwendet und die Trainingsdaten wurden vorverarbeitet, um die verdeckten Fußgängerinstanzen herauszufiltern. Valeo beteiligte sich am Release 4 und stellte den Code als Teil des Releases zur Verfügung. Tranche 7 von BIT-TS wurde ebenfalls während der Endphase des Projekts und des Berichtszeitraums freigegeben. Sie wurde jedoch nicht für Release 4 verwendet, da zu viele Anpassungen notwendig gewesen wären, da die LiDAR- und Kameradaten nicht synchron sind und der Datensatz Punktwolkendaten mit unterschiedlichen Mounting-Positionen der LiDARs enthält. Einige qualitative Ergebnisse des trainierten Modells werden hier als Referenz angegeben.

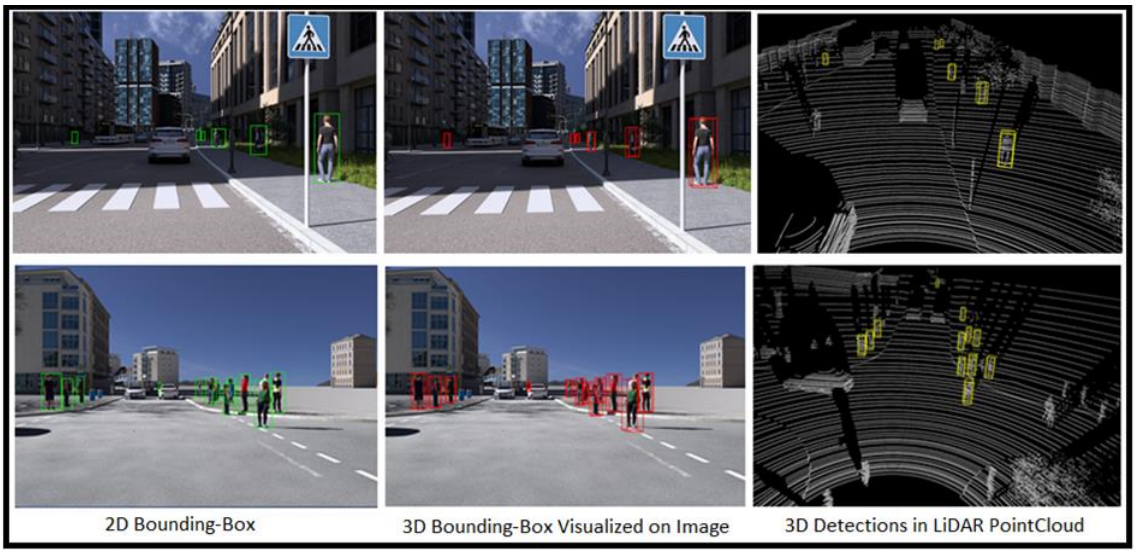


Abbildung 13: Qualitative Ergebnisse des trainierten Modells als Referenz

Außerdem übernahm Valeo weiterhin die Rolle des AP-Leiters in diesem AP. Es wurden regelmäßige dev-sync organisiert, um Ideen auszutauschen, die Veröffentlichung zu planen und Fragen im Zusammenhang mit den Daten und der endgültigen Veröffentlichung zu diskutieren. Während dieses Zeitraums wurde ausreichend Zeit für die Berichterstattung über die Endergebnisse verwendet und für die Veröffentlichung zur Verfügung gestellt. Es wurde auch ein abschließender Dev-Sync organisiert, bei dem jeder Partner, einschließlich Valeo, die Algorithmen und Ergebnisse seiner Arbeitspakete im AP vorstellte.

AP1.5 – Erweiterung um Posenschätzung (8 PM)

Aufgaben Valeo:

Angaben zum Stand der Wissenschaft (E1.5.1)

• Kontinuierliche Literatur-Recherche zu Algorithmen und Datensätzen.

Algorithmus Fusion von Posenschätzung und Personenerkennung (E1.5.5)

• Erforschung der Wechselwirkungen von Posenschätzung und Personenerkennung.

• Analyse Genauigkeit der von den Projektpartnern entwickelten Fußgänger-Posen-Schätzungen mit Skelett hinsichtlich der einfachen 3D Box Erkennung aus AP1.4. Aus diesem Grund werden 3D Boxen (aus der Definition von AP1.4) aus den 3D-Skelett Posen Schätzungen der Projektpartner (AP1.5) extrahiert.

• Ergebnisse aus AP1.4 werden mit zusätzlicher 3D-Skelett-Pose verbessert. Zu diesem Zweck werden die Posen-Schätzungen der Projektpartner als zusätzlicher Input in die Netze von AP1.4 gespeist.

Evaluation der Ergebnisse anhand der KPIs (E1.5.7)

• Evaluation der Ergebnisse anhand der KPIs aus AP1.2.

Stand der Arbeiten (30.06.2021):

Es wurde begonnen die geplanten Arbeiten unter den neuen Erkenntnissen des Projektes zu analysieren, um eine Richtungsänderung der Beiträge zu bewerten. Zur Bewertung der Absicherung von KI ist eine Betrachtung der gesamten System-Architektur erforderlich. Für den Use-Case der Fußgängernotbremssystems bedeutet dies, das die Perzeption nur ein Teil der System-Architektur darstellt. Im Anschluss an die Perzeption folgt die Posenschätzung bzw. die Intentionserkennung der Fußgänger. Die Intentionserkennung findet bisher keine Betrachtung im Projekt und stellt daher einen neuen Beitrag dar. Es wird aktuell mit weiteren Partnern aus den anderen Teilprojekten diskutiert inwieweit die Untersuchung einer Intentionserkennung für den Projekterfolg beitragen kann. Der Ausgang der Diskussionen ist Voraussetzung für den exakten Beitrag in AP1.5.

Teilprojekt 2: Generieren von synthetischen Lern- und Testdaten

TP2 ist in fünf Arbeitspakete mit folgender thematischer Struktur unterteilt:

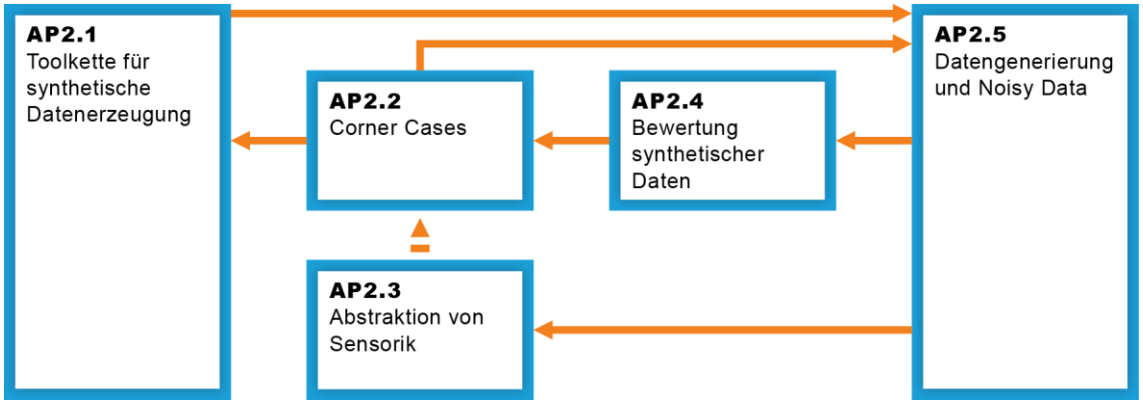


Abbildung 14: TP2-AP-Struktur

In AP2.1 wird eine technische Verarbeitungskette entwickelt, die die Synthese von Trainings- und Testdaten verfolgt. Die hierfür aufzubauende Werkzeugkette umfasst alle Verarbeitungsschritte, um von einer formalen Spezifikation der Szene, im Allgemeinen als eine oder mehrere Dateien gegeben, zu den simulierten Sensordaten zu gelangen.

AP2.2 hat das Ziel, möglichst breit nach Corner Cases zu suchen. 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. Des Weiteren fungiert AP2.2 als zentraler Ansprechpartner (so genannter Gate-Keeper) für die Anforderungen zwecks Datengenerierung im gesamten Projekt. In dieser Rolle werden alle eingehenden Datenanforderungen konsolidiert und priorisiert an AP2.5 weitergeleitet.

AP2.3 beschäftigt sich mit der Übertragbarkeit von KI-Funktionen bezüglich ihrer Absicherbarkeit bei Änderung der Sensorik. Weiterhin soll eine Übertragbarkeit der KI-Funktionen bei Änderung der Sensorik durch Methoden des Transfer Learnings untersucht sowie Konzepte zur Anpassung der Transfer Learning Methoden erarbeitet werden. Zur Evaluierung dienen die Performance-KPIs und insbesondere die Absicherungs-KPIs, die im Laufe des Projektes im TP3 entwickelt werden.

AP2.4 erarbeitet eine Methode zur vergleichenden Bewertung von Datensätzen unterschiedlicher Qualität unter Berücksichtigung von KPIs aus den Domänen Synthese-Qualität, KI-Performance, Modell-Qualität und leitet daraus eine Guideline für den Einsatz von synthetischen Daten unterschiedlicher Qualitäten für Training und Test von KI-basierten Algorithmen ab.

Im AP2.5 werden synthetische Trainings- und Validierungsdaten unter Einbezug von bewussten Verunreinigungen für bestimmte kritische Verkehrssituationen generiert.

AP2.1 Toolkette für synthetische Datenerzeugung (9 PM)

Aufgaben Valeo:

Spezifikation der Eigenschaften der synthetischen Datenerzeugungs-Pipeline (E2.1.2)

• Mitwirkung Sensormodelle und deren Fehler, Daten- und Fileformatspezifikation.

Implementierung Rendering weiterer Sensormodelle (E2.1.8)

• Entwicklung eines Plug-In Lidar-Modell.

• Untersuchung der zu modellierenden sensortypischen Eigenschaften eines einfachen *Velodyne* Lidar-Modells.

• Implementierung dieser Eigenschaften als Plug-In Modul für die Integration in die Rendering Pipepline.

Stand der Arbeiten (31.12.2019):

In Koordination mit E1.2.2 und AP2.1 wurde vom Sensortyp ‘Velodyne’ abgewichen zugunsten eines generischen Laser Scanners, angelehnt an Valeo Hardware. Die technische Machbarkeit und Anbindung an den Raytracer(Intel) wurden im Rahmen von AP2.1 evaluiert und vorbereitet. Seitens E2.1.8 sind erste erfolgreiche Schritte zur Inbetriebnahme der Raytracers erfolgt.

Stand der Arbeiten (30.06.2020):

Es wurde mit der Portierung eines synthetischen Laserscanner Modells in das Intel OSPRay Raytracer Code Projekt begonnen. Hierbei wurden die Spezifikation des Laserscanner Modells und die Schnittstelle zum Raytracer fortlaufend mit den Projektpartnern, insbesondere Innerhalb des AP2.1, abgestimmt. Aktuell existiert ein lauffähiges Raytracer plugin mit hinreichender Laufzeit-Performanz, welches jedoch noch nicht alle Spezifikationen erfüllt. Ein Baseline Modell des synthetischen Laserscanners als OSPRay Modul kann hieraus vorhersehbar in Kürze abgeleitet und mit den Projektpartnern geteilt werden.

Stand der Arbeiten (31.12.2020):

Nachdem das von Valeo implementierte synthetische LiDAR Modul in seinen internen Parametern weiter modifiziert und damit (noch nicht abschließend) an die Spezifikation aus UAP1.2.2 angenähert wurde, ist das Modul in Form eines ‘OSPRay module’ an den Projektpartner Intel (E2.1) zur ersten Evaluation übergeben und dort erfolgreich in Betrieb genommen worden. Die Übergabe erfolgte in Form eines kompilierten, binären Objektcodes auf GNU/Linux Platform (black box, Valeo background IP) und

zugehörigem C++ interface und Beispiel-Anwendung für Intel OSPRay (white code, Projekt foreground IP).

Stand der Arbeiten (30.06.2021):

Das synthetische LiDAR Modul wurde weiterentwickelt. Hierbei wurde die Konfiguration flexibler gestaltet, so dass jetzt der Nutzer verschiedene Parameter beispielsweise des Field of View aber auch die Pulslänge und Parameter für die Echoberechnung setzen kann. Des Weiteren gibt es einen Modus in dem ein realistischer Mobility Kit LiDAR von Valeo simuliert wird. Für den realistischen Modus wurde das Scanpattern des tatsächlichen Sensors einschließlich der Kippung des rotierenden Spiegels nachgebildet. Es ergeben sich schließlich drei Modi:

- Generic, bei dem der Nutzer die Konfiguration bestimmt,
- Standard, bei dem die Beispielkonfiguration aus UAP1.2.2 implementiert wird,
- Realistic, bei dem der realistische LiDAR simuliert wird.

Außerdem wurde die Ausgabe der Instanzsegmentierung implementiert. Hierbei wird eine zusätzliche PCD Datei generiert, die die Instanz IDs der Echos enthält. Da die Punktwolke durch Sensoreffekte verzerrt sein kann wird als Instanz ID die Instanz zugeordnet, die am nächsten an dem gemessenen Echo liegt. Die Zahl 0 wird für Echos reserviert, denen keine Instanz ID zugeordnet werden konnte.

Der Objektcode wurde auf Gitlab beziehungsweise Bitbucket zur Verfügung gestellt und in Ospray beziehungsweise Ospray Studio integriert.

Weiterhin wurde an der Idee des LiDAR als Lichtquelle gearbeitet. Hierbei soll der LiDAR als Gaussian Beam simuliert werden. Es wurde ein Script entwickelt, um in Abhängigkeit von horizontaler und vertikaler Divergenz eine ies Datei zu generieren, die die entsprechende Intensitätsverteilung darstellt. Die Integration in die Simulationsumgebung steht hierfür noch aus, ist jedoch zeitnah geplant.

Stand der Arbeiten (31.12.2021):

Valeo hat in diesem AP die Arbeit an dem Ospray LiDAR Plugin fortgesetzt. Hierbei sind die folgenden Punkte insbesondere hervorzuheben.

- Ideale Punktwolke

Für die Ground Truth-Ausgabe ist eine zusätzliche ideale Punktwolke erstellt worden. Dies ist notwendig, da durch Sensoreffekte Verzerrungen auftreten und dabei auch Punkte auf das „falsche“ Objekt abgebildet werden können. Die ideale Punktwolke unterliegt keinen Effekten wie dem Rollingshutter oder dem Sensitivitätsprofil. Um auch hier mehrere Echos zu ermöglichen wird nicht nur die direkte Richtung genutzt,

sondern auch vier Nachbarn entsprechend dem Öffnungswinkel des Empfängers. Es werden die kartesischen Koordinaten sowie die Instanz ID der Echos in einer pcd-Datei abgespeichert.

- Rolling Shutter

In Zusammenarbeit mit Intel wurde der Rolling Shutter implementiert. Somit ist es jetzt möglich die Spiegelrotation sinnvoll abzubilden. Der Rolling Shutter läuft auf der Ebene der upgesampelten Punktwolke aus der dann die simulierten Echos entstehen.

- Parameter Tuning

Interne Sensorparameter, u.a. solche, die die interne Beschreibung des optischen Pulses beschreiben, wurden getuned um einen realistischeren Output zu ermöglichen. Auch wurden beispielsweise die Öffnungswinkel der Empfangseinheiten angepasst.

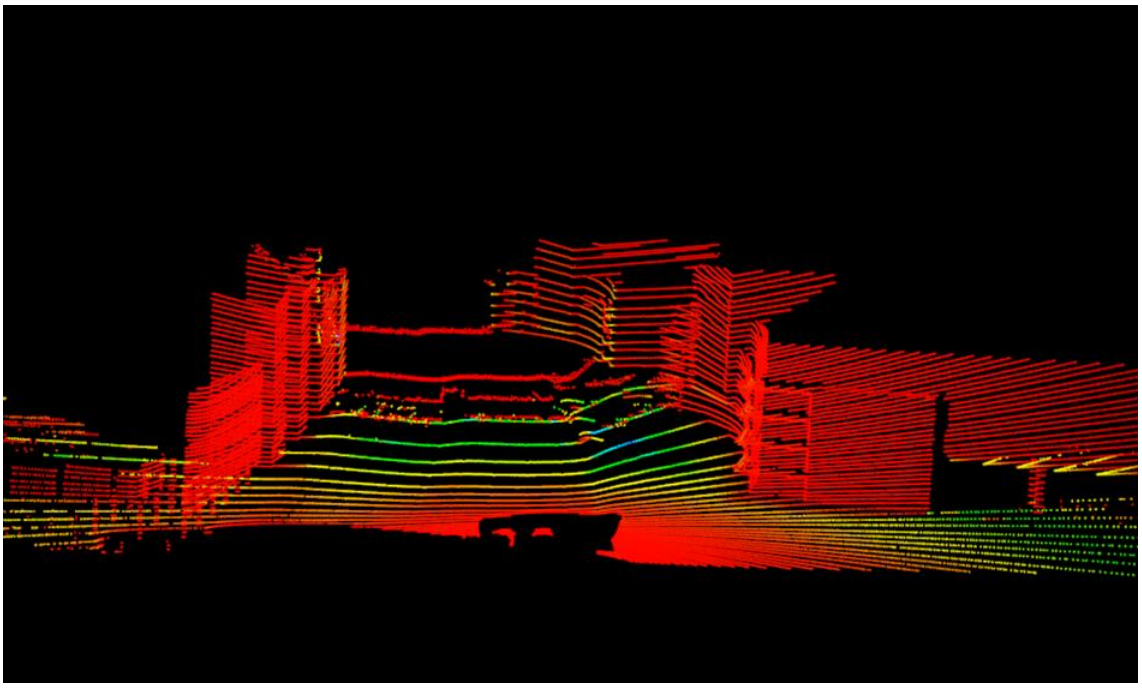


Abbildung 15: Output Parameter Tuning

Der Objektcode wurde erstellt und auf Bitbucket hochgeladen. Die Integration in die Ospray Renderingumgebung ist erfolgt.

AP2.2 Corner Cases (4 PM)

Aufgaben Valeo:

Bewertete Corner Cases (E2.2.7)

• Einbringen umfangreicher methodischer Erfahrungen im Test von DNN-basierten Algorithmen • Untersuchung der Corner Case-Variationen in Bezug auf ihre Auswirkung auf das Ergebnis des Trainings (direkte Unterstützung durch Arbeiten in AP3.5).

Bewertete Wissensbasis (E2.2.8)

• Analyse der Wissensbasis anhand weitere Quellen (bspw. Erkenntnisse aus Online-Performance bei parallelen DNN).

Stand der Arbeiten (30.06.2022):

Auf der Grundlage der in AP3.5 identifizierten Corner Cases wurde eine Analyse potenzieller Metadaten durchgeführt, die diese Corner Cases beschreiben. Die Identifizierung der Corner Cases basiert auf einem Unterschied in den DNN-Ausgaben auf der Grundlage verschiedener Modalitäten (LiDAR und Kamera). Die Metadaten können in zwei Kategorien unterteilt werden:

i) Globale Metadaten: Dazu gehören z. B. die Anzahl der Fahrspuren, die Geschwindigkeitsbegrenzung, der Straßentyp, die Umgebung (Stadt, Autobahn, Landstraße usw.) und weitere.

ii) Lokale Metadaten: Diese basieren auf einer visuellen Analyse der Szene und umfassen die Objektlokalisierung im Fahrzeugkoordinatensystem, die aktuellen Lichtverhältnisse, den Objekttyp (Auto, Lieferwagen, Fußgänger usw.) sowie durch einen Algorithmus zur Dimensionsreduktion (DNN-Encoder) berechnete Low-Level-Merkmale und das Wetter. Die automatische Aufzeichnung, Speicherung und Verarbeitung lokaler und globaler Metadaten für die Aufnahmen ist Teil des Ergebnisses (E.2.2.8) und kann zur Aufzeichnung neuer Corner Cases verwendet werden.

Diese Ergebnisse wurden in UAP2.2.7 weiterverarbeitet und eine Analyse der Metadaten der gefundenen Corner Cases durchgeführt. Die ermittelten Metadaten (lokal und global) wurden in einen Automatisierungsprozess für zukünftige Aufnahmen eingebunden, so dass sie automatisch erfasst, gespeichert und verarbeitet werden. Mit Hilfe der automatisierten Erfassung lokaler und globaler Metadaten können bei Aufzeichnungen spezifische Corner Cases erfasst werden. Durch die Analyse der aufgezeichneten Metadaten können geeignete Recoding-Szenen ermittelt werden. Beispiel: Treten die gefundenen Corner Cases bei bestimmten Geschwindigkeiten gehäuft auf, könnte eine bestimmte Bewegungsunschärfe die Ursache für die Corner Cases sein. Folglich würden die Aufnahmen häufiger bei höheren Geschwindigkeiten durchgeführt werden. Mit diesem Ansatz wird eine Richtung vorgegeben, welche

Kriterien für die Aufnahmen wichtig sind, und damit werden die aufzunehmenden Szenen stark eingeschränkt. Um die Vorhersageleistung der DNNs zu verbessern, werden sie mit den aufgenommenen Daten feinabgestimmt. Wenn überwachte Algorithmen verwendet werden, müssen diese zuvor gelabelt werden. Der Labeling-Prozess sowie die Feinabstimmung und Re-Analyse der DNN-Outputs in Bezug auf die Corner Cases lagen außerhalb des Rahmens dieses Arbeitspakets.

AP2.3 Abstraktion von Sensorik (27 PM)

Aufgaben Valeo:

Definition von vier geeigneten Use Cases und Identifikation der Einflussgrößen auf ein verändertes Sensor-Setup (E2.3.1)

• Spezifikation von relevanten Einflussgrößen des LiDAR Sensors.

Quantifizierung der Auswirkungen der primären Einflussgrößen auf mögliche KPI-Veränderungen (E2.3.3)

• Erzeugen von Vorhersagen mit einer KI-Funktion aus AP1.4, welche basierend auf den Evaluierungsergebnissen konkret ausgewählt werden soll.

• Evaluierung der Vorhersagen basierend auf Qualitäts-KPIs und den zu entwickelnden Absicherungs-KPIs in TP3.

• Vergleich der KPI-Werte mit dem Baseline-Sensorsetup.

Fine Tuning-Ansatz inkl. Quantifizierung notwendiger Menge an Trainingsdaten (E2.3.4)

• Durchführung von Fine Tunings mit Trainingsdaten des neuen Sensor-Setups. Entwicklung einer Strategie zur systematischen Auswahl der Trainingsmenge. Entwicklung einer Korrelation zwischen verwendeter Trainingsmenge und der resultierenden Annäherung der KPI-Werte.

Auf Domain-Adaptation beruhende Ansätze, mit dem die KPI-Werte erhöht werden können (E2.3.5)

• Implementierung der Self-Ensembling Methode in den Trainingsprozess der KI-Funktionen. Anpassung der Methode an die Netzwerk-Architektur aus TP1 und der Aufgabe der Fußgängererkennung. Variieren von Parametern in der Self-Ensembling Methode und des Trainingsprozesses zur Untersuchung der Auswirkung auf die Annäherung von Qualitäts-KPIs und Absicherungs-KPIs. Methoden-Erweiterung zur Annäherung von gezielten Absicherungs-KPIs.

Stand der Arbeiten (31.12.2019):

Erste Absprachen mit den beteiligten Partnern über die Spezifikation der verschiedenen Sensor-Setups wurden angestoßen. Es wurde mit ersten „Stand der Technik“-Recherchen begonnen.

Stand der Arbeiten (30.06.2020):

Das Arbeitspaket wurde offiziell und wie geplant in Monat 10 gestartet. Dabei hat aufgrund der Abwesenheit von Personen bei BMW Valeo (in der Rolle als Co-Lead) temporär den Lead übernommen. Die ersten Arbeiten im AP haben sich auf das

Ergebnis E2.3.1 konzentriert, da dies die Grundlage für alle anderen Arbeiten darstellt. Dafür wurden die verschiedenen Einflussgrößen für Kamera wie Lidarsensoren gemeinsam mit allen Partnern gesammelt und zusammengefasst. Valeo hat sich dabei auf die Einflussgrößen für Lidarsensoren konzentriert. Das Ergebnis wurde wie geplant abgeschlossen und die Anforderung werden an die Datenerzeugung weitergegeben.

Die inhaltlichen Themen von Valeo benötigen alle simulierte Lidardaten die im Rahmen von TP2 generiert werden soll. Da jedoch noch kein Lidarmodell für die Datenerzeugung existiert, wurden weitere inhaltliche Arbeiten pausiert, bis die entsprechenden Daten vorliegen. Die Organisation wird von Valeo weitergeführt, bis BMW wieder übernehmen kann. Eine Verwendung von offenen Datensätzen wurde geprüft, aber von einer Verwendung abgesehen, da in diesem AP konkrete Unterschiede in der Sensorik oder Verbauposition untersucht werden sollen. Dafür wären Daten von gleichen Sequenzen mit unterschiedlichen Lidarsensoren oder Einbaupositionen notwendig. Da solche Daten in keinem offenen Datensatz existieren, muss bei der Thematik auf die projekteinternen Daten gewartet werden.

Stand der Arbeiten (31.12.2020):

In AP 2.3. wurde zuerst ein Fokus auf die domain gap Analyse von Fußgängern in Lidar Daten gelegt. Speziell auf die Verteilung von Punkten die Lidare auf Fußgängern generieren, wenn die Position des Sensors angepasst werden sollte.

Dies ist motiviert durch einen der meist größten aufkommenden domain gaps, der in der Sensorposition gefunden werden kann: einem Wechsel der Einbauposition von Höhe des Dachs in die Höhe der Stoßstange. Diese Änderung ist von großem Interesse, da die typischen Datensätze, z.B. Kitti, Nuscenes, einen Lidar auf dem Dach nutzen, wohingegen die Position in Stoßstangenhöhe keinen extra Aufbau auf dem Dach benötigt und somit für Anwendungen attraktiver sind.

Für die Analyse wurde eine Annotations Box basiertes Koordinatensystem verwendet, das es erlaubt alle Punkte in der Annotation eines Fußgängers in ein Positions-, Rotations- und Skalierungs-invariantes Koordinatensystem zu transformieren, um ein quantitative Analyse der Punktwolken Dichte zu gestatten. Das Koordinatensystem stammt aus der Arbeit von Scheel et. al https://arxiv.org/pdf/1711.03799.pdf.

Aus Mangel der Verfügbarkeit von Projektdaten wurde der Lyft Lvl5 Datensatz verwandt. Dieser beinhaltet einen LiDAR Sensor auf dem Dach des Fahrzeugs und 2 Lidare in der Front auf Stoßstangenhöhe.

Die Akkumulation dieser Daten und die respektive Dichten sind daraufhin auf die Hauptachsen und Sensoren konditioniert aufgetragen und speziell im Unterschied zwischen den Sensor Höhen kann eine signifikant unterschiedliche Abdeckung der Person erkannt werden. Der LiDAR auf dem Dach hat beispielsweise eine deutlich stärkere Abdeckung des Oberkörpers als der Beine, wohingegen die Sensoren in Stoßstangen den gesamten Körper abdecken.

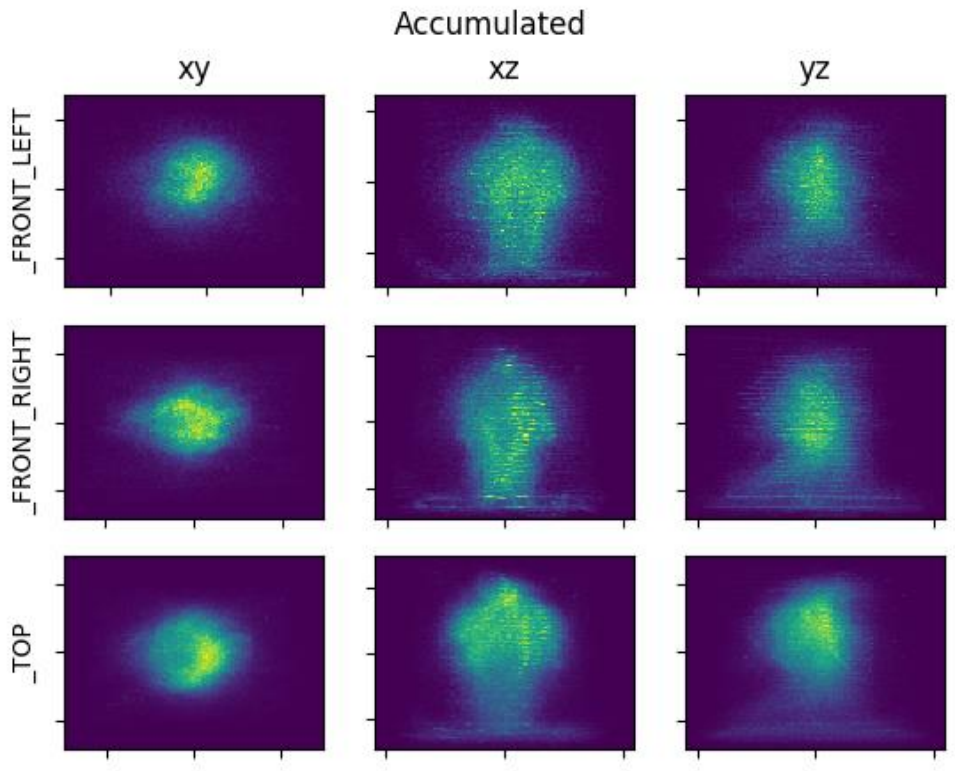


Abbildung 16: Abbildung der akkumulierten Punktwolken als Heatmap in dem invarianten Koordinatensystem. aufgetragen sind Blickwinkel (xy Top-, xz Front- und yz Seitenansicht) gegenüber des Sensor typs (Sensor auf dem Dach _TOP, Sensor vorne links, Sensor vorne rechts).

Um einer Verzerrung in den Daten vorzubeugen wurde des weiteren eine Konditionierung der Daten auf den relativen Winkel unter dem sich die Annotations Box zum Sensor befand, erstellt, welche in den folgenden Beispielbildern verdeutlicht werden soll:

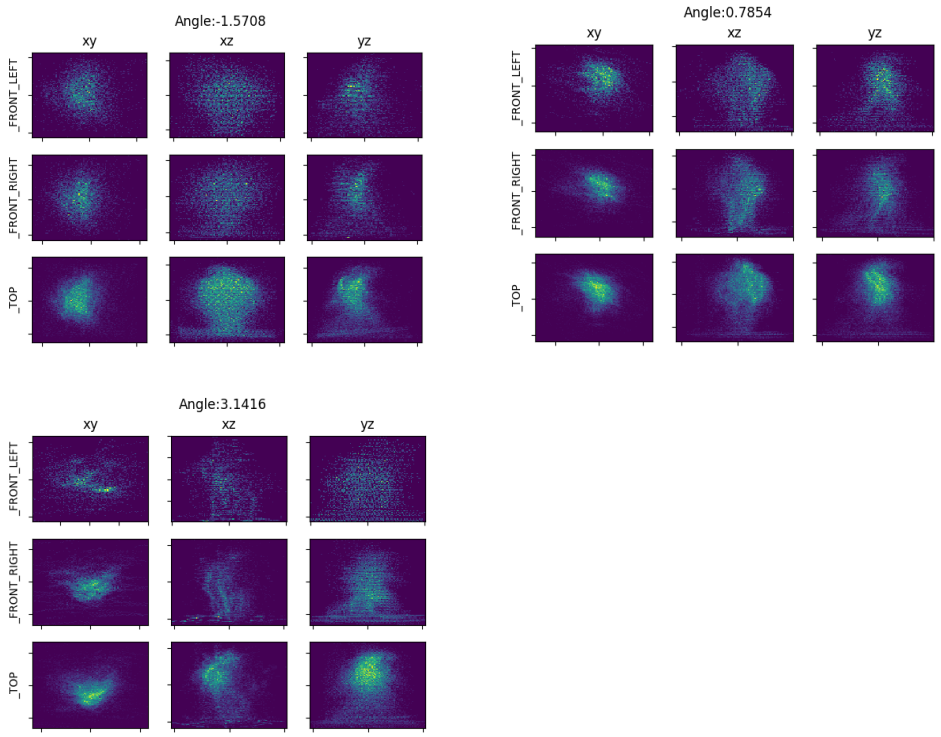


Abbildung 17: Konditionierte Heatmaps auf den Einfallswinkel. Es werden drei von acht gleich großen Segmenten der Winkelbereiche gezeigt wobei der mittlere Winkel als Indikator angegeben wurde. Die Bilder entsprechen aus Sicht des Fußgängers: 1.) Von der rechten Seit, 2.) Von Links Vorne 3.) Von Rechts

Diese Analyse beschreibt den domain shift zwischen den Sensor Einbaupositionen quantitative und seine Nutzung wird im weiteren Verlauf des Arbeitspakets analysiert.

Neben den inhaltlichen Tätigkeiten hat Valeo weiterhin die Lead Rolle im AP übernommen, wurde dabei unterstützt von BMW (eigentlich Lead von AP2.3) in allen Themen die Berichte oder AP Vorstellungen, z.B. vor dem Steuerkreis, betreffen. Im Rahmen von AP2.3 wurde eine zweiwöchentliche Telko abgehalten und die Tickets für die Datengenerierung wurden erstellt und in Kooperation mit BMW mit den Daten Erzeugern diskutiert.

<u>Stand der Arbeiten (30.06.2021):</u>

Die akkumulierten Daten wurden genutzt um ein Gauß Mischmodell in dem skalierten Raum zu trainieren. Das Modell wurde über die x,y,z-Dimension sowie den aspect Winkel als modulierbaren Parameter erstellt. Wir Illustrieren diese in den folgenden Abbildungen pro Sensor, speziell zeigen wir äquidistante Schnittebenen über den Raum, in welchem die über alle Schnittebenen normalisierte Likelihood des Modells aufgezeigt wird.

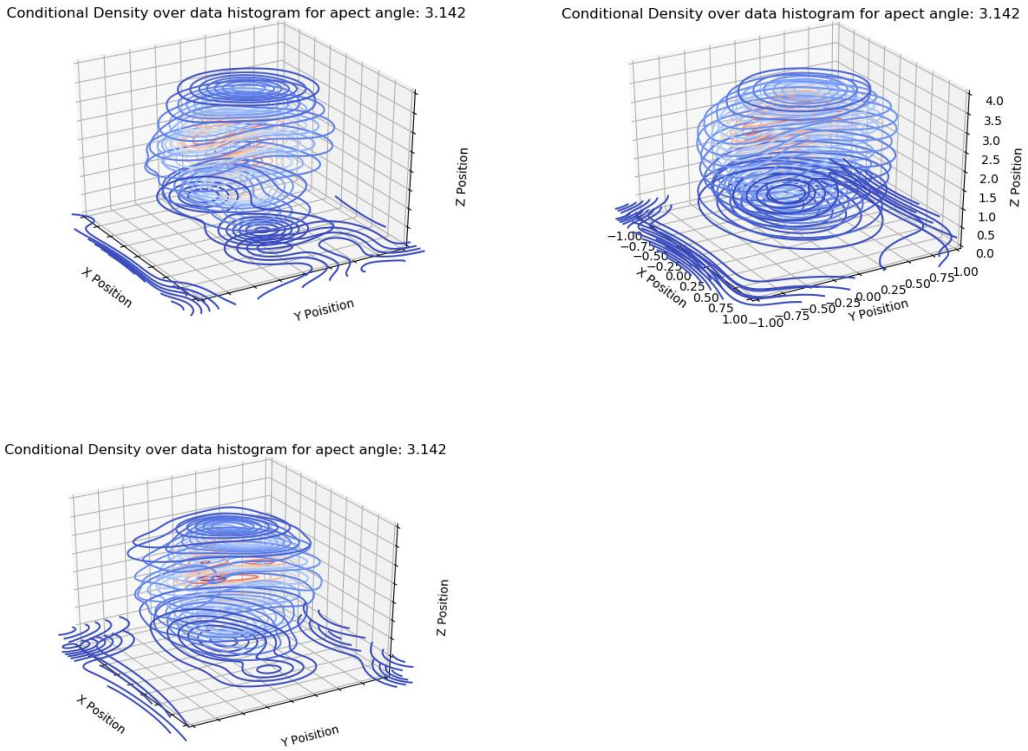


Abbildung 18: Oben links: Lidar Top; Oben rechts: Lidar Front Left; Unten links: Lidar Front Right

- Wir realisieren, dass die Auflösung des Oberkörpers deutlicher in den Daten des sich auf dem Dach befindlichen LiDARs repräsentieren wird, wie es in der Geometrie der erhöhten Einbauposition zu erwarten ist
- Für flexiblere Körperteile wie die Beine wird jedoch eine große räumliche Unsicherheit in der gelernten Verteilung aufgezeigt, speziell in den tiefer eingebauten Sensoren. Daher ist die Wahrscheinlichkeit die Klasse “Beine” zu sehen reduziert, obwohl die Sensoren sie besser auflösen können

Fortschritt:

- Das Training eines Initialen Mixture Models für die verschiedenen Sensor-Modalitäten sowie eine Problem Analyse dieser.

Folgende Schritte werden nun weiterverfolgt:

- Die Daten werden versucht basierend auf der Beinstellung zu clustern, um genauere Modelle auf dieser Bedingung zu erstellen.

<u>Stand der Arbeiten (31.12.2021):</u>

Basierend auf dem letzten Inkrement von AP 2.3 wurde ein Graph basiertes clustering für die akkumulierten LiDAR Daten von Fußgängern durchgeführt. Das Ziel ist es hier

ähnlich aufgebaute Frames zu finden und eine mögliche Datenorganisation zu erstellen. Es wurde ein Spanngraph auf Einzelframes der akkumulierten Daten erstellt. In diesem Graph repräsentiert jeder Knoten einen Frame und die Kanten ein Gewicht zwischen den Frames. Das Gewicht wird hier als Chamferdistanz gewählt, welche eine gängige Wahl in vielen Aufgaben der Punktwolkenverarbeitung ist. Die Gewichte werden in dieser Darstellung als mechanische Federn interpretiert und sorgen für eine Selbstorganisation der Punkte basierend auf den Gewichten, sprich der Ähnlichkeit der Frames.

Ein Beispiel eines solchen Graphen ist in Abbildung 2.3.1 zu sehen welcher testweise auf einen kleinen subset von 1000 Frames verwandt wurde. Der Graph zeigt die Knoten in bunt und die Kanten in schwarz, wobei durch die volle Verbindung von jedem Knoten zu jedem ein dichtes Netz entsteht. Die Farben des Knoten impliziert die Zugehörigkeit eines Frames zur selben Instanz, sprich die Zugehörigkeit zum selben Target.

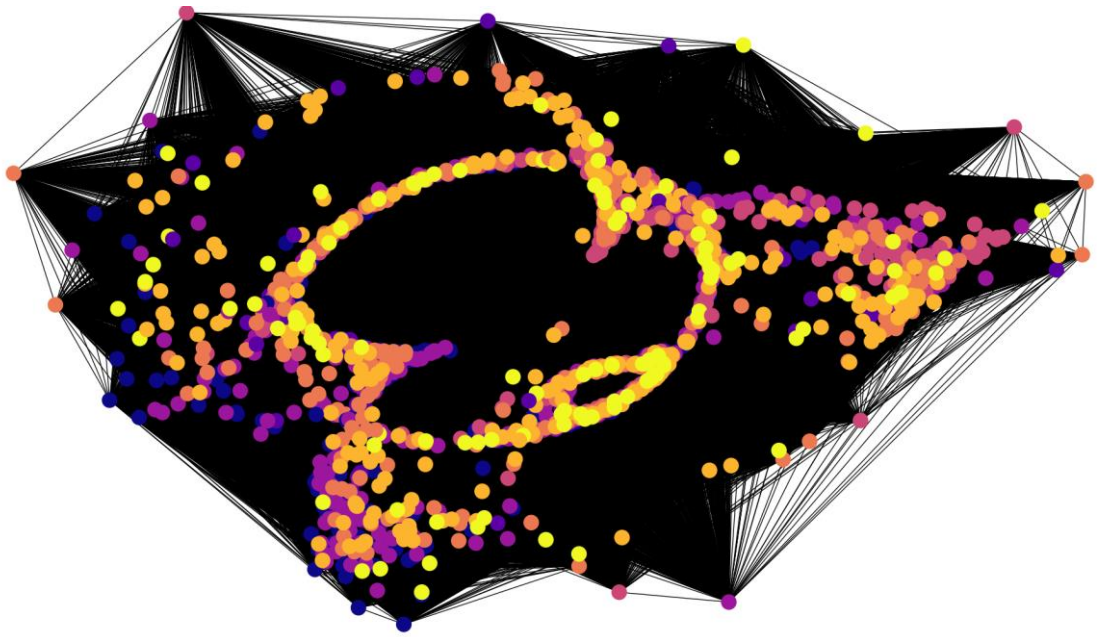


Abbildung 19: Beispiel Graph für ein subset der frames

Grundsätzliche Beobachtung sind, dass sich Strukturen ausbilden und dass keine Cluster einzelner Instanzen entstehen. Ein erweitertes Erstellen eines Spanngraphen auf der 10-fachen Menge an Daten wurde als Folge durchgeführt. Eine Analyse der gebildeten Struktur ist nun das nächste Ziel.

Fortschritt:

- Das Aufbauen der Spanngraphen wurde durchgeführt

Folgende Schritte werden nun weiterverfolgt:

- Auswertung der trainierten Spann Graphen, nearest neighbour Klassifikation und Clusterauswertung
- Auswertung von Trainingsdatenmengen die durch domain-gaps entstehen, sobald der Datensatz zur Verfügung steht

Stand der Arbeiten (30.06.2022):

Basierend auf dem letzten Inkrement von AP 2.3 wurde der Graph basierte Ansatz der Arbeiten auf einer um das 10-fache erhöhten Trainingsmenge durchgeführt und ausgewertet. Diese Aufgabe galt weiterhin der Organisation der Daten um gegebenenfalls zusätzliche Informationen über die Positionierung der Fußgänger zu generieren.

Die Grundsätzliche Idee der Auswertung bestand dabei in einer gridbasierten Abtastung des Raum in dem sich der Graph befindet.
Um dies darstellen zu können wird zu einem ausgewählten Punkt im Raum eine k-nearest-neighbour-suche durchgeführt und die Lidarpunkte der nachbarknoten werden akkumuliert dargestellt.

Dies erlaubt ein eine Inspektion der Organisation des Graphen durch eine Inspektion. Ein von 4 Testpunkten für diese Inspektion ist in Abbildung 2.3.2 gegeben. Während eine gesamte Abtastung in einer Animation in Abbildung 2.3.3 dargestellt wird (sofern das Betrachtungsformat dies erlaubt)

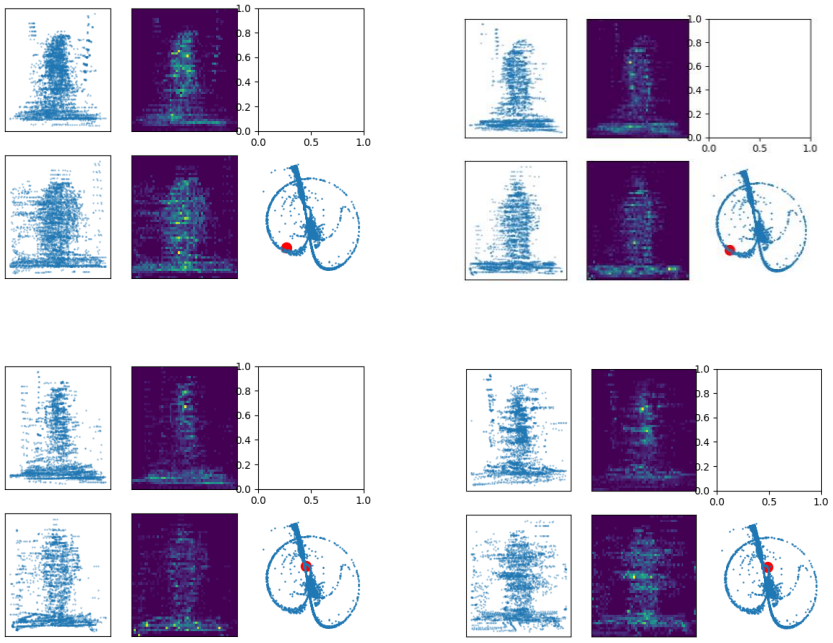


Abbildung 20: Abbildung 2.3.2 k-nearest-neighbour Darstellung spezifischer Positionen innerhalb des Graphen. Mit k=20. Die linken 4 Panel zeigen die Ansicht auf die akkumulierten punkte in x-y und x-z Koordinaten als Punkt und Heatmap Repräsentation. Die Position im Graph welche zu dieser Akkumulation führte ist im Rechten, unteren Panel gezeigte. Dabei ist der rote Punkt der Testpunkt und die blauen Punkte sind die Knoten des Graphen.

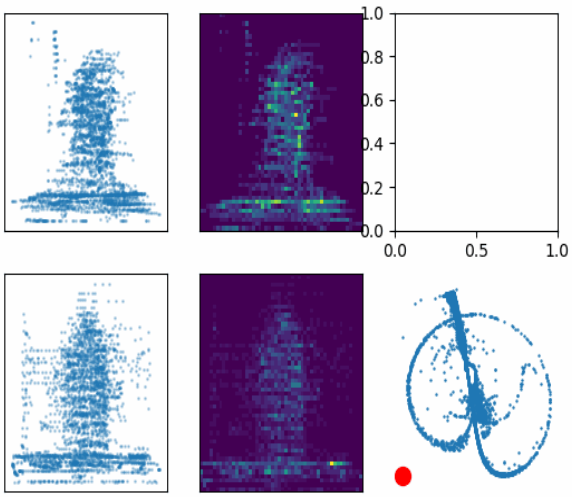


Abbildung 21: Gesamtabtastung des Raum

Die Analyse ergab, eine lokal Sichtbarkeit Ähnlichkeit der Knoten innerhalb des Graphen. Jedoch wurde keine stringente globale Ordnung erkannt, welche beispielsweise ein soft Labeling der Daten in spezifische Posen erlauben würde.

Der Graphen basierte Organisationsansatz ergab somit auf der Naiven ebene keine einfach zu nutzende Struktur. Jedoch sind weitere Verbesserungsmöglichkeiten noch offen:

- Der Ansatz verwendet in der implementierten Form die Chamferdistanz, welche durch eine andere Distanz ausgetauscht werden könnte
- Andere Verfahren für die Organisation könnten erprobt werden, wie die Organisation im Embedding raum eines Autoencoders

Abschluss Fortschritt:

- Die Abtastung des Graphen wurde erfolgreich durchgeführt.
- Eine Visuelle Analyse der Strukturen
  wurde durchgeführt.

Anmerkung:

- Durch Verzögerungen in der Bereitstellung des Datensatzes und Probleme mit der Transformation, sowie der Zeitsynchronität Sensor Instanzendaten wurde die Entwicklung eines Finetuning Ansatz für Lidare in verschiedenen Mountingposition in AP 2.3 abgebrochen.

AP2.4 Bewertung der Qualität synthetischer Daten (18 PM)

Aufgaben Valeo:

Vergleich der Auswirkung unterschiedlicher Datensätze auf KI-Netzwerk-Performance (E2.4.3)

• Messen der Auswirkung der Domänenverschiebung (mit synthetischen Daten unterschiedlicher Qualität) in Bezug auf die Performance-KPI und insbesondere auf die Absicherungs-KPI (aus TP3).

• Auswerten des DNN mit Testdatensatz Synthetische Daten basierend auf Performance-und Absicherungs-KPI

• Reduzierung der Domänenverschiebung durch Finetuning mit photorealistischen synthetischen Trainingsdaten (HQ Daten) (Benötigte Daten: Idealerweise 500 Bilder in HQ-Format)

• Finetuning des DNN mit photorealistischen Daten in mehreren Stufen (Ziel ist die Ermittlung der Anzahl der benötigten HQ Bilder)

• Auswerten der Performance und Absicherung KPI für jede Stufe

Optimierte Parametrierung mit einem Vergleich der Auswirkung unterschiedlicher Datensätze (E2.4.4)

• Anwendung von Domain Adaptation Methoden zur Annäherung der KPIs

• Reduzierung der Domänenverschiebung durch Domain Adaptation mit GAN's Implementierung und Anwendung auf.

• Anpassung auf das in TP1 entwickelte DNN-Analyse und Bewertung des Erfolgs dieses Ansatzes vor allem im Hinblick auf Absicherungs-KPI. Entwicklung einer Metrik, die Datenverteilungen zwischen intermediären Merkmalskarten vor und nach dem GAN-Ansatz misst (Erstellung einer Feature-Similarity-KPI).

Stand der Arbeiten (31.12.2020):

Die Spezifikation von Perzeptions-DNNs für das automatisierte Fahren sieht eine Beschreibung des Operational Design Domains (ODD) vor. Die ODD beschreibt unter anderem den Eingaberaum für das DNN. Je weiter man den Eingaberaum an die Realität annähert, desto größer die Vielzahl der darin enthaltenen Domänen. Domänen definieren sich über unterschiedliche Erscheinungsformen in den Sensordaten, d.h. einer Änderung der Datenverteilung. Man spricht hier von einem sogenannten “Domain Shift”. Verursacht wird dieser “Domain Shift” beispielsweise unterschiedliche Jahreszeiten, Wetter- und Beleuchtungseinflüsse verursacht werden.

Mit zunehmender Zahl der Domänen, mit der ein DNN trainiert wird, steigen die Anforderungen an das DNN, da es mit einer großen Varianz der Datenverteilung zurechtkommen muss. In dieser Arbeit vergleichen wir Generalist DNNs und Spezialist

DNNs bezüglich ihrer Prädiktionsgenauigkeit in den einzelnen Domänen. Als Generalist DNN bezeichnen wir ein DNN, dass mit einer Vielzahl von Domänen trainiert wurde und als Spezialist DNN bezeichnen wir ein DNN, das mit lediglich mit einer Domäne trainiert bzw. finegetunet wurde.

Im Hinblick auf sicherheitskritische Anwendungen, wie es das Erkennen von Fußgänger darstellt, wird eine möglichst hohe und zuverlässige Prädiktionsgenauigkeit vorausgesetzt, weshalb diese Fragestellung eine hohe Praxisrelevanz hat.

Wir führen die Experimente mit DeepLabV3+ zur semantischen Segmentierung durch und nutzen den Synthia Video Sequenzen Datensatz.

Wir nutzen 4 Sequenzen (SEQ), 2 SEQ mit Highway und 2 SEQ mit New York Szenen. Jede Sequenz ist weiter unterteilt in die folgenden Domänen: dawn, fog, night, spring, summer, sunset, winter.

Es sind noch mehr Domänen enthalten, jedoch konzentrieren wir uns nur auf die genannten Domänen, da nur diese in allen vier Sequenzen in einer ähnlichen Anzahl vorliegen. Wir verwenden die Sequenzen 1 und 2 für das Training und 5 und 6 für das Testen.

Zuerst wurde das Generalist DNN trainiert. Hierzu wurde das DeepLabV3+ mit den Domänen dawn, fog, night, spring, summer, sunset, winter trainiert. Folgende Abbildung zeigt die Prädiktionsgenauigkeit in “mean intersection over union” (MIoU) über die Epochen des Trainings. Die Evaluierung wurde auf jede einzelne Domäne durchgeführt (siehe Legende in Abbildung 22 und Abbildung 23).

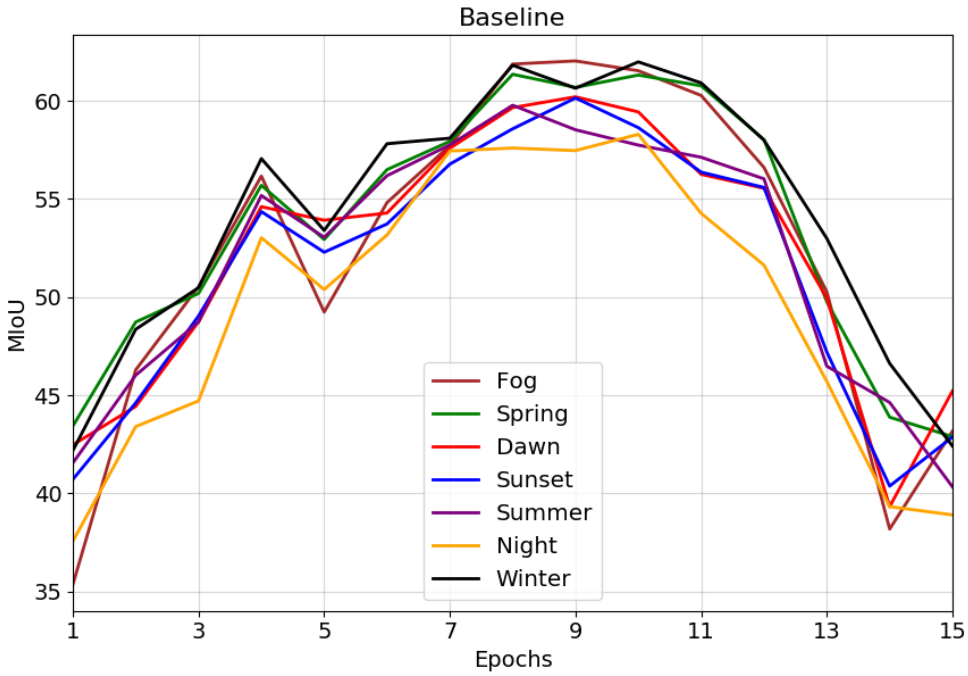


Abbildung 22: Prädiktionsgenauigkeit MIoU über die Epochen des Trainings

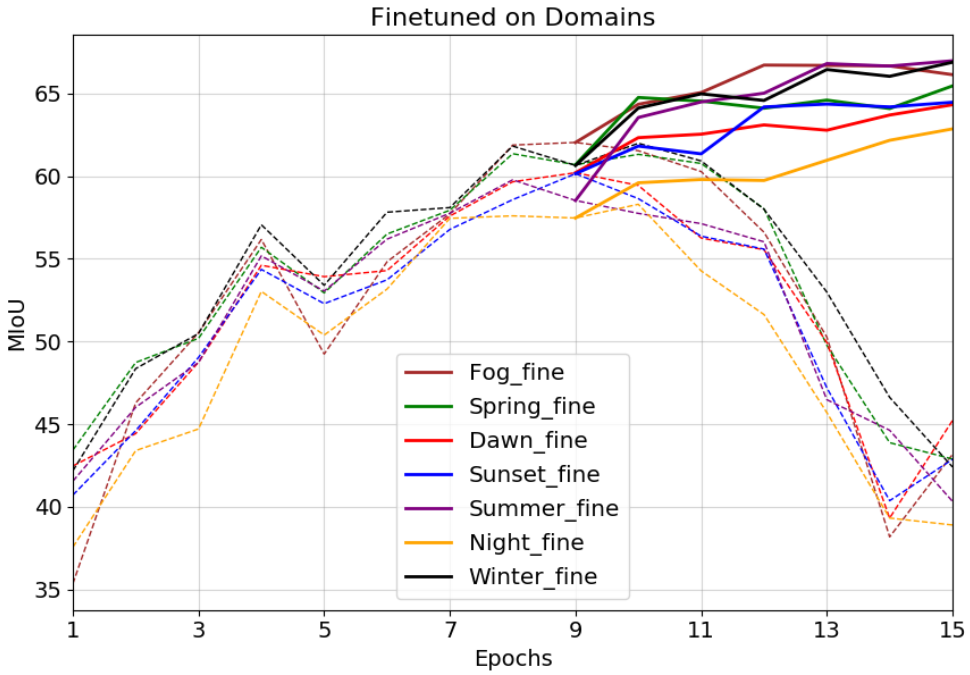


Abbildung 23: Prädiktionsgenauigkeit MIoU über die Epochen des Trainings - Finetuned

Es ist zu erkennen, dass die Prädiktionsgenauigkeit nach 9 Epochen seinen Höchstwert erreicht und danach abfällt. Daher wurde für das Training der Spezialisierten DNNs dieser Wert als Startpunkt verwendet. Folgende Abbildung zeigt die Ergebnisse für die Spezialisierten DNNs. Die Resultate des Generalist DNN wird gestrichelt dargestellt und die Resultate der Spezialist DNNs wird einer durchgezogenen Linie dargestellt. Die Abb. zeigt einen deutlichen Anstieg des MIoU durch das Finetuning auf die einzelnen Domänen. Dies bestätigt die Annahme, dass Spezialisierte DNNs in ihrer Domäne eine bessere Prädiktions-Genauigkeit liefern als das Generalist DNN.

Qualitative Ergebnisse werden in folgenden zwei Abbildungen gezeigt. Das “Baseline” bezeichnet das Generalist DNN und das “Finetuned” das Spezialist DNN.

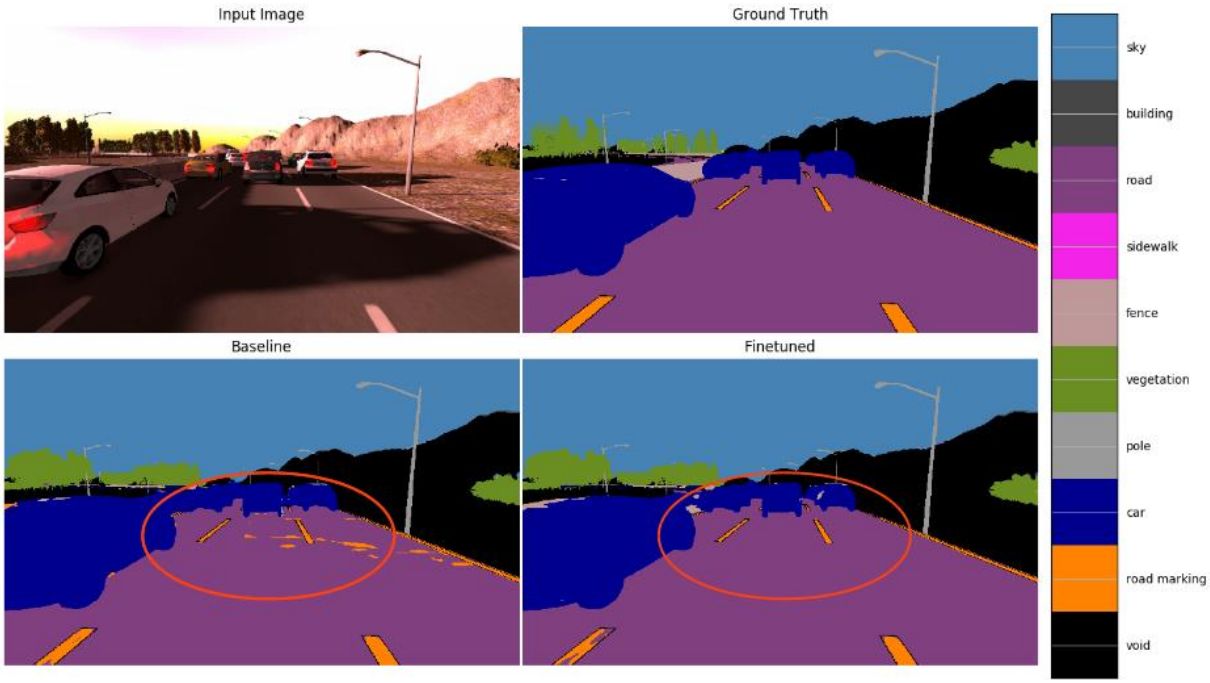


Abbildung 24: Baseline Modell

Es ist zu sehen, dass das Baseline Modell teile des Schattens als Straßenmarkierung segmentiert (siehe rote Ellipse). Offensichtlich hat das Finetuned Modell diese Probleme nicht, da es auf Daten der Domäne “dawn” finegetuned wurde.

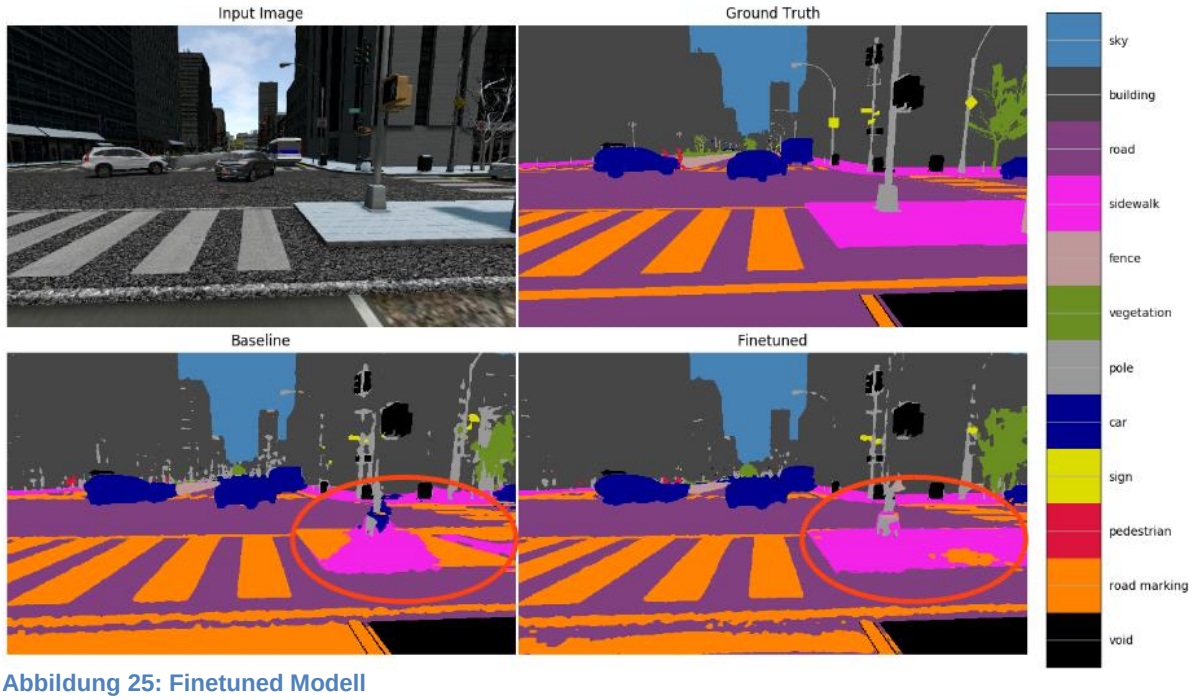


Abbildung 25: Finetuned Modell

Das Baseline Modell segmentiert große Teile des Bürgersteigs als Straßenmarkierung (siehe rote Ellipse). Vermutlich da der Bürgersteig mit Schnee bedeckt ist und dadurch ähnlich aussieht, wie die Straßenmarkierung. Das Finetuned Modell weist diese Probleme nicht auf.

Die gezeigten Ergebnisse bestärken uns in unserer Annahme, dass Spezialisierte DNNs den Generalist DNN in den einzelnen Domänen bezüglich der Prädiktions Qualität überlegen ist. Als nächsten Schritt sollen die Experimente mit einem anderen DNN wiederholt werden, um zu überprüfen, ob die Idee von Generalist und Spezialist DNN auf andere DNNs übertragbar ist.

Stand der Arbeiten (30.06.2021):

Die durchgeführten Experimente bauen auf den Ergebnissen von UAP2.4.3 auf und dienen dazu, generalistische DNNs und spezialisierte DNNs in Bezug auf ihre Vorhersagegenauigkeit in jeder Domäne zu untersuchen. Als generalistisches DNN bezeichnen wir ein DNN, welches mit einer großen Anzahl von Domänen trainiert wurde, und als spezialisiertes DNN bezeichnen wir ein DNN, das nur mit einer Domäne trainiert oder feinabgestimmt wurde. Im Hinblick auf sicherheitskritische Anwendungen, wie z.B. das automatisierte Fahren, ist eine möglichst hohe und zuverlässige Vorhersagegenauigkeit erforderlich, weshalb diese Problemstellung eine hohe praktische Relevanz hat. Wir führen unsere Experimente mit zwei DNNs (DeeplabV3+ und SAN) zur semantischen Segmentierung durch und verwenden den Synthia Video Sequences-Datensatz, der die Domänen Dämmerung, Nebel, Nacht, Frühling, Sommer, Sonnenuntergang, Winter aufweist.

Die exakt gleichen Experimente wurden mit beiden DNNs durchgeführt. Für das Training wurde eine Lernrate von 0,01 für SAN und 0,001 für DeeplabV3+ verwendet. Bei DeeplabV3+ konnten bei einer Lernrate von 0,01 starke Schwankungen des Verlustes beobachtet werden, weshalb die Lernrate auf 0,001 reduziert wurde. Für beide DNNs wurde die "poly"-Policy verwendet, um die Lernrate zu reduzieren: lr = self.lr * pow((1 - 1.0 * T / self.N), 0.9)

Zunächst wurden beide DNNs mit allen Domänen für 50 Epochen trainiert. Dann wurde ein Finetuning der DNNs mit den Daten der jeweiligen Domäne durchgeführt.

Das Ergebnis sind 7 durchgeführte Finetunings mit jeweils 7 Domänen. Um einen fairen Vergleich mit dem All-in-One-DNN zu erreichen, wurde ein achtes Finetuning mit allen Daten durchgeführt. Dieses wird fortlaufend als Baseline bezeichnet. Es ist zu erwähnen, dass das achte Finetuning die meiste Trainingszeit benötigt, da hier die Daten aus den 7 Domänen zusammen verwendet werden.

Die Ergebnisse sind in den folgenden Tabellen zu finden. Die Tabellen zeigen die Ergebnisse für SAN und DeeplabV3+ mit 100 und 150 Epochen. Baseline bedeutet, dass das DNN auf allen Domänen trainiert wurde (Generalist). Die Spalte Finetuning beschreibt die spezialisierten DNNs, d. h., es wurde nur mit einer einzigen Domäne trainiert. Die Spalte Difference zeigt die Leistungsunterschiede zwischen Baseline und Finetuning in Mean Intersection over Union (MIou). Generell ist zu erkennen, dass für beide DNNs die spezialisierten DNNs in ihrer Domäne besser abschneiden als die generalistischen DNNs.

Tabelle 6: Ergebnisse für SAN und DeeplabV3+

	Deeplabv3+ 100 epoch			Deeplabv3+ 150 epoch		
Domains	Baseline	Finetuned	Difference	Baseline	Finetuned	Difference
dawn	61.93	65.27	+3.34	68.91	70.79	+1.88
fog	66.58	68.44	+1.86	71.71	72.75	+1.04
night	63.51	64.64	+1.13	69.01	70.16	+1.15
spring	64.47	66.95	+2.48	70.15	70.84	+0.69
summer	65.02	67.42	+2.40	70.99	72.39	+1.40
sunset	64.05	65.71	+1.66	69.04	70.15	+1.11
winter	65.03	66.82	+1.79	69.91	71.50	+1.59

Stand der Arbeiten (31.12.2021):

Um die Auswirkungen von leistungsbegrenzenden Faktoren (PLFs) auf verschiedene Datensätze zu untersuchen, stellt sich die Frage nach der statistischen Relevanz von Studien mit einem einzigen DNN und einer einzigen Gewichtungsdatei. D.h. ein DNN (z.B. DeeplabV3+, einmal trainiert auf z.B. BDD100k). Diese Arbeit ist als Vorarbeit zu sehen und kann als Beleg für die durchgeführten PLFs-Studien dienen.

Dass die statistische Relevanz bei der Betrachtung eines DNN (mit einer Gewichtsdatei) fraglich ist, zeigen unsere durchgeführten Experimente mit dem DeeplabV3+ und dem BDD100k-Datensatz.

Zu diesem Zweck haben wir ein neues Testverfahren entwickelt, das die Diversität von zwei DNN-Ausgängen messen kann und das wir "Orakeltest" nennen. Die Grundidee ist eine Fusion von zwei oder mehr Vorhersagen entsprechend der Übereinstimmung mit den Grundwahrheitsdaten. Je besser die Vorhersagen übereinstimmen, desto höher ist der mIoU des Orakeltests. Je höher der mIoU der Fusion, desto vielfältiger sind die DNN-Vorhersagen. Wenn wir nun ein einzelnes DNN betrachten und es mehrmals auf denselben Daten trainieren, können wir beobachten, dass das mIoU auf demselben Testdatensatz für alle Gewichtsdateien sehr ähnlich ist. Daraus könnte man schließen, dass die Diversität zwischen den DNNs sehr gering ist und die Fehlermodi (PLFs, bei denen das DNN versagt) sehr ähnlich sind.

Die Anwendung von Orakeltests zeigt jedoch, dass dies mit zunehmender Anzahl der für die Fusion verwendeten DNNs zunimmt.

Single weights:

Model A: Acc:0.926394901439326, Acc_class:0.6766385850120534, mIoU:0.565924100204785, fwIoU: 0.8705210207415233

Model B: Acc:0.9268251601916934, Acc_class:0.6962266287253496, mIoU:0.576756125791772, fwIoU: 0.8718260905995445

Model C: Acc:0.9267747528734702, Acc_class:0.6976599971238758, mIoU:0.5771777216553012, fwIoU: 0.8717732584201114

Model D: Acc:0.9270017513669702, Acc_class:0.6950419994589709, mIoU:0.5746009098119959, fwIoU: 0.872242335295981

Oracle testing:

Model A + B: Acc:0.9446270861398401, Acc_class:0.7386071292093997, mIoU:0.6420553099273455, fwIoU: 0.8997195786536527

Model A +B + C: Acc:0.9505230566548976, Acc_class:0.7581816682613032, mIoU:0.667009805599257, fwIoU: 0.9095667584868402

Model A + B + C + D: Acc:0.9540160101523951, Acc_class:0.7706982672914732, mIoU:0.6812629155402113, fwIoU: 0.915582030373858

Daraus lässt sich schließen, dass die Modelle, obwohl es sich um dasselbe DNN (DeeplabV3+) handelt, das mit demselben Trainingsprotokoll und denselben Daten trainiert wurde, unterschiedliche PLFs aufweisen.

Ein weiteres Experiment, bei dem alle Kontrollpunkte innerhalb eines Trainingsruns für Orakeltests verwendet wurden, zeigt einen PLF in der Zaunregion am Rand des Bildes. Mit anderen Worten: Kein einzelner Kontrollpunkt ist in der Lage, den Zaun als "Zaun" zu segmentieren. Stattdessen wurde der Zaunbereich als "Mauer" segmentiert.

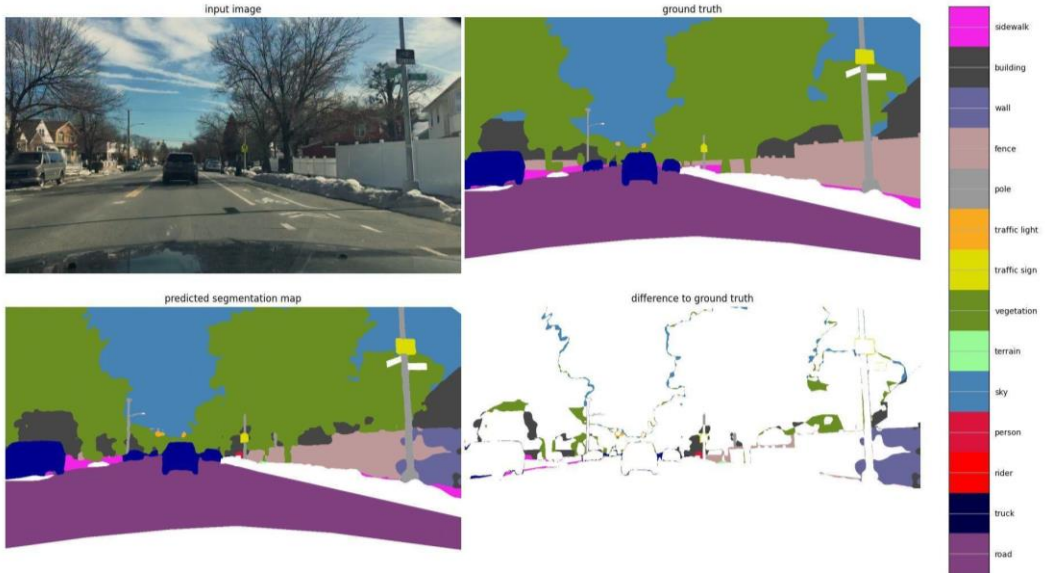


Abbildung 26: Beispiel eines Trainingsruns „Zaun“

Die Experimente von UAP2.3.3 zeigen die Vielfalt der DNN-Vorhersagen trotz desselben Trainingsprotokolls und derselben Trainingsdaten. Ein Ansatz zur Ausnutzung dieser Vielfalt ist die Deep-Ensemble-Methode, bei der die einzelnen DNNs als Mitglieder des Ensembles fungieren. Eine Möglichkeit, die Vorhersagen im Deep Ensemble zu fusionieren, ist die Fusion auf Basis der Logits (vor der Softmax/Argmax-Schicht). Dazu wird das Eingangsbild an alle Mitglieder des Ensembles weitergeleitet und ein gewichteter Durchschnitt der Logits berechnet.

Member A: Acc:0.926394901439326, Acc_class:0.6766385850120534, mIoU:0.565924100204785, fwIoU: 0.8705210207415233

Member B: Acc:0.9268251601916934, Acc_class:0.6962266287253496, mIoU:0.5767566125791772, fwIoU: 0.8718260905995445

Member C: Acc:0.9267747528734702, Acc_class:0.6976599971238758, mIoU:0.5771777216553012, fwIoU: 0.8717732584201114

Deep Ensemble (logits fusion with A, B, C): Acc:0.9309540955987105, Acc_class:0.7015837278766553, mIoU:0.591113255199448, fwIoU: 0.8779293000011158

Dadurch kann der mIoU von 57,72 (höchster mIoU bei den Mitgliedern) auf 59,11 mIoU beim BDD100k-Testdatensatz erhöht werden. Ein Nachteil dieses Ansatzes ist der erhöhte Berechnungsaufwand während der Inferenzzeit. Ein Ansatz, bei dem der Rechenaufwand konstant bleibt, ist die Verschmelzung der Gewichte (anstelle der Vorhersagen), was mit der Methode des Stochastic Weight Averaging (SWA, https://arxiv.org/abs/1803.05407) durchgeführt werden kann. Dazu wurde das DNN mit einer zyklischen Lernrate und einem Cosinus-Annealing-Lernratenplan feinabgestimmt und die letzten Gewichte jedes Zyklus mit einem Durchschnittsgewicht verschmolzen. Zu diesem Zweck wurden 3 Zyklen verwendet, so dass die endgültige Gewichtungsdatei den Mittelwert von insgesamt 4 Gewichtungsdateien darstellt.

Stochastische Gewichtsmittelung (SWA): Acc:0.9277321744794863, Acc_class:0.723084682236998, mIoU:0.5899846708957989, fwIoU: 0.874244710522533

Mit dieser Methode kann in unserem Experiment eine sehr ähnliche Leistung wie mit dem klassischen Deep Ensemble erzielt werden, ohne dass sich die Rechenzeit erhöht. Allerdings entspricht die beobachtete Leistungssteigerung nicht der Steigerung, die potenziell möglich wäre (siehe Orakeltest in UAP2.3.3). Weitere Ansätze zur Verschmelzung der vorhandenen Vielfalt von DNNs sind hierdurch motiviert, bleiben aber im Rahmen dieser Arbeit in WP2.4 außerhalb unseres Aufgabenbereichs.

Schließlich wird die Generalisierungsfähigkeit des beschriebenen SWA- und Deep-Ensembles-Ansatzes durch Tests z.B. mit dem ACDC-Datensatz überprüft, in dem Daten aus verschiedenen Bereichen (Nebel, Schnee, Nacht) verfügbar sind.

<u>Stand der Arbeiten (30.06.2022):</u>

Aufbauend auf die Arbeiten des vorherigen Quartals wurden die beschriebenen Deep Ensembles (DE) und SWA Ansätze auf den ACDC-Datensatz evaluiert. Es wurden keine Anpassungen auf diesen Daten durchgeführt, sodass die Evaluierung eine Messung der Generalisierungsfähigkeit ermöglicht. Die Ergebnisse zeigen, dass das DE mit 3 Member und mehr eine höhere Generalisierungsfähigkeit aufweist als der SWA Ansatz. Hingegen kann der SWA Ansatz das DE mit nur 2 Member in der predictive Performance übertreffen. Die erzielten Ergebnisse lassen darauf schließen, dass eine PLF Untersuchung mit einem SWA Ansatz eine Höhere Aussagekraft besitzt als die Untersuchung mit lediglich einem einzelnen DNN. Der SWA Ansatz kann hier als eine Approximation zur Verwendung von DE mit vielen Members betrachtet werden.

Teilprojekt 3: Methoden und Maßnahmen zur Absicherung von KI

TP3 ist in sechs Arbeitspakete mit folgender thematischer Struktur unterteilt:

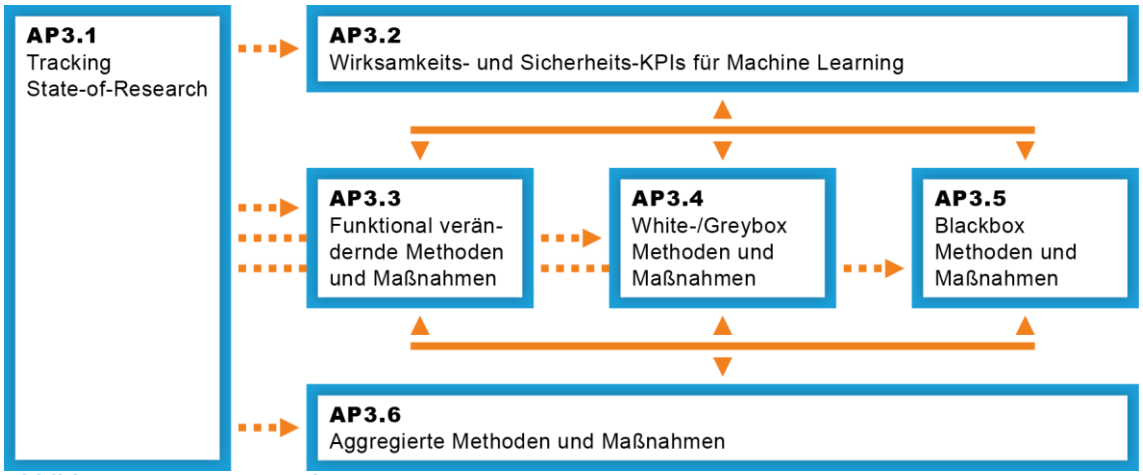


Abbildung 27: TP3-AP-Struktur

Beim Aufbau des Werkzeugkastens soll auf bekannte und unabhängig vom Projektgeschehen entstehende Technologien zurückgegriffen und diese erweitert werden. Dafür ergeben sich drei grundsätzliche Anforderungen, die in diesem Teilprojekt bearbeitet werden müssen:

- Das hochdynamische Forschungsfeld im Bereich Deep Learning (bezogen auf die relevanten Funktionen), Erklärung und Verständnis von KI soll während der gesamten Projektlaufzeit Basis und Absprungplattform für die Aktivitäten im TP3 sein. Dafür werden vorangegangene Ergebnisse in einer Umfeldanalyse evaluiert und neu hinzukommende Ansätze und Ergebnisse kontinuierlich im Rahmen des Umfeldmonitorings mit in den Projektprozess eingespeist.

- Um die im Projekt entwickelten Modelle, Mechanismen und Daten hinsichtlich der Absicherbarkeit von KI im automatisierten Fahren bewerten zu können, müssen akzeptierte und relevante Kennzahlen für diese Modelle, Mechanismen und Daten erstellt werden. Diese Kennzahlen (KPI) müssen zudem mit der Beschreibungssprache, die in TP4 entwickelt und für die Absicherungsargumentation verwendet wird, kompatibel sein.

- Es müssen die bestehenden Methoden und Maßnahmen für die Plausibilisierung, Bewertung Sicherstellung von (nicht) funktionalen Eigenschaften von KI bewertet und erweitert werden.

Diesen Bedarfen wird in der Strukturierung des Teilprojektes in sechs Arbeitspaketen Rechnung getragen.

Für die Bewertung bestehender und Entwicklung neuer Methoden und Maßnahmen wird eine Unterteilung in vier Paradigmen vorgenommen wurde (AP3.3-3.6). Motiviert ist diese Unterteilung zum einen durch die Komplexität der Aufgabe, zum anderen durch die verschiedenen benötigten Kompetenzen:

- Kenntnisse über Training und Optimierung (AP3.3)

- Kenntnisse in der Analyse und Bewertung von Strukturen und Dynamiken von tiefen neuronalen Netzen (AP3.4)
- Datenanalysekenntnisse (AP3.5)

AP3.1 Tracking des State-of-Research in den Bereichen Bewertung, Plausibilisierung und Erklärung von KI-Methoden (13 PM)

Aufgaben Valeo:

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

• Mitwirken bei der Erstellung des Kriterienkatalogs und Gliederung zur Definition des relevanten Forschungsfelds.

Initialer State-of-Research-Report (E3.1.2)

• Beitrag zum initialen State-of-Research-Report mit Schwerpunkt auf introspektive Methoden (AP3.4). Hierzu zählen Methoden, die eine Erklärbarkeit von CNNs zum Ziel haben, jedoch nicht die funktionalen Eigenschaften ändern. Zudem soll relevanter, öffentlich verfügbarer Code mit verlinkt werden.

Auflistung und Kategorisierung relevanter State-of-Research (E3.1.3)

• Fortsetzung der State-of-Research-Analyse gemäß der in E3.1.1 festgelegten Themenbereiche. Mitwirken bei der Identifizierung von neuen Akteuren und Themenbereichen, die für KI-Absicherung relevant sind.

Öffentlicher Zugang zu E3.1.3 (E3.1.4)

• Veröffentlichung der Forschungsergebnisse in Form von Publikationen im Themenbereich KI-Absicherung auf einschlägigen Konferenzen im Bereich Machine Learning und Computer Vision.

Aktiver Austausch mit wissenschaftlicher Community (E3.1.5)

• Gastvorträge auf entsprechenden Konferenzen / Workshops über KI-Absicherung. Vorstellung von Forschungsergebnissen der projektinternen Summer School.

Stand der Arbeiten (31.12.2019):

Der aktive Austausch mit wissenschaftlicher Community (**E3.1.5**) wurde vorangetrieben, in dem ein Workshop auf der „Conference on Computer Vision and Pattern Recognition“ (CVPR) organisiert wurde. Die CVPR gehört zu den bedeutendsten Konferenzen im Bereich Computer Vision und künstliche Intelligenz und hatte eine Rekordbesucherzahl von 9000 Teilnehmern im Jahre 2019 zu verzeichnen. Der organisierte Workshop trägt den Namen „Safe Artificial Intelligence for Automated Driving“ (SAIAD; https://sites.google.com/view/saiad-wscvpr19). Der Workshop wurde von zahlreichen Industrie- und Wissenschaftspartner co-organisiert und dies unter der Leitung von Valeo. Der Workshop erreichte großes mediales Interesse, wie auf der Website unter „Workshop Day“ zu sehen ist. Der Workshop

erstreckte sich über einen kompletten Tag und war ganztägig ausgebucht. Wir konnten hochkarätige Redner für unseren Workshop gewinnen und erhielten zahlreiche Paper Einreichungen zum Thema Safe AI, sodass wir mit einem hochqualitativen Review Prozess die besten Papers auswählen konnten. Aufgrund des überragenden Feedbacks entschieden wir uns, den SAIAD Workshop für das Jahr 2020 fortzusetzen. Trotz einer drastischen Reduzierung der angenommenen Workshops, (ca. 33 %) gegenüber 2019, konnten wir den CVPR Workshop Chair von der Qualität und Relevanz unseres Workshops überzeugen. Die Vorbereitungen für den zweiten SAIAD Workshop laufen (https://sites.google.com/view/saiad2020/).

Weiterhin wurde im Rahmen des Ergebnis **E3.1.4** ein Paper zum Thema „Strategy to Increase the Safety of a DNN-based Perception for HAD Systems“ geschrieben und auf der Hauptkonferenz der CVPR eingereicht. Das Paper entstand unter Führung von Valeo und in Kooperation mit Peter Schlicht und Fabian Hüger von Volkswagen.

Weiterhin wurde im Rahmen von **E3.1.1** mit einer thematischen Clusterung des State of the Art begonnen. Der konsolidierte Stand des Clusterings wurde mit dem sogenannten KPI-Graphen aus AP3.2 zusammengeführt und mit allen Teilnehmern in TP3 abgestimmt. Für die Durchführung der State of the Art Recherche wurden Themenverantwortlichkeiten jedes Partners in AP3.1 vereinbart.

E3.1.3: Für die State of the Art Recherche wurden geeignete Tools (Confluence vs. Mendeley vs. Zotero) evaluiert und man hat sich schließlich auf Confluence geeinigt. Eine geeignete Confluence Struktur wurde zudem erarbeitet.

Stand der Arbeiten (30.06.2020):

Auch im Jahre 2020 wurde der Workshop mit dem Titel: Safe Artificial Intelligence for Automated Driving (SAIAD, https://sites.google.com/view/saiad2020/) angenommen **(E3.1.5)**. Aufgrund der Corona Beschränkung wurde der Workshop nicht in Seattle, sondern komplett virtuell abgehalten. Die erneute Annahme des Workshops zeigt das Interesse am Thema: „Sichere KI“. Im Vergleich zum Vorjahr wurden auch Opel und Bosch in das Organisationsteam aufgenommen, um eine noch größere Wirkung zu erzielen. Es konnten 13 Publikationen angenommen werden, die jeweils 3 hochkarätige Reviews erhielten. Aus den 13 Publikationen wurden aufgrund der Bewertung der Reviewer 6 Orals und 7 Poster ausgewählt. Der Best-Paper-Award wurde während des Workshops festgelegt und ging an Andreas Bär von der TU Braunschweig. Die Publikationen wurden erneut in die IEEE-Proceedings aufgenommen. Die Videos der Hauptredner und der Autoren der Publikationen wurden auf die Website hochgeladen, um sie für die Nachwelt zu erhalten (https://sites.google.com/view/saiad2020/videos?authuser=0).

Für den initialen Report des **E3.1.2** wurden mehrere Unterkapitel geschrieben. Die Unterkapitel “Temporal Consistency”, “Measuring Safety: Safety Metrics for DNN”, sowie “Multi-Task Networks”. Der initiale Report soll als Survey Paper auf der Website

von KI-Absicherung veröffentlicht werden. Eine Einreichung an eine Konferenz/Journal ist zusätzlich im Gespräch.

Im Rahmen von **E3.1.4** wurde an einem Paper mitgearbeitet (dritter Autor) mit dem Titel:

“Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications”. Federführend war Continental. Das Paper wurde beim WAISE Workshop auf der SafeComp 2020 angenommen und wird ab September 2020 in den entsprechenden Proceedings veröffentlicht.

Stand der Arbeiten (31.12.2020):

Nach bereits zwei erfolgreich durchgeführten SAIAD Workshops wurde beschlossen, den aktiven Austausch mit der wissenschaftlichen Community der CVPR weiter fortzuführen. Daher wurde eine Bewerbung für die dritte Episode des SAIAD Workshops eingereicht. Erneut wurde unsere Bewerbung akzeptiert, sodass es einen dritten SAIAD Workshop im Jahre 2021 geben wird.

Der Fokus des Workshops liegt erneut auf sichere KI für die Wahrnehmung im automobilen Umfeld. Für die praktikable Umsetzung sind Rahmenbedingungen notwendig, die in Standardisierungen beschrieben werden. Diese müssen unter Beteiligung von KI-Experten geschaffen werden. Daher werden wir den Themenhorizont erweitern und aktuelle Arbeiten zur Erstellung von Standardisierungen diskutieren.

Die intensive Beschäftigung mit Sicherer KI bringt zwangsläufig ethische Aspekte mit sich, denen wir eine Bühne geben wollen. Darüber hinaus ist Safe AI für uns auch in anderen Bereichen als der Automobilindustrie interessant, um eventuell die Übertragbarkeit zu diskutieren.

Zusammengefasst ist das Hauptthema des diesjährigen SAIAD Workshops Safe AI in der Wahrnehmung plus drei angrenzende Bereiche Standardisierung in Safe AI, Ethik und die Safe AI aus anderen Bereichen als der Automobilindustrie und deren Übertragbarkeit. Um den Austausch über diese Themen zu fördern, haben wir geeignete Redner eingeladen.

Website: https://sites.google.com/view/saiad2021

Im Rahmen von **E3.1.4** wurde an einem Short Paper gearbeitet mit dem Titel: Online Out-of-Domain Detection for Automated Driving

Dieses wurde auf dem Machine Learning in Certified Systems Workshop (https://mlcertifiedsystems.deel.ai/) eingereicht und angenommen.

Stand der Arbeiten (30.06.2021):

Erneut wurde der SAIAD-Workshop auf der CVPR angenommen (https://sites.google.com/view/saiad2021). Es wurden hochkarätige Referenten wie Zico Kolter, Patrick Perez, Mike Wagner, Eric Hilgendorf, Been Kim und Bernt Schiele eingeladen, um die neuesten Erkenntnisse im Bereich der sicheren KI zu präsentieren. Darüber hinaus wurden Podiumsteilnehmer eingeladen, die sich den kritischen Fragen der Moderatoren und des Publikums stellten. Alex Kendall, Fisher Yu, Beipeng Mu, Markus Enzweiler und Peter Schlicht haben sich dazu bereit erklärt. In diesem Jahr wurde ein neuer Rekord bei der Anzahl der eingereichten und angenommenen Papers aufgestellt: Es wurden 5 (“long orals”) und 10 (“short orals”) angenommen. Es gab drei bis vier Reviews pro Paper, um einen hohen Qualitätsstandard der Vorträge zu gewährleisten. Die Teilnehmerzahl war im Peak bei 150. Anbei ein Screenshot aus dem Talk von Patrick Perez:

Abbildung 28: Auszug aus einer Präsentation beim SAIAD-Workshop

Stand der Arbeiten (31.12.2021):

Die Bewerbung für eine vierte Auflage des SAIAD-Workshops wurde geschrieben und eingereicht. Es fanden mehrere Sitzungen statt, um sich über das Format und die Ausrichtung der vierten Ausgabe auszutauschen. Unter anderem wurde die Option einer Herausforderung (Anomalie-Segmentierung) diskutiert, aber vorerst auf Eis gelegt.

Stand der Arbeiten (30.06.2022):

Die vierte Auflage des SAIAD-Workshops wurde auf der European Conference on Computer Vision (ECCV) 2022 angenommen. Es wurden entsprechende Vorbereitungen getroffen, die das Paper Review und Keynote Talks betreffen. Da es sich um ein hybrides Modell handelt, d.h. vor Ort sowie online, ergibt sich ein höherer Organisationsaufwand im Vergleich zu den Jahren davor.

AP3.2 Höherwertige Funktion KPIs für KI Funktionen (5 PM)

Aufgaben Valeo:

Strukturierte Übersicht der verfügbaren Basis-KPIs (E3.2.2)

• Erstellen einer strukturierten Übersicht der verfügbaren KPIs.

• Kommunikation der Annahmen und des Verständnisses der verfügbaren KPIs (von AP3.4).

Stand der Arbeiten (31.12.2019):

E.3.2.2: Es wurde eine erste strukturierte und kategorisierte Übersicht über Key Performance Indicator (KPI) erstellt. Diese Übersicht wurde in einem Graphen abgebildet. Der Graph umfasst einen Umfang von ca. 16 DIN A4 Seiten und wurde über Monate stetig von mehreren Partner (hauptsächlich Valeo) weiterentwickelt[6].

Eine textuelle Beschreibung von wesentlichen Punkten dieses Graphen ist im Paper „Strategy to Increase the Safety of a DNN-based Perception for HAD Systems“ zu finden[7].

Stand der Arbeiten (30.06.2020):

Der beschriebene Graph aus dem **E3.2.2** (Stand: 31.12.2019) wurde mit einer GSN Struktur und modifizierter Auflistung der DNN Unzulänglichkeiten erweitert.

Die Erweiterungen des Graphen wurden unter anderem in dem Paper “Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications” beschrieben.

[6] https://sharepoint.cloud4partner.com/websites/KI-ML/Freigegebene%20Dokumente/TP3%20Methoden%20und%20Maßnahmen/02_AP3.2/MM-KPI-Bride-Graph.drawio

[7] https://sharepoint.cloud4partner.com/websites/KI-ML/Freigegebene%20Dokumente/07%20Publikationen/2020-06%20CVPR%20Seattle%20-%20Saemann%20et%20al/Saemann%20et%20al%202020%20-%20Strategy_to_Increase_the_Safety_of_a_DNN-based_Perception_for_HAD_Systems.pdf

Stand der Arbeiten (31.12.2020):

Aufgrund der bereits geleisteten Arbeit in diesem AP wurden in diesem Halbjahr lediglich Dokumente zur Beschreibung der DNN Unzulänglichkeiten begutachtet und an Diskussionen hierzu teilgenommen.

Stand der Arbeiten (30.06.2021):

Für die Erstellung von Benchmarks ist es notwendig, geeignete Metriken zu definieren. Insbesondere wurden bestehende Metriken wie MIoU, Recall, Precision, etc. hinsichtlich ihrer Eignung als Benchmark-Metriken analysiert. Auch die Verknüpfung mit den Machine Learning Safety Requirements aus AP4.3 war Teil der laufenden Diskussion.

Stand der Arbeiten (31.12.2021):

Um die Signifikanz der Benchmark-Metriken (Basis-KPIs) zu erhöhen, wurden eine k-fache Kreuzvalidierung und ein Nested-Cross-Validation-Verfahren diskutiert. Diese Verfahren sehen vor, die statische Signifikanz der metrischen Ergebnisse u.a. durch Variation der Trainings- und Testsplits zu erhöhen. Aufgrund des erheblichen Mehraufwands in der Rechenzeit einigte man sich auf den Kompromiss, diese Verfahren nicht anzuwenden.

Wir haben uns bei der Diskussion über die Datenauswahl für die Realdatenuntersuchung beteiligt. Die Position bei der Datenauswahl war, eine Kombination aus CityPersons und BDD100k zu verwenden, um die Signifikanz der Ergebnisse der Metrik zu erhöhen.

Stand der Arbeiten (30.06.2022):

Die Entwicklung des Benchmark Tools wurde unterstützt. Die eigenen Arbeiten hinsichtlich der Strukturierung der Basis-KPIs wurden finalisiert und in Form des Ergebnissteckbriefs dokumentiert.

AP3.3 Funktional verändernde Methoden und Maßnahmen (28 PM)

Aufgaben Valeo:

Netzwerk-Optimierung (E3.3.3) > Multi-Task-Netzwerk

• Erweiterung der Netzwerk-Architektur zu einem Multi-Task Netzwerk (z.B. Bounding Box + Semantische Segmentierung von Fußgängern).

• Vergleich der Netzwerkrobustheit des Multi-Task Netzes im Vergleich zum Single-Task Netz.

• Vergleich der Netzwerkrobustheit bei harter und weicher Parameterteilung.

• Implementierung und Evaluierung von Fehlerfunktionen basierend auf die Ausnutzung von zeitlicher Konsistenz in Videos.

Netzwerk-Optimierung (E3.3.3) > Erhöhung der Robustheit durch Ausnutzung von zeitlicher Konsistenz in Videos

• Kalibrierte Konfidenz-Werte aus AP3.4 dienen als Maßstab zur Nutzung von berechneten Merkmalskarten aus vorherigen Zeitschritten. Entwicklung und Implementierung von Fusionskonzepten der Merkmalskarten aus verschiedenen Zeitschritten, die zur Laufzeit durchgeführt werden können. Diese Fusionskonzepte schließen die Betrachtung von LSTMs sowie die Integration des optischen Flusses mit ein. Der optische Fluss dient als Maß für die räumliche Änderung von Merkmalen zwischen Eingangsbilder aufeinanderfolgender Zeitstufen und soll auf die Merkmalskarten innerhalb des DNN angewendet werden. Die dadurch generierten Merkmalskarten werden mit den neu berechneten Merkmalskarten fusioniert. Weiterhin soll eine Transformation des optischen Flusses in die Merkmalskartenebene untersucht werden, um die Anwendbarkeit auf die Merkmalskarten zu erhöhen. Diese Transformation sowie der Einsatz von LSTMs erfordert ein erneutes Training des DNN, um entsprechende Parameter zu lernen.

• Evaluierung der Robustheit bezüglich Performanz-KPIs bei gezielte, kurzzeitig manipulierte Eingangsdaten. Gegenüberstellung verschiedener Fusionskonzepte bzw. Parametrierungen bezüglich der Performanz-KPIs und stärke der manipulierten Eingangsdaten.

Optimierte Datensatz-Selektion (E3.3.1) > Iteratives Corner Case-Training

• Trainieren mit aus AP3.5 gefundenen Corner Cases in einem iterativen Prozess (konkrete Szenen oder auch durch generative Ansätze erzeugte Datenverteilungen) und vergleichende Analyse der Robustheit.

Stand der Arbeiten (31.12.2019):

Netzwerk-Optimierung (E3.3.3) > Erhöhung der Robustheit durch Ausnutzung von zeitlicher Konsistenz in Videos: Konzeption von Fusionskonzepten der Merkmalskarten aus unterschiedlichen Zeitstufen. Es wurden Merkmalskarten aus unterschiedlichen Zeitstufen für ein Deep Neural Network (DNN) zur semantischen Segmentierung begutachtet und evaluiert. Es wurden mehrere Konzepte für eine Fusionierung erstellt, implementiert und ausgewertet. Beispielsweise wurde die Fusionierung durch Mittelwertbildung von Merkmalskarten aus hintereinander folgenden Zeitschritten realisiert. Diese führen in manchen Szenen zu einer verbesserten/robusteren semantischen Segmentierung. Dieses Konzept der Fusionierung setzt jedoch eine hohe Framerate der Kamera und eine geringe Geschwindigkeit des Fahrzeugs voraus. Je niedriger die Framerate und je höher die Geschwindigkeit des Fahrzeugs desto größer die semantischen Unterschiede von hintereinander folgenden Zeitstufen. Aus diesem Grund folgen weitere Anpassungen und Untersuchungen bzgl. des Fusionskonzeptes.

Netzwerk-Optimierung (E3.3.3) > Multi-Task-Netzwerk: Arbeiten haben noch nicht gestartet. Laut Plan beginnen die Arbeiten in diesem AP im Monat 7.

Optimierte Datensatz-Selektion (E3.3.1): Arbeiten haben noch nicht gestartet. Laut Plan beginnen die Arbeiten in diesem AP im Monat 19.

Stand der Arbeiten (30.06.2020):

Moderne selbstüberwachte Lernansätze zur monokularen Tiefenschätzung leiden in der Regel unter einer Skalenmehrdeutigkeit. Sie lassen sich nicht gut verallgemeinern, wenn sie auf die Entfernungsschätzung für komplexe Projektionsmodelle wie bei Fischaugen- und Rundumkameras angewendet werden. Wir stellten eine neuartige Multi-Task-Lernstrategie vor, um die selbstüberwachte monokulare Entfernungsschätzung bei Fischaugen- und Lochkamerabildern zu verbessern. Zunächst stellen wir eine neuartige Netzwerkarchitektur zur Entfernungsschätzung vor, die einen auf Selbstbeobachtung basierenden Encoder in Verbindung mit einer robusten semantischen Merkmalsführung zum Decoder verwendet, der einstufig trainiert werden kann. Zweitens integrieren wir eine verallgemeinerte robuste Verlustfunktion, die die Leistung erheblich verbessert und gleichzeitig die Notwendigkeit einer Hyperparameter-Abstimmung mit dem Reprojektionsverlust beseitigt. Schließlich reduzieren wir die Artefakte, die durch dynamische Objekte verursacht werden, die die statische Weltannahme verletzen, indem wir eine semantische Maskierungsstrategie verwenden. Wir verbessern den RMSE-Wert früherer Arbeiten über Fischauge erheblich, indem wir den RMSE-Wert um 25% reduzieren. Da es nur begrenzte Arbeiten an Fischaugen-Kameras gibt, haben wir die vorgeschlagene Methode am KITTI mit einem Lochblendenmodell evaluiert. Wir erreichten den neuesten Stand der Technik unter den selbstüberwachten Methoden, ohne dass eine externe Maßstabsschätzung erforderlich ist.

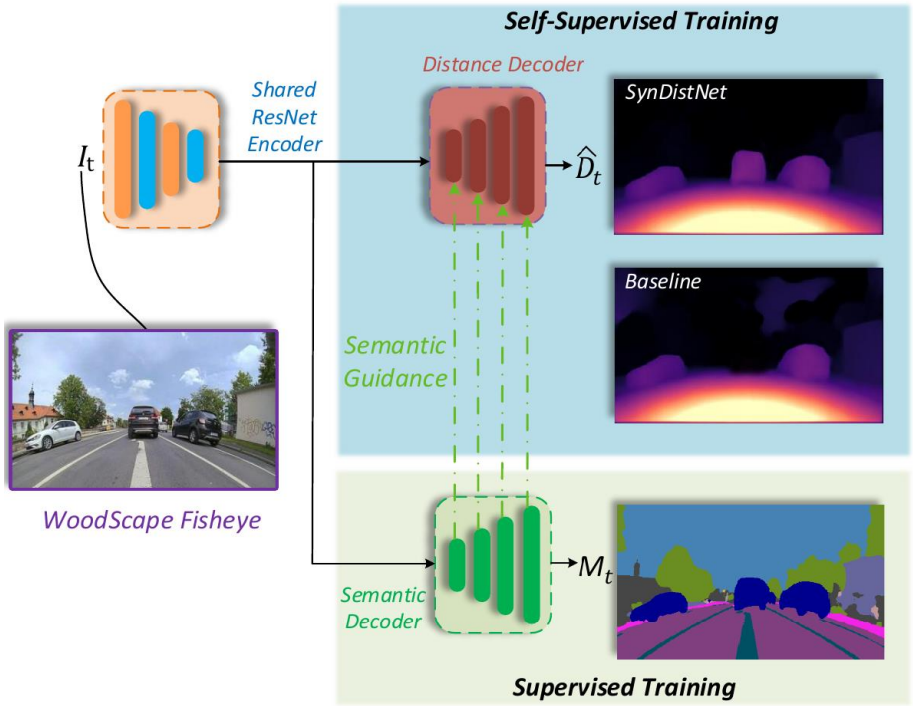


Abbildung 29: Überblick über die gemeinsame Vorhersage des Abstands D_t und der semantischen Segmentierung M_t aus einem einzigen Eingabebild I_t. Im Vergleich zu früheren Ansätzen erzeugt unsere semantisch geführte Abstandsabschätzung schärfere Tiefenkanten und vernünftige Abstandsschätzungen für dynamische Objekte.

Selbstüberwachte Tiefenschätzung:

Garg und *Zhou* zeigten, dass es möglich ist, Netzwerke selbstüberwacht zu trainieren, indem man die Tiefe als Teil einer geometrischen Projektion zwischen Stereobildern bzw. sequentiellen Bildern modelliert. Das ursprüngliche Konzept wurde durch die Berücksichtigung verbesserter Verlustfunktionen, die Anwendung von generativen kontradiktorischen Netzen (GANs) oder generierten Proxy-Labels aus traditionellen Stereo-Algorithmen oder synthetischen Daten erweitert. Andere Ansätze schlugen vor, spezialisierte Architekturen für die selbstüberwachte Tiefenschätzung zu verwenden. Sie wenden Lehrer-Schüler-Lernen an, um Testzeitverfeinerungsstrategien zu verwenden, wiederkehrende neuronale Netze einzusetzen oder die Kameraparameter vorherzusagen, um ein Training über Bilder von verschiedenen Kameras hinweg zu ermöglichen.

Ein neuerer Ansatz von *Ravi Kumar* stellt einen erfolgreichen Proof-of-Concept für die Anwendung selbstüberwachten Tiefenschätzverfahrens bei der Aufgabe der Entfernungsschätzung aus Fischaugen-Kamerabildern vor, der als Basis für diese Arbeit verwendet wird. Auch in neueren Ansätzen wurde die Anwendung der selbstüberwachten Tiefenschätzung auf 360°-Bilder untersucht. Abgesehen von diesen Arbeiten wurde die Anwendung der selbstüberwachten Tiefenschätzung auf weiter fortgeschrittenen Geometrien, wie z.B. Fischaugen-Kamerabilder, jedoch noch nicht eingehend untersucht.

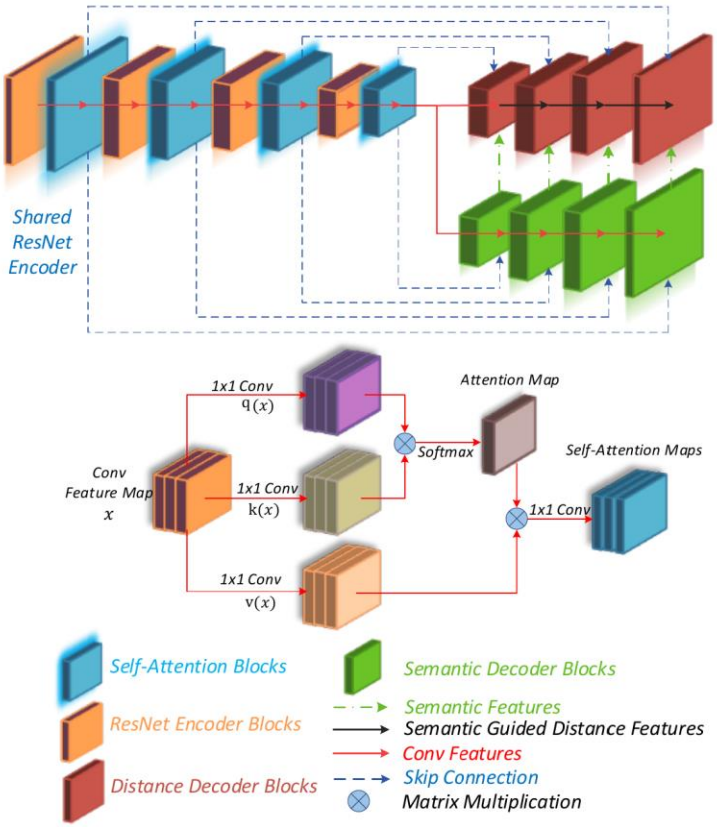


Abbildung 30: Visualisierung der von uns vorgeschlagenen Netzwerkarchitektur}, um die Tiefenabschätzung semantisch zu leiten. Wir verwenden einen auf Selbstbeobachtung basierenden Encoder und einen semantisch geführten Decoder mit pixeladaptiven Faltungen.

Multi-Task-Lernen:

Im Gegensatz dazu, ein Netzwerk nur eine einzelne Aufgabe vorhersagen zu lassen, ist es auch möglich, ein Netzwerk so zu trainieren, dass es mehrere Aufgaben auf einmal vorhersagen kann, was nachweislich Aufgaben wie z.B. semantische Segmentierung, Domänenanpassung und Tiefenschätzung verbessert. Während anfängliche Arbeiten Verluste oder Gradienten durch einen empirischen Faktor gewichtet haben, können aktuelle Ansätze diesen Skalenfaktor automatisch schätzen. Wir übernehmen die auf Unsicherheit basierende Gewichtung von Kendall, um verschiedene Aufgaben abzuwägen.

Viele neuere Ansätze zielen darauf ab, den optischen Fluss in das selbstüberwachte Training der Tiefenschätzung zu integrieren, da diese zusätzliche Aufgabe ebenfalls selbstüberwacht trainiert werden kann. In diesen Ansätzen werden beide Aufgaben gleichzeitig vorhergesagt. Dann werden Verluste angewendet, um aufgabenübergreifende Konsistenz zu erzwingen, um bekannte geometrische Einschränkungen durchzusetzen oder um eine modifizierte Rekonstruktion des verzerrten Bildes zu induzieren. Obwohl der typische Ansatz die Kompensation durch optischen Fluss ist, schlagen wir aus zwei Gründen eine alternative Methode vor, stattdessen semantische Segmentierung zu verwenden. Erstens ist die semantische Segmentierung eine ausgereifte und gängige Aufgabe beim autonomen Fahren, die

genutzt werden kann. Zweitens ist der optische Fluss rechnerisch komplexer und schwieriger zu validieren, da es schwierig ist, die Grundwahrheit zu erhalten.

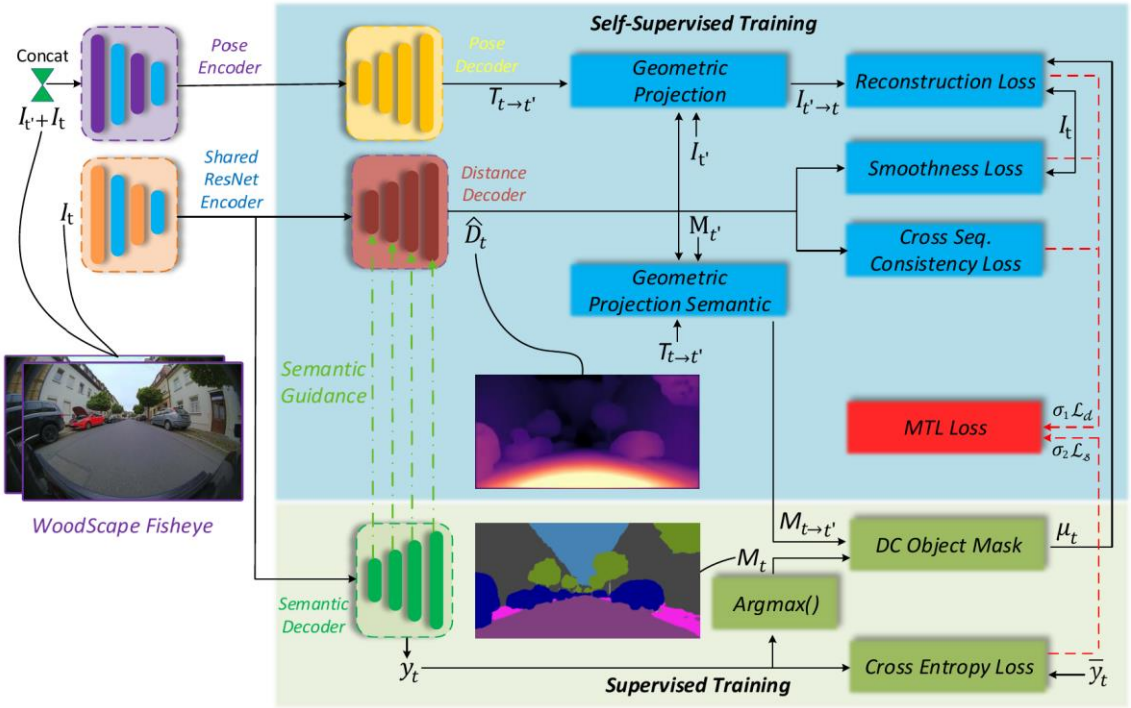


Abbildung 31: Überblick über den von uns vorgeschlagenen Rahmen für die gemeinsame Vorhersage von Entfernung und semantischer Segmentierung. Der obere Teil (blaue Blöcke) beschreibt die einzelnen Schritte für die Tiefenschätzung, während die grünen Blöcke die einzelnen Schritte beschreiben, die für die Vorhersage der semantischen Segmentierung erforderlich sind. Beide Aufgaben werden innerhalb eines Multi-Task-Netzes unter Verwendung des gewichteten Gesamtverlustes optimiert.

Semantisch-gesteuerte Tiefenschätzung:

Mehrere neuere Ansätze verwendeten auch semantische oder Instanz-Segmentierungstechniken, um sich bewegende Objekte zu identifizieren und sie innerhalb des photometrischen Verlusts entsprechend zu behandeln. Zu diesem Zweck werden die Segmentierungsmasken entweder als zusätzliche Eingabe in das Netzwerk gegeben oder dazu verwendet, die Posen für jedes Objekt separat zwischen zwei aufeinanderfolgenden Bildern vorherzusagen und für jedes Objekt eine separate starre Transformation anzuwenden. Um ein ungünstiges zweistufiges (Vor-)Trainingsverfahren zu vermeiden, werden bei anderen Ansätzen beide Aufgaben in einem Multi-Task-Netzwerk gleichzeitig trainiert, wodurch die Leistung durch aufgabenübergreifende Führung zwischen diesen beiden Facetten des Szenenverständnisses verbessert wird. Darüber hinaus können die Segmentierungsmasken zwischen Einzelbildern projiziert werden, um semantische Konsistenz zu erzwingen, oder die Kanten können so erzwungen werden, dass sie in beiden Vorhersagen in ähnlichen Regionen erscheinen. In dieser Arbeit schlagen wir vor, das sog. „Warping“ zu verwenden, um Rahmen mit bewegten Objekten zu entdecken und ihre Tiefe aus diesen Rahmen durch Anwendung einer einfachen semantischen Maskierungstechnik zu lernen. Darüber hinaus schlagen wir einen

neuartigen, auf Selbstbeobachtung basierenden Encoder vor, zusammen mit semantischen Merkmalen, die den Decoder mittels pixeladaptiver Faltungen führen. Durch diese einfache Änderung können wir ein einstufiges Training anwenden, so dass das Vortraining eines semantischen Segmentierungsnetzes entfällt.

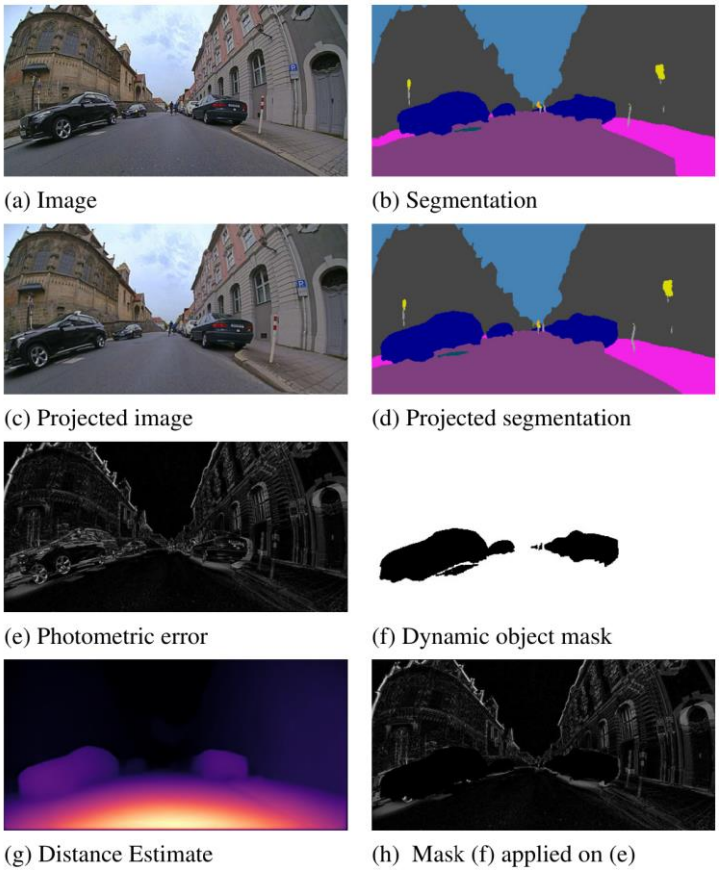


Abbildung 32: Anwendung unserer semantischen Maskierungsmethoden, um potenziell dynamische Objekte zu behandeln. Die dynamischen Objekte innerhalb der Segmentierungsmasken aus aufeinanderfolgenden Bildern in (b) und (d) werden zu einer dynamischen Objektmaske akkumuliert, die zur Maskierung des photometrischen Fehlers (e) verwendet wird, wie in (h) gezeigt.

Geometrie und Aussehen sind zwei entscheidende Anhaltspunkte für das Verständnis von Szenen, z. B. in Automobilszenen. Wir entwickeln ein Multi-Task-Lernmodell, um den metrischen Abstand und die semantische Segmentierung auf synergetische Weise abzuschätzen. Insbesondere nutzen wir die semantische Segmentierung potenziell sich bewegender Objekte, um falsch projizierte Objekte innerhalb des Ansichtssyntheseschritts zu entfernen. Wir schlagen auch eine neuartige Architektur vor, um die Entfernungsschätzung semantisch zu steuern, die in einer einstufigen Weise trainierbar ist, und führen die Anwendung einer robusten Verlustfunktion ein. Unser Hauptaugenmerk liegt auf der Entwicklung des von uns vorgeschlagenen Modells für weniger erforschte Fischaugen-Kameras auf der Grundlage des WoodScape-Datensatzes. Wir demonstrieren die Wirkung jedes vorgeschlagenen Beitrags einzeln und erhalten hochmoderne Ergebnisse sowohl für WoodScape- als auch für KITTI-Datensätze zur Selbstüberwachung Entfernungsschätzung.

Stand der Arbeiten (31.12.2020):

Die aktuelle Vorgehensweise besteht darin, ein MTL-Netzwerk mit verschiedenen Wahrnehmungsaufgaben wie Tiefenschätzung, semantische und Bewegung Segmentierung gefolgt von 2D-Objekterkennung zu implementieren, die Robustheit der Methode an unserem internen Datensatz zu testen und die Ergebnisse zu überprüfen. Sobald wir ein robustes Encoder-Decoder-Netzwerk mit gemeinsamen Merkmalen für alle Aufgaben, werden wir ein Multi-Task-Modell für semantische und Objekterkennung aufbauen und die semantischen Merkmale an den Objekterkennungs-Decoder weitergeben. Dann führen wir eine Ablation Studie durch und überprüfen die Auswirkungen des Designs auf den KIA-Datensatz. Wir werden dann den Code des Modells zusammen mit dem Trainingscode den anderen Projektpartnern zur Verfügung stellen.

Stand der Arbeiten (30.06.2021):

Das Multi-Task-Learning (MTL) der Objekterkennung und semantischen Segmentierung unter Verwendung der semantischen Führung wurde an den KIA-Datensätzen der Tranche 3, 4 und 5 durchgeführt. Die in der Tabelle dargestellten Ergebnisse zeigen, dass die MTL dieser Aufgaben die Trainingsergebnisse der Einzelaufgaben übertrifft. Wir haben eine Ablation Studie durchgeführt und die Auswirkungen des Designs auf den KIA-Datensatz überprüft. Wir stellen den MTL-Code des Modells zusammen mit dem Trainings Code den anderen Projektpartnern zur Verfügung.

Tabelle 7 Trainingsergebnisse im Vergleich

Model	Backbone	Epochs	mAP	mIoU
Detection	ResNet 18	15	0.3791	x
Semantic	ResNet 18	15	x	0.7685
MTL	ResNet 18	15	0.3830	0.7836
Detection	ResNet 50	15	0.3852	x
Semantic	ResNet 50	15	x	0.7861
MTL	ResNet 50	15	0.4193	0.8049

Stand der Arbeiten (31.12.2021):

Der Code wurde auf das KIA interne Bitbucket hochgeladen und dokumentiert. Die finalen quantitativen und qualitativen Ergebnisse wurden im AP3.3 Workshop vorgestellt. Die Arbeiten im Arbeitspaket 3.3 sind damit abgeschlossen.

Stand der Arbeiten (30.06.2022):

Die durchgeführten Arbeiten wurden für das finale Release 5 aufbereitet und veröffentlicht. Hierzu wurde eine Vielzahl von Release Requirements erfüllt, die z.B. das Ausfüllen des Mechanismen-Katalogs sowie das Erstellen einer Mechanismen Beschreibung, welche in 4 Blöcke unterteilt wurde, beinhalten.

AP3.4 White-/Greybox-Methoden und -Maßnahmen (34 PM)

Aufgaben Valeo:

Plausibilisierung (E3.4.1) > Erweiterte Untersuchung und Anpassung von Heatmap-Methoden (E3.4.1b)

• Auswahl geeigneter, konkreter Heatmap-Verfahren, die auf unterschiedliche Prinzipien beruhen, z. B. Gradienten- und Dekompositionsverfahren.

• Entwicklung eines Heatmap-Verfahrens durch Kombination unterschiedlicher Prinzipien.

• Implementieren und Kombinieren von bestehenden Heatmap-Verfahren.

• Evaluierung der entwickelten Heatmap basierend einer Verdeckungsmetrik, welche die relevanten Bildregionen innerhalb der Heatmap verdeckt und die folgende Verringerung der Konfidenz misst.

Plausibilisierung (E3.4.1) > Gelernte Repräsentation (E3.4.1c)

• Interpretation von Merkmalskarten: Ermittlung eines Ähnlichkeitsmaßes zwischen Merkmalskarten entlang der Zeitachse. Annahme: Die Ähnlichkeit der Merkmalskarten zwischen einzelnen Zeitschritten ist hoch. Je größer die Zeitschritte von Bild zu Bild, desto niedriger die Ähnlichkeit zwischen Merkmalskarten (Ähnlichkeitsmaß). Die Interpretation der Merkmalskarten soll an unterschiedlichen Tiefen im Netzwerk erfolgen und auf Vergleichbarkeit ausgewertet werden.

• Erhöhung der Interpretierbarkeit von Merkmalskarten basierend auf ihren Einfluss auf die Netzwerkvorhersage. Untersuchung einzelner Merkmalskarten auf ihre Aussagekraft und Auswirkung auf die Netzwerkausgabe.

• Untersuchung einzelner Merkmalskarten visuell, z. B. durch eine t-SNE Visualisierung oder einem Deconvolutional Netzwerk • Ermittlung einer Metrik, die bedeutungsvolle Unterschiede zwischen den Merkmalskarten abbilden kann.

Unsicherheitsmodellierung (E3.4.2) > Netzwerkkalibrierung (E3.4.2b)

• Kalibrierung der Netzwerk-Konfidenzen (Fußgängererkennung mit semantischer Segmentierung).

• Ausgabe eines relativen Konfidenz-Wertes pro Pixel.

• Ermittlung von Konfidenz-Wertebereichen, die eine Aussage über die Zuverlässigkeit der Netzwerkvorhersage zum aktuellen Zeitpunkt haben.

Offline-Validierung (E3.4.5) > Plausibilisierungsmaß (E3.4.5b)

• Auswahl eines geeigneten Klassifikators zur Bewertung der Heatmaps (Metadaten).

• Aktive Interpretation der kombinierten Heatmaps durch Training eines Meta-klassifikators und Formulierung eines Plausibilisierungsmaßes.

• Ableiten einer Metrik, die die Bedeutsamkeit von Merkmalskarten innerhalb eines DNNs wiedergibt.

Stand der Arbeiten (31.12.2019):

E3.4.2: Die Konfidenzkalibrierung beschreibt das Problem der Vorhersage von Wahrscheinlichkeitsschätzungen, die repräsentativ für die wahre Wahrscheinlichkeit der Richtigkeit sind [Guo et al. 2017]. Mit anderen Worten, das Ziel der Vertrauenskalibrierung ist es, die bestmögliche Konsistenz bei der Vorhersage von Vertrauen und Genauigkeit zu erreichen. Wenn beispielsweise das Vertrauen eines Bildes zu 90% resultiert, sollte die Genauigkeit dieses Bildes ebenfalls 90% betragen. [Guo et al. 2017] haben festgestellt, dass moderne Netzwerke dazu neigen, sich bei der Vorhersage von Vertraulichkeiten zu sehr zu trauen. Der Grund für das Übervertrauen moderner Netzwerke ist die erhöhte Netzwerkkapazität, die Verwendung der Batch-Normalisierung und der Gewichtsabfall. Eine Metrik, die angibt, wie gut das Netzwerk kalibriert ist, ist der Expected Calibration Error (ECE). Die ECE-Metrik wurde unseres Wissens bisher nur für die Bildklassifizierung verwendet. Im Gegensatz zur Bildklassifikation berechnen wir bei der semantischen Segmentierung die Lücke zwischen acc und conf nicht pro Bild, sondern pro Pixel. Diese Änderung erfordert eine zusätzliche Schleife über alle Bilder, die den ECE mitteln.

Zur Kalibrierung der DNN verwenden wir eine Temperaturskalierung, die aus einem einzigen Wert besteht, der der Softmax-Schicht hinzugefügt wird. Es wurde in [Guo et al. 2017] festgestellt, dass diese Art der Kalibrierung die einfachste und gleichzeitig effektivste ist. Die Erweiterung sieht eine Teilung der Eingabe der Softmax-Schicht z mit einem skalaren T vor. Der optimale Temperaturskalierungsparameter wurde durch Grid Search auf unserem Validierungsdatensatz ermittelt. Dadurch konnten wir den ECE von 1,9 auf 1,1 reduzieren. Relevante Referenzwerte aus der Literatur konnten nicht gefunden werden. Ein direkter Vergleich mit den ECE-Werten aus der Aufgabe der Bildklassifikation ist nicht möglich, da die Berechnung unterschiedlich ist.

Stand der Arbeiten (30.06.2020):

E3.4.1

Für die Untersuchung der Interpretierbarkeit von Merkmalskarten wurde die ENet Architekur verwendet, ein ResNet-basiertes Netzwerk, das eine sehr geringe Laufzeit hat und dabei eine respektable Qualität bietet. Das Modell wurde auf einem internen

Fischaugen-Trainingsdatensatz mit 17 Klassen trainiert (semantischer Segmentierung), da der Datensatz aus TP2 noch zu wenige Bilder aufweist. Nach der letzten Schicht beträgt die Anzahl der Feature-Maps 17.

Bei der Untersuchung der intermediären Merkmalskarten wurde eine starke zeitliche Abhängigkeit festgestellt. D.h. die Werte der Merkmalskarten ändert sich von Frame zu Frame nur geringfügig und ist in den meisten Fällen für den Menschen interpretierbar, d.h. das Aussehen der Merkmalskarten entspricht zu Teilen, der resultierenden Merkmalskarten am Ausgang des Netzwerkes.

Basierend auf dieser Erkenntnis wurde der ein Ansatz zur Nutzung von zeitlicher Konsistenz entwickelt, die im Folgenden beschrieben wird:

Die gesamte Pipeline unseres Ansatzes ist für einen einzigen Zeitschritt t0 dargestellt:

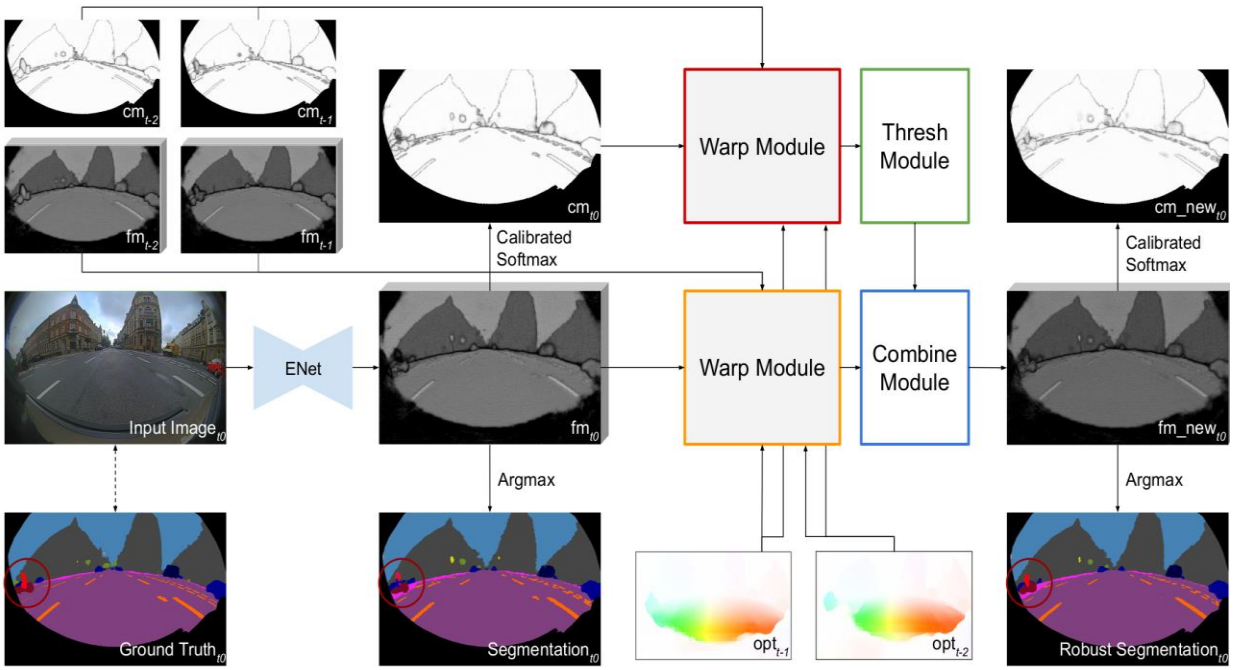


Abbildung 33: Pipeline des im Text dargestellten Ansatzes

Die Konfidenzkarte des aktuellen Zeitschritts cm_t0 wird durch die Wahrscheinlichkeiten aus den Softmax-Verteilungen bestimmt. Um zuverlässige Konfidenzwerte zu erhalten, werden die Konfidenzwerte durch Modifizierung der Softmax-Schicht kalibriert (siehe Stand 30.06.2020). Außerdem verglichen wir diese Konfidenzkarten mit der epistemischen Unsicherheit, die durch Monte Carlo Dropout erhalten wurde. Es hat sich herausgestellt, dass der Unterschied in der Regel ziemlich gering ist, sodass wir die Softmax als eine laufzeiteffiziente Alternative betrachten.

Die Vertrauens- und Merkmalskarten werden in einem sogenannten “Warp-Modul” verzerrt (siehe Kasten mit rotem Rand). Die Funktion des “Warp-Moduls” besteht darin, die Feature- oder Konfidenz-Maps von vergangenen Zeitschritten in den aktuellen Zeitschritt zu „Warpen“, um eine ausgerichtete Darstellung zu erhalten. Für das Warpen verwenden wir den optischen Fluss, den wir mit FlowNet2 erstellen.

Diese ausgerichteten Konfidenzkarten werden im sogenannten “Thresh-Modul (siehe Kasten mit grünem Rand) mit Schwellenwerten und einer Gewichtung verarbeitet.

Die resultierenden Konfidenzkarten können als eine Maske betrachtet werden, die für die Multiplikation mit den Merkmalskarten im Combine-Modul (siehe Kasten mit blauem Rand) verwendet wird. Im “Combine-Modul” werden die Feature-Maps aus dem “Warp-Modul” mit den Schwellenwert-Vertrauenskarten aus dem “Thresh-Modul” multipliziert. Die Ausgabe des “Combine-Modul” sind 17 Feature-Maps, die Pixel für Pixel aus den Feature-Maps der Zeitschritte t_0 bis t_-n zusammengesetzt werden. Die neue Konfidenzkarte heißt cm_new_t0 und die robuste semantische Segmentierung RobustSegmentation_t0.

Mit dieser Vorgehensweise ist es möglich eine semantische Videosegmentierung zu erreichen, um die Konsistenz in Videodaten zu nutzen und die Vorhersage wesentlich robuster zu machen. Im Hinblick auf plötzlich auftretende Störungen in den Eingabedaten kann unser Ansatz die Robustheit der Vorhersage drastisch erhöhen.

Stand der Arbeiten (31.12.2020):

Aufgrund der Verschiebung von Aufwänden nach AP3.6 (siehe “Änderung der Zielsetzung” im letzten Zwischenbericht) wurden die Arbeiten im Rahmen des AP3.4 beendet. Es bleibt die Weiterführung der Leitungsfunktion des AP3.4. Hierzu gehören wöchentliche Telkos und Begutachtung der Partner Beiträge (Dokumentation und Code).

Stand der Arbeiten (30.06.2021):

Aufgrund der Verschiebung von Aufwänden nach AP3.6 (siehe “Änderung der Zielsetzung” im letzten Zwischenbericht) wurden die Arbeiten im Rahmen des AP3.4 beendet. Es bleibt die Weiterführung der Leitungsfunktion des AP3.4. Hierzu gehören wöchentliche Telkos und Begutachtung der Partner Beiträge (Dokumentation und Code).

Stand der Arbeiten (31.12.2021):

Aufgrund der Verschiebung von Aufwänden nach AP3.6 (siehe “Änderung der Zielsetzung” im letzten Zwischenbericht) wurden die Arbeiten im Rahmen des AP3.4 beendet. Es bleibt die Weiterführung der Leitungsfunktion des AP3.4. Hierzu gehören wöchentliche Telkos und Begutachtung der Partner Beiträge (Dokumentation und Code).

Stand der Arbeiten (30.06.2022):

So wie in AP3.3 wurden die durchgeführten Arbeiten für das finale Release 5 aufbereitet und veröffentlicht. Hierzu wurden eine Vielzahl von Release Requirements erfüllt, die z.B. das Ausfüllen des Mechanismen-Katalogs sowie das Erstellen einer Mechanismen Beschreibung, welche in 4 Blöcke unterteilt wurde, beinhalten. Weiterhin wurde die Rolle des AP Leads wahrgenommen und für den Release 5 sämtliche Requirements der Partner überprüft und gegebenenfalls nachgefordert.

AP3.5 Externe Methoden und Maßnahmen (17 PM)

Aufgabe Valeo:

Implementierung von Basis-Methoden und Basis-Maßnahmen für die Ableitung von KPI Bewertungen (E3.5.1)

• Implementierung von Rauschverfahren in Absprache mit den weiteren Partnern im Arbeitspaket, insbesondere beispielsweise Bilddaten (Sensor-Rauschen vom CCD-Sensor: Hintergrundrauschen, Photonenrauschen, Diskretisierung) sowie LIDAR-Daten (Blockade durch Veränderung von Intensitätswerten und Entfernung von Punkten in der Punktwolke in den Eingangsdaten).

Ausführbare komplexer Methoden und Maßnahmen für die Ableitung v. KPI Bewertungen (E3.5.4)

• Plausibilisierung- und Robustheitsprüfung des KI-Ausgangs durch Veränderung der Eingangsdaten (Blackbox-Betrachtung) für Eingabedaten (z.B. Bild, LiDAR, Absicht).

Gefundene Corner Cases sowie Verfahren zur Identifikation von Corner Cases (E3.5.5)

• Corner Case Detektion mit DeepXplore (Kamera und LIDAR): Mehrere Netzwerke parallel ausführen und Eingabedaten variieren, mit dem Ziel den Unterschied zwischen den Ausgaben der Netzwerke zu maximieren. Wenn sich die Ausgaben der Netzwerke unterscheiden, kann dies als Maß für die Erkennung eines Corner Case betrachtet werden. Die gefundenen Corner Cases werden an AP3.3 geleitet, welche diese Daten für ein Finetuning nutzen und die neu trainierten Parameter wieder an AP3.5 zurückgeben. Hierfür sind zwei Iterationsschleifen vorgesehen.

Methoden und Maßnahmen für die Ableitung von KPI-Bewertungen (E3.5.6)

• Plausibilisierung- und Robustheitsprüfung des KI-Ausgangs durch Veränderung der Eingangsdaten (Blackbox-Betrachtung) für Kombination von Bild- und Entfernungsdaten.

Stand der Arbeiten (31.12.2019):

• **(E3.5.1)** Es wurden Verfahren zur Addition von Rauschen und Nebel auf Bilddaten implementiert. Erste Ergebnisse sind in den folgenden Bildern zu sehen

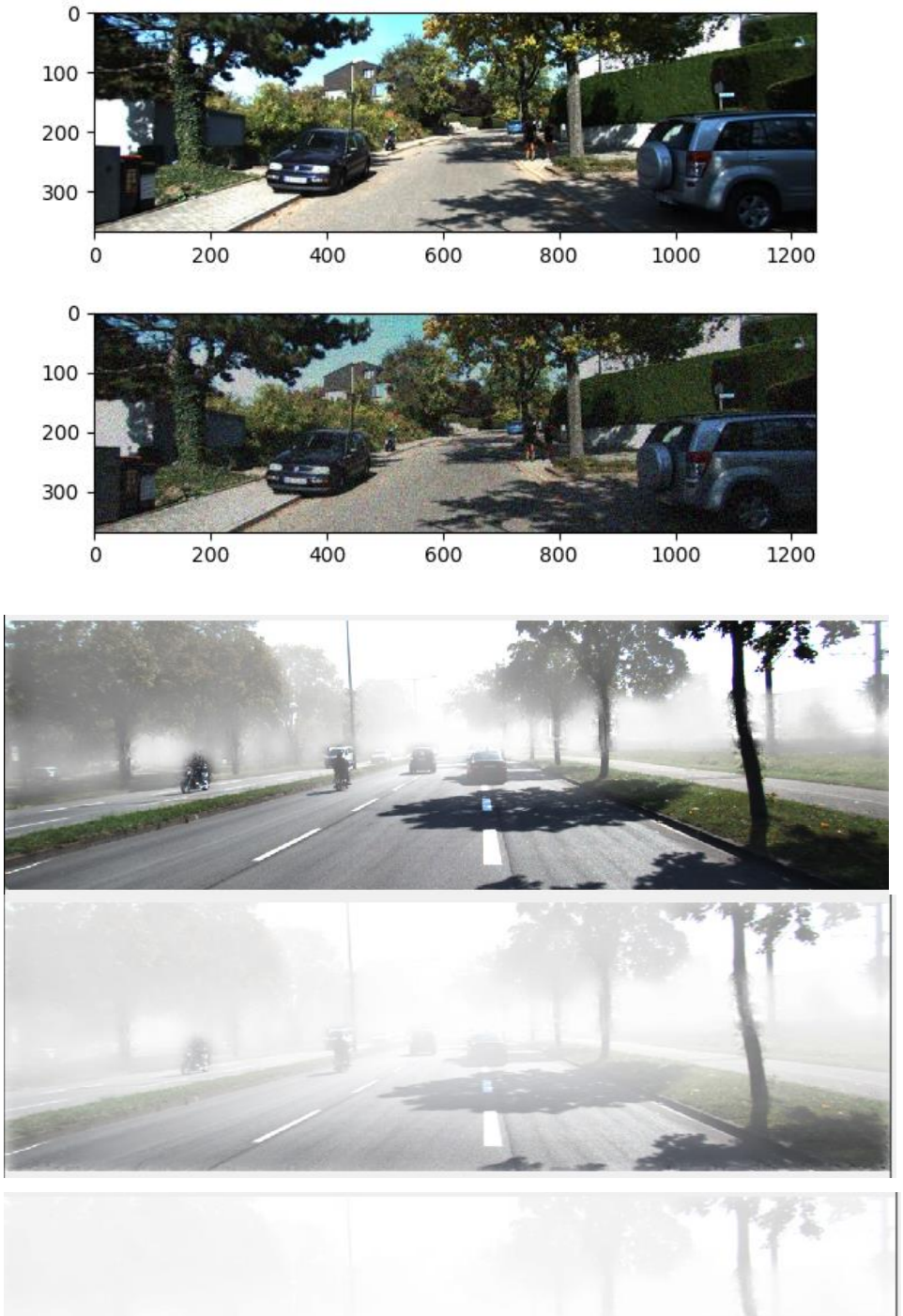


Abbildung 34: Illustration von “salt and pepper” noise sowie der Generierung von Nebel anhand von KITTI Daten

• **(E3.5.4)** Es werden aktuell Konzepte zur Plausibilisierung- und Robustheitsprüfung erarbeitet

• **(E3.5.5)** Arbeiten haben noch nicht begonnen. Laut Plan beginnen die diese in diesem AP im Monat 12

• **(E3.5.6)** Arbeiten haben noch nicht begonnen. Laut Plan beginnen die diese in diesem AP im Monat 33

Stand der Arbeiten (30.06.2020):

• **(E3.5.1)** Die Verfahren zur Addition von Rauschen und Nebel auf Bilddaten wurden optimiert und ein Regelmeeting zum Thema Cluster "Data augmentation techniques" ins Leben gerufen, um den Austausch zwischen den Partnern zu fördern. Aufgrund der aktuellen Lage mit dem Coronavirus sind weitere Arbeiten aufgrund von Kurzarbeit nicht möglich gewesen.

• **(E3.5.4)** Es wurde mit der Erarbeitung von Konzepten zur Plausibilisierung- und Robustheitsprüfung begonnen. Aufgrund der aktuellen Lage mit dem Coronavirus und Kurzarbeit sind weitere Arbeiten nicht möglich gewesen.

• **(E3.5.5)** Laut Plan beginnen diese Arbeiten im Monat 12. Aufgrund der aktuellen Lage mit dem Coronavirus und Kurzarbeit konnte mit diesen Arbeiten nicht begonnen werden.

• **(E3.5.6)** Arbeiten haben noch nicht begonnen. Laut Plan beginnen die diese in diesem AP im Monat 33

Stand der Arbeiten (31.12.2021):

Um robuste Objekterkennungsalgorithmen zu entwickeln, ist es wichtig, sie mit sogenannten “Corner Cases” zu trainieren. Das bedeutet, dass Trainingsdaten für Situationen gefunden werden müssen, die im bisherigen Training nicht abgedeckt wurden. Da die Erkennung einer fehlenden Situation im Trainingsdatensatz aufgrund ihrer Vielzahl nahezu unmöglich ist, bietet sich die Variante an, “Corner Cases” während der Datenerfassung unter Verwendung von Objekterkennungsmodellen zu erkennen. Um dies zu erreichen, sollen die aufgenommenen Daten mit einem Bilderkennungsalgorithmus sowie einer Objekterkennung auf Punktwolken verarbeitet werden. Die Sensorfusion vergleicht dann die folgenden Ergebnisse. Weichen sie stark voneinander ab, bedeutet dies, dass für diese spezielle Situation ein Training erforderlich ist. Diese Szene wird also als "Corner Case" markiert.

Zunächst sollte eine gute Objekterkennung auf Kamerabildern implementiert werden. Zu diesem Zweck wurden Instanzsegmentierungs-Algorithmen ausgewählt, trainiert und getestet. Da es sich um einen Nachbearbeitungsschritt handelt und die Bearbeitungszeit daher nicht so streng ist wie bei einer Live-Anwendung, wurde der "Swin-Transformer" ausgewählt, da er der derzeit beste Erkennungsalgorithmus ist. Dieser Algorithmus wurde dann mit verschiedenen Straßendatensätzen trainiert und anschließend mit Bildern des aufnehmenden Fahrzeugs perfektioniert. Die Erkennung wurde auf Autos, Lastwagen, Busse, Fußgänger, Fahrräder und Motorräder

beschränkt. Außerdem wurden entfernte Objekte aus den Datensätzen entfernt, so dass nur nahe Objekte erkannt werden.

Abbildung 35: Im Bild erkannte Objekte

Die Ergebnisse werden verwendet, um die Erkennung anhand von Kameradaten mit der Erkennung anhand von LIDAR-Daten zu vergleichen. Parallel zum "Swin-Transformer" wird "PointPillars" verwendet, um die Autos rund um das Fahrzeug anhand von LIDAR-Punktwolken zu erkennen. Die daraus resultierenden 3D-Boxen werden dann auf die Kamerabilder projiziert, so dass ein Vergleich durchgeführt werden kann. Da der verwendete PointPillars-Algorithmus nur Fahrzeuge erkennt und diese häufig vorkommende Klasse für den Vergleich völlig ausreichend ist, werden nur die Autos aus den Ergebnissen des "Swin-Transformers" herausgehalten.

Abbildung 36: Im Bild erkannte Objekte und 3D-Bounding-Boxen

Die projizierten LIDAR-erfassten Objekte werden dann gefiltert, um einen optimalen Vergleich zu ermöglichen. Zum einen werden entfernte Objekte weggelassen, um den gleichen Umgebungsradius des Fahrzeugs zu erfassen. Zum anderen wird der Winkel des Objekts zur Kamerarichtung berechnet. Objekte jenseits 45 Grad werden entfernt, da sie sich nicht im Sichtfeld der Kamera befinden und somit vom "Swin-Transformer" nicht erkannt werden können.

Dann werden Paare gebildet, die die nächstgelegenen LIDAR-erfassten Fahrzeuge mit den kameragefundenen Fahrzeugen verbinden. Dies geschieht mithilfe der “Ungarische Methode”, die die optimalen Paare unter Berücksichtigung aller Entfernungen zwischen den Objekten ermittelt.

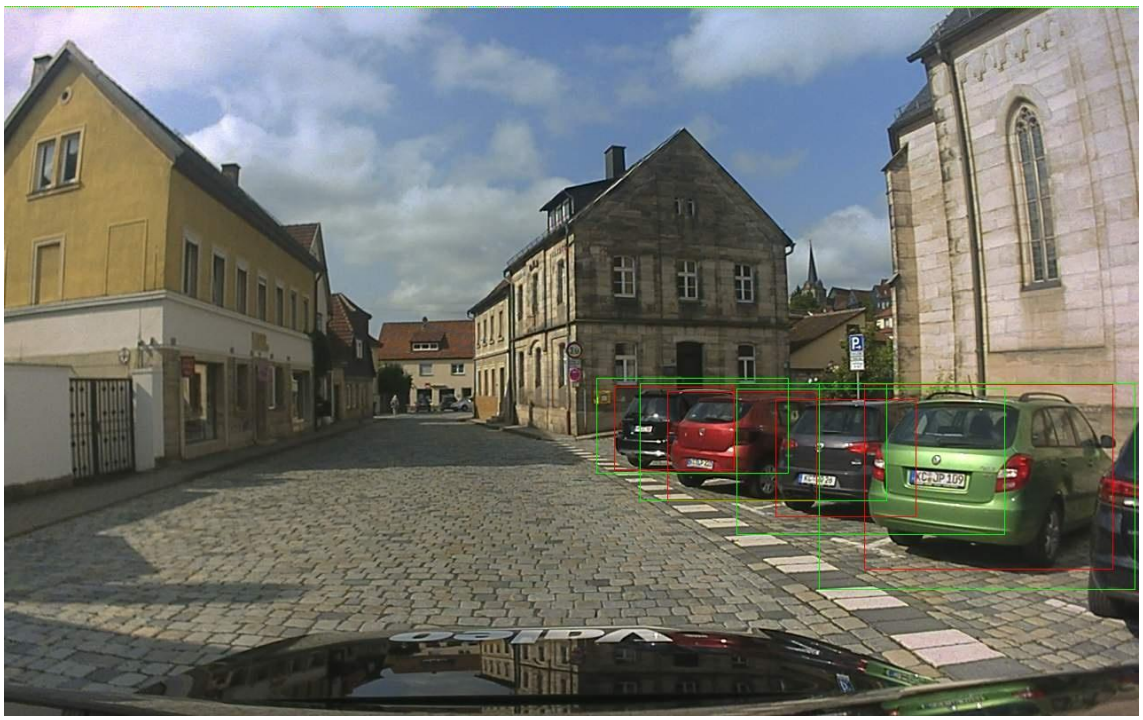


Abbildung 37: Projizierte LIDAR-Erkennung, Kameraerkennung

Nachdem die “Ungarische Methode” die optimalen Paare gebildet hat, werden die verbleibenden Boxen pro Erkennung gezählt. So erhalten wir für jedes Bild die Anzahl der Falscherkennungen, d. h. der Objekte, die nur von einem Sensor erkannt wurden. Außerdem wird für jedes Paar der Abdeckungsgrad der Boxen berechnet. Von diesem Prozentsatz hängt die Einstufung des Paares in eine von drei Kategorien ab: "Gute Übereinstimmung", "Schlechte Übereinstimmung", "Keine Übereinstimmung". Ab 50% gehört ein Paar in die erste Kategorie, unter 50% in die zweite und bei 0% in die dritte. So können wir die folgenden Informationen bereitstellen und die Frame mit ihnen versehen: Die Differenz der erkannten Objekte zwischen beiden Sensoren und die Anzahl der Paare in jeder Kategorie.

Durch diese Informationen: zum einen die zusätzlich erkannten Objekte im Vergleich zum anderen Sensor und zum anderen die Übereinstimmung der erkannten Boxen, kann in einem weiteren Schritt je nach Voraussetzungen bestimmt werden, ob es sich bei der Szene um einen “Corner Case” handelt. Da sowohl Aufzeichnungsprobleme als auch Erkennungsprobleme auftreten können, werden die oben genannten Werte nicht sofort ausgewertet, sondern es wird ein Mittelwert aus den Werten der nahegelegenen Frames berechnet. Die Anzahl der Frames, die in die Durchschnittsberechnung einbezogen werden, hängt von der Geschwindigkeit ab. Je schneller sich das Fahrzeug bewegt, desto weniger Frames werden für die Durchschnittsberechnung verwendet.

Stand der Arbeiten (30.06.2022):

So wie in AP3.3 und AP3.4 wurden die durchgeführten Arbeiten für das finale Release 5 aufbereitet und veröffentlicht. Hierzu wurde eine Vielzahl von Release Requirements

erfüllt, die z.B. das Ausfüllen des Mechanismen-Katalogs sowie das Erstellen einer Mechanismen Beschreibung, welche in 4 Blöcke unterteilt wurde, beinhalten.

AP3.6 Aggregierte Methoden (6 + 17 PM)

Aufgaben Valeo:

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

• Auswahl von Heatmapping-Verfahren die sich für eine Kombination eignen.

• Training eines Metaklassifikators basierend auf die kombinierten Heatmaps zur Entwicklung einer Metrik (Absicherungs-KPI) • Bewertung hinsichtlich der in AP3.2 definierten KPIs.

Stand der Arbeiten (31.12.2020):

Einleitung:

Für ein DNN wird die Operational Design Domain (ODD) durch die zugrundeliegende Datenverteilung der Trainingsdaten definiert. Wenn die Testdaten aus der gleichen Verteilung gezogen werden, ist es wahrscheinlich, dass das DNN in Bezug auf die Genauigkeit recht gut abschneidet. Für Daten außerhalb der ODD ist die DNN-Genauigkeit in der Regel deutlich reduziert. Das DNN ist nicht in der Lage, über seine Trainingsdatenverteilung hinaus zu generalisieren. Neben der Steigerung der Generalisierungsfähigkeit ist die Erkennung der ODD zur Laufzeit ein wichtiger Mechanismus. Ziel ist es, zur Laufzeit erkennen zu können, ob die Eingabedaten in ihrer Verteilung den Trainingsdaten ähnlich sind und somit als "in-domain" klassifiziert werden können oder ob sie sich stark unterscheiden und somit als "out-of-ODD" klassifiziert werden müssen. Eine korrekte Klassifizierung von in-ODD und out-of-ODD kann als Maß für die Unsicherheit oder das Vertrauen in die DNN-Ausgabe dienen. Wenn die Eingabedaten außerhalb der ODD liegen, kann die Ausgabe des DNN nicht mehr als zuverlässig eingestuft werden. Mögliche Maßnahmen sind in diesem Fall die Verwendung von Redundanzzweigen, die aufgrund unterschiedlicher Trainingsdaten oder Sensormodalitäten eine andere ODD haben, oder der Übergang in einen Notfallmodus, der die Anwendung so schnell wie möglich beendet.

Ansatz:

Wir haben DeepLabV3+ als Baseline für die semantische Segmentierung verwendet und die Architektur um einen zweiten Decoder (siehe grüne Schichten in der Abbildung) erweitert, der lernt, das Eingabebild zu rekonstruieren. Beim Training des zweiten Decoders werden alle lernbaren Parameter von DeepLabV3+ eingefroren, so dass die Genauigkeit der semantischen Segmentierung in keiner Weise beeinträchtigt wird. Die Rekonstruktion wurde mit Hilfe des Mean Scare Error (MSE) und des Kullback Leibler Divergence (KLD) Verlustes gelernt. Der KLD-Verlust wird

üblicherweise zwischen Encoder und Decoder gemessen. Da die lernbaren Parameter des Encoders eingefroren sind und somit keine Auswirkung des KLD-Verlustes bestehen würde, wird eine weitere Faltungsschicht vor den zweiten Decoder gelegt (siehe gelbe Schicht). Beide Verluste werden addiert (Verlust = 0,1 * KLD + 1 * MSE). Bei einer Eingangsbildgröße von 3x768x1280 px beträgt die Feature-Map-Größe nach dem Encoder 256x48x80. Dividiert man beide Werte, so ergibt sich ein Kompressionsfaktor von 0,33.

Die Annahme ist, dass der zweite Decoder nur bestimmte Merkmale lernt, so dass ein Bild rekonstruiert werden kann, das der Trainingsdatenverteilung ähnlich ist. Bei abweichenden Eingabebildern ist die Rekonstruktion schwieriger. Das PSNR ist ein Maß dafür, wie gut die Rekonstruktion ist. Ein hohes PSNR bedeutet einen niedrigen Rekonstruktionsfehler und ein niedriges PSNR bedeutet einen hohen Rekonstruktionsfehler. Durch die Messung des PSNR zur Laufzeit kann überwacht werden, ob das Eingangsbild innerhalb oder außerhalb der ODD liegt, d. h. über oder unter einem bestimmten Schwellenwert.

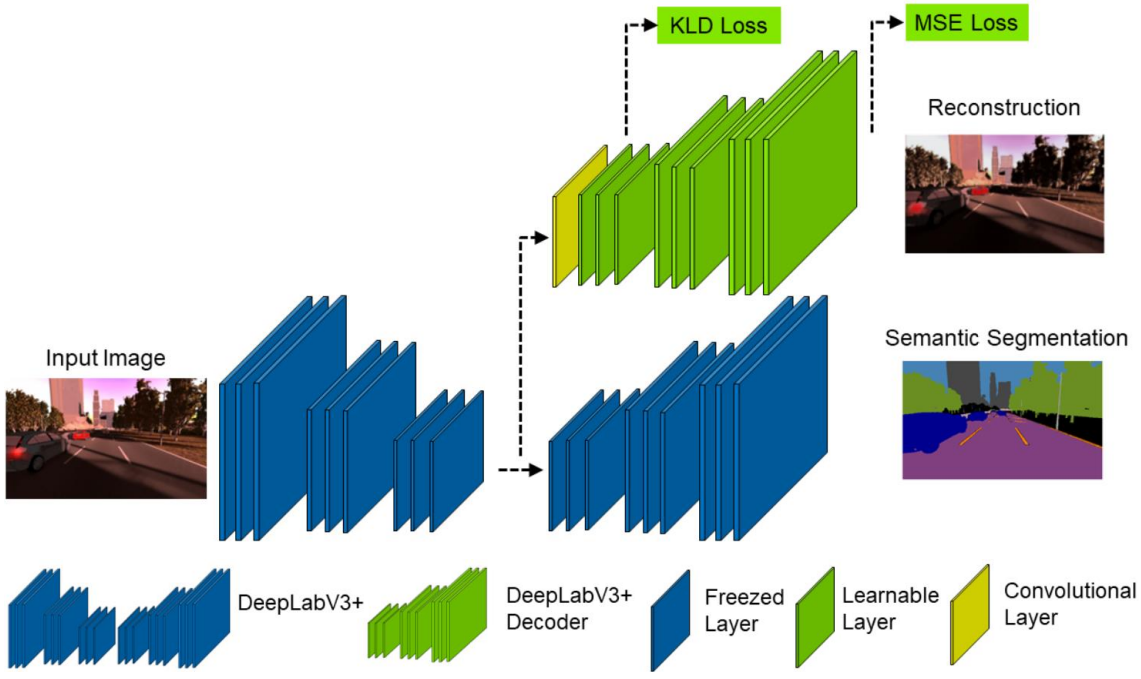


Abbildung 38: Entwickelter Ansatz zur OoD-Erkennung

Dataset:

Aufgrund der fehlenden Vielfalt in den Projektdaten in Bezug auf Wetter- und Lichtbedingungen, die die Realisierung verschiedener Domänen ermöglichen würden, wurde der Synthia-Sequenzdatensatz verwendet. Er enthält z. B. die Domänen Dämmerung, Nebel, Regen, Winter, Sommer, Frühling, Herbst, Nacht usw. Zum Trainieren wurde die Sequenz (SEQ) 1+2 und zum Testen 5+6 verwendet. Die Auflösung des Synthia beträgt 768, 1280 Pixel. Das Training wurde mit der Domain "dawn" (SEQ 1+2) durchgeführt, d.h. die Domain "dawn" ist in-domain und alle anderen Domains sind out-of-domain. Zur Validierung mit In-Domain-Daten wurde die

SEQ 5+6 (dawn) verwendet. Die Testbilder (Out-of-Domain-Daten) wurden aus anderen Domänen als der Morgendämmerung von SEQ 5+6 verwendet. Das Training unseres Ansatzes wurde in 15 Epochen durchgeführt.

Training data, 2392 images:

- Dawn (SEQ1+2)

Test data (in ODD), 1775 images:

- Dawn (SEQ5+6)

Test data (out-of-ODD), 13921 images:

- Winter (SEQ5+6)
- Sunset (SEQ5+6)
- Summer (SEQ5+6)
- Spring (SEQ5+6)
- Night (SEQ5+6)
- Fog (SEQ5+6)
- Rain (SEQ5)
- Rainnight (SEQ5)
- Softrain (SEQ5)
- Winternight (SEQ5+6)

Ergebnisse:

Gemessen wurde der PSNR für In-Domain- und Out-of-Domain-Daten. Die Einteilung der Daten in In-Domain und Out-of-Domain kann im vorherigen Kapitel "Datensatz" nachgelesen werden. Das Ergebnis ist im ersten Histogramm dargestellt. Die x-Achse stellt den PSNR dar und ist in 50 Bins unterteilt. Die y-Achse stellt die Häufigkeit dar, also die Anzahl der ausgewerteten Bilder. Die Auswertung zeigt, dass die In-Domain-Daten hauptsächlich zwischen 18 und 22 dB liegen. Die Out-of-Domain-Daten liegen hauptsächlich zwischen 13 und 20 dB. Zwischen 18 und 20 dB ist eine deutliche Überlappung zwischen In-Domain- und Out-Of-Domain-Daten zu erkennen. Eine klare Trennung zwischen In-Domain- und Out-of-Domain-Daten ist also in einer Einzelbildanalyse nicht ohne weiteres möglich. Aus diesem Grund wurde ein neuer Wert tau eingeführt, der angibt, wie viele Bilder zu einer Sequenz zusammengefasst werden, bevor ein einzelner PSNR-Wert ermittelt wird. D.h. es wird der durchschnittliche PSNR über die Sequenz mit der Länge tau ermittelt. tau ist 10 für das zweite Histogramm, 50 für das dritte und 100 für das vierte. Diese Mittelung der Werte verkleinert die Varianz der einzelnen Bereiche, da Ausreißer am Rand des Spektrums geglättet werden. Je höher das Tau ist, desto geringer ist die Varianz. Eine vollständige Trennung von In- und Out-of-Domain-Daten ist bereits mit tau = 50 möglich. Da die Daten mit einer Frequenz von 5 Hz erzeugt wurden, bedeutet dies,

dass zur Laufzeit alle 10 Sekunden eine zuverlässige Abschätzung vorgenommen werden kann, ob die Eingangsbilder in oder out-of-domain sind. Bei einer höheren Bildrate kann wahrscheinlich schon nach kürzerer Zeit eine zuverlässige Schätzung vorgenommen werden. Für den Fall, dass eine kürzere Zeit benötigt wird, z. B. 2 Sekunden wie im Fall tau = 10, kann der Aussage in- oder out-of-ODD eine Wahrscheinlichkeit hinzugefügt werden, so dass im konkreten Fall für tau=10 der Bereich des PSNR zwischen 18 und 19 mit einer Wahrscheinlichkeit überlagert würde. Liegt der Wert näher an 18 dB, ist es wahrscheinlicher, dass es sich um Out-of-Domain handelt, liegt der Wert näher an 19, ist es wahrscheinlicher, dass es sich um In-Domain handelt.

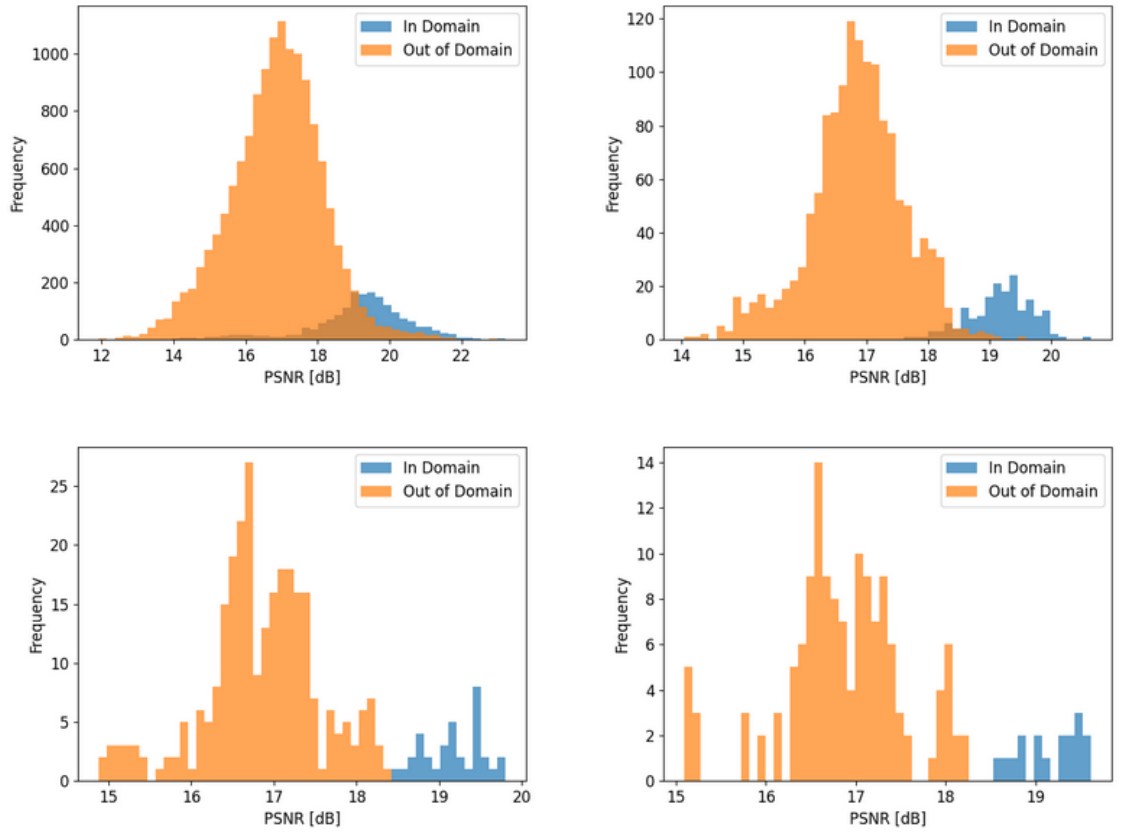


Abbildung 39: Ergebnisse zur In- und Out of Domain Erkennung

Zu diesem Ansatz wurde ebenfalls ein “Short Paper” geschrieben mit dem Titel: Online Out-of-Domain Detection for Automated Driving

Dieses wurde auf dem Machine Learning in Certified Systems Workshop (https://mlcertifiedsystems.deel.ai/) eingereicht und angenommen.

Stand der Arbeiten (30.06.2021):

Die Arbeit wird in Kooperation mit Opel (Ahmed) durchgeführt und beinhaltet die Kombination von Deep Ensemble, aleatorischer und epistemischer Unsicherheitsmodellierung.

Das Ziel ist es, ein tiefes Ensemble für die Aufgabe der semantischen Segmentierung zu erstellen, in dem die Mitglieder unterschiedliche Fehlermodi haben. Durch die unterschiedlichen Fehlermodi sollen sich die Mitglieder gegenseitig ergänzen, so dass schließlich eine höhere Genauigkeit erreicht wird. Um einen möglichst unterschiedlichen Fehlermodus unter den Mitgliedern zu realisieren, wird eine Unsicherheitsmodellierung (aleatorisch und epistemisch) in den Trainingsprozess des tiefen Ensembles integriert. Zur Laufzeit werden die Ausgaben der Deep-Ensemble-Mitglieder auf Feature-Map-Ebene mittels Unsicherheitsmodellierung fusioniert. Für die Durchführung der Experimente werden das Deeplabv3+ für semantische Segmentierung und der BDD100k-Datensatz verwendet.

Trainingsprozess:

1.DNN A (bestehend aus Deeplabv3+ und Unsicherheitsmodellierung) wird auf dem Train Set 1 trainiert. Die Unsicherheitsmodellierung kann aus aleatorischem oder MC-Dropout oder Deep Ensemble oder Rekonstruktionsfehler oder einer Mischung aus den vorgenannten bestehen.

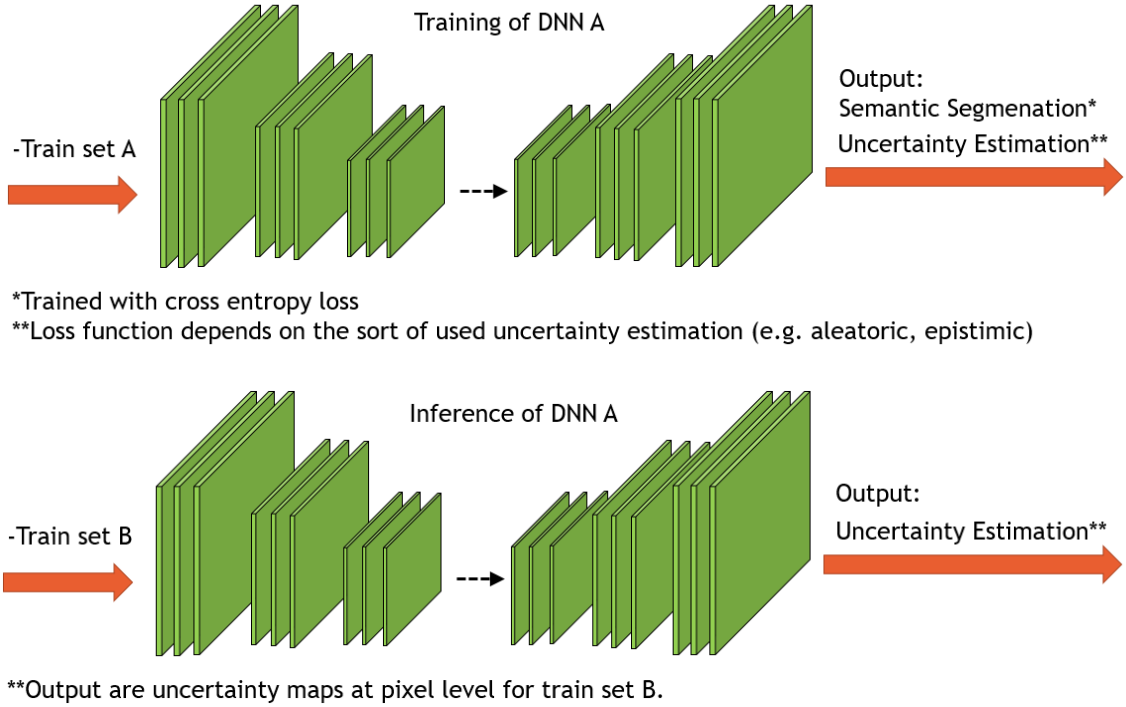


Abbildung 40: Darstellung des Trainingsprozesses

3. DNN B (bestehend aus Deeplabv3+) wird auf dem Train-Set 2 zusammen mit den generierten Unsicherheitskarten aus Schritt 2 trainiert. Der Cross-Entropie-Verlust wird mit der Unschärfekarte multipliziert, so dass der Verlust für große Unschärfewerte größer wird und umgekehrt (ähnliche Idee wie beim fokalen Verlust).

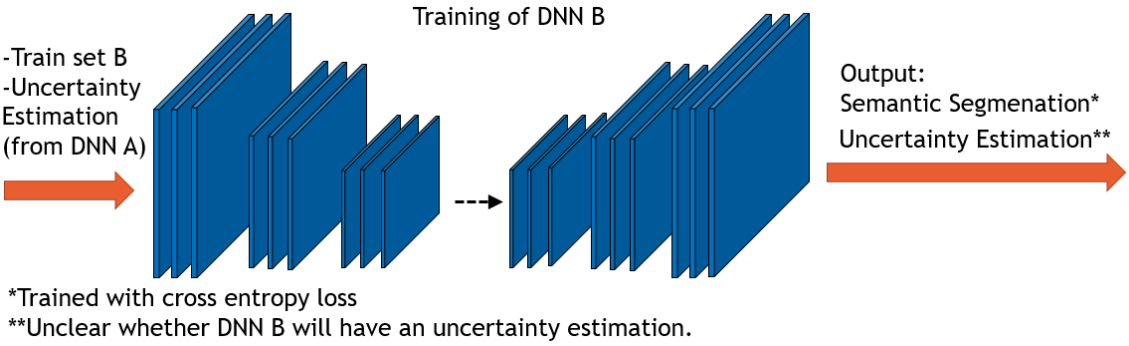


Abbildung 41: Prinzipielle Darstellung der Unsicherheit

$$CE(p_t) = -\log(p_t) * alpha * (e^{uncertainty} - 1)$$

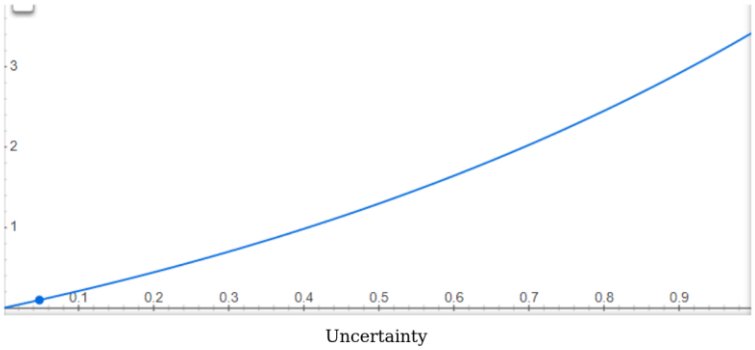


Abbildung 42: Grafische Darstellung der Unsicherheit

Inferenz:

DNN A und DNN B werden parallel ausgeführt. Bereiche, in denen DNN A eine hohe Unsicherheit aufweist, werden mit den Ausgaben von DNN B fusioniert. Die Fusion findet auf Feature-Map-Ebene vor dem Argmax-Layer statt.

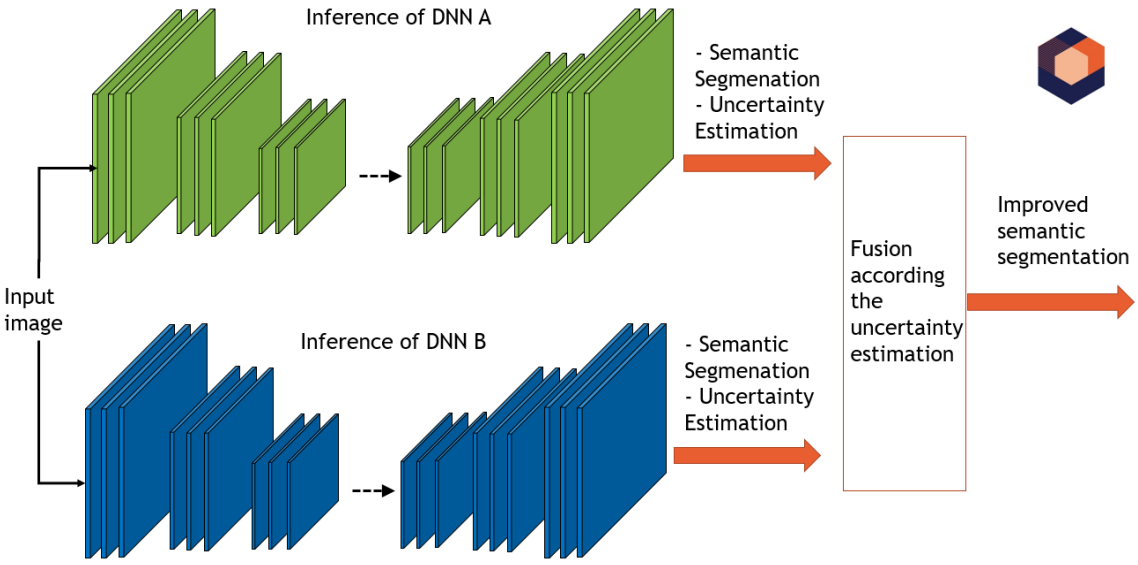


Abbildung 43: Inferenz von DNN A und B

Stand der Arbeiten (31.12.2021):

Während dieser Berichtsphase arbeitete Valeo an der Konzeption eines Fallback System, das auf der Blockierung von Sensoren und der Fusion von 3D-Erkennungen basiert. Die Verfügbarkeit mehrerer Sensoren im System könnte genutzt werden, um eine 3D-Objekterkennung für den Fall zu erreichen, dass einer der Sensoren blockiert ist oder seine Funktion vollständig eingestellt hat. Die Blockierung eines Sensors kann verschiedene Ursachen haben, wie z. B. Frost, Schnee, Nebel oder Regenspritzer, ohne darauf beschränkt zu sein.

Valeo hat ein konzeptionelles System implementiert, bei dem der LiDAR-Sensor der Hauptsensor für die 3D-Objekterkennung ist. Wir gehen davon aus, dass der Sensor ein Blockage Module enthält, welcher die Blockadeinformationen liefert. Ein 3D-Objekterkennungsalgorithmus nimmt die Punktwolke als Eingabe und gibt 3D-Bounding-Boxen von Fußgängern aus. Die 3D-Parameter des Fußgängers werden mit einem monokularen 3D-Objekterkennungsalgorithmus geschätzt, der nur Kamerabilder verwendet. Wenn die Blockierungsinformationen des Blockierungsmoduls auf eine teilweise Blockierung des LiDAR-Sensors hinweisen, geht das System in einen Fallback-Modus über und die Erkennungen des monokularen Kameramoduls für den blockierten Teil des FOV werden mit den Erkennungen des LiDAR-Sensors für den nicht blockierten Teil in der Punktwolke verschmolzen.

Bei der derzeitigen Implementierung wird davon ausgegangen, dass die Kamera voll funktionsfähig ist und nicht blockiert wird. Ein Modul zur Erkennung von Blockierungen in der Kamera kann verwendet werden, um festzustellen, ob die Kamera für das Fallback-System verfügbar ist oder nicht. Wenn die Kamera ebenfalls blockiert ist, sollte das System idealerweise eine Warnung aussenden, die besagt, dass beide Sensoren blockiert oder ausgefallen sind. Die DNNs für die LiDAR only und die monokulare Kamera-basierte 3D-Objekterkennung wurden mit den BIT-TS-Daten der Tranche 3 und 4 trainiert.

Um die Generalisierungsfähigkeit zu erhöhen, wird eine Fusion der Gewichtsdateien der DNNs durchgeführt. Zu diesem Zweck werden der DeepLabV3+ und der BDD- oder Cityscapes-Datensatz verwendet. Dieser Ansatz basiert auf der Annahme, dass durch die Fusion der Gewichte ein flacheres lokales Minimum in der Verlustlandschaft erreicht wird, das im Allgemeinen bessere Generalisierungsfähigkeiten aufweist als ein vergleichsweise scharfes Minimum. Aktuelle Ansätze (https://arxiv.org/abs/1803.05407, https://arxiv.org/pdf/1806.05594.pdf) verwenden ein strenges Trainingsprotokoll, um die Gewichte zu generieren, die dann in einem Verhältnis von 50 zu 50 fusioniert werden. Unsere Forschung geht darüber hinaus, die Eigenschaften der Gewichte zu beleuchten, die Voraussetzung für eine erfolgreiche Fusion sind, und damit bessere Ergebnisse in Bezug auf Leistung und Kalibrierung zu erzielen als SWA und Vertreter.

Die Leistungsanalyse der Fusion von zwei Gewichtungsdateien mit Alpha und Beta auf der x-Achse ist in der folgenden Abbildung dargestellt.

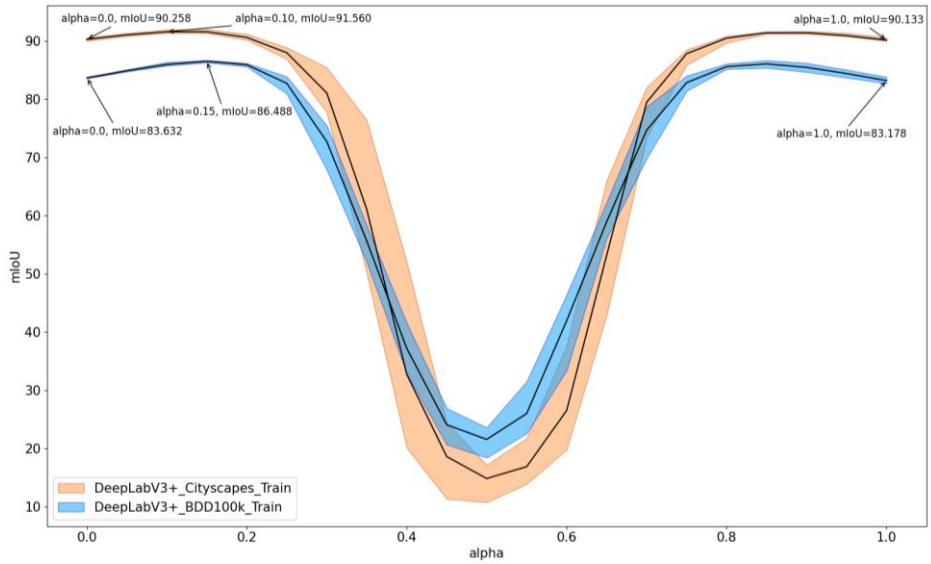


Abbildung 44: Leistungsanalyse der Fusion von zwei Gewichtungsdateien mit Alpha und Beta

Zunächst wurden 4 Gewichtungsdateien durch 4 unabhängige Trainings von Deeplabv3+ auf dem BDD-Datensatz erzeugt. Anschließend wurden 2 Gewichtungsdateien miteinander fusioniert, was zu 6 fusionierten Gewichten führte. Die Fusion basiert auf gewichteter Mittelung und wurde mit dem Parameter alpha oder beta = 1 - alpha in 0,05 Schritten durchgeführt. Die obere Abbildung zeigt die gemittelte Leistung der 6 fusionierten Gewichte in Abhängigkeit vom Parameter alpha bzw. beta. Das Fehlerband zeigt die maximale und minimale Leistung der 6 fusionierten Gewichte im jeweiligen Alpha/Beta-Verhältnis. Es ist zu erkennen, dass ein Alpha von 0,15 oder 0,85 zu einem erhöhten mIoU führt. Im Gegensatz dazu führt eine Fusion mit alpha und beta = 0,5 zu einer starken Reduktion. Die Grafik zeigt zunächst die Ergebnisse für die Trainingsdaten des BDD-Datensatzes. Die folgende Abbildung zeigt sie nun für die Validierungs- und Testdatensätze von BDD bzw. Cityscapes. (Bitte beachten Sie, dass sich die x-Achse geändert hat und die Leistungswerte von alpha und 1-alpha gemittelt wurden, um die statistische Aussagekraft der Ergebnisse zu erhöhen). Es ist zu erkennen, dass die erhöhten mIoU-Werte auf die Validierungs-/Testdatensätze verallgemeinert werden und somit die Gewichtsfusion auf diese verallgemeinert wird.

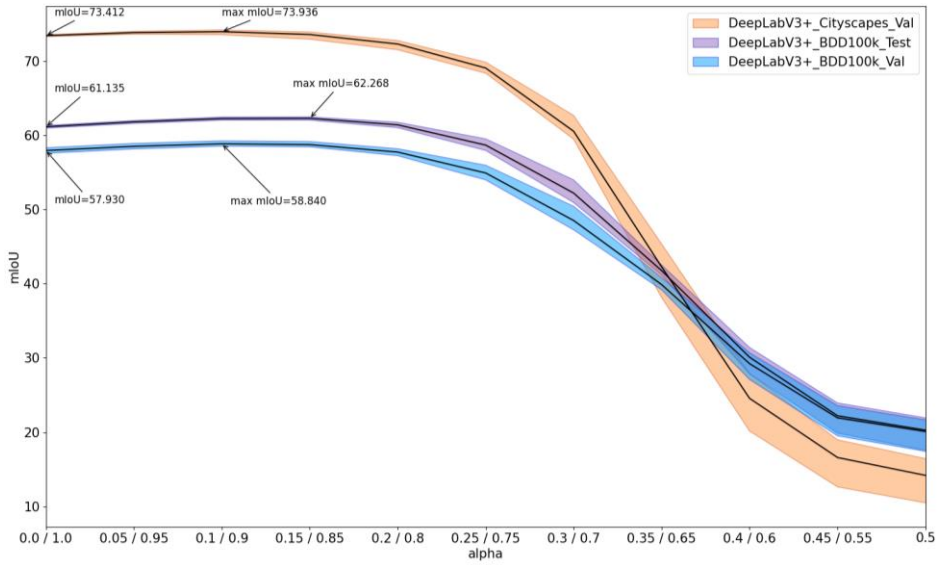


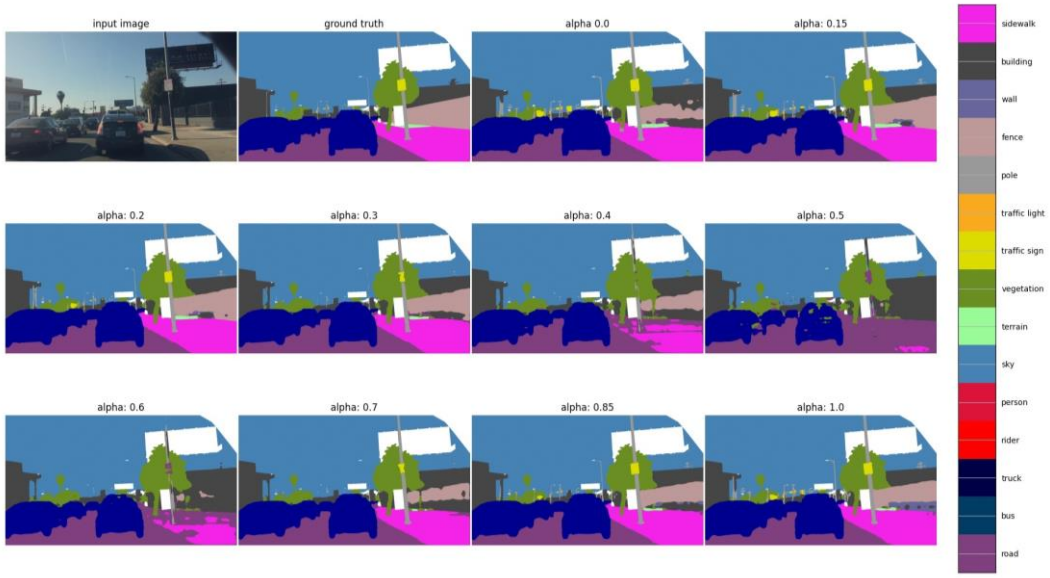
Abbildung 45: Ergebnisse für die Trainingsdaten des BDD-Datensatzes

Die folgende Tabelle wurde für die quantitative Auswertung erstellt. Sie zeigt die durchschnittliche Leistung für DeepLabV3+ auf den BDD-Testdaten (Spalten 1 und 2) und auf den Cityscapes-Validierungsdaten (Spalten 3 und 4). Alpha = 0/1 bezieht sich auf die individuelle Leistung der Gewichte und stellt somit die Basislinie für die Gewichtsfusion dar. Interessanterweise ist ein Anstieg der mIoU-Werte vor allem bei Klassen mit geringer Pixeldichte, wie z. B. Menschen, zu beobachten. Insgesamt kann mIoU um mehr als 1 % für BDD und 0,5 % für Cityscapes gesteigert werden. Beachten Sie, dass es sich hierbei um Durchschnittswerte handelt, um die Aussagekraft der Ergebnisse zu erhöhen, und dass der Parameter alpha für alle fusionierten Gewichte gleich eingestellt wurde (alpha = 0,15), obwohl einzelne Verhältnisse (z. B. alpha = 0,1) bei einigen Fusionen zu besseren Ergebnissen führen.

Tabelle 8:Durchschnittliche Performanz für DeepLabV3+ auf BDD100k Testdaten (Spalten 1 und 2) und City Scapes val Daten

Classes	alpha = 0/1	alpha = 0.15 / 0.85	alpha = 0/1	alpha = 0.1 / 0.9
Road	94.57	94.71	97.70	97.73
Sidewalk	64.89	65.40	82.93	82.92
Building	85.42	85.77	91.14	91.39
Wall	29.27	29.29	52.28	52.72
Fence	51.00	51.14	57.0	56.77
Pole	50.05	**52.91**	59.31	60.09
Traffic light	53.66	**58.33**	64.1	**67.03**
Traffic sign	52.83	**56.29**	73.97	**75.64**
Vegetation	86.37	86.49	91.39	91.56
Terrain	50.61	51.17	62.81	62.95
Sky	95.13	95.32	93.46	94.06
Person	63.63	**65.73**	78.16	**79.41**
Rider	**47.69**	46.36	59.62	60.62
Car	90.31	90.70	93.47	93.48
Truck	57.15	**58.88**	64.61	64.25
Bus	79.11	**81.41**	80.42	80.67
Train	0.0	0.0	60.56	**61.55**
Motorcycle	54.45	**55.87**	58.66	58.86
Bicycle	53.63	**57.30**	73.23	**74.59**
Mean IoU	61.13	**62.27**	73.41	**73.94**

Eine Visualisierung der verschiedenen Abstufungen von Alpha ist unten zu finden. Es ist zu erkennen, dass die Annäherung von alpha an 0,5 zu einem starken Anstieg der Präzision und einem starken Rückgang des Recalls führt. Dies erklärt den verminderten mIoU-Wert bei alpha = 0,5. Im Gegensatz dazu ist bei alpha = 0,15 eine Verbesserung der Segmentierung zu erkennen. Interessanterweise ist alpha = 0,15 nicht einfach eine Verschmelzung der Vorhersagen von alpha = 0 und alpha = 1,0, sondern eine völlig neue Funktion. Dies ist daran zu erkennen, dass z. B. der Zaun in der ersten Abbildung nur bei Alpha = 0,15 korrekt segmentiert wird. Etwas Ähnliches ist in der zweiten Abbildung mit dem Fußgänger und dem Gehweg zu sehen.



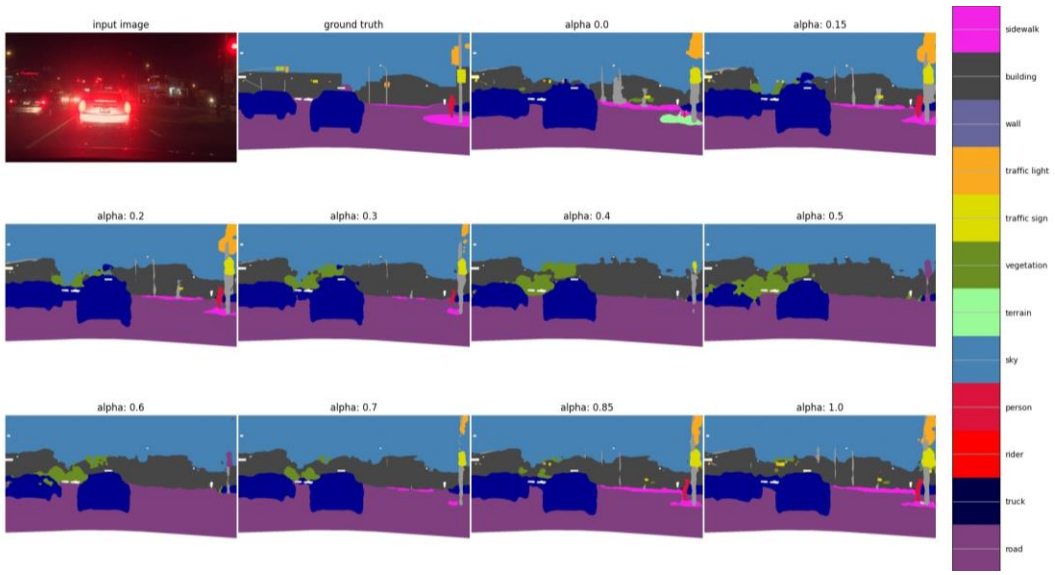


Abbildung 46: Visualisierungen der Segmentierungsqualität mit verschiedenen Alpha-Werten

Ein Vergleich mit einer anderen Basislinie, dem Deep Ensemble mit zwei Mitgliedern, ist weiter unten zu finden. Für das Deep Ensemble wurden die Logits der beiden Mitglieder gemittelt. Die Tabelle zeigt die Leistungswerte, die für das Deep Ensemble (einfache Fusion) und die Gewichtsfusion auf den BDD-Trainings- und Testdaten erzielt wurden. Der mIoU-Wert auf den Testdaten ist vergleichbar, was bedeutet, dass mit der Gewichtsfusion eine ähnliche Leistung erzielt werden kann wie mit dem Deep Ensemble, obwohl die Gewichtsfusion nur die Ausführung eines einzigen DNN erfordert und keinen zusätzlichen Rechenaufwand erfordert. Dies steht im Gegensatz zum Deep Ensemble, das für zwei Mitglieder die doppelte Laufzeit benötigt, was aus praktischen Gründen oft ein KO-Kriterium für den Einsatz in Produkten ist.

Tabelle 9: Einfache Fusion; gewichtete Fusion mit Vielfachen zwischen 0 und 1

	simple fusion		checkpoint fusion	
Checkpoint ID	train data	test data	train data	test data
09_48_00, 17_70_59	84,12	62,11	86.48	61.93
09_48_00, 21_53_59	84.06	62,25	86.62	62.38
09_48_00, 14_50_40	83.61	62.45	86.27	62.31
17_70_59, 21_53_59	**84.18**	**62,04**	86.64	62.18
17_70_59, 14_50_40	83.70	62,13	86.50	61.89
21_53_59, 14_50_40	83.65	62,44	**86.68**	**62.17**

Eine Voraussetzung für die erfolgreiche Fusion der Gewichtsdateien ist eine hohe Kosinusähnlichkeit, die als Ähnlichkeitsmaß der Gewichtsdateien angesehen werden kann. Die restriktiven Trainingsprotokolle z.B. in (https://arxiv.org/abs/1803.05407, https://arxiv.org/pdf/1806.05594.pdf) führen automatisch zu einer hohen Kosinusähnlichkeit. Durch Änderung des Trainingsprotokolls sind niedrigere (aber immer noch hohe) Cosinus-Ähnlichkeiten möglich (wie in unseren Experimenten

erhalten), was sich auf die Parameter alpha bzw. beta auswirkt. Weitere Untersuchungen und Vergleiche mit Baselines werden folgen.

Stand der Arbeiten (30.06.2022):

Es wurden weitere Experimente zur weight fusion durchgeführt. Die Ergebnisse zur weight fusion wurden zusammengefasst und in Form eines Papers aufbereitet. Das Paper hat den Titel “Improving Predictive Performance and Calibration by Weight Fusion in Semantic Segmentation” und wurde bei der ACCV eingereicht.

Teilprojekt 4: Gesamtheitliche KI-Absicherungsstrategie

TP4 ist in fünf Arbeitspakete mit folgender thematischer Struktur unterteilt:

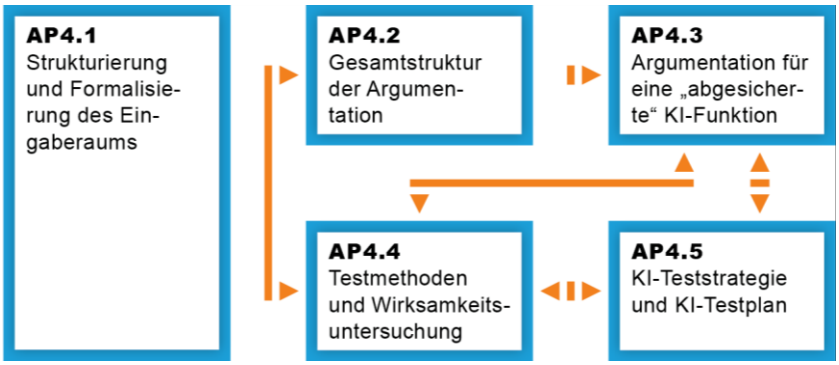


Abbildung 47: TP4-AP-Struktur

Im Folgenden wird ein kurzer Abriss der einzelnen Arbeitspakete gegeben.

AP4.1: Strukturierung und Formalisierung des Eingaberaums (Kontext)

Grundlage einer robusten ➲ Absicherungsstrategie für eine KI-Funktion ist eine genaue und formale Beschreibung der Umgebung (Kontext), in der die KI-Funktion eingesetzt werden soll.

In AP4.1 wird dieser Grundkontext für die Absicherung der Erkennung von Fußgängern im urbanen Kreuzungsbereich definiert und unter Verwendung einer geeigneten Beschreibungssprache beschrieben. Zudem wird 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 erarbeitet. Außerdem werden physikalische Effekte und bekannte Zusammenhänge zwischen Einflussfaktoren formuliert („A-priori-Wissen“).

Der in AP4.1 verankerte Prozess zur TP-übergreifenden Definition der Beschreibungssprache (siehe auch Kapitel 5.3.1) soll die Anforderungen an absicherungsrelevante Kontextelemente und Kontextdimensionen aus allen AP berücksichtigen und zusammenführen. Ein in AP4.1 zu entwickelndes Frontend zur Auswahl und Kombination möglicher Kontextelemente soll die systematische Datenanforderung und Datengenerierung vereinheitlichen.

AP4.2: Gesamtstruktur Argumentation und Sicherheitsanforderungen für KI-Funktion

Basierend auf der KI-Funktionsspezifikation aus AP1.2 werden unter anderem in AP4.2 die Sicherheitsziele der eingesetzten Funktion (z.B. jeder 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 – definiert.

Die Sicherheitsziele als auch die Zielgrößen für die in AP1.2 und AP3.2 zu definierenden KPIs leiten sich aus den übergeordneten Sicherheitszielen des Systems (z.B. keinen Fußgänger verletzen) und der übergeordneten Systemarchitektur ab. Zudem werden in AP4.2 die Sicherheitsziele in eine „Contract-based Design and Validation“-Struktur überführt, mit deren Hilfe der Nachweis über das Einhalten der Sicherheitsziele erbracht werden kann. Abbildung 9 zeigt eine Übersicht des „Contract-based Design and Validation“-Ansatzes unter Verwendung von Sicherheitsverträgen (Safety Contracts). Ein Vertrag ist eine Verknüpfung von Annahmen (Assumptions) und Garantien (Guarantees). Garantien sind in diesem Fall als Zusicherungen zu verstehen, die ein Experte bereit ist auf Grundlage der notwendigen Annahmen zu geben. Eine Reihe von hinreichenden Evidenzen wird genutzt um die Garantien zu bestätigen. Ab wann die Evidenzen hinreichend sind, liegt unter anderen im Ermessen von Experten. Im Kontext des Projekts müssen die Annahmen und Garantien unter der in AP4.1 entwickelten Beschreibungssprache darstellbar sein. Ein Sicherheitsvertrag bedient sich dem gleichen Prinzip, enthält jedoch vor allem sicherheitsrelevante Aspekte. Wird ein Sicherheitsvertrag eingehalten, so sollte eine sichere Funktionsweise gewährleistet sein.

Ein weiteres Ziel in AP4.2 ist die Formulierung der Gesamtstruktur eines gesamtheitlichen Arguments (Assurance Case), welches die Ansprüche an den zu erbringenden Nachweis zur Erreichung des akzeptablen Restrisikos und zur Einhaltung der Sicherheitsziele erbringen soll. Damit soll sichergestellt werden, dass eine KI-Funktion kein unangemessenes Risiko hinsichtlich der Systemsicherheit darstellt. Der Nachweis wird in Form von einer absicherungswirksame Kombinationen verschiedener KI-Analyse- und Plausibilisierungs-Methoden sowie KI-Absicherungsmaßnahmen aus TP3 formuliert.

Der Assurance Case bedingt das Vorhandensein der Mechanismen (Methoden & Maßnahmen) aus AP3.4 bis AP3.6 sowie eine Einschätzbarkeit deren Wirksamkeit. Der Assurance Case enthält auch die Teststrategie, die im folgenden AP4.5 beschrieben wird. Eine tiefergehendere Argumentation als die Gesamtstruktur findet in AP4.3 statt.

AP4.3: Argumentation für eine „abgesicherte“ KI-Funktion

Das Konsortium nimmt die Arbeiten in TP4 unter der Annahme auf, dass eine KI-Funktion aufgrund der intrinsischen Nichtlinearitäten schwierig als alleinstehende Funktion ohne zusätzliche Maßnahmen als sicher argumentiert werden kann. Vielmehr muss eine KI-Funktion um zusätzliche KI-spezifische und KI-unspezifische Maßnahmen erweitert werden.

Folgende Nachweisstrategien gehören zu einem systematischen und gesamtheitlichen Vorgehen zur Formulierung einer gesamtheitlichen Argumentation (Assurance Case) zur Absicherung einer KI-Funktion. Dabei soll das Erreichen eines akzeptablen Restrisikos, welches von funktionalen Unzulänglichkeiten einer KI-Funktion ausgehen könnte, dargelegt werden.

- Nachweisstrategie für eine hinreichende Spezifikation der Funktion Fußgängererkennung (basierend unter anderen auf den Ergebnissen der Domänenanalyse - siehe AP4.1)
- Nachweisstrategie für eine hinreichende Datenbasis als Grundlage für die trainierte KI-Funktion
- Nachweisstrategie für eine korrekte Umsetzung der Funktion (inklusive des trainierten Modells und weiterer externer Maßnahmen) mit hinreichender Performanz entsprechend der Spezifikation

Im Anschluss wird eine Argumentation für ein allgemeines methodisches Vorgehen zur Erhebung von KPI sowie Sicherheitszielen einer KI-Funktion auf Basis der drei zuvor genannten Nachweisstrategien erarbeitet werden. Dabei sollen Lücken in der Argumentation aufgezeigt und geschlossen werden. Darüber hinaus soll dieser Beitrag in enger Zusammenarbeit mit AP4.2 in einen Assurance Case münden und als Ausgangspunkt für einen Industrie-Konsens dienen.

Die Kommunikation und Diskussion des Assurance Case mit relevanten öffentlichen, zertifizierenden Stellen soll in AP5.3 erfolgen. Die Erarbeitung des fachrelevanten Inputs kommt jedoch aus AP4.2 und AP4.3.

AP4.4: Testmethoden und Bestätigung der Wirksamkeit der Projektergebnisse

Die zu entwickelnden Testmethoden sollen bezogen auf die in AP4.1 durchgeführte Domänenanalyse und die in AP4.2 definierten Sicherheitsziele zum einen eine hinreichende Abdeckung des Eingaberaums und zum anderen eine hinreichende Wirksamkeit der Einzelmaßnahmen sicherstellen. Ausgangspunkt für die Überprüfung der Abdeckung des Eingaberaums sind die in AP2.5 erzeugten Test- und Validierungsdaten. Mit ihnen soll geprüft werden, mit welchen in den AP3.3 bis 3.6 entwickelten Methoden und Maßnahmen Metriken entwickelt werden können, die eine Aussage darüber ermöglichen, welche Testdaten für eine sinnvolle Abdeckung des Eingaberaum bezogen auf ein vorliegendes, trainiertes neuronales Netzwerk zusätzlich notwendig sind. Aus dieser Betrachtung sollen iterativ Spezifikationen von Bildinhalten unter Verwendung des Domänenmodells erzeugt werden, aus denen dann fehlende Bilder beauftragt werden können. Parallel dazu soll für die in AP3.3 bis 3.6 entwickelten Methoden und Maßnahmen überprüft werden, welche Wirksamkeit sie für den Nachweis der in AP4.2 definierten Sicherheitsziele haben. Hierzu sollen gezielt neuronale Netzwerke mit Defekten bzw. Unzulänglichkeiten trainiert werden. Anschließend soll überprüft werden, welche der entwickelten Absicherungsmaßnahmen zum Aufdecken welcher Defekte bzw. Unzulänglichkeiten geeignet sind und wie aussagekräftig die von diesen Methoden berechneten KPI bezogen auf die definierten Sicherheitsziele und Performance-Anforderungen sind.

Beispiele für künstliche erzeugte Defekte sind fehlende Domänenelemente in den Trainingsdaten (z.B. nur Tagbilder in den Trainingsdaten) und das Hinzufügen von Bias (z.B. alle Fußgänger tragen gelbe Kleidung und weitere gelbe Objekte gibt es

nicht in den Trainingsdaten). Hieraus entstehende Erkenntnisse bzgl. der Wirksamkeit, die an das AP3.6 zurückgespielt werden sollen. Zusätzlich soll in AP4.4 die statistische Extrapolierbarkeit der Testergebnisse untersucht werden. Bezogen auf die Abdeckung des Eingaberaums (Blackbox-Coverage) soll mit diesen Methoden zusätzlich untersucht werden, ob auch ein Abdeckungsmaß für das neuronale Netzwerk definiert werden kann (Whitebox-Coverage). Bezogen auf die Wirksamkeit der einzelnen Absicherungsmaßnahmen ist hier zu betrachten, in wie weit von den Resultaten der Absicherungsmaßnahmen auf dem verwendeten Testdatensatz auf die korrekte Funktion des neuralen Netzwerkes in der realen Welt in Bezug auf ein definiertes Sicherheitsziel geschlossen werden kann.

AP4.5: KI-Teststrategie und KI-Testplan als Ausgangspunkt für eine Produktfreigabe

Unter Verwendung der Ergebnisse von TP3 und AP4.4 soll eine KI-Teststrategie entwickelt werden, die definiert, welche Methoden und Maßnahmen für den Test einer KI-Funktion in welcher Kombination anzuwenden sind, um die im Assurance Case geforderten Garantien hinreichend nachzuweisen. Die KI-Teststrategie ist dabei als ein allgemeines Vorgehen zum Testen einer KI-Funktion definiert, das noch unabhängig von der konkreten getesteten KI-Funktion und vom in AP4.3 entwickelten Assurance Case für die Fußgängererkennung ist. Darauf aufbauend soll für die entwickelte Fußgängererkennung auf Basis des konkreten Assurance Case unter Verwendung der KI-Teststrategie ein KI-Testplan instanziiert werden. Auf Basis des KI-Testplans soll dann die im Projekt beispielhaft entwickelte KI-Funktion getestet werden und die Testergebnisse sollen in den Assurance Case zurückgespielt werden.

AP4.1 Strukturierung und Formalisierung des Eingaberaums (16 PM)

Aufgaben Valeo:

Definition von Grundkontexts (E4.1.1)

• Herausarbeiten von relevanten Grundkontexten, insbesondere im Hinblick auf Anforderungen; Erarbeitung eines Vorgehens zur Beschreibung der Grundkontexte durch Kontextdimensionen.

Strategie zur Analyse des absicherungsrelevanten Eingaberaums (E4.1.2a)

• Definition von Begrifflichkeiten, die in der Domänenanalyse vorkommen.

Strukturierung des Eingaberaums (E4.1.2b)

• Identifikation von sicherheitsrelevanten Eigenschaften und Einflussfaktoren im Eingaberaum insbesondere durch Integration der Ergebnisse aus bottom-up Ansatz und Erarbeitung einer Methodik, um eine vollständige Kontextbeschreibung zu ermöglichen.

Strukturierung physikalischer Zusammenhänge (E4.1.3)

• Betrachtung und Formulierung bekannter Zusammenhänge und physikalische Effekte, die im Eingaberaum bzw. in den Szenarien berücksichtigt werden müssen, um

beim Einsatz von LIDAR-Sensoren möglichst reale Bedingungen in den synthetischen Daten vorzufinden.

• Formulierte bekannte Zusammenhänge und physikalische Effekte vor allem im Hinblick auf LIDAR Sensoren.

Beschreibungssprache und Ontologie der Dimensionen (E4.1.4a)

• Definition der Beschreibungssprache und Spezifizieren der Datenform einzelner Dimension; Erweiterungen um statistische Verteilung einzelner Parameter, um eine automatische Generierung von Variationen zu ermöglichen.

Stand der Arbeiten (31.12.2019):

Zeitnah nach dem Projektstart wurden die Arbeiten in AP4.1 gestartet und eine aktive Zusammenarbeit vorangetrieben. In Zusammenarbeit mit Bosch fand am 11.7 in Abstatt ein AP4.1 Kickoff bzw. Workshop zur Beschreibungssprache (P1 Prozess). In der ersten Phase hat sich Valeo darauf konzentriert, eine Übersicht bestehender Ansätze zu Kontextbeschreibungen zu sammeln und zu analysieren, um daraus mögliche Strategien zum weiteren Vorgehen zu generieren. Nach Abstimmung mit den Projektpartnern wurde entschieden insbesondere auf die Vorarbeit aus Pegasus (6-Ebenen-Modell) und auf *OpenDrive*® bzw. *OpenScenario*® aufzubauen. Ebenso hat sich Valeo mit Ansätzen bei konventionellen Algorithmen auseinandergesetzt und eine Übersicht zu NCAP Szenarien erstellt, deren Inhalte hauptsächlich auf Unfallstatistiken basieren. Darauf aufbauend wurde versucht durch eine Sammlung von Corner-Cases eine Sammlung von kritischen Szenarien für eine KI-Funktion zusammenzustellen. Da die ersten Ergebnisse aus AP2.2 noch ausstanden, wurde auf das Wissen der Corner-Cases innerhalb des APs zurückgegriffen. Valeo hat hier insbesondere sein Wissen bezüglich Corner-Cases bei LiDAR mit eingebracht. Die daraus entstandenen Vorschläge für Grundkontexte flossen in den ersten Ergebnisbericht für E4.1.1 mit ein, der im Monat 3 fertiggestellt wurde.

Parallel dazu begannen die Arbeiten an einer umfassenden Domänenanalyse, die auf die entsprechenden Partner eingeteilt wurde. Valeo hat hier insbesondere bezüglich eine Struktur für eine Fußgängerbeschreibung mit aufgebaut und einen Einblick in die sicherheitsrelevanten Ansprüche (etwa aus UL4600) geliefert. Die Ergebnisse hierzu, als auch ein erster Vorschlag zu der Strukturierung des Eingaberaums ist im E4.1.2 Bericht festgehalten.

Bei E4.1.3 hat sich Valeo insbesondere mit den verschiedenen Möglichkeiten zur Dimensionsreduktion auseinandergesetzt und verschiedene Verfahren exemplarisch aufgezeigt, u.a. durch SCODE, um darauf eine spätere Strukturierung aufbauen zu können.

Ebenso wie in E4.1.1 flossen in E4.1.4 die Voruntersuchung in vorhandenen Ansätzen zur Beschreibungssprache ein. Darüber hinaus hat Valeo einen intensiven Austausch mit TP2 angestrebt, um frühzeitig ein abgestimmtes Dateiformat beitragen zu können.

Über den Status der jeweiligen Ergebnisse wurde auf verschiedenen TP4 Veranstaltung (z.B. Kick-Off in Renningen am 15.10) oder Projekttreffen (z.B. in Sankt Augustin am 4.12) Einblick im Rahmen eines Vortrags gegeben.

Stand der Arbeiten (30.06.2020):

Im zweiten Halbjahr wurde die Arbeit an AP4.1 planmäßig fortgeführt. Am 14. Januar fand ein Workshop in Wolfsburg zu AP4.1.1, AP4.1.2 und AP4.1.3 statt. In den anschließenden wöchentlichen Telefonmeetings von AP4.1, als auch vom P1 Prozess wurden für E4.1.1 die Anforderungen bzgl. des Grundkontexts präzisiert und priorisiert. Valeo hat weiterhin zwei existierende Grundkontexte aus dem Raum Kronach im OpenDrive-Format zur Verfügung gestellt. In Zusammenarbeit mit den anderen Partnern wurden diese und 17 weitere Grundkontexte anhand der zuvor gesammelten Kriterien analysiert und bewertet, von denen zwei für die nächste Datentranche ausgewählt wurden.

Abbildung 48: Von Valeo bereitgestellter Grundkontext Stöhrstraße / Bürgermeister-Mertel-Straße in vereinfachter OpenDrive® Ansicht und Vogelperspektive

Für E4.1.2 wurde basierend auf den Grundkontexten der Eingaberaum analysiert und relevante Kontextdimensionen semantisch beschrieben, strukturiert und in verschiedene Cluster klassifiziert. Dazu diente ein von Valeo organisierter dreitägiger Hackathon am 16.3., 23.3. und 24.3., bei dem auch erste Entwürfe für die Ontologie (E4.1.4) erarbeitet wurden. Im weiteren Projektverlauf wurden die Cluster in Zusammenarbeit verfeinert und mit weiteren Informationen angereichert. Das Ergebnis wurde insbesondere für die Arbeit im P1 Prozess genutzt, bei der die Operational Design Domain (ODD) für die Datentranche #3 definiert und priorisiert wurde.
In E4.1.3 hat Valeo Methoden entwickelt, um a-priori Wissen und physikalische Zusammenhänge in der Ontologie zu berücksichtigen. Diese wurden zu Dokumentationszwecken beispielhaft angewendet und zusammengetragen.

Im Rahmen von E4.1.4 hat Valeo neben dem Hackathon im März einen zusätzlichen Workshop am 23.6. mitorganisiert. In einem ersten Schritt wurden verschiedene

Teilontologien entwickelt. Valeo hat durch einen Top-Down Ansatz aus diesen Ontologien einen ersten Vorschlag für eine Gesamtontologie ausgearbeitet, der momentan gereviewed wird. Durch den Ausfall von ZF aufgrund von Kurzarbeit hat Valeo kurzfristig stellvertretend die Organisation und Dokumentation von E4.1.4 übernommen, um einem größeren Verzug entgegenzuwirken. Dazu wurde u.a. eine Anleitung für die Ontologie-Entwicklung verfasst, um die Zusammenarbeit zu verstärken.

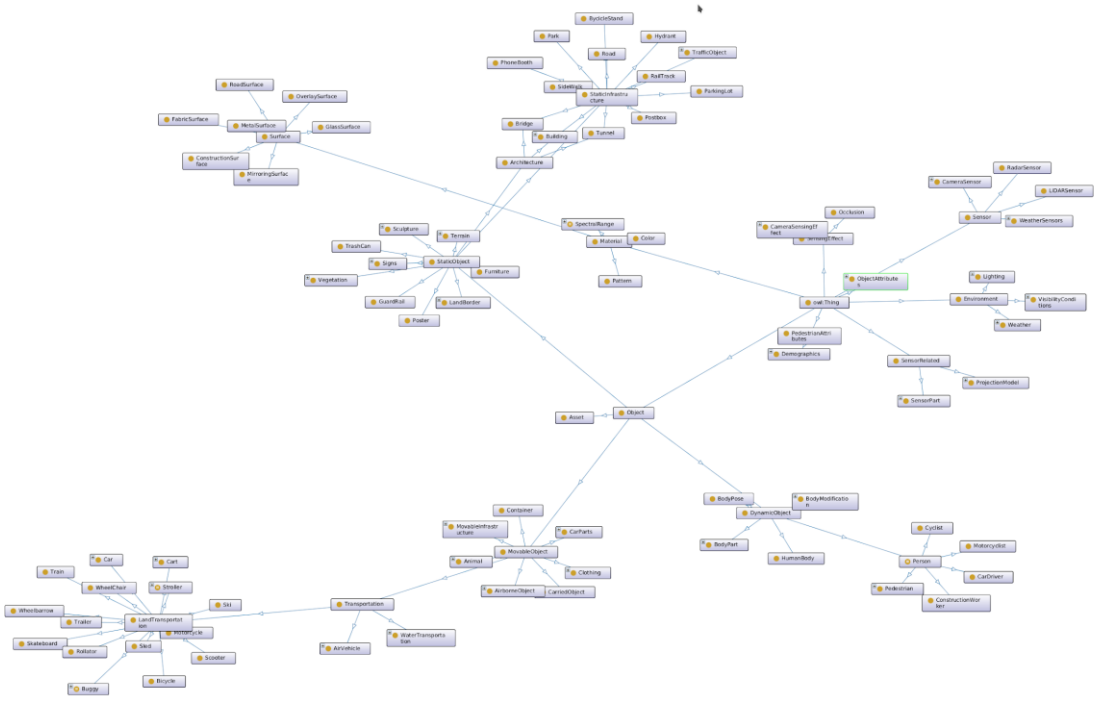


Abbildung 49: Schematische Darstellung der von Valeo entwickelten Ontologie für den Use Case Fußgängererkennung im Kreuzungsbereich

Weiterhin hat Valeo zum einen die bisherigen Ergebnisse beim Steuerkreistreffen am 13.2. in Garching vorgestellt. Ebenso wurde aktiv die Zusammenarbeit mit VV Methoden und ein Austausch hinsichtlich der Ontologie vorangetrieben.

<u>Stand der Arbeiten (31.12.2020):</u>

Die Arbeiten in AP4.1 wurden weiterhin planmäßig fortgeführt. Valeo leitet hierbei die wöchentliche Telefonkonferenz und vertritt das Arbeitspaket im ebenfalls wöchentlich stattfindenden P1 Prozess. Ebenso treibt Valeo die Zusammenarbeit mit verschiedenen Schwesterprojekten, wie V&V Methoden und KI Data Tooling beim Thema Ontologie.

Für P4.1.1 wurden die synthetischen Daten evaluiert. Aus den gewonnenen Erkenntnissen wurden Verbesserungsvorschläge für weitere Grundkontexte herausgearbeitet. Diese betrafen insbesondere die räumliche Variation von Fußgängern in generierten Bildern und Effekte die in Zusammenhang mit Fußgängern auftreten können. Durch die Diskussionen mit den Projektpartnern wurden

Anforderungen an weitere Grundkontexten erarbeitet, die diese Probleme in zukünftigen synthetischen Daten verringern sollen.

Für E4.1.2 wurden die bisher erarbeiteten Kontextdimensionen verfeinert. Das bisherige SCODE Modell wurde iterativ verbessert. Insbesondere wurden das Feedback aus E4.1.2 evaluiert und entsprechende neue Anforderungen ausgearbeitet. Diese haben maßgeblich die Struktur und Maschinenlesbarkeit des SCODE Modells beeinflusst. Valeo hat insbesondere zur Formalisierung beitragen können, wodurch die einzelnen Elemente des SCODE Ansatzes auf die Ontologie übertragen werden konnten.

In E4.1.3 wurde gemeinsam mit den Projektpartnern an einem Objektkatalog gearbeitet, der verschiedene physikalische Zusammenhänge auflistet. In Zusammenarbeit mit understand.ai wurden außerdem Methoden entwickelt, um das erarbeitete Wissen in die Ontologie zu übernehmen. Dazu werden auf die Elemente von OWL bzw. auf die Semantic Web Rule Language (SWRL) zurückgegriffen. Die gewonnenen Erkenntnisse wurden außerdem benutzt, um die Ontologie als auch das SCODE Modell zu erweitern und auf die Verwendung eines Reasoners vorzubereiten.

In E4.1.4 wurde seitens Valeo mittels eines Top-Down Ansatz eine Ontologie entworfen. Die dabei gewonnenen Erkenntnisse flossen später in einem Exporter ein, der von Bosch entwickelt wurde, um eine Ontologie aus dem bereits vorhandenen SCODE Modell zu erzeugen. Die nötigen Voraussetzungen und Anforderungen wurden gemeinsam in verschiedenen Workshops herausgearbeitet. Valeo hat maßgeblich dazu beigetragen die verschiedenen Arten von Dimensionen des SCODE Modells auf Klassen und Properties der Ontologie zu übertragen. In Abstimmung mit den Projektpartnern wurde iterativ die exportierte Ontologie verbessert, so dass am Ende 10 Subontologien enstehen konnten, die das Wissen aus dem SCODE Modell beinhalten. Valeo hat diese Subontologien in eine gemeinsame Hauptontlogie zusammengeführt und die entsprechenden notwendigen Tools den Projektpartnern zur Verfügung gestellt. Darüber hinaus wurde die Ontologie seitens Valeo mehrfach gereviewed und auf Praktikabilität untersucht. Dazu wurde die Ontologie genutzt, um die Assets aus TP2 zu beschreiben. Die Erkenntnisse die aus der Beschreibung der Assets gewonnen werden konnten, soll in den kommenden Monaten genutzt werden, um die Ontologie weiter zu verfeinern und insbesondere maschinenlesbar zu gestalten.

<u>Stand der Arbeiten (30.06.2021):</u>

Im vierten Halbjahr sind die Arbeiten seitens Valeo planmäßig fortgeführt. Valeo leitet die zweiwöchentliche Telefonkonferenz und vertritt das Arbeitspaket im P1 und P3 Prozess. Ebenso wurden die Arbeiten aus AP4.1 von Valeo in verschiedenen Schwesterprojekten, wie KI Data Tooling und auf der Halbzeitpräsentation vorgestellt.

Die Arbeiten an E4.1.1 wurden in M18 abgeschlossen und die Dokumentation der Ergebnisse wurden abgeschlossen.

In E4.1.2 hat Valeo die Methodik, die zur Erstellung der vorhandenen Ontologie genutzt wurde, ausführlich beschrieben und dokumentiert. Die absicherungsrelevanten Kontextdimensionen des Eingaberaums wurden gemeinsam weiter verfeinert und in mehreren Treffen mit den Projektpartnern in eine Ontologie überführt. Valeo hat insbesondere dazu beigetragen, dass die Ergebnisse des SCODE Modell mit in die Ontologie überführt werden konnten und zusätzliche Annotationen erarbeitet, die für eine bessere Weiterverarbeitung notwendig sind.

In E4.1.3 hat Valeo im vergangenen Halbjahr mit den Projektpartnern an Herangehensweisen gearbeitet, um physikalisches Wissen mit in der Ontologie zu formulieren. Dazu wurde insbesondere das Beispiel eines Fußgängers mit einer Sonnenbrille und die sich daraus ergebenden physikalischen Folgen erarbeitet. Darüber hinaus trug Valeo zur Formulierung von Wahrscheinlichkeitsketten bei, die im Rahmen des Unterarbeitspakets entstanden, als auch zu der Weitergabe der Ergebnisse in anderen Unterarbeitspakete.

In E4.1.4a wurden der gewählte Bottom-Up Ansatz weiter von Valeo vorangetrieben. Dazu wurden Anforderungen erarbeitet, die einen erneuten Export aus SCODE benötigten. Anschließend wurden mit den anderen Partnern die verschiedenen Subontologien weiter verfeinert und gegenseitig überprüft.

Um den Asset Katalog aus AP2.5 mit in die Ontologie einzuarbeiten und dort mit detaillierten Beschreibungen zu ergänzen, wurde von Valeo ein Tool zu entwickelt. Dieses generiert aus der Ontologie automatisiert eine GUI, die das einfache Anlegen neuer Assets erlaubt. Neben der GUI wurden verschiedene Python Module entwickelt, die das einlesen, verarbeiten und speichern der Asset und Master Ontologie in Python ermöglicht und wiederverwendet werden kann. Das Tool wurde nachfolgend um weitere Funktionen ergänzt, wie der Auswahl von mehreren Alternativen für dieselbe Dimension oder dem Hinzufügen von Filtern.

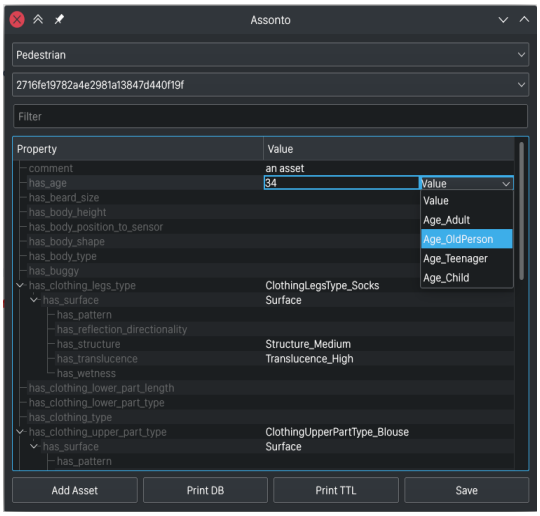


Abbildung 50: Beschreibung eines Fußgänger Assets in Assonto

Durch die automatische Generierung der GUI sind weitere Anforderungen an die Ontologie entstanden, die sicherstellen, dass die Ontologie automatisiert verarbeitet werden kann und etwa der Typ verschiedener Dimensionen hinterlegt ist. Das Tool wurde in verschiedenen Austauschrunden mit z.B. KI-Daten Tooling vorgestellt und es ist abzusehen, dass es dort eine Weiterverwendung findet.

Valeo hat darüber hinaus die Arbeiten in E4.1.4b mit übernommen und einen gemeinsamen Workshop mit den Datenproduzenten organisiert und das weitere Vorgehen erarbeitet. Valeo hat außer dem ein Teil der daraus entstandenen Arbeitspunkte umgesetzt. Insbesondere hat Valeo beigetragen, die Anforderungen für ein maschinelles Beschreibungsformat "Data Specification and Data Description Format" (DSDF) zu formulieren.

Stand der Arbeiten (31.12.2021):

In diesem Berichtszeitraum sind die Arbeiten wie geplant fortgeführt wurden. Valeo hat die Leitung des APs weitergeführt und eine zweiwöchentliche Telefonkonferenz zur Organisation der Arbeiten geleitet.

In E4.1.2 wurden vor allen die bisher ausgearbeiteten strukturierten Eingaberäume durch Experten gereviewed und Anpassungen vorgenommen. Valeo hat insbesondere die Beschreibung der Sensoreffekte und Lichtquellen überarbeitet und mit fehlenden Informationen ergänzt. Daneben hat Valeo eine Verknüpfung zwischen dem Eingaberaum und der ODD herausgearbeitet und in das Ergebnis einfließen lassen.

In E4.1.3 wurden die Arbeiten an dem ausgewählten Beispiel der Wahrscheinlichkeitsbasierten Beschreibung des Tragens einer Sonnenbrille fortgeführt und vorgestellt.

In E4.1.4a hat Valeo die Ontologie anhand der neuen Anforderungen der Projektpartner ergänzt, etwa für die Erweiterte Beschreibung von MoCap Assets und anderer Metadaten, die teilweise durch Projektpartner wie Bosch in ihren Tools weitergenutzt wurden.

Darüber hinaus wurde das im vergangenen Halbjahr begonnene Assonto Tool für die Beschreibung des Asset Katalogs weiterentwickelt und mit neuen Features angereichert. Diese umfassen die Ergänzung von Filtermöglichkeiten, dem Kopieren oder Löschen von Assets, sowie verschiedener Export Funktionen. Dadurch ist es möglich die Asset Ontologie auch in Formaten wie JSON oder CSV bereitzustellen, wodurch tiefergehende Auswertungen ermöglicht werden. Des Weiteren wurde eine Python API bereitgestellt, um die KIA Ontologie, als auch die Asset Ontologie in Skripten durch einfache Funktionen auszulesen und anhand der UUID weitere Informationen zu den Assets zu erlangen.

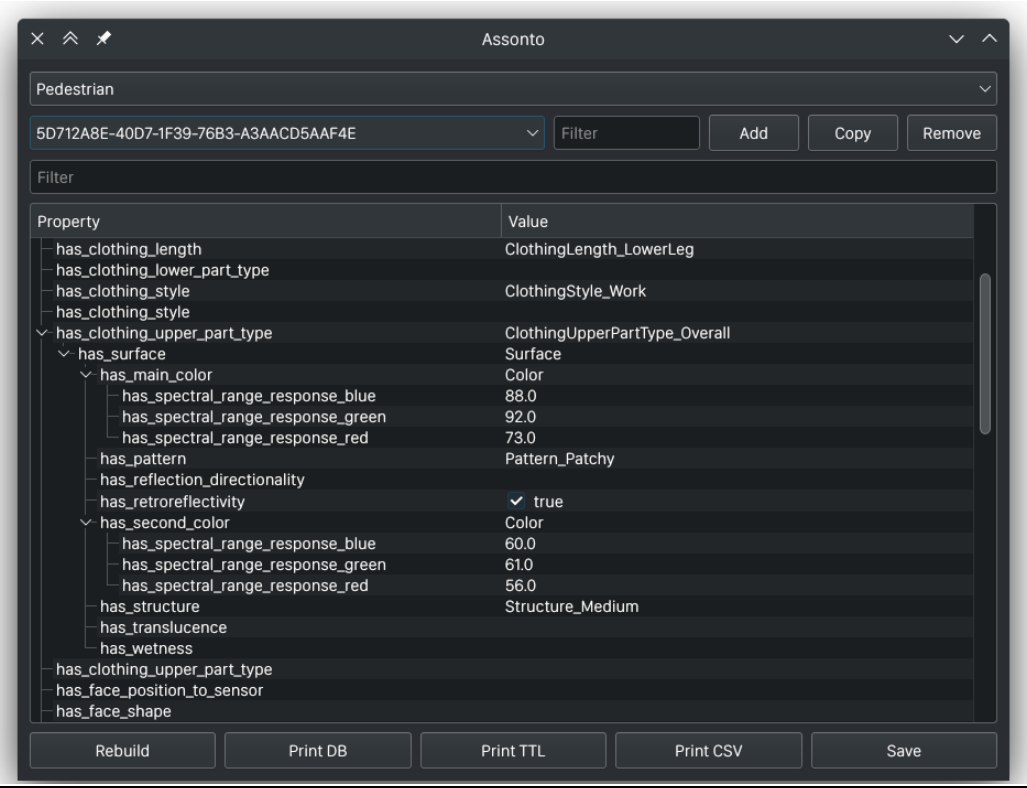


Abbildung 51: Python API

In Zusammenarbeit mit verschiedenen Projektpartnern ist weiterhin eine Veröffentlichung entstanden. Das Paper “Using ontologies for data set engineering in automotive AI applications” von M. Herrmann, C. Witt, L. Lake, S. Guneshka, C. Heinzemann, F. Bonarens, P. Feifel und S. Funke im März auf der DATE vorgestellt werden und gibt eine übersichtliche Zusammenfassung der in AP4.1 erarbeiteten Erkenntnisse und angewendeten Methoden.

Stand der Arbeiten (30.06.2022):

Die Arbeiten in AP4.1 wurden im letzten Berichtszeitraum nahezu vollständig abgeschlossen. In diesem Halbjahr stand daher die Zusammenschrift und Dokumentation der erzielten Ergebnisse im Vordergrund. Valeo hat weiterhin die Leitung des APs und die Organisation in Telkos fortgeführt.

In E4.1.4a wurde die Ontologie an die letzten Tranches angepasst, so dass auch diese vollumfänglich beschrieben werden konnten. Außerdem sind kleinere Fehler in dem Assonto Tool behoben worden.

Für die Abschlusspräsentation wurden für AP4.1 zwei Poster von Valeo erstellten, die insbesondere die Arbeiten in E4.1.2 und E4.1.4 (Methodology of the Ontology) und E4.1.3 (Application of the Ontology) wiedergeben. Diese wurden von Valeo und Opel auf der Abschlusspräsentation vorgestellt

Valeo hat in dem Projektzeitraum die geplanten Ergebnisse in AP4.1 erzielen können und für die Weiterverwendung in anderen Projekten aufbereiten können.

AP4.2 Ableitung von Sicherheitszielen für die KI-Funktion und Gesamtstruktur der Argumentation (11 PM)

Aufgaben Valeo:

Argumentation der Sicherheitsziele sowie der geforderten Zielgrößen auf Ebene der KI-Funktion (E4.2.5)

• Unterstützung bei der Argumentation der Sicherheitsziele mit den Schwerpunkten auf den Eigenheiten verwendeter Sensorik und den Architekturvarianten, welche in TP1 gewählt werden.

Zusammenstellung von KPI für die KI-Funktion als Black-Box im Kontext der Annahmen zu Gesamtfunktionskontext und Systemarchitektur (E4.2.6)

• Koordinieren der Analyse von KPIs zur Anwendbarkeit bei der Argumentation von Sicherheitszielen und identifizieren weiterer KPIs die hierfür verwendet werden können.

Safety Contract für KI-Funktion (E4.2.7)

• Unterstützung bei der Erstellung des Safety Contract für KI-Funktion mit dem Schwerpunkt auf Anwendbarkeit für KI Entwicklungsingenieure.

Stand der Arbeiten (31.12.2019):

Im Hinblick auf E4.2.5 und insbesondere den übergreifenden Prozess P3 wurden beginnend mit einem Workshop am 03.09. in Garching und weitergeführt in wöchentlichen Telkos die benötigten Annahmen zur Systemarchitektur und zum Gesamtfunktionskontext mit den beteiligten Partnern diskutiert und konsolidiert. In der Folge wurden außerdem Vorstellungen und Ideen dazu besprochen, wie aus den Sicherheitszielen der Gesamtfunktion die Sicherheitsanforderungen auf Ebene der KI-Funktion abgeleitet werden können. Im Rahmen eines Workshops am 02. und 03.12. in Sankt Augustin wurde weiterhin eine gemeinsame Vorstellung darüber erarbeitet, wann ein Fußgänger im Sinne des Projektes als relevant gilt.

In Vorbereitung von E4.2.6 hat Valeo die prinzipielle Herangehensweise für die Analyse der KPIs hinsichtlich ihrer Anwendbarkeit auf die Sicherheitsargumentation aus E4.2.5 entwickelt und den entsprechenden Steckbrief verfasst.

Stand der Arbeiten (30.06.2020):

Im Rahmen des übergreifenden Prozesses **P3** hat Valeo zur Entwicklung der ersten Version einer konkreten Systemarchitektur beigetragen, die als Grundlage für die Sicherheitsargumentation und Ableitung der Sicherheitsanforderungen auf Ebene der KI-Funktion innerhalb von AP4.2 dient.

Obwohl Valeo gemäß VHB eigentlich nicht an **E4.2.3** beteiligt ist, ergab sich die Notwendigkeit einer Beteiligung aus der Abwesenheit und dem späteren kompletten Ausscheiden des Partners AID. Basierend auf dem “Graphical Modeling Framework” von Eclipse wurde für das Projekt ein Tool entwickelt, mit dem eine Sicherheitsargumentation in Form der “Goal Structuring Notation” (GSN) veranschaulicht werden kann. Eine solche Veranschaulichung ist unerlässlich für die verständliche Darstellung der Ergebnisse von AP4.2 und AP4.3. Valeo war durch umfangreiche Tests, sowohl der Installation und Anpassung als auch der Funktionalität, und damit verbundenes Feedback wesentlich an der Entwicklung des GSN-Editors beteiligt.

Die Aufgabe von **E4.2.5** ist es, Sicherheitsanforderungen auf der Ebene der KI-Funktion aus übergeordneten Sicherheitszielen schrittweise abzuleiten. Nachdem die erste Version einer solchen Sicherheitsargumentation vorlag, hat Valeo als Resultat eines ausführlichen Reviews zahlreiche Verbesserungs- und Ergänzungsvorschläge eingebracht und damit zur Weiterentwicklung der Argumentation beigetragen.

Im Rahmen von **E4.2.6** sollen geeignete Black-Box Metriken zusammengestellt und in die Sicherheitsargumentation von E4.2.5 eingebaut werden. Da die Argumentation und insbesondere die Sicherheitsanforderungen auf Ebene der KI-Funktion noch nicht bzw. nur in vorläufiger Fassung vorlagen, wurde als Ausgangspunkt für ein erstes Ergebnis eine Sammlung von Sicherheitsanforderungen aus dem *SafeComp2020* Paper von *Lydia Gauerhof* (Bosch) gewählt. Valeo hat eine Präsentation am 17.06. organisiert, in der *Lydia* interessierten Teilnehmern aus AP4.2 und AP4.3 ihre Forschungsergebnisse vorgestellt hat. Darauf aufbauend hat Valeo die für E4.2.6 relevanten Inhalte im Ergebnisbericht dokumentiert und den noch laufenden Prozess der Zusammenstellung von Metriken durch Erstellung einer Übersicht und Eintragung erster Beispiele angestoßen.

Da sich die Sicherheitsargumentation aus E4.2.5 noch in der Entwicklung befindet, war die Ableitung konkreter projektbezogener Safety Contracts für **E4.2.7** bisher noch nicht möglich. Deshalb wurde als erstes Resultat unter Beteiligung von Valeo ein umfassender Übersichtsbericht zu Contract-Based Design und Safety Contracts erstellt, in dem der State of the Art sowie Ansätze zur Anwendung auf maschinelles Lernen und funktionale Schnittstellen dargestellt sind. Insbesondere hat Valeo ein Konzept zur Kombination von Evidenzen unter Verwendung der Dempster-Shafer Evidenztheorie entwickelt und dokumentiert.

<u>Stand der Arbeiten (31.12.2020):</u>

Im Rahmen von **E4.2.2** hat Valeo durch konstruktives Feedback zur Erstellung eines Musters für die Gesamtargumentation aus im Projekt vorhandenen Teilaspekten der Absicherung beigetragen. Außerdem wurden in Diskussionen mit den anderen Projektpartnern Kriterien identifiziert, die eine wirkungsvolle sowie übersichtliche Sicherheitsargumentation ausmachen und in der “Argumentation Structuring Strategy” (ASS) zusammengefasst sind.

Die Sicherheitsargumentation von **E4.2.5** hat Valeo zusammen mit den anderen Projektpartnern weiterentwickelt. Dabei lag der Fokus insbesondere darauf, Sicherheitsanforderungen an die KI-Funktion aus technischen Sicherheitsanforderungen abzuleiten und dabei die Eindeutigkeit (klare und ausreichend detaillierte Formulierungen) und Überprüfbarkeit (konkrete Aussagen und Zahlenwerte) sicherzustellen sowie Redundanzen zu vermeiden.

Im Rahmen von **E4.2.6** hat Valeo für verschiedene Sicherheitsanforderungen aus dem “Proof of Project Concept” (PoPC) geeignete Metriken identifiziert und diskutiert, die Evidenzen für deren Erfüllung liefern können. Ziel war es dabei auch, die Sicherheitsanforderungen mithilfe der quantitativen Metriken mess- und somit überprüfbar zu machen. Neben den Metriken "reconstruction error" und "latent likelihood", die spezifisch für den verwendeten Mechanismus “variational autoencoder” (VAE) sind, lag der Fokus auch auf den folgenden allgemeineren Metriken: Detektionsraten (“false positives”, “false negatives”, “mean average precision”); Metriken zur Messung der Korrelation zwischen VAE und KI-Funktion; Metriken zur Quantifizierung von Verdeckungen, Rauschen, "adversarial attacks" und ODD-Abdeckung. Obwohl der PoPC nur einen kleinen, speziellen Teilaspekt der gesamten Sicherheitsargumentation beinhaltet, wurden in dessen Diskussion auch viele allgemeinere Konzepte beleuchtet, die für zukünftige Untersuchungen nützlich sein werden.

Ziel von **E4.2.7** ist es, an geeigneten Stellen in der Sicherheitsargumentation sogenannte Safety Contracts zu definieren. Valeo hat hierfür ein Konzept entwickelt und vorgestellt. Die Grundidee besteht darin, dass ein Teilbereich der Argumentation, der die Erfüllung einer bestimmten Sicherheitsanforderung mithilfe von Strategien und Evidenzen nachweist, in einen Safety Contract “verpackt” und die Gesamtargumentation somit modularisiert wird. Dabei wird durch den Safety Contract “garantiert”, dass die entsprechende Sicherheitsanforderung erfüllt ist. Für die darüber liegende Argumentation reicht diese “Garantie” aus und es muss nicht bekannt sein, wie sie im Detail begründet wird. So können Teilaspekte der Gesamtargumentation abgetrennt und von einem speziellen Team, z.B. mithilfe einer konkreten Methode, separat betrachtet werden, was zu einer besseren Übersichtlichkeit führt. Einige bereits vorhandene Teilaspekte der Absicherung wurden hinsichtlich der Möglichkeit zur Einführung derartiger Safety Contracts untersucht. Wie die “Garantie” eines Safety Contracts angesichts der inhärenten Unsicherheit von KI-Funktionen aussehen kann, ist Thema einer noch laufenden Diskussion.

<u>Stand der Arbeiten (30.06.2021):</u>

Im Rahmen des übergreifenden Prozesses **P3** hat Valeo zur Weiterentwicklung der Systemarchitektur beigetragen, die schließlich in der zweiten offiziellen Version mündete. Außerdem wurden basierend auf Vorschlägen und zugehörigem Feedback in zahlreichen Diskussionen Annahmen zum Kontext der Systemarchitektur formuliert,

die die Sicherheitsargumentation in einen Gesamtzusammenhang bringen und somit deren konkrete Ausprägung definieren.

Für die im Rahmen von **E4.2.5** entwickelte Sicherheitsargumentation auf Ebene der KI-Funktion hat Valeo Anknüpfungspunkte identifiziert, an denen die konkreten Dimensionen und Parameter der Operational Design Domain relevant sind. Des Weiteren wurden mit einem Vertreter von TP3 das in E3.2.1 gesammelte Feedback zu einzelnen Machine Learning Safety Requirements (MLSR) diskutiert und gemeinsame Verbesserungsvorschläge hinsichtlich Inhalt und Formulierung ausgearbeitet. Durch umfangreiche Anmerkungen bezüglich der Verifizierbarkeit und darauf aufbauende Diskussionen hat Valeo zur Verbesserung der MLSR beigetragen. In diesem Zusammenhang wurde auch ein Vorschlag ausgearbeitet, der die zugrundeliegende Struktur und das Zusammenspiel der MLSR vereinfacht, wodurch Unklarheiten sowie Wiederholungen vermieden werden können.

Das Hauptziel von **E4.2.6** ist es, geeignete Metriken für die Sicherheitsargumentation aus E4.2.5 zu identifizieren. Hierfür hat Valeo Metriken vorgeschlagen und näher spezifiziert, die eine Konkretisierung sowie Verfeinerung der MLSR aus E4.2.5 ermöglichen, wodurch sie mess- und damit überprüfbar werden. Außerdem wurden zahlreiche Diskussionen mit den anderen beteiligten Partnern geleitet, in denen das Team basierend auf dem vorhandenen Input konkrete Verbesserungsvorschläge für die MLSR entwickelt und anschließend an E4.2.5 kommuniziert hat. Im nächsten Schritt sollen die zu den Metriken gewonnen Erkenntnisse mit dem Feedback aus E3.2.1 verglichen werden, um eventuell vorhandene Unterschiede sowie Lücken zu identifizieren, eine gemeinsame Vorstellung zu etablieren und schließlich die Metriken so weiter bzw. neu zu entwickeln, dass sie optimal in die Sicherheitsargumentation passen und diese bestmöglich unterstützen.

In Fortführung der vorherigen Arbeiten wurde im Rahmen von **E4.2.7** die probabilistische Natur von KI-Funktionen und der entsprechenden Sicherheitsargumentation näher untersucht. Einerseits hat Valeo zusammen mit den anderen beteiligten Partnern die verschiedenen Quellen von Unsicherheiten identifiziert, diskutiert und kategorisiert, zunächst am Beispiel des GSN Graphen für den “variational autoencoder” und anschließend allgemein. Diese Untersuchung resultierte in der Formulierung von Anforderungen für die Darstellung und Einbeziehung von Unsicherheiten in die Sicherheitsargumentation. Andererseits war Valeo wesentlich an der Untersuchung und Diskussion zweier unterschiedlicher mathematischer Herangehensweisen für die probabilistische Inferenz basierend auf vorhandener Fachliteratur beteiligt. Hierbei handelt es sich um Bayesian Belief Networks und die Dempster-Shafer Evidenztheorie. Letztere beschreibt Evidenzen nicht mit Einzelwahrscheinlichkeiten, sondern mit Wahrscheinlichkeitsintervallen, deren Länge die Unsicherheit einer Evidenz direkt widerspiegelt. Dies verdeutlicht, dass die Dempster-Shafer Evidenztheorie gut geeignet ist, um die inhärente Unsicherheit der Sicherheitsargumentation für KI-Funktionen handzuhaben. Konkret

hat Valeo ein Konzept dafür erarbeitet, wie der Glaube an die Erfüllung eines Sicherheitsziels auf Grundlage einer gegebenen Evidenz und der getroffenen Annahmen quantitativ beurteilt werden kann. Außerdem wurde ein einfaches Beispiel konstruiert, anhand dessen das Team die Anwendung des Konzepts erprobt und evaluiert hat. Als nächste Schritte sind die mathematische Betrachtung der Kombination von Evidenzen bzw. Sicherheitszielen und die Anwendung der beiden Herangehensweisen auf konkrete Beispiele aus verschiedenen Arbeitspaketen des Projekts geplant.

Stand der Arbeiten (31.12.2021):

Im Rahmen der Überarbeitung der Machine Learning Safety Requirements (MLSRs) von **E4.2.5** hatte Valeo zuvor einen Vorschlag zur Vereinfachung der Struktur und Erhöhung der Übersichtlichkeit entwickelt. Dieser führt sogenannte atomare Aspekte ein, die jeweils nur einen Sicherheitsaspekt beinhalten und möglichst konkret formulieren. Diese atomaren Aspekte wurden mit den beteiligten Partnern näher diskutiert und verfeinert, was schließlich zu einer finalen Version geführt hat, die an TP3 und AP4.3 übergeben wurde. Die verschiedenen atomaren Aspekte und ihre zugehörigen MLSRs sind in der Abbildung am Ende dieses Abschnitts zu finden. Außerdem hat Valeo dazu beigetragen, die Vollständigkeit der für E4.2.5 vorgenommenen STPA-Analyse sicherzustellen, indem angesichts der erarbeiteten Kontrollstruktur die Ursachen weiterer “loss scenarios” auf funktionaler Ebene identifiziert wurden. Dabei stand die “unsafe control action” im Mittelpunkt, die besagt, dass die Bremse nicht betätigt wird, obwohl sich ein Fußgänger in der für die aktuelle Geschwindigkeit sicherheitskritischen Zone befindet. Mithilfe der “loss scenarios” wurden weitere zu berücksichtigende Sicherheitsbeschränkungen und -anforderungen abgeleitet sowie die Kontrollstruktur angepasst.

The 2D bounding box detector shall detect pedestrians ...

atomic aspect	corresponding MLSR	FuSi / SOTIF	corresponding metrics from E4.2.6 (see also the different Benchmarks)
... with the bounding box position differing by at most <suitable tolerance> from the position of the minimum sized box capable of including the entirety of the pedestrian.	MLSR02	FuSa	• Euclidean distance between BB centers • BB center = intersection of diagonals
... with the bounding box being at most <suitable max. tolerance> smaller or larger in any dimension than the minimum sized box capable of including the entirety of the pedestrian.	MLSR03	FuSa	• width of predicted BB divided by width of ground truth BB • height of predicted BB divided by height of ground truth BB
... with a processing time per frame of at most 40 ms (due to camera frame rate of 25 Hz (40 ms) (camera frame rate)).	MLSR05	FuSa	• processing time of one frame
... if they are relevant according to 2021-07-27 Consolidation and definition of detection zone - KI Absicherung - Confluence (vdali.de)	MLSR07	Definition	
... if they are partially occluded (The amount of occlusion will be defined in the "Benchmark 1 WS" @Christian Hellert)	MLSR08	SOTIF	quantification of pedestrian occlusion: • based on body joint model: weighted sum of occluded joints divided by weighted sum of all joints • number of occluded pixels divided by total number of pixels information necessary to calculate both is available in meta data of the synthetic images
... if they (as well as the present scenery and environmental conditions) are inside the ODD (Generalization-capabilities), including corner-cases , ethical considerations, light conditions	MLSR09	SOTIF	• for test data: dataset metrics measuring the coverage of the ODD, including corner case coverage (in other contexts also coverage of perturbations, adversarial attacks, ...)
... with <suitable performance metric and threshold> → Individual metrics for false negatives and false positives or a combined one? The individual hazards H01 and H02 might require the former.	Deprecated, due to decision on WS with TuV MLSR09		• recall & precision • F1-score as established combination of recall and precision
... at least 4 times in a sequence of 5 consecutive frames	MLSR10	FuSa	• number of false negatives for a particular pedestrian in every sequence of 5 consecutive frames • equivalent assuming constant frame rate: min time between false negatives for a particular pedestrian
... if a foreseeable type of perturbation is present (e.g. strong rain, dirt on sensor, noise due to electro magnetic interference below high voltage lines. ...) → This aspect could be included several times for different types of noise.	MLSR12	SOTIF	• in general: robustness metric = correlation between performance metric & image metric measuring the degree of perturbation (specifics depend on perturbation type) • metric measuring the probability of aleatoric perturbations
... if known adversarial attacks based on expert judgement are present.	MLSR13	SOTIF	• see perturbations above
... with at most one occurrence of false positives inside the detect zone in a sequence of 5 consecutive frames	MLSR19	FuSa	• number of frames with a false positive in the detect zone in every sequence of 5 consecutive frames • equivalent assuming constant frame rate: min time between false positives in detect zone
... with the error of the output (shape offset & class membership) having the form of an additive Gaussian noise with mean zero and a standard deviation of at most <suitable value> (No residual systematic errors)	MLSR20	FuSa	• metrics for investigation of Gaussian nature: MTRC-818714, MTRC-828849, MTRC-838194

Abbildung 52: Atomare Aspekte der Sicherheitsargumentation aus E4.2.5 und zugeordnete Metriken, die im Rahmen von E4.2.6 identifiziert und konsolidiert wurden

Die zuvor im Rahmen von **E4.2.6** erarbeiteten Metriken für die Sicherheitsargumentation aus E4.2.5 wurden mit dem Feedback von TP3 konsolidiert, um die Sicherheits- und KI-Perspektive zusammenzubringen und somit eine einheitliche Sichtweise im Projekt zu etablieren sowie eine optimale Unterstützung der Sicherheitsargumentation zu gewährleisten. Valeo hat die dafür benötigten Diskussionen organisiert und geleitet. Damit ist die Arbeit an E4.2.6 abgeschlossen, die finalen Ergebnisse sind in der rechten Spalte der oben abgebildeten Tabelle zusammengefasst.

Nachdem im Rahmen von **E4.2.7** zuvor ein Konzept zur quantitativen Einschätzung des Glaubens an die Erfüllung von Sicherheitszielen basierend auf der Dempster-Shafer Evidenztheorie entwickelt wurde, hat Valeo dieses Konzept auf die Kombination von Evidenzen zur Argumentation für darüber liegende Sicherheitsziele erweitert. Außerdem hat Valeo eine vereinfachte Sicherheitsargumentation in Form eines GSN-Graphen entworfen und das Konzept darauf exemplarisch angewendet. Diese Sicherheitsargumentation, die in der Abbildung am Ende dieses Abschnitts zu sehen ist, befasst sich mit der Robustheit des neuronalen Netzes gegenüber zwei verschiedenen Störungstypen, wobei jeweils Tests mithilfe zweier verschiedener Datensätze durchgeführt werden. Die konkrete quantitative Betrachtung umfasst drei Schritte:

1. Expertenbeurteilung von Teilzielen und Umwandlung in Wahrscheinlichkeitsmassen
2. Kombination der Teilzielmassen zu Massen für das oberste Sicherheitsziel
3. Umwandlung der obersten Massen in eine besser verständliche Beurteilung

Die Beurteilung trifft eine Aussage darüber, ob die Erfüllung des obersten Sicherheitsziels angesichts der vorliegenden Evidenzen akzeptiert werden kann bzw. abgelehnt werden muss und wie sicher diese Erkenntnis ist. Als weitere Aktivität hat Valeo als Reaktion auf das Feedback zum vorherigen Meilensteinbericht zur Klarstellung des E4.2.7-Verständnisses von “safety contracts” und zur Einordnung von E4.2.7 in den Projektkontext von KI-Absicherung beigetragen.

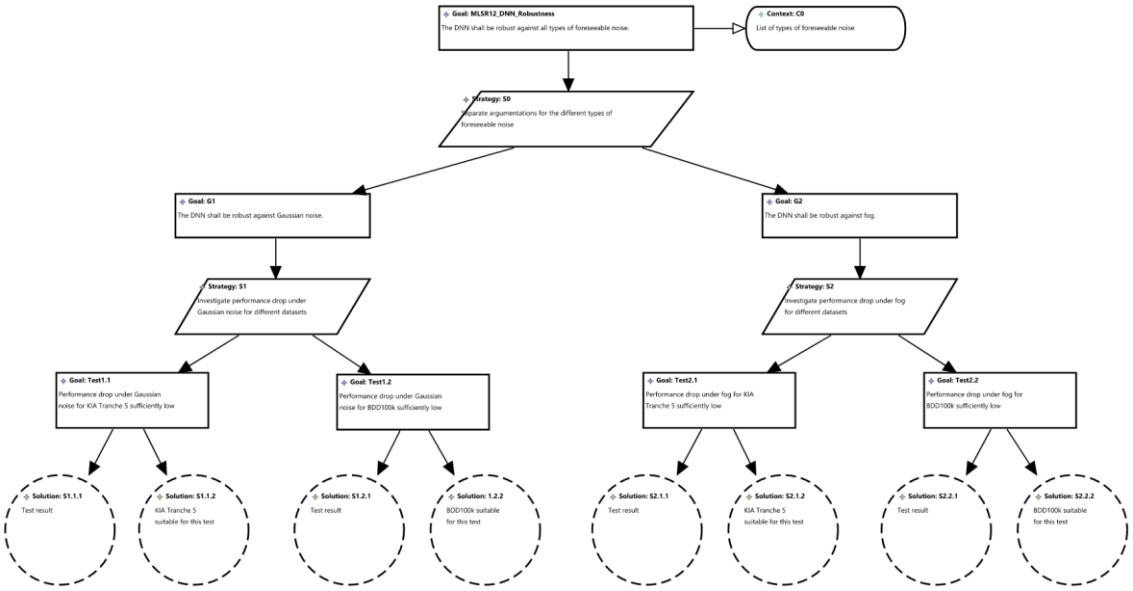


Abbildung 53: Vereinfachte Sicherheitsargumentation der DNN-Robustheit, auf die die Dempster-Shafer Evidenztheorie exemplarisch angewendet wurde

Stand der Arbeiten (30.06.2022):

Im Rahmen von E4.2.5 bestand noch die offene Frage, wie die aus Sicherheitsüberlegungen stammenden Machine Learning Safety Requirements (MLSR) in der Sicherheitsargumentation abgeleitet und mit den DNN-spezifischen Safety Concerns in Zusammenhang gebracht werden können. Dafür hat Valeo ein Konzept zur mehrstufigen Ableitung von MLSRs entwickelt:

- Basierend auf der definierten Kontrollstruktur sollen die Unsafe Control Actions mitigiert werden, die zu den identifizierten Hazards führen können.
- Argumentation über die Kontrollstruktur und ihre Wirkkette: Die Pedestrian-Detection-Komponente garantiert für ihren Output, der von der Perception-Komponente weiterverarbeitet wird, dass sämtliche Loss Scenarios, die eine Fehlfunktion der Perception-Komponente verursachen können, vermieden werden.
- Die Loss Scenarios werden beschrieben durch Causal Factors (Safety Concerns, Performance Limiting Factors), Triggering Event und mögliche Konsequenzen sowie ein daraus abgeleitetes MLSR.
- Die Vermeidung der definierten Loss Scenarios erreicht man durch eine Erfüllung der entsprechenden MLSRs, was gleichzeitig eine Mitigation der zugehörigen Causal Factors bewirkt.

Außerdem hat Valeo zur notwendigen Aktualisierung der Unsafe Control Actions und Loss Scenarios beigetragen, dieses Konzept in Goal Structuring Notation umgesetzt und die entsprechende Überarbeitung der Gesamtargumentation vorgenommen. Dabei stellte sich heraus, dass eine Unterscheidung zwischen der AI-Component-Ebene und der AI-Algorithm-Ebene nicht mehr sinnvoll ist, weshalb letztere entfernt wurde. Weitere Valeo-Beiträge zu E4.2.5 bestanden einerseits in der Klarstellung von

Begriffen aus den Bereichen SOTIF (Safety Of The Intended Functionality) und STPA (System-Theoretic Process Analysis), um deren Verwendung in der Gesamtargumentation zu konsolidieren, und andererseits in einer ausführlichen Überprüfung des E4.2.5-Berichts inklusive notwendiger Überarbeitung.

Als Abschluss von E4.2.7 hat Valeo das zuvor entwickelte Konzept zur quantitativen Einschätzung des Glaubens an die Erfüllung von Sicherheitszielen basierend auf der Dempster-Shafer Evidenztheorie auf den Assessment-GSN-Graphen des Evidence Workstreams “Brittleness of DNNs” angewendet. Dabei wurde die Robustheit zweier neuronaler Netze gegenüber zwei verschiedenen Störungstypen mit jeweils drei Stärkegraden untersucht und basierend auf experimentellen Ergebnissen eine anschauliche Bewertung formuliert.

Im Zusammenhang mit dem letzten Projekttreffen und der Abschlussveranstaltung hat Valeo wesentlich zur Erstellung von Postern für AP4.2 und E4.2.5 beigetragen.

AP4.3 Argumentation für eine „abgesicherte“ KI-Funktion (8 PM)

Aufgabe Valeo:

Nachweisstrategie für eine hinreichende Datenbasis als Grundlage für die trainierte KI-Funktion (E4.3.4)

• Der Schwerpunkt der Argumentation liegt auf einer Analyse der synthetischen Datenbasis im Vergleich zu realen Referenzdaten (diese ggf. aus AP2.4 bzw. VVMethoden). Untersuchungen zur Vergleichbarkeit und Ergänzbarkeit aus AP2.4 sollen in die Argumentation eingebunden werden.

Nachweisstrategie für eine korrekte Umsetzung der Funktion mit hinreichender Performanz entsprechend der Spezifikation (E4.3.5)

• Untersuchung des Einflusses einer Erweiterung der Sensorik (Tiefendaten, LiDAR Sensorik) auf die Nachweisstrategie für eine korrekte Umsetzung der Funktion.

Argumentation für ein allgemeines methodisches Vorgehen zur Absicherung von KI-Funktionen (E4.3.6)

• Transition der Argumentation für die konkrete KI-Funktion auf ein allgemeines Vorgehen für KI-Funktionen.

• Review mit Fokus aus Sicht der KI-Entwickler (Rückverfolgbarkeit von Anforderungen und entwickelter Funktion, Checklisten zu best practices in der KI-Funktionsentwicklung, etc.).

Stand der Arbeiten (30.06.2020):

Die Arbeit in **AP4.3** begann am 10.03. mit einem online durchgeführten Kick-Off Workshop, in welchem fundamentale Themen, wie z.B. Begriffsdefinitionen und

Lieferbeziehungen, besprochen wurden. Im Anschluss daran wurde ein “Minimum Viable Product” (MVP) ausgewählt, anhand dessen die verschiedenen Nachweisstrategien beispielhaft erarbeitet werden können. Das MVP beinhaltet konkret eine mögliche Fehlklassifikation von Fußgängern basierend auf Pixeln, die außerhalb der eigentlichen Bounding Box liegen. Um die Zusammenarbeit zwischen TP3 und TP4 im Hinblick auf die Einbindung der entwickelten Methoden und Maßnahmen in die Sicherheitsargumentation zu stärken, startete am 16.06. eine Workshop-Reihe, in der konkrete vielversprechende Mechanismen von den Entwicklern und zugeordneten “Safety Buddies” aus AP4.3 auf ihren möglichen Absicherungsbeitrag hin untersucht werden.

Als Startpunkt für die zukünftige Arbeit an **E4.3.4** wurden die verschiedenen Strategien bzw. Ansätze der beteiligten Partner zusammengetragen und priorisiert.

Um ein erstes Beispiel für die Nachweisstrategie **E4.3.5** und Ansätze für die zukünftige Vorgehensweise zu erhalten, wurde anhand des MVP ein vertikaler Durchstich durchgeführt: auf jeder Ebene der Sicherheitsargumentation wurde beispielhaft eine konkrete Strategie für eine spezifische Sicherheitsanforderung ausgewählt, wodurch ein linearer Abstieg vom übergreifenden Sicherheitsziel hin zur Detektion eines einzelnen Fehlerereignisses mithilfe einer konkreten Metrik ermöglicht wird. Auf der untersten Ebene wurde die Nachweisstrategie mit bestimmten Zielen und Evidenzen ausgearbeitet, deren aktueller Stand in der folgenden Abbildung zu sehen ist.

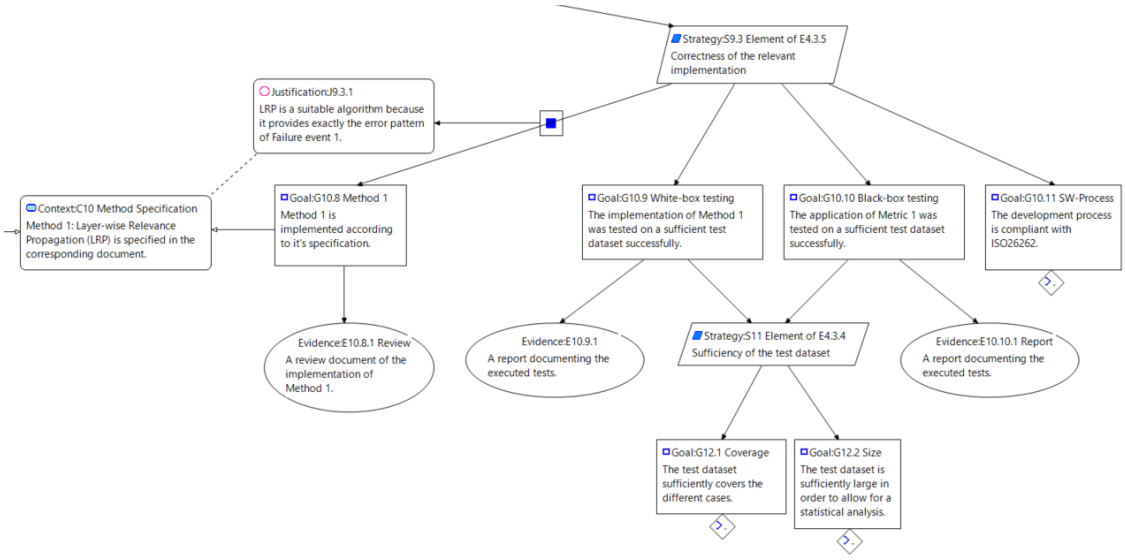


Abbildung 54: Aktueller Stand des GSN-Graphen der Nachweisstrategie E4.3.5 am Beispiel des MVP

Valeo hat nicht nur die oben genannten Tätigkeiten koordiniert und dokumentiert, sondern auch das MVP und die Nachweisstrategie in wesentlichen Punkten weiterentwickelt und den GSN-Graphen entsprechend erweitert.

Die Arbeit an **E4.3.6** beginnt eigentlich erst in M13, bei der oben beschriebenen Beschäftigung mit dem MVP konnten aber bereits erste Erkenntnisse zum allgemeinen Zusammenspiel der verschiedenen Nachweisstrategien gewonnen werden.

Stand der Arbeiten (31.12.2020):

Um die Einbettung der verschiedenen Methoden und Maßnahmen aus TP3 in die Sicherheitsargumentation und deren Wirksamkeit hinsichtlich der Absicherung von KI-Funktionen näher zu untersuchen, wurden im Rahmen des übergreifenden Prozesses **P4** sogenannte Evidence Workshops durchgeführt. Valeo hat am Evidence Workshop des “variational autoencoder” teilgenommen und durch konstruktive Diskussionen zur Weiterentwicklung des zugehörigen Teilaspekts der Sicherheitsargumentation beigetragen. Außerdem hat Valeo für den Evidence Workshop der Methode “Meta LRP” die Rolle des sogenannten Safety Buddies übernommen. In dieser Funktion wurde nach Rücksprache mit den Methodenentwicklern ein GSN-Graph, der geeignete Absicherungsstrategien enthält, angefertigt, den Teilnehmern des Workshops vorgestellt und im Anschluss daran diskutiert sowie weiterentwickelt.

Im Rahmen des **AP4.3** Boxenstopps hat Valeo nicht nur die Ergebnisse des Jahres 2020 präsentiert, sondern zusammen mit den anderen Projektpartnern auch die Weiterarbeit im kommenden Jahr geplant. Dabei liegt der Fokus insbesondere auf dem Zusammenspiel und der Verallgemeinerung der verschiedenen Nachweisstrategien.

Die Arbeit an **E4.3.4** und der damit verbundenen Nachweisstrategie für eine hinreichende Datenbasis hat sich bisher auf eine spezielle Sicherheitsanforderung bezogen: der Trainingsdatensatz soll die Entwicklung der KI-Funktion für Fußgängererkennung so unterstützen, dass bestimmte Performanz-Ziele erreicht werden. In diesem Zusammenhang hat Valeo eine Strategie entwickelt, die sicherstellt, dass die Trainingsdaten ausgewogen (“balanced”) sind, also sich über die semantischen Dimensionen der ODD korrekt verteilen angesichts der tatsächlichen Verteilung in der Betriebsumgebung. Die Argumentation umfasst separat für jede semantische Dimension sowohl den quantitativen Vergleich der beiden Verteilungen als auch die Bestimmung der “wahren” Verteilung. Nachdem sich herausgestellt hat, dass die Strategie für einen ausgewogenen Trainingsdatensatz nicht von der für einen vollständigen Datensatz (ODD is komplett abgedeckt) getrennt werden kann, hat Valeo zur Vereinigung der beiden Strategien beigetragen. Während der Diskussion wurde eine wichtige Frage identifiziert, die aktuell noch ungeklärt ist: wie kann ein Gleichgewicht zwischen der Gesamtperformanz der KI-Funktion und der Berücksichtigung seltener Fälle gefunden werden? Ferner hat Valeo durch ausführliches Feedback auch zur Weiterentwicklung der dritten relevanten Strategie (Korrektheit des Datensatzes) beigetragen.

Im Rahmen von **E4.3.5** hat Valeo die Betrachtung des MVP-Beispiels aus AP4.3 fortgeführt, die entsprechende Nachweisstrategie für eine korrekte Implementierung vertieft und den zugehörigen GSN-Graphen weiterentwickelt. Konkret wurde die Nachweisstrategie in vier zu erreichende Ziele aufgeteilt, wobei der Fokus auf demjenigen Ziel liegt, die Eignung der Metrik "Heat Map Distribution" für die Detektion des Fehlerevents (= die Klassifikation eines Objektes ist durch Pixel außerhalb seiner

Bounding Box beeinflusst, was zu einer Fehlklassifikation führen könnte) mithilfe geeigneter Tests nachzuweisen. Dabei kommen Äquivalenzklassen und vollständige Induktion zum Einsatz, um die Zahl der zu untersuchenden Fälle stark zu reduzieren. Während der Betrachtung des MVP-Beispiels hat Valeo verschiedene Konzepte identifiziert, die verallgemeinerbar sind und somit auch bei der zukünftigen Arbeit an der Nachweisstrategie für eine korrekte Implementierung verwendet werden können.

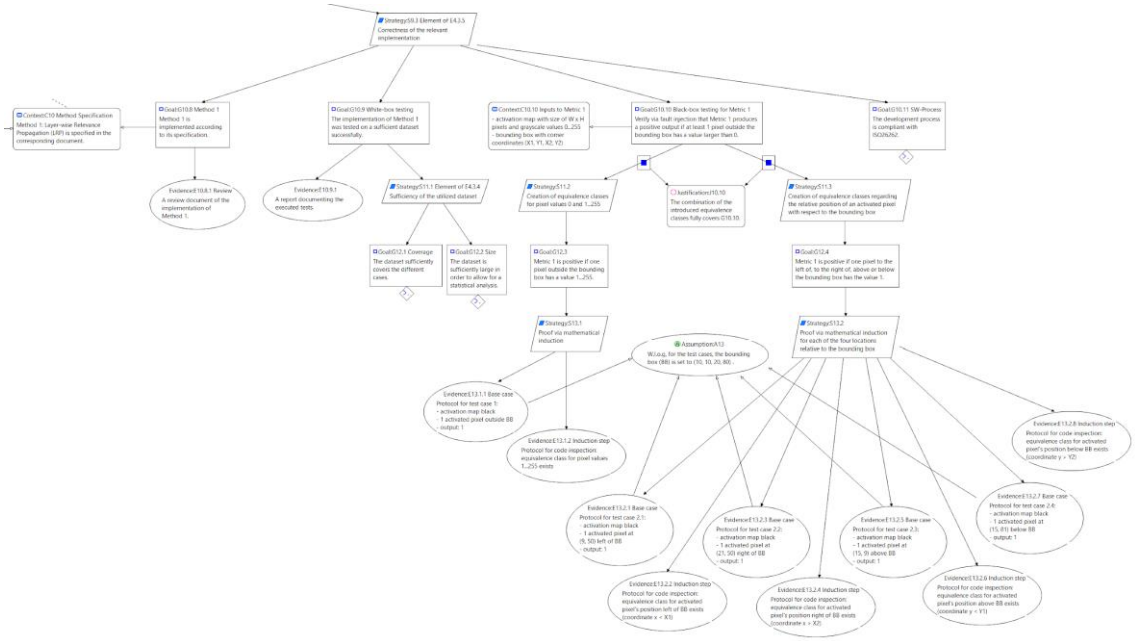


Abbildung 55: GSN-Graph mit dem für das MVP-Beispiel entwickelten Teil der Nachweisstrategie E4.3.5

Das Ziel von **E4.3.6** ist es, die einzelnen in AP4.3 entwickelten Nachweisstrategien in die Gesamtargumentation einzubinden und dabei ein allgemeines Vorgehen für die Absicherung zu entwerfen. In diesem Kontext hat Valeo zusammen mit den anderen Projektpartnern Anknüpfungspunkte für die Nachweisstrategien im Gesamtargumentationsmuster aus E4.2.2 identifiziert. Außerdem wurde die Idee entwickelt, verallgemeinerte Versionen der Nachweisstrategien, sogenannte Safety Case Patterns, zu erstellen, die an verschiedenen Stellen der Gesamtargumentation verwendet und dabei individuell spezifiziert werden können.

<u>Stand der Arbeiten (30.06.2021):</u>

Im Rahmen des übergreifenden Prozesses **P4** hat Valeo an drei Boxenstopp-Treffen teilgenommen, in denen die Vorgehensweise und Ergebnisse der zurückliegenden Evidence Workshops zusammengefasst und bewertet wurden. Auf dieser Grundlage wurde zusammen mit den anderen Projektpartnern entschieden, dass in Zukunft die sogenannten Evidence Workstreams in einem ähnlichen Format durchgeführt werden sollen. Im Gegensatz zu den Evidence Workshops konzentrieren sich diese nicht auf einen einzelnen TP3 Mechanismus, sondern untersuchen mehrere davon, die über ein gemeinsames Konzept, z.B. die Mitigation eines bestimmten Safety Concerns, in

Verbindung stehen. Valeo fungiert als Safety Buddy für den Evidence Workstream “Brittleness of DNNs”, der die Robustheit neuronaler Netze gegenüber verschiedenartigen Störungen auf Bildebene thematisiert, und hat in dieser Funktion beim Kick-Off-Treffen aus den präsentierten, potentiell interessanten TP3 Mechanismen die relevantesten und somit im Weiteren zu untersuchenden ausgewählt. Außerdem hat Valeo einen ersten Entwurf für die Sicherheitsargumentation hinsichtlich der Robustheit entwickelt (siehe auch unten bei E4.3.4) und beim ersten Workshop präsentiert.

Die Arbeit in **AP4.3** wurde zu Beginn der zweiten Hauptarbeitsphase unter Mitwirkung von Valeo neu ausgerichtet. Der bisherige Fokus auf die drei Nachweisstrategien, die durch E4.3.3, E4.3.4 und E4.3.5 abgedeckt waren, hat sich dabei als weniger vielversprechende Herangehensweise herausgestellt. Stattdessen stehen künftig die folgenden Aspekte im Mittelpunkt: Erzeugung von Bestandteilen der Sicherheitsargumentation; Zusammensetzung von Teilargumentationen und Einbettung in die Gesamtargumentation; enge Einbindung der Evidence Workstreams und Tätigkeit als Safety Buddies.

Im Rahmen von **E4.3.4** wurden weitere Sicherheitsaspekte untersucht, die einen hinreichenden Datensatz benötigen. Speziell hat Valeo das folgende allgemeine Sicherheitsziel vorgeschlagen und das daran arbeitende Team geleitet: “Eine bestimmte Teilmenge des Assurance-Datensatzes, die verwendet wird, um eine Sicherheitsmetrik zu messen bzw. einen Mechanismus aus TP3 zu überprüfen, muss für diese Anwendung geeignet sein.” Die Arbeit konzentrierte sich auf eine bestimmte Klasse von TP3-Metriken und -Mechanismen, nämlich diejenigen, die sich mit der Robustheit eines DNN gegenüber Störungen wie Wettereinflüssen, Sensoreffekten oder “adversarial attacks” befassen. Da die Sicherheitsargumentationen für die Robustheit gegenüber verschiedenen Arten von Störungen sehr ähnlich sind, hat Valeo einen GSN-Graphen für ein allgemeines “Safety Case Pattern” entwickelt, das auf eine Vielzahl von Störungsarten angewendet werden kann. Darin spielt Datenaugmentierung basierend auf einem Störungsmodell eine entscheidende Rolle für die Erzeugung von Assurance-Daten. Der entwickelte GSN-Graph ist in der folgenden Abbildung zu sehen.

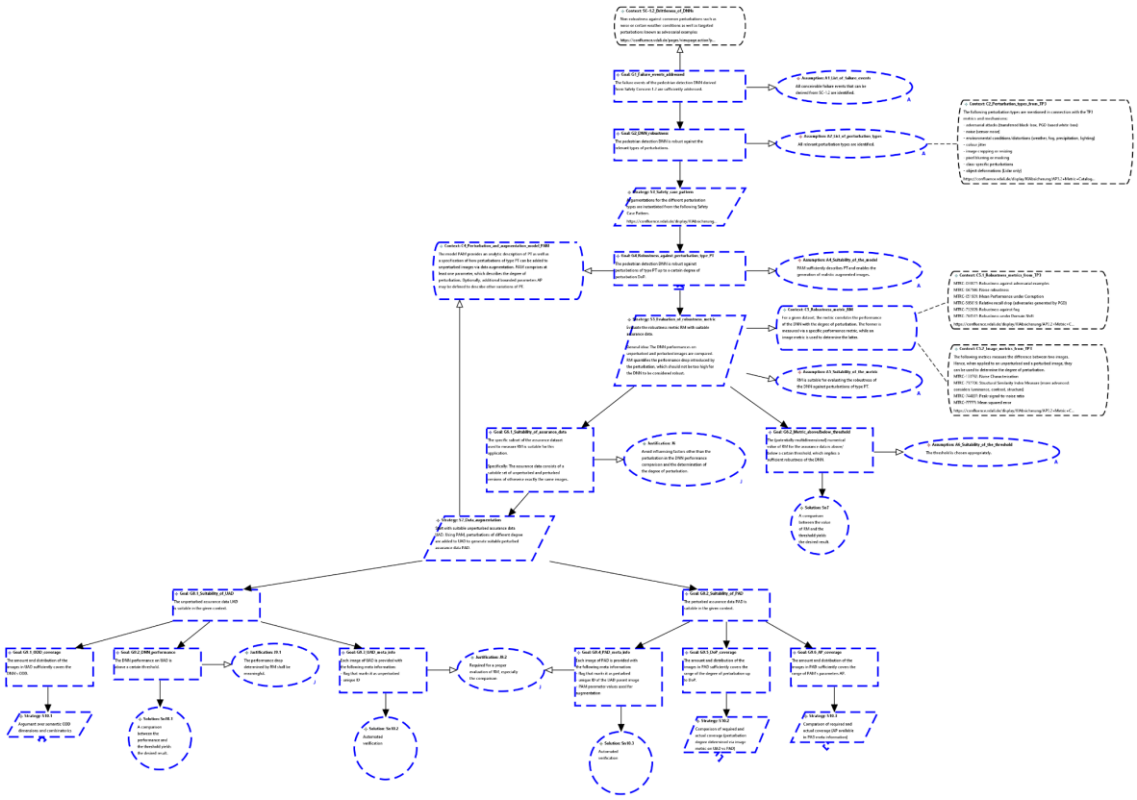


Abbildung 56: Allgemeiner GSN-Graph für die Robustheit eines DNN gegenüber verschiedenen Störungen

Die Robustheit von DNNs und die entsprechende Sicherheitsargumentation werden im Evidence Workstream “Brittleness of DNNs”, zu dem Valeo als Safety Buddy beiträgt, weiter untersucht. Ferner hat Valeo durch Feedback und Diskussionen zur Weiterentwicklung eines GSN-Graphen beigetragen, der die hinreichende Abdeckung von “corner cases” behandelt.

Im Zusammenhang mit der Neuausrichtung von AP4.3 wurde entschieden, dass **E4.3.5** nicht in der ursprünglich vorgesehenen Form fortgesetzt, sondern die Kapazität auf neue Aufgaben umverteilt wird (siehe auch oben bei AP4.3). Die bereits vorhandenen Ergebnisse aus der ersten Hauptarbeitsphase werden derzeit abschließend dokumentiert.

Bei der Arbeit an **E4.3.6** lag der Fokus auf der Integration der Ergebnisse der Evidence Workshops in die Struktur der Gesamtargumentation aus E4.2.2. In diesem Zusammenhang hat Valeo den GSN-Graphen zum TP3-Mechanismus “Meta LRP” aus der Arbeit als Safety Buddy des entsprechenden Evidence Workshops vorgestellt und eingehend erläutert. Außerdem wurden mit den anderen beteiligten Partnern die weiteren GSN-Fragmente besprochen und mögliche Anknüpfungspunkte in der Gesamtargumentation identifiziert. Für die Integration des Mechanismus “Local Uncertainty Realism” hat Valeo ein konkretes Konzept erarbeitet, das aus zwei Hauptaspekten besteht. Einerseits muss die Frage beantwortet werden, wie die Konfidenzinformation eines DNN zuverlässig gemacht bzw. richtig kalibriert werden kann. Dieser Frage geht der Evidence Workstream “Unreliable Confidence

Information" nach. Andererseits muss näher betrachtet werden, wofür die Konfidenzinformation in der Systemarchitektur konkret verwendet wird und ob diese Verwendung sicher ist. Dies setzt den ersten Aspekt, also eine ausreichende Zuverlässigkeit der Konfidenzinformation, voraus. Die beschriebene Herangehensweise ermöglicht die Verknüpfung eines DNN-spezifischen Safety Concerns mit der klassischen Sicherheitsargumentation, bei der sich Gefahren und Fehlerereignisse schrittweise durch die Systemarchitektur verbreiten.

Stand der Arbeiten (31.12.2021):

Als Beitrag zur Finalisierung der Neuausrichtung von **AP4.3** hat Valeo im Rahmen eines Meetings den aktuellen Stand des Evidence Workstreams "Brittleness of DNNs" präsentiert und die enge Verbindung zu den Aktivitäten in AP4.3 aufgezeigt.

Das neu definierte **Cluster "Creating Things"** umfasst sämtliche Aktivitäten, bei denen Teile der Sicherheitsargumentation in Form von GSN-Graphen erzeugt werden. Valeo hat insbesondere zu zwei Sicherheitsaspekten beigetragen: zur Robustheit von neuronalen Netzen im Rahmen des **Evidence Workstreams "Brittleness of DNNs"**, der zum Prozess **P4** gehört, und zur Datenargumentation, die im Rahmen von **E4.3.4** entwickelt wird. Im Evidence Workstream übernimmt Valeo die Rolle als Safety Buddy und hat ein Konzept zur Instanziierung des zuvor entwickelten Safety Case Patterns erstellt, das die Robustheit eines gegebenen DNNs basierend auf Tests mit augmentierten Daten beurteilt, sowie einen zusätzlichen GSN-Graphen konzipiert, in dem es um die Robustifizierung während der Entwicklung geht. Für E4.3.4 hat Valeo den zuvor erstellten GSN-Graphen für Trainingsdaten zu einem Safety Case Pattern verallgemeinert, das sich mit der Datensatzeigenschaft der Repräsentativität beschäftigt (siehe Abbildung unten). Hierdurch wird sichergestellt, dass die Datenverteilung entlang der semantischen Dimensionen der Zieldomäne (z.B. ODD) geeignet ist. Dabei spielen zwei Aspekte eine Rolle: die Grundabdeckung bis zu einem definierten Schwellwert und ein angemessenes Gleichgewicht zwischen Ausprägungen mit unterschiedlichen Auftrittshäufigkeiten. Der Grundgedanke hinter dem Safety Case Pattern liegt in der Beobachtung begründet, dass die Sicherheitsargumentation hinsichtlich Datenrepräsentativität für verschiedene Arten von Daten (Training, Validierung, Test) und ihre jeweiligen Domänen sehr ähnlich ist und somit allgemein erfasst werden kann. Das Pattern kann vielfach instanziiert werden, jeweils für verschiedene spezifische Datensätze, wodurch die vereinfachte Erstellung einer modularisierten Sicherheitsargumentation ermöglicht wird. Ferner hat Valeo im Rahmen von E4.3.4 sichergestellt, dass die aus der Literatur bekannten "Data Management Safety Requirements" ausreichend in der Argumentation berücksichtigt sind.

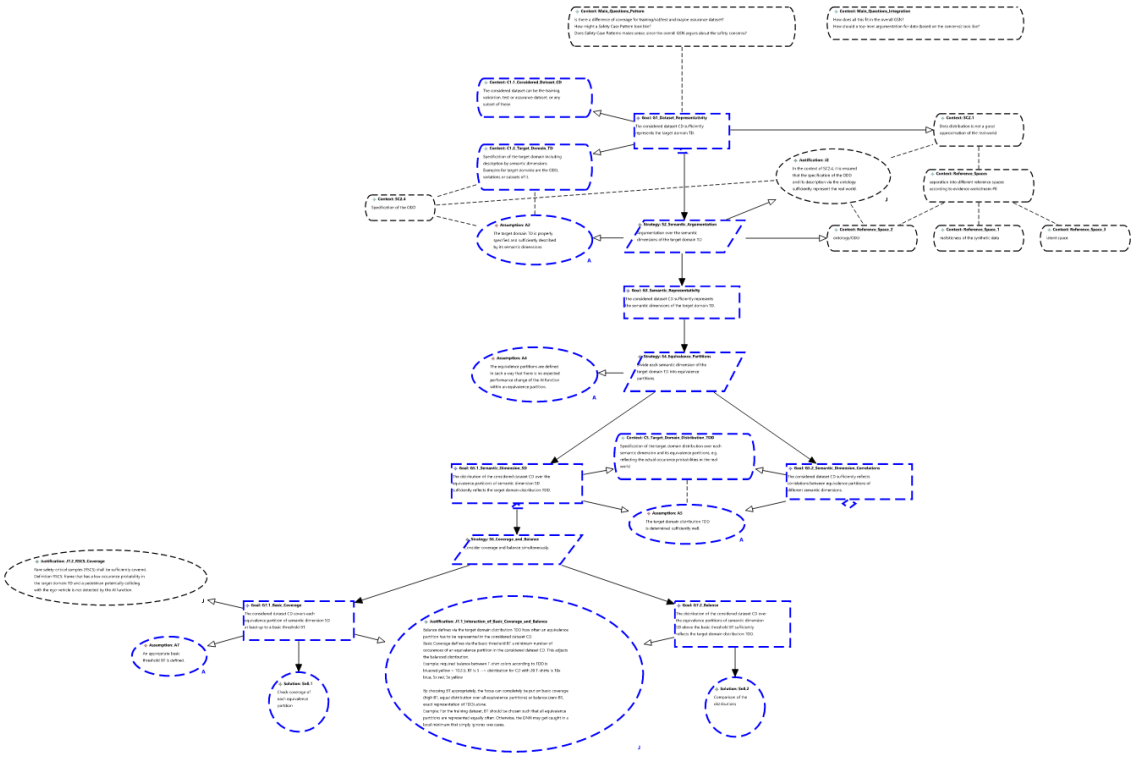


Abbildung 57: Übersicht des Safety Case Patterns für die Repräsentativität eines Datensatzes

Das neu definierte **Cluster “Putting Things Together”** kümmert sich um die Integration der im Projekt entwickelten Teilsicherheitsargumentationen in Form von GSN-Graphen in die Gesamtargumentation. Dafür müssen geeignete Anknüpfungsstellen identifiziert werden. Valeo hat die Integration mehrerer GSN-Argumentationen untersucht. Für die Umsetzbarkeit des zuvor von Valeo im Evidence Workshop “Meta LRP” entwickelten Argumentationskonzepts gibt es leider bisher keine experimentelle Evidenz, weshalb der zugehörige GSN-Graph nicht in die Gesamtargumentation integriert wird. Für den im Rahmen von E4.3.5 weiterentwickelten GSN-Graphen des AP4.3 MVP wurde zwar eine geeignete Anknüpfungsstelle in der Gesamtargumentation gefunden, das Team hat sich aber dennoch aus mehreren Gründen gegen eine Integration entschieden. Erstens ist der betrachtete Fall sehr vereinfacht und spezifisch, zweitens konzentriert er sich auf eine Monitorfunktion statt auf die eigentliche KI-Funktion für Fußgängererkennung und drittens basiert die Argumentation auf einer Nachweisstrategie, die im Zuge der Neuausrichtung von AP4.3 als veraltet eingestuft wurde. Für die Integration der E4.3.3-Ergebnisse hat Valeo ausführliches Feedback an die Entwickler erarbeitet und somit wesentlich zum Endergebnis beigetragen.

Für die bereits beendeten Arbeiten an **E4.3.5** hat Valeo den Abschlussbericht verfasst und die Ergebnisse sowohl in einem AP4.3 Meeting als auch in einem Projekttreffen präsentiert.

Stand der Arbeiten (30.06.2022):

Im Rahmen des Clusters “Creating Things” hat Valeo die Arbeit an den GSN-Graphen für die bereits zuvor betrachteten Sicherheitsaspekte fortgesetzt. Dazu wurden für den Evidence Workstream “Brittleness of DNNs” in der Funktion als Safety Buddy vier Workshops mit den Entwicklern der relevanten Mechanismen organisiert und durchgeführt, um die jeweiligen Sicherheitsbeiträge zu ergründen. Basierend auf den gewonnenen Erkenntnissen hat Valeo einen Development-GSN-Graphen erstellt, in dem es um die Verbesserung der Robustheit von DNNs gegenüber relevanten Störungstypen durch die Anwendung von Mechanismen zur Robustifizierung während des Entwicklungsprozesses geht. Dabei wird für jeden Mechanismus separat argumentiert, dass er die Gesamtrobustheit tatsächlich erhöht, während negative Nebenwirkungen vernachlässigbar sind. Konkret einbezogen wurden die Mechanismen Robustness via Data Augmentation (VW AugMix) und Robustness via Adversarial Training (FZI). Neben der Erstellung des GSN-Graphen wurden dessen Hauptaspekte und Evidenzen dokumentiert. Außerdem hat Valeo für den Evidence Workstream “Brittleness of DNNs” einen Assessment-GSN-Graphen erstellt, für den das bereits vorliegende Safety Case Pattern dreifach instanziiert wurde. Die Grundidee besteht darin, die Robustheit eines gegebenen DNNs gegenüber verschiedenen Störungstypen zu beurteilen, indem dessen Performanz auf ungestörten Bildern mit der auf durch Datenaugmentierung generierten gestörten Bildern verglichen wird. Konkret betrachtet wurden das Merantix Testing Framework für den Störungstyp Sun/Brightness, Opel Distorted Images Assessment für Störungen vom Typ Random Noise und das Neurocat Corruption Framework für den Störungstyp Local Motion Blur. Basierend auf experimentellen Ergebnissen für ein Basis-DNN und ein durch AugMix robustifiziertes DNN wurde die Robustheit jeweils beurteilt und verglichen. Die bei der Instanziierung gewonnenen Erkenntnisse halfen dabei, das zugrundeliegende Safety Case Pattern zu verbessern. Die folgende Abbildung zeigt eine Übersicht des erstellten GSN-Graphen.

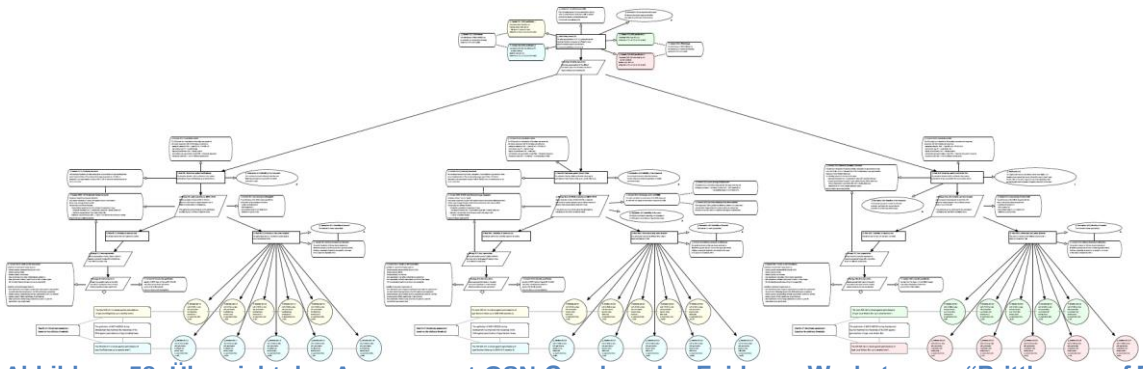


Abbildung 58: Übersicht des Assessment-GSN-Graphen des Evidence Workstreams “Brittleness of DNNs”

Als weitere Aktivität im Cluster “Creating Things” hat Valeo im Rahmen von E4.3.4 zur Verbesserung aller vier Safety Case Patterns (übergeordnetes Dataset Properties Pattern sowie GSN-Graphen zu den Eigenschaften Accurateness, Fidelity und

Representativity) und der beiden Instanziierungen (Fidelity und Representativity) durch umfangreiches Feedback und Umsetzung vieler notwendiger Änderungen beigetragen. Außerdem hat Valeo als Grundlage für eine Richtlinie zur Aufteilung von Datensätzen (Training, Validierung, Test) eine Übersicht über Best Practices erstellt, die während des Projekts Anwendung fanden.

Das Cluster “Putting Things Together” verfolgte das Ziel, alle im Projekt erstellten Teilargumentationen in Form verschiedener GSN-Graphen in die Gesamtargumentation zu integrieren und dabei die schrittweise Ableitung von Sicherheitsanforderungen (Blickwinkel der Sicherheit) mit der Mitigation von DNN-spezifischen Safety Concerns (Blickwinkel der KI) zu verbinden. Letzteres hat Valeo durch die Erarbeitung und Umsetzung eines geeigneten Konzepts wesentlich vorangebracht (siehe E4.2.5 oben). Für die Einbindung der GSN-Graphen, die hauptsächlich aus den Evidence Workstreams stammen, erweiterte und etablierte Valeo eine systematische Methodik, die auf Safety Contracts beruht. Für die Umsetzung war Valeo an der Identifizierung der Anknüpfungsstellen für sämtliche Teilargumentationen beteiligt. Außerdem hat sich Valeo um die Konsolidierung der verschiedenen GSN-Graphen in einem gemeinsamen Repository gekümmert. Die folgende Abbildung zeigt eine Übersicht der AI-Component-Ebene der Gesamtargumentation, in der die Anknüpfungsstellen verschiedener Evidence Workstreams farblich gekennzeichnet sind.

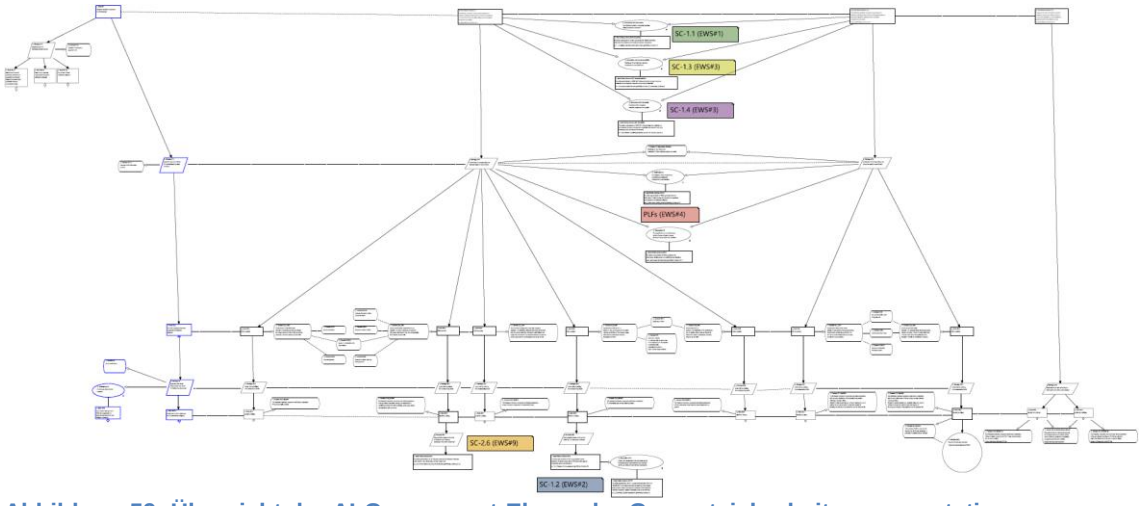


Abbildung 59: Übersicht der AI-Component-Ebene der Gesamtsicherheitsargumentation

Im Rahmen des Clusters “Expert Knowledge” hat sich Valeo an der Faltung der Experten-Leitfragen mit den GSN-Fragmenten beteiligt. Dabei erfolgte eine Untersuchung jedes im Projekt erstellten GSN-Graphen hinsichtlich Datenanforderungen und -aufteilung sowie Machine Learning Safety Requirements und deren Metriken. Die Ergebnisse wurden als konstruktives Feedback an die Ersteller der GSN-Graphen übergeben.

Im Zusammenhang mit dem letzten Projekttreffen und der Abschlussveranstaltung hat Valeo wesentlich zur Erstellung von Postern für die Gesamtargumentation, den

Evidence Workstream “Brittleness of DNNs” und E4.3.4 beigetragen sowie die beiden letztgenannten Poster präsentiert.

AP4.4 Entwicklung von Testmethoden zur Bestätigung der Eingaberaumabdeckung, der Wirksamkeit der Absicherungsmaßnahmen und der Extrapolierbarkeit von Testergebnissen innerhalb des Use Cases (16 PM)

Aufgaben Valeo:

Ausführbarer und dokumentierter Methodensatz zur Identifikation fehlender Testdaten (E4.4.1a)

• Hier gibt es bereits Ideen über Schätzungen von Dichteverteilungen von Datensätzen im N-Dimensionalen-Raum. Dabei sollen redundante bzw. fehlende Stellen identifiziert werden. Daraus sollen Lücken in den Testdaten aufgedeckt werden. Durch die Erzeugung von weiteren Testdaten sollen die Lücken mittels eines iterativen Prozess geschlossen werden.

Nachweis der Wirksamkeit der einzelnen Maßnahmen in Bezug auf die Sicherheitsziele (E4.4.2)

• Untersuchung und Evaluierung von Wirksamkeit der Methoden & Maßnahmen anhand von fotorealistischen Daten. Bei Verfügbarkeit von Realdaten im Projekt werden diese anstelle der fotorealistischen Daten verwendet. Die geplanten Aufwände bleiben konstant.

Extrapolation der Testergebnisse in Bezug auf Sicherheitsziele und Guarantees im Assurance Case (E4.4.3b)

• Analyse der Extrapolation bzw. Abdeckung im Eingaberaum in Bezug auf die Garantien des Safety Contracts.

Stand der Arbeiten (31.12.2019):

Ausführbarer und dokumentierter Methodensatz zur Identifikation fehlender Testdaten (E4.4.1a)

• Die Arbeiten haben begonnen und befassen sich im ersten Schritt mit Konzepten zur Identifikation von fehlenden Testdaten. Anschließend werden die Konzepte implementiert

Nachweis der Wirksamkeit der einzelnen Maßnahmen in Bezug auf die Sicherheitsziele (E4.4.2)

• Im Zusammenhang mit 4.4.1a muss die Wirksamkeit bei der Konzepterstellung berücksichtigt werden.

Extrapolation der Testergebnisse in Bezug auf Sicherheitsziele und Garantien im Assurance Case (E4.4.3b)

• Arbeiten haben noch nicht begonnen

Stand der Arbeiten (30.06.2020):

E.4.4.2: Es wurde eine Teststrategie erarbeitet, die basierend auf Unsicherheitsmetriken die Daten extrahiert, die vermehrt in den Testdaten präsent sein sollten. Hierzu werden die DNN Modelle während der Anwendung mittels Unsicherheitsmodellierung überwacht. Überschreitet die Unsicherheitsmodellierung einen gewissen Schwellwert, wird dies als Maß verwendet, um darüber zu entscheiden, ob diese Daten vermehrt in den Testdaten präsent sein sollen oder nicht. Für die Bewertung der Unsicherheiten werden keine Labels benötigt, wodurch es möglich ist, dem DNN eine Vielzahl von Eingangsdaten zu präsentieren. Diese Teststrategie befindet sich noch in der Konzeptphase und Änderungen/Erweiterungen können noch berücksichtigt werden.

E.4.4.2: Im Rahmen des Proof of Project (PoP) wurde ein GSN Graph erstellt, der einen Variational Auto Encoder (VAE) als Evidenz nutzt. Die Wirksamkeit dieses VAEs soll im Rahmen dieses UAP überprüft werden, um den VAE für den PoP freigegeben zu können. Konkret soll die Korrelation zwischen Reconstruction Error des VAE und der Genauigkeit der DNN Ausgabe überprüft werden. Hierzu muss jedoch erst der VAE implementiert werden, was im TP3 bis zum Oktober 2020 erfolgen soll, bevor hiermit praktisch begonnen werden kann.

Stand der Arbeiten (31.12.2020):

Valeo hat die Wirksamkeit des Variational Autoencoders (VAE) von Bosch (aus dem AP 3.5) auf deren Wirksamkeit überprüft. Im Folgenden werden die Ergebnisse beschrieben:

Wie im GSN-Diagramm Testen des PoPC beschrieben, sollte die VAE zur Laufzeit verwendet werden, um Out-of-Domain-Daten zu erkennen.

Out-of-Domain-Daten sind Daten, die eine bestimmte andere Datenverteilung im Vergleich zu den Trainingsdaten aufweisen und somit einen negativen Einfluss auf die DNN Accuracy haben. Bislang gilt es als nicht bewiesen, dass die VAE dazu in der Lage ist. Die Wirksamkeit wurde in dieser Arbeit untersucht. Als Basis wurde die VAE von Bosch verwendet[8].

Aufgrund der fehlenden Diversität der Projektdaten in Bezug auf Wetter- und Lichtverhältnisse, die die Realisierung verschiedener Domänen erlauben würde, wurde der Synthia-Sequenzdatensatz verwendet. Er enthält z.B. die Domänen Morgendämmerung, Nebel, Regen, Winter, Sommer, Frühling, Herbst, Nacht usw. Zum Trainieren wurde die Sequenz (SEQ) 1+2 und zum Testen 5+6 verwendet.

[8] https://gitlab.com/kia2/tp3/ap3.5/e3.5.1_e3.5.4_e3.5.6_vae_reconstruction-error_likelihood_bosch

Die Auflösung des Synthia beträgt 768, 1280 Pixel, daher musste die VAE für das Padding der Conv-Schicht und der Upsample-Schicht leicht angepasst werden. Der latente Raum ist 512x29x48 für mean bzw. var. Die VAE wurde mit MSE und KLD (Gewichtung 1 zu 0,1) trainiert. Das Training wurde mit der Domain "dawn" (SEQ 1+2) durchgeführt, d. h. die Domain "dawn" ist in-domain und alle anderen Domains sind out-of-domain. Zur Validierung mit In-Domain-Daten wurde die SEQ 5+6 (dawn) verwendet. Die Testbilder (Out-of-Domain-Daten) wurden aus anderen Domänen als der Morgendämmerung von SEQ 5+6 verwendet. Die VAE wurde für 15 Epochen trainiert.

Die Auswertung über den kompletten Datensatz zeigt ein ähnliches Bild. Das linke Histogramm wurde über die kompletten Trainingsdaten (Morgendämmerung SEQ 1 und 2) erstellt. Das mittlere über die kompletten Validierungsdaten (Morgendämmerung SEQ 4 und SEQ5) und das rechte über Testdaten der Domänen Winternight, Rain, Fall, Softrain der Sequenzen 5 und 6 (d.h. Domänen, die sich signifikant von der Morgendämmerung unterscheiden). Aufgrund des MSE ist es nicht möglich zu bestimmen, aus welcher Domäne das Eingangsbild stammt. Bitte beachten Sie, dass die Anzahl der Bins pro Histogramm ungewollt unterschiedlich gewählt wurde.

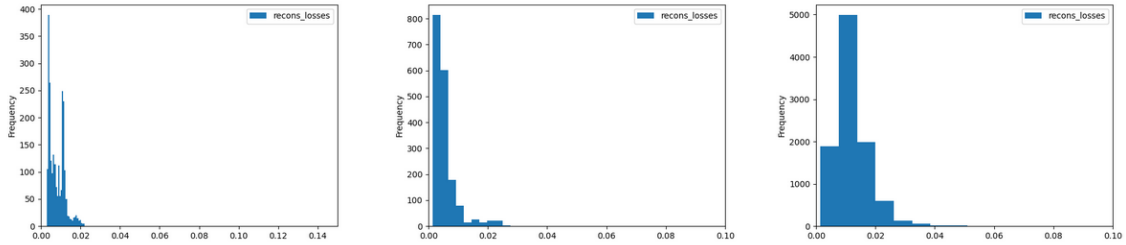


Abbildung 60: Histogramm-Auswertungen

Die Auswertung der KLD (anstelle der MSE) zeigt ein ähnliches Bild. Auch die binäre Kreuzentropie wurde getestet (F.binary_cross_entropy) (anstelle des MSE), aber auch hier das gleiche Bild.

Weiterhin wurde die z-Score-Distanz ausgewertet, die von FZI verwendet wird[9].

Sie nutzt den latenten Raum nur zur Laufzeit, so dass der Decoder nicht ausgeführt werden muss, was letztlich Laufzeit spart. Blau sind die In-Domain-Daten (Morgendämmerung, SEQ5 und 6) und orange die Out-of-Domain-Daten (wie oben beschrieben). Bitte beachten Sie, dass jeweils nur 1000 Bilder verwendet wurden (d.h. nur eine Teilmenge im Vergleich zu den obigen Daten), um die Berechnungszeit zu verkürzen. Dunkelorange zeigt die Überlappung, die idealerweise nicht vorhanden wäre, weil man dann die beiden Domänen leicht trennen könnte. Wie Sie sehen

[9] https://gitlab.com/kia2/tp3/ap3.5/e3.5.1_fzi_ap3.5_object_centric_domain_shift

können, ist dies auch mit dem z-Score-Abstand nicht möglich. Die Unterschiede der Histogramme sind die Anzahl der Nachbarn: von links nach rechts: 1, 5, 10, 20.

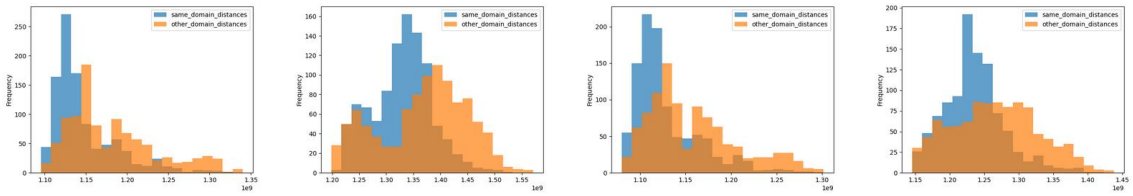


Abbildung 61: z-Score-Abstand

Alternativ zum z-Score-Abstand haben wir auch den Manhattan- und euklidischen Abstand (Anzahl der Nachbarn 10) ausgewertet:

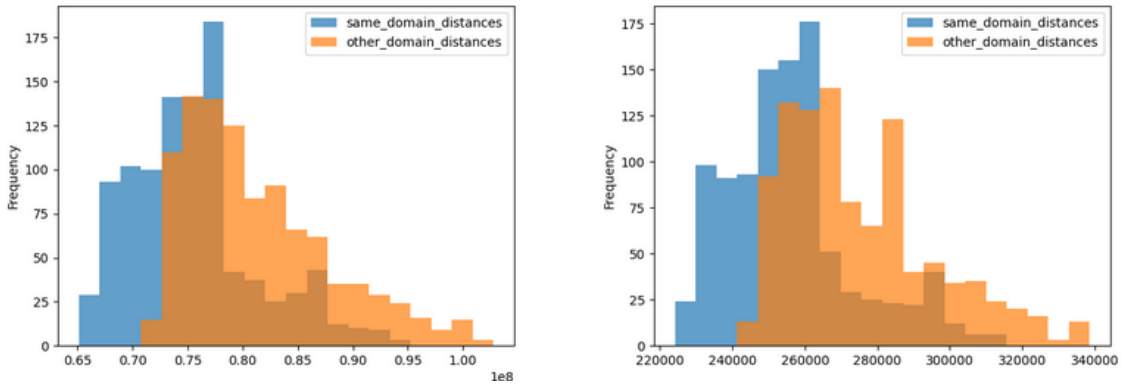


Abbildung 62: Manhattan- und euklidischer Abstand

Zusätzlich haben wir visuell beurteilt, ob es eine Korrelation zwischen der DNN-Genauigkeit und dem Rekonstruktionsfehler (auf Pixelebene) gibt. Wir haben die DeepLabV3+ für die semantische Segmentierung auf Dawn-Daten (In-Domain-Daten / Seq. 1 + 2) trainiert, die die gleichen Daten sind, auf denen die VAE trainiert wurde. Die ersten drei Abbildungen zeigen die In-Domain-Daten und die anderen drei Abbildungen zeigen die Out-of-Domain-Daten. Bitte beachten Sie den Text für jedes Bild innerhalb der Abbildung. Das zweite Bild in jeder Abbildung bedarf möglicherweise einer zusätzlichen Erklärung: Es handelt sich um die MSE-Karte → (Bild - Aufklärung)^2 * x. wobei x nur ein Faktor zur Erhöhung der Sichtbarkeit ist.

In-domain Daten:

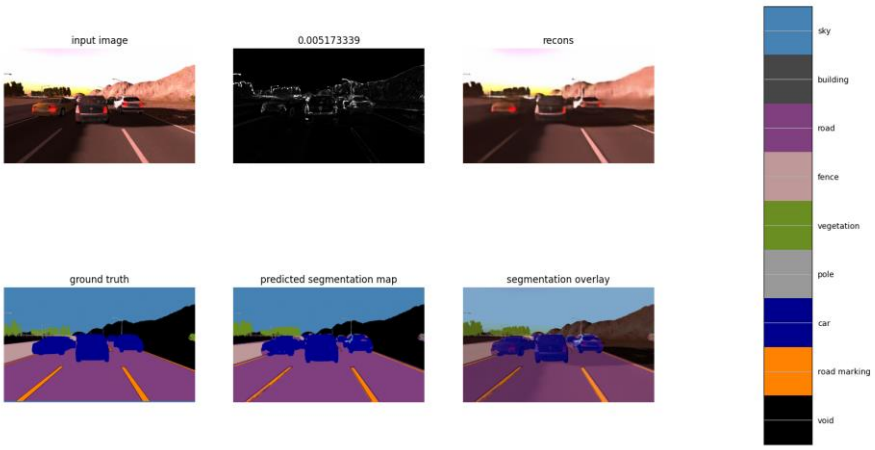


Abbildung 63: Visuelle Darstellung des MSE Errors (mitte), des rekonstruierten Eingangsbildes (rechts), sowie der sich ergebenden semantischen Maske (untere Reihe) für in domain Daten

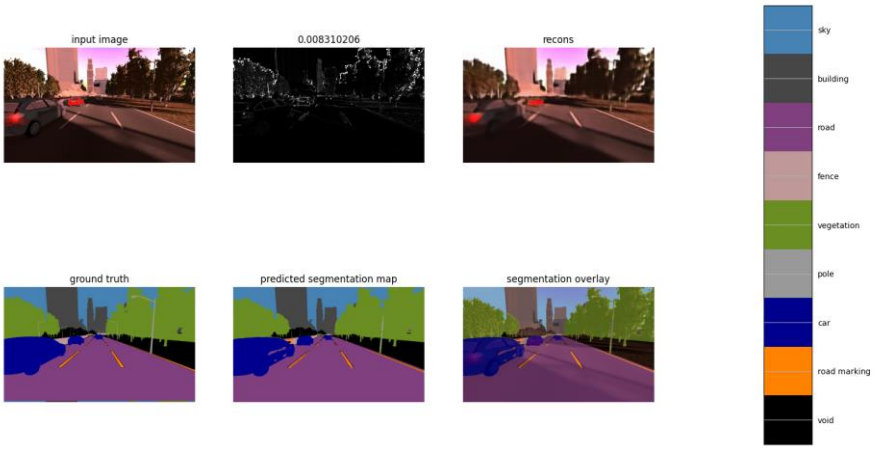


Abbildung 64: Visuelle Darstellung des MSE Errors (mitte), des rekonstruierten Eingangsbildes (rechts), sowie der sich ergebenden semantischen Maske (untere Reihe) für in domain Daten

Out-domain Daten:

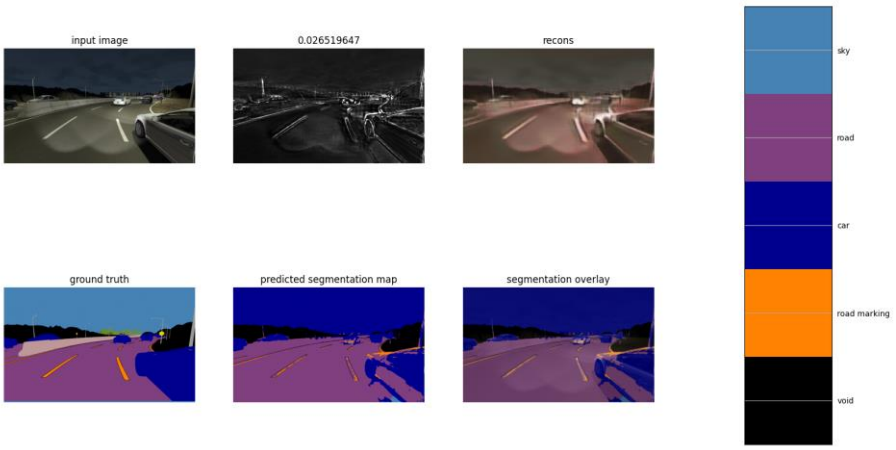


Abbildung 65: Visuelle Darstellung des MSE Errors (mitte), des rekonstruierten Eingangsbildes (rechts), sowie der sich ergebenden semantischen Maske (untere Reihe) für out of domain Daten

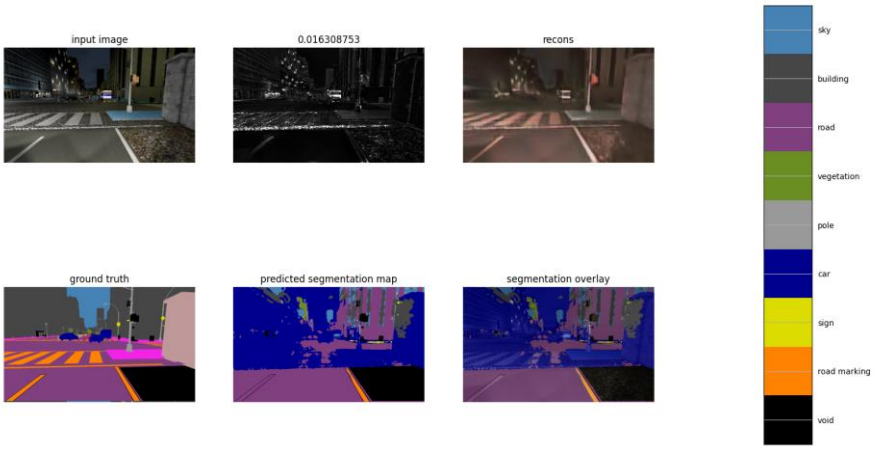


Abbildung 66: Visuelle Darstellung des MSE Errors (mitte), des rekonstruierten Eingangsbildes (rechts), sowie der sich ergebenden semantischen Maske (untere Reihe) für out of domain Daten

Es ist zu erkennen, dass der Rekonstruktionsfehler keine Korrelation zu den Segmentierungskarten hat. Wir haben etwa 50 weitere Bilder analysiert und nach unserem besten Wissen keine Korrelation gefunden. Bei den Out-of-Domain-Beispielen sind die Segmentierungskarten aufgrund der Domänenverschiebung überhaupt nicht verwendbar. Bei diesen Beispielen sind die Rekonstruktionsfehler für Out-of-Domain-Daten sogar höher als für In-Domain-Daten (die Bilder wurden zufällig ausgewählt).

Schlussfolgerung: In unseren Experimenten haben wir Dawn (Synthia-Sequenzdatensatz) als in-domain und Winternight, Rain, Fall, Softrain als out-of-

domain verwendet. Weder mit MSE, KLD, binärer Kreuzentropie noch mit dem z-Score, Manhatten und euklidischer Distanz ist es möglich zu erkennen, ob die Eingabedaten aus in-domain (Trainingsdomäne) oder out-of-domain (andere Domäne als Trainingsdomäne) stammen. Außerdem haben wir das DeepLabV3+ auf denselben Daten wie die VAE trainiert, aber es konnte keine Korrelation zwischen IoU und Rekonstruktionsfehler (auf Pixelebene) festgestellt werden. Somit kann die Effektivität für den vorgesehenen Einsatz der VAE (siehe GSN Graphen hier Testen des PoPC) nicht bestätigt werden.

Stand der Arbeiten (30.06.2021):

4.4.2

Um die Wirksamkeit der Mechanismen aus TP3 nachzuweisen, hat sich gezeigt, dass eine Einzeluntersuchung mit einem hohen Aufwand verbunden ist. Aus diesem Grund sollte es möglich sein, die Mechanismen im Rahmen eines Benchmarks zu vergleichen. Der Beitrag war die Erstellung und der Aufbau von Benchmarks.

4.4.4

Die Hochrechnung der Testergebnisse wird im Kontext der im Projekt betrachteten Gesamtfunktion bewertet. Dazu muss eine Gesamtfunktion erstellt werden, mit der die Realisierung eines Fußgänger-Notbremssystems (Anwendungsfall im Projekt) möglich ist. Die Struktur der Gesamtfunktion lässt sich grob in 3 Module unterteilen: Wahrnehmen, Planen und Steuern.

Für die Arbeit in diesem UAP konzentrieren wir uns auf Perception und Planning. Für Perception muss eine Detektion des Fußgängers durchgeführt werden. Diese Detektion beinhaltet mindestens eine Erkennung des Fußgängers im Bildbereich sowie eine Entfernungsinformation. Mit Hilfe der Kameraparameter kann eine 3D-Position des Fußgängers berechnet werden. Noch genauer wäre die Verwendung einer 3D-Bounding-Box-Erkennung, wie sie in AP1.4 entwickelt wurde. Ein weiterer Teil des Wahrnehmungsmoduls ist die Intentionserkennung, die oft eine Pose-Erkennung als Vorläufer verwendet. Die Intentionserkennung ermöglicht die Schätzung einer Trajektorie des Fußgängers, die mit der Trajektorie des Ego-Fahrzeugs abgeglichen werden muss (Planung). Wird eine Überschneidung dieser Trajektorien berechnet, muss das Notbremssystem aktiv werden. Die bisherigen Testergebnisse im Projekt beziehen sich hauptsächlich auf die Fußgängererkennung in der Bilddomäne (2D). Die anderen beschriebenen Komponenten des Moduls Wahrnehmen und Planen werden weitgehend ignoriert. Die Annahme von fiktiven Algorithmen zur Erfüllung der fehlenden Komponenten ist Teil der Untersuchung der Hochrechnung der Testergebnisse. Die Ergebnisse erlauben eine Beurteilung der Hochrechnung der Testergebnisse.

Stand der Arbeiten (31.12.2021):

Um die Wirksamkeit der einzelnen Sicherheitsmechanismen nachzuweisen, ist eine gewisse statistische Signifikanz erforderlich. Benchmarks berücksichtigen dies in der Regel nicht, da nur einzelne Leistungswerte angegeben werden und keine Mittelwertbildung über mehrere Läufe stattfindet. So kann ein vermeintlich positiver Effekt eines Sicherheitsmechanismus lediglich ein statistischer Ausreißer nach oben sein. Das Hauptproblem bei der Betrachtung der statistischen Validität ist ein praktisches. Das erneute Trainieren von DNNs erfordert einen nicht unerheblichen zusätzlichen Aufwand an Rechenleistung und Zeit. Aus diesem Grund untersuchen wir die Möglichkeit der Trunkierung durch Verwendung einer zyklischen Lernrate als eine Art Finetuning am Ende des Trainings. Indem die Lernrate abrupt erhöht und dann langsam verringert wird, wird ein neues lokales Optimum in der “loss landscape” gefunden, ähnlich wie bei einem Training von Grund auf. Der Vergleich mit Mittelwert und Varianz gibt Aufschluss über die Effektivität dieser Methode.

Wir führen unsere Experimente mit dem DeepLabV3+ und dem BDD100k-Datensatz durch. 4 Modelle aus dem Scratch wurden auf dem Trainingsdatensatz von BDD trainiert und auf den Validierungsdaten evaluiert. Die mIoU-Werte sind in der ersten Spalte angegeben. Die Spalten 2 und 3 zeigen die mIoU-Werte für das Finetuning von Modell 1 bzw. Modell 2. Das Finetuning wurde mit 15 Epochen durchgeführt, wobei ein Zyklus 5 Epochen dauert. Nach den 5 Epochen wurde die Lernrate wieder auf die ursprüngliche Lernrate von 0,004 erhöht. Während des 5-Epochen-Zyklus wurde die Lernrate mit Hilfe des polynomialen Lernraten-Schedulers reduziert. Die Kontrollpunkte am Ende eines jeden Zyklus wurden gespeichert und in der Tabelle unter "Modell 1 - Modell 4" angezeigt.

Tabelle 10: Vergleich von Mittelwert und Varianz für 4 Trainings- und Epochenläufe mit einer zyklischen Lernrate und dem Polynomlernratenplaner mIoU

	Training from scratch	Finetuned from model 1	Finetuned from model 2
Model 1	57.38	57.38	58.49
Model 2	58.61	57.35	58.61
Model 3	57.73	57.59	58.07
Model 4	58.00	57.80	57.86
mean	**57.93**	**57.53**	**58.26**
var	**0.2025**	**0.0328**	**0.0930**

Aus den bisherigen Ergebnissen lässt sich schließen, dass die Varianz der feinabgestimmten Modelle geringer und zum Teil deutlich geringer ist als beim Training von Grund auf. Daher kann das vorgestellte Finetuning-Verfahren nur als erste Annäherung an die Anwendung als Beweis für die Wirksamkeit von Sicherheitsmechanismen dienen. Durch eine Änderung der Parameter des Finetuning-Protokolls (z. B. Änderung des Zeitplans für die Lernrate, der Dauer des Zyklus, der Anzahl der Zyklen und der Ursprungs-Lernrate) kann die Varianz wahrscheinlich kontrolliert werden. Experimente zur Bewertung der Auswirkungen dieser Parameter sind geplant.

Stand der Arbeiten (30.06.2022):

Um die Wirksamkeit der einzelnen Sicherheitsmechanismen nachzuweisen, ist eine gewisse statistische Signifikanz erforderlich. Benchmarks berücksichtigen dies in der Regel nicht, da nur einzelne Leistungswerte angegeben werden und keine Mittelwertbildung über mehrere Läufe stattfindet. So kann ein vermeintlich positiver Effekt eines Sicherheitsmechanismus lediglich ein statistischer Ausreißer nach oben sein. Das Hauptproblem bei der Betrachtung der statistischen Validität ist ein praktisches. Das Umlernen von DNNs erfordert einen nicht unerheblichen zusätzlichen Aufwand an Rechenleistung und Zeit. Aus diesem Grund untersuchen wir die Möglichkeit der Trunkierung durch Verwendung einer zyklischen Lernrate als eine Art Finetuning am Ende des Trainings. Indem die Lernrate abrupt erhöht und dann langsam verringert wird, wird ein neues lokales Optimum in der Verlustlandschaft gefunden, ähnlich wie bei einem Training von Grund auf. Der Vergleich mit Mittelwert, Varianz und Orakeltest gibt Aufschluss über die Effektivität dieser Methode. Wir führen unsere Experimente mit dem DeepLabV3+ und dem BDD100k-Datensatz durch.

Tabelle 11: Vergleich von Mittelwert, Varianz und Oarcle-Test-Werten für verschiedene Lernraten und Zykluslängen. Mittelwert und Oacle Test sind mIoU-Werte in %. Baseline bezieht sich auf die 5 Trainings von Grund auf

Learning rate	cycle length	mean	variance	oracle test
0.004	1	58.29	0.008	64.04
0.005	1	54.64	2.911	69.92
0.006	1	54.61	0.465	67.57
0.004	2	58.35	0.045	64.76
0.005	2	56.08	0.340	67.18
0.006	2	55.63	0.202	67.90
0.003	4	58.57	0.001	63.11
0.004	4	57.47	0.108	67.16
0.005	4	57.73	0.821	67.22
0.006	4	56.77	0.265	67.33
0.003	6	58.58	0.134	65.43
0.004	6	57.85	0.149	66.46
0.005	6	56.97	0.208	66.39
0.006	6	57.39	0.054	66.88
Baseline	-	57.836	0.197	69.25

Tendenziell lässt sich aus den Ergebnissen ableiten, dass die Orakeltestwerte mit einer höheren Lernrate zunehmen, was naheliegend ist, da lokale Minima weiter voneinander entfernt gefunden werden. Eine Erhöhung der Zykluslänge verringert das Orakeltestergebnis, woraus geschlossen werden kann, dass mit zunehmender Zykluslänge ein ähnliches lokales Minimum gefunden wird.

Der Mittelwert sinkt mit zunehmender Lernrate und steigt mit einer höheren Zykluslänge. Wie erwartet, ist die Varianz bei der niedrigsten Lernrate sehr gering. Es ist schwierig, aus den Daten einen Trend über das Verhalten der Varianz bei wechselnder Lernrate und Zykluslänge abzuleiten. Wahrscheinlich ist die Anzahl der Stichproben mit jeweils 5 zu gering, um statistische Ausreißer zu glätten. Eine direkte Übereinstimmung mit der Baseline in Bezug auf Mittelwert, Varianz und Orakeltest ist

nicht zu erkennen. Aus den abgeleiteten Trends ergibt sich, dass eine weitere Erhöhung der Lernrate (zur Erhöhung des Orakeltests) und der Zykluslänge (zur Erhöhung des Mittelwerts) notwendig wäre, um sich der Baseline anzunähern. Dies kann als Ausblick oder Empfehlung für weitere Arbeiten gesehen werden.

AP4.5 KI-Teststrategie & KI-Testplan als Ausgangspunkt für Produktfreigabe (6 PM)

Aufgabe Valeo:

Testplan für die Evaluierung der Fehlerrate der KI-Funktion aus TP1 (E4.5.2)

• Erstellung eines Testplans, insbesondere bzgl. der Tiefensensoren-relevanten Garantien des Assurance Cases und Anwendung der spezifizierten Tests auf die KI-Funktion aus AP1.4.

Stand der Arbeiten (30.06.2021):

Valeo hat bei der Herstellung eines Testplans mitgewirkt. Dabei wurde, unter anderem, das Testobjekt spezifiziert, eine Teststrategie definiert, Testmethoden ausgewählt, einen Prüfstand entworfen und einen groben Zeitplan herstellt.

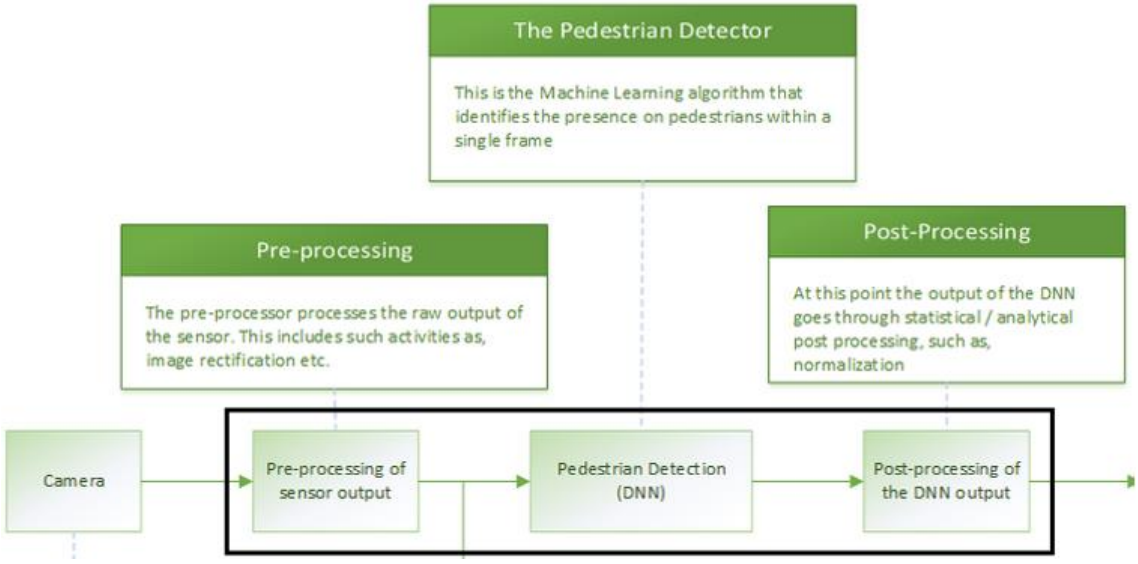


Abbildung 67: Prozesspipeline

Das Testobjekt umfasst eine Vor- und Nachbearbeitung und das KI-System (siehe Bild oben). Der Test zielt dazu, die Effizienz des KI-Systems nachzuweisen. Das Testen basiert vornehmlich auf anforderungsbasiertes Testen und baut auf den Testdatensatz auf, der klassisch für maschinell-lernende Systeme verwendet wird.

Darüber hinaus, hat Valeo dazu beigetragen, den Informationsfluss mit anderen Teilprojekt abzugleichen.

Stand der Arbeiten (31.12.2021):

Die Analyse des “Stand der Technik” und die Bewertung der Performance von Testmethoden standen im Mittel der Arbeit während dieses Halbjahres. Hierfür wurde folgende Punkte betrachtet:

- die Probleme, die bei der Anwendung der Methode gefunden wurden,
- die Klassen von Defiziten/Fehlern, die prinzipiell durch das Verfahren gefunden werden können

Insbesondere würden von Valeo folgende Methoden untersucht:

1. Unsicherheiten für die Erkennung von Anomalien
2. Photometrische Robustheitsschätzung

Im Gegensatz zu der ersten Methode, ist die 2. Methoden eine Testmethode und konnte für die weiteren Schritte der Arbeitsgruppe in Betracht gezogen werden.

Stand der Arbeiten (30.06.2022):

Mit dem Projektabschluss in Sicht lag der Schwerpunkt während dieses Halbjahres auf das Reviewen des Testplanes und der Teststrategie. Darüber hinaus, wurde auf die Erstellung des Posters für die Abschlusspräsentation hingearbeitet.

Teilprojekt 5: Projektmanagement und Dissemination

TP5 gliedert sich in drei Arbeitspakete und umfasst Management- und Koordinationsaufgaben (AP5.1), die Ergebnisverbreitung (AP5.2) sowie die Kommunikation mit Normungsgremien (AP5.3).

AP5.1: Projektmanagement

Übergeordnete Aufgabe der technischen Gesamtkoordination ist die inhaltliche und zeitliche Organisation der technischen Arbeiten sowie die Sicherstellung der Qualität der Ergebnisse. Hierzu werden die Arbeiten der einzelnen Arbeitspakete inhaltlich abgestimmt und nachverfolgt sowie der Arbeitsfortschritt hinsichtlich der definierten Meilensteine (siehe auch Kapitel 4.7) bestimmt werden. Bei inhaltlichen Konfliktsituationen und Planabweichungen werden Vorschläge für adäquate Gegenmaßnahmen erarbeitet.

Diese Aufgaben werden durch das Projektmanagementteam (siehe Kapitel 5.2.2) mit Unterstützung eines Projektbüros (siehe GUA1) wahrgenommen.

Der aktuelle Stand der Arbeiten und Ergebnisse (sowie im Falle von inhaltlichen oder zeitlichen Abweichungen) vorgeschlagene Gegenmaßnahmen werden periodisch für den Steuerkreis und den Fördermittelgeber aufbereitet und vorgestellt.

Der Austausch im Projektmanagementteam findet im Kern in Form von physischen (ca. 10 p.a.) und virtuellen Projekttreffen (Web-/Telefonkonferenzen) statt. Die Organisation und Koordination dieser Projekttreffen umfasst Aufgaben der Terminplanung und -abstimmung, die Unterstützung der inhaltlichen Vorbereitung der Treffen (Erstellen einer Agenda) sowie deren Nachbereitung (Erstellen von Protokollen inklusive Formulierung von Aufgaben sowie deren Nachverfolgung).

Die gesamte Ergebnisdokumentation, die Erstellung von Berichten sowie die Delegation von Aufgaben – innerhalb des Projektmanagementteams sowie für das gesamte Konsortium – wird technisch über eine Projektmanagement-Plattform ermöglicht.

AP5.2: Ergebnisverbreitung

Ziel von AP5.2 ist die Kommunikation des Vorhabens und dessen Ergebnissen nach außen.

Eine Hauptaufgabe besteht in der zielgruppengerechten Aufbereitung und Zur-Verfügung-Stellung von relevanten Projektinformationen durch geeignete Kommunikationsmittel.

Zum anderen werden Veranstaltungen (Halbzeit- und Abschlusspräsentation) konzipiert, vorbereitet und umgesetzt, zu denen das Projektkonsortium unterschiedliche Adressatengruppen aus dem projektexternen Umfeld einlädt und Projektergebnisse vorstellt.

Die Formulierung der inhaltlichen Anforderungen an die Ergebnisverbreitung findet in starker Abstimmung mit dem AP5.1 an statt. Die konkrete Umsetzung von AP5.2 soll durch ein Projektbüro erfolgen (siehe GUA2). Die formale Verantwortung für den GUA, die Priorisierung der Aufgaben und die Abnahme der Leistungen erfolgt durch den Projektkoordinator.

AP5.3: Kommunikation mit Normungsgremien

AP5.3 unterstützt das Vorhaben im Allgemeinen und TP4 im Besonderen bei der Kommunikation mit relevanten Normungsgremien und Zertifizierungsstellen.

Hierfür gilt es zunächst, relevante Gremien, Aktivitäten und Organisation zu identifizieren. Ausgangspunkt hierfür sind die zu Vorhabenbeginn bekannten, für das Vorhaben KI-Absicherung relevanten Normen (siehe Tabelle 33). In Anbetracht der immer greifbareren Markteinführung automatisierter Fahrfunktionen ist davon auszugehen, dass sich über die Projektlaufzeit hinweg relevante Normierungs- und Standardisierungsaktivitäten – sowohl im Themenbereich der Absicherung als auch im Technologiefeld KI – ausweiten werden. Diese gilt es zu beobachten.

Dem AP5.3 trägt darüber hinaus die Verantwortung dafür, dass das „Zugehen“ und „Einbinden“ der relevanten Akteure projektweit abgestimmt stattfindet. Hierzu wird AP5.3

- in Abstimmung mit TP4 die relevanten Projektergebnisse identifizieren,
- die Ergebnisse über die AP hinweg (soweit sinnvoll) bündeln und in Zusammenarbeit mit AP5.2 für Dritte geeignet aufbereiten sowie
- einen Freigabe- und Veröffentlichungsprozess für diese Ergebnisse aufsetzen und diesen Prozess über die Projektlaufzeit hinweg betreuen.

Gegeben die Inhalte, die es aus dem Vorhaben KI-Absicherung in Richtung von Normungsgremien, Zertifizierungsstellen und Standardisierungsaktivitäten zu kommunizieren gibt und unter Berücksichtigung der Anforderung an die Art und Aufbereitung der Ergebnisse aus KI-Absicherung, die dort eingebracht werden können, wird AP5.3 eine Kommunikations-Roadmap erstellen und fortschreiben. Soweit möglich wird AP5.3 auch Impulse zur Gründung neuer Gremien im Bereich der Normung und Standardisierung geben.

Mit Blick auf die entgegengesetzte Kommunikationsrichtung kommt dem AP5.3 auch die Rolle des neutralen Erst-Ansprechpartners für Gremien und Zertifizierungsstellen, die an den Ergebnissen des Vorhabens KI-Absicherung interessiert sind, zu.

Die inhaltliche Erarbeitung und fachliche Kommunikation der zu kommunizierenden Ergebnisse erfolgt durch die Experten aus den jeweiligen Arbeitspaketen - maßgeblich AP4.2, AP4.3 und AP4.5.

Die definierten Aktivitäten sollen im Wesentlichen – unter Anleitung jeweils eines Partners aus dem Konsortium aus der Gruppe der OEM sowie der Tier-1 – durch ein Projektbüro (GUA1 und GUA2) erfolgen.

AP5.1: Projektmanagement (3 PM)

Stand der Arbeiten (31.12.2019)

Im Rahmen des Projektmanagements wurden zur Koordination von technischen und organisatorischen Inhalten zahlreiche Treffen (monatlich) und Telefonkonferenzen (wöchentlich) durchgeführt. Für den Projekt-Übergreifenden Prozess „Entwicklung und Bereitstellung von DNN Modellen“ (P2) für den Valeo verantwortlich ist, wurden erste Abstimmungen durchgeführt. Hierzu zählt die Koordination der Aufgaben, die für die Veröffentlichung von DNN-Modellen, aus TP1 an andere Partner des Projektes, benötigt werden. Dies umfasst die Strukturierung der Repositories in *Gitlab,* sowie die Vereinheitlichung von Code-Templates zum Trainieren und Testen von DNN Modellen.

Stand der Arbeiten (30.06.2020):

Neben den Aufgaben im Projektmanagement, wozu auch zahlreiche Telefonkonferenzen gehören, wurde der übergreifende Prozess P2 „Entwicklung und Bereitstellung von DNN Modellen“ vorangetrieben. Es wurden viele Koordinierungsaufgaben durchgeführt und den Austausch zwischen den APs, vor allem den Entwicklungs-APs 1.3 bis 1.5, vorangetrieben. Im Rahmen des P2 wurde das Release Management zum M12 umgesetzt, in dem erste entwickelte DNN Modelle mit trainierten Gewichten weitergegeben wurden. Für die Weitergabe der Modelle wurde eine ausführliche Dokumentation erstellt, um eine einheitliche und einfache Anwendbarkeit zu gewährleisten. Diese Dokumentation beinhaltet Vorgaben bei der Ordnerstruktur, Aufbau des ReadMe Files sowie zahlreiche Python Skripte zur Verwendung der Daten aus TP2 sowie zum Ausführen der Inferenz und des Trainings der DNN Modelle: https://gitlab.com/kia2/code/templates

Stand der Arbeiten (31.12.2020):

Im Fokus stand der Prozess P2 „Entwicklung und Bereitstellung von DNN Modellen“. Zur Gewährleistung der Vergleichbarkeit der DNN Prädiktions-Qualität wurde eine Einteilung der verfügbaren Daten in Trainingsdaten, Validierungsdaten und Testdaten vorgenommen. Die Aufteilung beinhaltet die Daten aus der Tranche 2 und 3.

Für den Code Release 2 des P2 Prozesses sollen die Daten aus der Tranche 3 (kombiniert mit Tranche 2, sofern möglich) zum Einsatz kommen. Hierzu sollen Erweiterungen bei der Automatisierung der Tests angeboten werden, um den Entwicklern die Arbeit zu erleichtern und eine hochwertige Qualität der Repositories sicherzustellen. Ein großer Zeitaufwand wurde für die Dokumentation aufgewendet, die für die DNN-Entwickler als Anleitung zur Durchführung des Releases dient. Der Code Release 2 konnte mit Hilfe des automatisierten Testtools erfolgreich durchgeführt werden, zeigte jedoch Schwachstellen bei der Benutzung, sodass eine Verfeinerung angedacht ist. Weiterhin wurde ersichtlich, dass die KIA Whitelist

erweitert werden muss, um die Benutzung nötiger Code Libraries verwendbar zu machen. Dieser Prozess wurde angestoßen.

Stand der Arbeiten (30.06.2021):

Mit der Veröffentlichung der Tranche 4 wurden 3D Bounding Box Labels geliefert, 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. Aus diesem Grund und dem Wunsch aus TP3, auf KIA Daten trainierte DNNs zur 3D BB Erkennung zu haben, wurde ein Release 2* durchgeführt. In vorherigen Releases wurde auf öffentlich verfügbaren Datensätzen zurückgegriffen. Die Tabelle mit den Links zu den Repositories des Release 2* kann hier gefunden werden: Release #2 and Release #2*

Für die Daten Tranche 4 wurde ein Datensplit vorgestellt, der die Daten in Trainings-, Validierung und Testdaten teilt. Weiterhin wurden die Splits zu den Splits der Tranche 3 hinzugefügt. Bei der Einteilung wurde darauf geachtet, dass nur wenige Daten der Tranche 4 zu den Testdaten zugeordnet werden, da die Tranche 5 als reiner Testdatensatz vorgesehen ist, und damit eine Erhöhung der Trainingsdaten für sinnvoller erachtet wurde. Im Unterschied zu vorherigen Daten Lieferungen, die hauptsächlich von BIT TS kamen, wurde dieses Mal auch Daten von Mackevision veröffentlicht. Diese wurden ebenfalls aufgeteilt und zu den bisherigen Splits hinzugefügt.

Aufgrund der Datenlieferung von zwei Partnern im Projekt (BIT TS und Mackevision) kam es zu verschiedenen Konflikten und Schwierigkeiten in Bezug auf Datenkonsistenz und Kompatibilität. Seit November 2019 gibt es mit E1.2.3 eine gemeinsame Spezifikation des Annotationsformats, die es ermöglicht, verschiedene Datentranchen und -Lieferanten zusammenzuführen. Da diese bisher nicht vollständig umgesetzt wurde, können die Tranchen nicht vollständig kombiniert werden. Aus diesem Grund wurden Fix-Skripte für die Daten implementiert. Diese Fix-Skripte befinden sich im kia_dataset Repository in Bitbucket und korrigieren alles (spezifiziert in E1.2.3) von 2d Bounding Boxen bis zu den 3d Koordinatensystemtransformationen. Das wurde umgesetzt:

- Automatisches Herunterladen und Extrahieren aller Dateien, Fixes und Fixes für Fixes
- Kompatibilität der Dateinamen für alle Tranchen
- 2D Bounding Box Kompatibilität (BIT T2, BIT T3, BIT T4, MV T2, MV T4)
- 3D Bounding Box Kompatibilität (BIT T3, BIT T4, MV T4)
    - BIT T2 ist teilweise kompatibel, da es einige nicht behebbare Probleme gibt
- Koordinaten System Kompatibilität (BIT T3, BIT T4)

- MV T4 hat Probleme mit der Projektionsmatrix (-> 3D-Boxen können nicht ins Bild projiziert werden)

Im weiteren Projektverlauf wird eine Vereinfachung der Datennutzung untersucht, sodass nicht jeder Nutzer die Fix Skripte eigenständig durchführen muss. Die Erstellung und die Bereitstellung eines fehlerfreien Datensatzes wird untersucht.

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. Die Umsetzung der Output Formate sowie das Hochladen der DNN Prädiktionen für den entsprechenden Testdatensplit auf dem DSP sind Teil der Release Anforderungen.

Aus lizenztechnischen Gründen musste ein Wechsel von Gitlab zu Bitbucket vollzogen werden. Da dies die automatisierte Testpipeline für die Software Releases betrifft, kam es bisher beim Release 3 zu Verzögerungen, sodass der Release 3 erst im Juli stattfinden wird.

Stand der Arbeiten (31.12.2021):

In diesem Berichtszeitraum wurde der dritte TP1 Code Release durchgeführt. 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 im 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, sollen die Inferenz- und Trainingstests eingeschränkt werden, sodass innerhalb der Testtoolchain 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+.

Der vierte und voraussichtlich letzte Release wird im 6ten Berichtszeitraum durchgeführt werden.

Stand der Arbeiten (30.06.2022):

Im letzten Berichtszeitraum wurde der Release 4 durchgeführt in dem sämtliche DNNs aus AP1.3 bis AP1.5 teilnahmen. Die Datenlage hat dazu geführt, dass nicht für alle APs bzw. Tasks (z.B. 3D Bounding Box etc.) neue Daten zur Verfügung standen, sodass diese weiterhin dem Stand des Release 3 entsprechen.

2.1. Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

Tabelle 12: Verwendung der Zuwendung von Valeo

Geplantes Ergebnis	Verwendung der Zuwendung	Erzieltes Ergebnis
Definition und Entwicklung von gemeinsamen Schnittstellen und Formaten zum Austausch und Testen von Code und Daten.	Personalmittel, Reisemittel, Sachmittel (Software, GPU-Server)	Definition von Dateiformaten bzgl. Ground Truth Daten für 2D/3D Bounding Box Erkennung sowie semantische und instance Segmentierung. Gemeinsame Entwicklungsumgebung in Form eines definierten Docker Containers. Umgesetzte Test Pipeline zum Testen des Entwickelten Codes in Form von “Smoke Tests”. Gemeinsames Auswertungskript zur Berechnung von 2D Bounding Box Metriken.
Entwicklung von DNNs zur Perzeption (3D BB Erkennung) basierend auf Kamera und Lidar Sensoren für Erfüllung von Redundanz Anforderungen.	Personalmittel, Reisemittel, Sachmittel (Software, GPU-Server)	Erfolgreiche Umsetzung (Implementierung, Training und Auswertung) eines single-stage DNN zur 3D BB Erkennung auf Monokular Kamera Basis. Erfolgreiche Umsetzung eines 3D BB Erkennung auf LiDAR Basis. Die Algorithmen wurden maßgeblich auf den KIA sowie KITTI Daten ausgewertet.
Generierung von synthetischen LiDAR Punktwolken mit dem Einsatz des Valeo LiDAR-Modells.	Personalmittel, Reisemittel, Sachmittel (Software)	Es wurde eine Vielzahl von synthetischen Daten erzeugt, die mittels des Valeo LiDAR Modells generiert wurden. Diese konnten zu Trainings- und Testzwecken der DNNs verwendet werden.
Entwicklung von Sicherheitsmechanismen, die die Sicherheit der DNNs erhöhen.	Personalmittel, Reisemittel, Sachmittel (Software, GPU-Server)	Es wurde eine Vielzahl von Sicherheitsmechanismen entwickelt. Darunter Mechanismen zur: • Erkennung von out-of-domain Daten. • Dateierweiterung durch noise und Nebel. • Erhöhung der Robustheit durch Ausnutzung zeitlicher Konsistenz.

		• Erhöhung der Genalisierungsfähigkeit durch Multi-Task-Learning. • Anpassung auf Domänen, die nicht im Trainingsdatensatz vorhanden waren. • Verbesserung der prädiktiven Performance und Kalibrierung von DNNs zur semantischen Segmentierung. • Fusion von LiDAR und Kamera DNNs zur Fußgängererkennung auf Objektebene.
Entwicklung einer Strategie zur Erhöhung der Sicherheit von KI Algorithmen.	Personalmittel, Reisemittel	Publikation, in der die Strategie ausführlich beschrieben wird.
Wissensaufbau im Bereich der Sicherheitsargumentation von KI Algorithmen.	Personalmittel, Reisemittel	Erfolgreicher Wissensaufbau über mögliche Strukturen von Sicherheitsargumenationen, deren Bestandteile sowie Voraussetzungen. Erstellung einer Vielzahl von GSN Graphen zur exemplarischen Demonstration von Sicherheitsaspekten bei KI Algorithmen.
Austausch mit der wissenschaftlichen Community.	Personalmittel, Reisemittel	Veröffentlichung von mehreren Papern auf nationalen und internationalen Konferenzen. Organisation von 3 Workshops auf der CVPR Konferenz. Titel des Workshops: Safe AI for Automated Driving. Zahlreiche Paper aus dem KIA Konsortium wurden im Rahmen dieses Workshops eine internationale Bühne gegeben.

2.2. Wichtigste Positionen des zahlenmäßigen Nachweises

Tabelle 13: Relativer zahlenmäßiger Nachweis von Valeo

Position	Benennung im Antrag (AZK/AZA)	Verwendung
F0823	FE-Fremdleistungen	Gemeinsamer Unterauftrag an Neurocat. 1. Ausführbarer und dokumentierter Methodensatz zur Identifikation fehlender Testdaten 2. Nachweis der Wirksamkeit der einzelnen Maßnahmen in Bezug auf die Sicherheitsziele 3. Empfohlene Testmethoden für KI-Funktionen im Bereich Objektdetektion
F0837	Personalkosten	Valeo hat mehr Personalkosten und Arbeitsstunden benötigt als ursprünglich geplant. Mehrstunden, soweit Sie nicht durch den Gesamtförderungsbetrag abgedeckt werden konnten, wurden auf eigene Kosten geleistet.
F0838	Reisekosten	Für Treffen der Partner untereinander, zu Veranstaltungen und zu den Testgeländen. Pandemiebedingt wurde in den Jahren 2020 und 2021 deutlich weniger gereist, als ursprünglich geplant. Es wurde nur ein Viertel der Reisekosten benötigt.
F0850	Sonstige unmittelbare Vorhabenkosten	Anschaffungen betrafen, das Projektbüro, die Ergebnisverbreitung, PyCharm-Lizenzen, der GUA 8 (Implementierung Werkzeugkette) und ein Leasingserver. Die Abweichung der Soll-/Ist-Kosten beträgt 2,40 %

Der zur Verfügung stehende Kostenrahmen wurde von Valeo voll ausgeschöpft. Die zeitliche Planung und die Erfüllung der Meilensteine MS 1 – 6 konnten vollumfänglich erreicht werden.

2.3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Notwendigkeit

Die Valeo Schalter und Sensoren GmbH ist einer der führenden OES. In manchen Bereichen sogar Marktführer. Die Konkurrenz und der allgemeine Fortschritt im Bereich Automotive Sensorik sind groß. Um den Anschluss nicht zu verlieren, gilt es stets innovative und marktgerechte Produkte zu entwickeln. Ohne Forschung ist das nicht möglich, weshalb Valeo auch jährlich große Summen des jährlichen Umsatzes in die Forschung investiert. Im Zuge von öffentlichen Forschungsprojekten ist es zudem möglich mit den einschlägigen OEM und OES zusammenzuarbeiten und zielgerichtet den Bedürfnissen des Marktes Rechnung zu tragen. Viele Themen die im Rahmen von KI-Absicherung bearbeitet wurden, können zudem nur gemeinschaftlich in dieser Größenordnung umgesetzt werden, wie z.B. der Zugriff auf gemeinsame Daten und die Beauftragung über gemeinschaftliche Unteraufträge.

Angemessenheit

KI Absicherung entwickelt und untersucht Mittel und Methoden zur Verifizierung KI-basierter Funktionen für hochgradig automatisiertes Fahren. Für den Anwendungsfall Fußgängererkennung entwickelte das Projekt eine exemplarische Sicherheitsargumentation und Methoden zur Verifizierung einer komplexen KI-Funktion. Die Ergebnisse des Projekts wurden im Austausch mit Standardisierungsgremien zur Unterstützung der Entwicklung eines Standards zur Absicherung von KI-basierten Funktionsmodulen genutzt. KI-Absicherung hat 28 Verbundpartner. Darunter OEMs, Zulieferer, Technologieprovider, Forschungseinrichtungen, Hochschulen und externe Technologiepartner und zählt in Deutschland zu den. sog. Leuchtturmprojekten, denn KI-Absicherung nimmt eine Schlüsselrolle unter den KI-Projekten ein. Ohne Absicherung der KI-basierten Funktionen wird eine Betriebserlaubnis für zukünftige Fahrzeuge nur schwer erreichbar sein und somit ein Must-Have, um dem internationalen Wettbewerb standzuhalten. Das Projekt KI-Absicherung trägt für Valeo dazu bei wirtschaftlich nicht den Anschluss im Bereich des automatisierten Fahrens oder den Anschluss am aktuellen Stand der Technik zu verlieren und sichert somit Arbeitsplätze bzw. hilft sogar dabei neue Arbeitsstellen zu schaffen. Aus der Sicht von Valeo ist das Projekt deshalb nicht nur angemessen, sondern absolut notwendig.

2.4. Voraussichtlicher Nutzen, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Die Projekterkenntnisse und Methoden, insbesondere hinsichtlich robusterer und zuverlässigerer Funktionen für KI-basierte Wahrnehmungsmodule, werden in verschiedene Entwicklungsprozesse im Bereich Hard- und Software einfließen.

Nachfolgende Tabelle zeigt hierzu die mögliche Verwertung der Projektergebnisse aus KI-Absicherung für 5 Hard- und Softwareprodukte auf. Lfd.Nr. 6 betrifft die Ergebnisverbreitung direkt. Farbig hinterlegt sind die Hardwareprodukte.

Tabelle 14: Verwertung von Valeo

Projektergebnis / Inhalt	**Nutzen / Verwertung / Zeitraum**
-Testpipeline zum Einchecken von Code. Tooling für “Smoke” Tests und automatisierter Validierung von prädiktiver Performance. -Sichere 3D Fußgängererkennung basierend ausschließlich auf Kamera-Sensoren.	Erkennung von Fußgängern und Hindernissen während des Parkvorgangs / Valeo Park4U / ab 2025
-Kenntnisse über Datenerweiterungen (data augmentation) zum Trainieren von Fußgängererkennungen in größeren Distanzen zum Ego-Vehicle. -Out-of-domain Erkennung in Bezug auf Wetterszenarien	Verbesserte Perception / Valeo Cruise4U / ab 2025
-Fusion von Kamera- und LiDAR Sensoren zur hochpräzisen 3D Fußgängererkennung im innerstädtischen Umfeld. -Multi-Task DNN zur Erhöhung der Generalisierbarkeit und Umgang mit einer Vielzahl von nötigen Perzeptionsaufgaben.	Verbesserte Perception und Fußgängererkennung / Valeo Drive4U / ab 2025
Kenntnisse zur Komplexität bei der Integration von LiDAR Sensor Eigenschaften in Datengenerierungs-Pipelines zur Erzeugung von synthetischen LiDAR Punktwolken.	Lidar-Integration / Lidar Scala / ab 2025
Entwicklungen und Erfahrungen von Multi-Task DNN Architekturen, die einfach auf Multi-Sensor DNNs übertragen werden können, um den Use Case von Surround View Kameras gerecht zu werden.	Sensorfusion / Front- und Surround View Kameras / ab 2025
Organisation des SAIAD Workshops.	Verbindungen zu Projektpartnern und ehemaligen Projektpartnern bleiben bestehen durch weitere Organisationen des SAIAD Workshops. So wurde bereits im Oktober 2022 ein vierter Workshop organisiert. Dadurch wird das Thema “Safe AI”

	weiterhin Bestandteil der wissenschaftlichen Community bleiben und den Stand der Technik in diesem Bereich vorantreiben.

Bei allen 5 Produktgruppen ist die Verwertung ab 2025 geplant. Hierzu werden in einem ersten Schritt die relevanten Arbeitsergebnisse für Hardwareprodukte ab 2023 an die Vorentwicklung am Standort Bietigheim-Bissingen übergeben. 2024 werden deren Ergebnisse dann an den Standort Wemding übergeben, wo letztendlich die Fertigung der Produkte stattfinden soll. Über die Generierung von A bis C-Muster erfolgt dann die stufenweise Überführung in ein marktfähiges Produkt. Bei den Softwareprodukten werden die Arbeitsergebnisse an unsere Mutterzentrale in Frankreich übergeben, wo diese in dem bestehenden Software-Stack integriert werden und im Jahr 2024 verifiziert und validiert werden.

Marktanteil und wirtschaftliche Abschätzung
Der Marktanteil und die wirtschaftliche Abschätzung sind Bestandteil des nicht öffentlich einsehbaren Erfolgskontrollberichts.

Weitere Verwertung von Ergebnissen

Zudem war in der Planung zu Projektbeginn folgende Verwertbarkeit der Ergebnisse eingeplant:

Ergebnisverbreitung
Siehe hierzu Kapitel 2.6, sowie https://www.ki-absicherung-projekt.de/veroeffentlichungen und https://www.ki-absicherung-projekt.de/news.

Folgeaktivitäten
Geplante Folgeaktivitäten betreffen das Einbringen von Wissen in weitere Projekte, wie z.B. mit dem EU-Förderprojekt namens “AIthena”, das sich mit trustworthiness und explainability von KI beschäftigt, was einen Teilbereich von Safe AI darstellt.
Es finden zudem Bestrebungen statt, einen ISO PAS 8800 zum Thema Safe AI zu verfassen, in dem Valeo sowie weitere Projektpartner aus KIA mitwirken.
Weiterhin ist geplant, den SAIAD Workshop (https://sites.google.com/view/saiad2022) weiterzuführen, um weiterhin auf das Thema Safe AI aufmerksam zu machen und Arbeiten in diesem Bereich eine Plattform zu geben.

2.5. Während der Durchführung des Vorhabens dem ZE bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Im Folgenden Ergebnisse von Valeo und von dritter Seite die im Laufe des Projektvorhabens entstanden oder bekannt geworden sind.

- Pitch-Vortrag auf der Fachtagung Automatisiertes Fahren „Mechanismen zur Erhöhung der Absicherbarkeit von Deep Learning basierten Wahrnehmungsfunktionen für das hochautomatiserte Fahren“
- Von Valeo organisierter Workshop auf der *Conference on Computer Vision and Pattern Recognition* (CVPR). Der organisierte Workshop trägt den Namen Safe Artificial Intelligence for Automated Driving (SAIAD). Siehe auch: https://sites.google.com/view/saiad-wscvpr19. In diesem Zusammenhang wurden zahlreiche Paper und Einreichungen zum Thema „Safe AI“ von externen publik. Die Anzahl der Publikationen ist zu umfangreich, um diese hier wiederzugeben.
  Die Publikationen sind hier zu finden: https://ieeexplore.ieee.org/Xplore/home.jsp
- Es wurden auch im Jahre 2020 und 2021 der SAIAD Workshop auf der CVPR durchgeführt und mit einem deutlichen Zuwachs an Beteiligung und Resonanz: https://sites.google.com/view/saiad2020/ und https://sites.google.com/view/saiad2021.
- Paper zum Thema „Strategy to Increase the Safety of a DNN-based Perception for HAD Systems“. Eingereicht auf der Hauptkonferenz des CVPR. Das Paper entstand unter Führung von Valeo und in Kooperation mit Peter Schlicht und Fabian Hüger von Volkswagen und befindet sich derzeit noch im Review-Prozess.
- Sämann, T; Vorstellung von Forschungsergebnissen der KI-Absicherung-projektinternen Summer School; Juli 2019.
- Safe Artificial Intelligence for Automated Driving (SAIAD, https://sites.google.com/view/saiad2020/)
- “Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications”
- Sämann, T., Schlicht, P., Hüger, F.; „Strategy to Increase the Safety of a DNN-based Perception for HAD Systems“; Feb. 2020; [https://arxiv.org/abs/2002.08935]
- Safe Artificial Intelligence for Automated Driving (SAIAD, https://sites.google.com/view/saiad2021/)

- Paper zum Thema “Online Out-of-Domain Detection for Automated Driving”. Das Paper wurde auf dem Machine Learning in Certified Systems Workshop (https://mlcertifiedsystems.deel.ai/) eingereicht und angenommen.
- Paper zum Thema “Improving Predictive Performance and Calibration by Weight Fusion in Semantic Segmentation”. Das Paper wurde auf der Asian Conference on Computer Vision (ACCV) eingereicht.

Es sind jedoch keine Ergebnisse publiziert worden, die das Themenspektrum von KI-Absicherung begrenzen oder obsolet werden lassen.

2.6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses

Veröffentlichungen, an denen mehrere Partner beteiligt waren:

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

https://www.ki-absicherung-projekt.de/news/news-detail/erfolgreicher-projektabschluss-von-ki-absicherung

https://www.ki-absicherung-projekt.de/final-event-results

https://www.youtube.com/watch?v=QdxmGtSb5_s

- Timo Sämann, Peter Schlicht, Fabian Hüger: Strategy to Increase the Safety of a DNN-based Perception for HAD Systems. In: arXiv preprint 20.02.2020

- Gesina Schwalbe, Bernhard Knie, Timo Sämann, Timo Dobberphul, Lydia Gauerhof, Shervin Raaftnia, Oliver Willers: Structuring the Safety Argumentation for Deep Neural Networks. In: SafeComp 2020, Computer Safety, Reliability and Security, Lissabon (Portugal), 15. – 18.09.2020

- Timo Sämann, Horst-Michael Gross: Online Out-of-Domain Detection for Automated Driving. In: Machine Learning in Certified Systems Workshop (https://mlcertifiedsystems.deel.ai/), 14.-15.01.2021

- Sebastian Houben, Stephanie Abrecht, Maram Akila, Andreas Bär, Felix Brockherde, Patrick Feifel, Tim Fingscheidt, Sujan Sai Gannamaneni, Seyed Eghbal Ghobadi, Ahmed Hammam, Anselm Haselhoff, Felix Hauser, Christian Heinzemann, Marco Hoffmann, Nikhil Kapoor, Falk Kappel, Marvin Klingner, Jan Kronenberger, Fabian Küppers, Jonas Löhdefink, Michael Mlynarski, Michael Mock, Firas Mualla, Svetlana Pavlitskaya, Maximilian Poretschkin, Alexander Pohl, Varun Ravi-Kumar, Julia Rosenzweig, Matthias Rottmann, Stefan Rüping, Timo Sämann, Jan David Schneider, Elena Schulz, Gesina Schwalbe, Joachim Sicking, Toshika Srivastava, Serin Varghese, Michael Weber, Sebastian Wirkert, Tim Wirtz, and Matthias Woehrle: Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety. KI Absicherung 2020

Anlage 01: Literaturverzeichnis

Verwendete Fachliteratur sowie benutzte Informations- und Dokumentationsdienste im Laufe des Projektvorhabens KI Absicherung:

- Pitch-Vortrag auf der Fachtagung Automatisiertes Fahren „Mechanismen zur Erhöhung der Absicherbarkeit von Deep Learning basierten Wahrnehmungsfunktionen für das hochautomatiserte Fahren“
- Von Valeo organisierter Workshop auf der *Conference on Computer Vision and Pattern Recognition* (CVPR). Der organisierte Workshop trägt den Namen Safe Artificial Intelligence for Automated Driving (SAIAD). Siehe auch: https://sites.google.com/view/saiad-wscvpr19. In diesem Zusammenhang wurden zahlreiche Paper und Einreichungen zum Thema „Safe AI“ von externen publik. Die Anzahl der Publikationen ist zu umfangreich, um diese hier wiederzugeben.
  Die Publikationen sind hier zu finden: https://ieeexplore.ieee.org/Xplore/home.jsp
- Es wurden auch im Jahre 2020 und 2021 der SAIAD Workshop auf der CVPR durchgeführt und mit einem deutlichen Zuwachs an Beteiligung und Resonanz: https://sites.google.com/view/saiad2020/ und https://sites.google.com/view/saiad2021.
- Paper zum Thema „Strategy to Increase the Safety of a DNN-based Perception for HAD Systems“. Eingereicht auf der Hauptkonferenz des CVPR. Das Paper entstand unter Führung von Valeo und in Kooperation mit Peter Schlicht und Fabian Hüger von Volkswagen und befindet sich derzeit noch im Review-Prozess.
- Safe Artificial Intelligence for Automated Driving (SAIAD, https://sites.google.com/view/saiad2020/)
- “Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications”
- Sämann, T., Schlicht, P., Hüger, F.; „Strategy to Increase the Safety of a DNN-based Perception for HAD Systems“; Feb. 2020; [https://arxiv.org/abs/2002.08935]
- Safe Artificial Intelligence for Automated Driving (SAIAD, https://sites.google.com/view/saiad2021/)
- Paper zum Thema “Online Out-of-Domain Detection for Automated Driving”. Das Paper wurde auf dem Machine Learning in Certified Systems Workshop (https://mlcertifiedsystems.deel.ai/) eingereicht und angenommen.
- Paper zum Thema “Improving Predictive Performance and Calibration by Weight Fusion in Semantic Segmentation”. Das Paper wurde auf der Asian Conference on Computer Vision (ACCV) eingereicht.

Verweise

[1] VDA. (2017). Positionspapier „Leitinitiative autonomes und vernetztes Fahren". Berlin.

[2] Bundesministerium für Wirtschaft und Energie, „Fachprogramm Neue Fahrzeug- und Systemtechnologien," Mai 2015.

[3] S. Zhang, R. Benenson, M. Omran, J. Hosang und B. Schiele, *How Far are We from Solving Pedestrian Detection?,* arXiv:1602.01237, 2016.

[4] W. Wachenfeld und H. Winner, „Die Freigabe des autonomen Fahrens.," in *Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte*, Springer Vieweg, 2015.

[5] S. R. Richter, V. Vineet, S. Roth und V. Koltun, „Playing for data: Ground truth from computer games," *European Conference on Computer Vision (ECCV),* pp. 102-118, 2016.

[6] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez und V. Koltun, „CARLA: An Open Urban Driving Simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017.

[7] X. Yue, B. Wu, S. A. Seshia, K. Keutzer und A. L. Sangiovanni-Vincentelli, „A LiDAR Point Cloud Generator: from a Virtual World to Autonomous Driving," 2018. [Online]. Available: https://arxiv.org/pdf/1804.00103.pdf. [Zugriff am 22 11 2018].

[8] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng und R. Yang, „The ApolloScape Open Dataset for Autonomous Driving and its Application," arXiv:1803.06184, 2018.

[9] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan und T. Darrell, „BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling," arXiv:1805.04687, 2018.

[10] P. Dollar, C. Wojek, B. Schiele und P. Perona, „Pedestrian Detection: An Evaluation of the State of the Art," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, pp. 743-761.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth und B. Schiele, „The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, arXiv:1604.01685, 2016, pp. 3213-3223.

[12] F. Flohr und D. M. Gavrila, „PedCut: An Iterative Framework for Pedestrian Segmentation Combining Shape Models and Multiple Data Cues," in *BMVC*, 2013.

[13] N. Schneider und D. M. Gavrila, „Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study," in *German Conference on Pattern Recognition*, Berlin, Heidelberg, Springer, 2013, pp. 174-183.

[14] S. R. Richter, Z. Hayder und V. Koltun, „Playing for benchmarks," in *International Conference on Computer Vision (ICCV)*, 2017.

[15] D. Kondermann, R. Nair, K. Honauer, K. Krispin, J. Andrulis, A. Brock, B. Gussefeld, M. Rahimimoghaddam, S. Hofmann, C. Brenner und B. Jahne, „The HCI Benchmark Suite: Stereo and Flow Ground Truth with Uncertainties for Urban Autonomous Driving," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* pp. 19-28, 2016.

[16] I. Kotseruba, A. Rasouli und J. K. Tsotsos, „Joint Attention in Autonomous Driving (JAAD),“ arXiv preprint, 2016.

[17] A. Geiger, P. Lenz und R. Urtasun, „Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite,“ in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354-3361.

[18] G. Neuhold, T. Ollmann, S. R. Bulò und P. Kontschieder, „The Mapillary Vistas Dataset for Semantic Unterstanding of Street Scenes,“ in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5000-5009.

[19] G. Ros, L. Sellart, J. Materzynska, D. Vazquez und A. M. Lopez, „The Synthia Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes,“ *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 3234-3243, 2016.

[20] A. Gaidon, Q. Wang, Y. Cabon und E. Vig, „Virtual worlds as proxy for multi-object tracking analysis,“ *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 4340-4349, 2016.

[21] A. Kendall und Y. Gal, „What Uncertainties Do We Need in Bayesian Deep Learning For Computer Vision?,“ *Neural Information Processing Systems,* 2017.

[22] Y. Gal, „Uncertainty in Deep Learning,“ Department of Engineering University of Cambridge, September 2016. [Online]. Available: mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf. [Zugriff am 24 April 2018].

[23] S. Liang, Y. Li und R. Srikant, Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks, arXiv:1706.02690, 2018.

[24] D. Hendrycks und K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, arXiv:1610.02136, 2018.

[25] A. Bendale und T. Boult, Towards Open Set Deep Networks, arXiv:1511.06233, 2015.

[26] P. Oberdiek, M. Rottmann und H. Gottschalk, Classification Uncertainty of Deep Neural Networks Based on Gradient Information, arXiv:1805.08440, 2018.

[27] J. Davis und M. Goadrich, The relationship between precision-recall and ROC curves, ICML, 2006.

[28] Y. Gal, I. Riashat und Z. Ghahramani, Deep bayesian active learning with image data, arXiv:1703.02910, 2017.

[29] W. Hu, T. Miyato, S. Tokui, E. Matsumoto und M. Sugiyama, Learning discrete representations via information maximizing self-augmented training, ICML PMLR Vol. 70: arXiv:1702.08720, 2017.

[30] D. P. Kingma, D. J. Rezende, S. Mohamed und M. Welling, Semi-supervised learning with deep generative models, arXiv:1406.5298, 2014.

[31] A. Rasmus, H. Valpola, M. Honkala, M. Berglund und T. Raiko, Semi-supervised learning with ladder network, arXiv:1507.02672, 2015.

[32] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio und X. Muller, The manifold tangent classifier, Advances in Neural Information Processing Systems 24: NIPS, 2011.

[33] J. Weston, Deep learning via semi-supervised embedding, 2012.

[34] M. Rottmann, K. Kahl und H. Gottschalk, Deep Bayesian Active Semi-Supervised Learning, arXiv:1803.01216, 2018.

[35] I. J. Goodfellow, J. Shlens und C. Szegedy, Explaining and harnessing adversarial examples, arXiv:1412.6572, 2014.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow und R. Fergus, „Intriguing properties of neural networks," arXiv:1312.6199, 2014.

[37] X. Yuan, P. He, Q. Zhu und X. Li, „Adversarial Examples: Attacks and Defenses for Deep Learning," 19 Dezember 2017. [Online]. Available: https://arxiv.org/abs/1712.07107. [Zugriff am 24 April 2018].

[38] X. Li und F. Li, Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics, arXiv:1612.07767, 2016.

[39] K. Grosse, P. Manoharan, N. Papernot, M. Backes und P. McDaniel, On the (Statistical) Detection of Adversarial Examples, arXiv:1702.06280, 2017.

[40] M. Rottmann, P. Colling, T.-P. Hack, F. Hüger, P. Schlicht und H. Gottschalk, Prediction Error Meta Classification in Semantic Segmention: Detection via Aggregated Dispersion Measures of Softmax Probabilities., arXiv:1811.00648, 2018.

[41] ISO, *ISO 26262 Road vehicles - Functional safety,* Geneva, Switzerland: ISO, 2011.

[42] K. Wagstaff, Machine Learning that Matters. Proceedings of the 29th International Conference on Machine Learning, Omnipress, 2012.

[43] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla und F. Herrera, A unifying view on dataset shift in classification Pattern Recognition,, 2012.

[44] S. Geman, E. Bienenstock und R. Doursat, Neural Networks and the Bias/Variance Dilemma Neural Computation, MIT Press, 1992.

[45] S. Burton, L. Gauerhof und C. Heinzemann, Making the Case for Safety of Machine Learning in Highly Automated Driving. In: Tonetta S., Schoitsch E., Bitsch F. (eds) Computer Safety, Reliability, and Security. SAFECOMP 2017. Lecture Notes in Computer Science, vol 10489, Springer, 2017.

[46] D. A. Leonhardi und S. Vasu, *SCODE Essential Analysis, Whitepaper,* ETAS GmbH Stuttgart: https://www.etas.com/download-center-files/products_RTA_Software_Products/Whitepaper_SCODE_2016_12_19.pdf , 2016.

[47] F. Zwicky, *Morphologische Forschung: Wesen und Wandel materieller und geistiger struktureller Zusammenhänge,* Fritz Zwicky-Stiftung, 1989.

[48] L. Gauerhof, P. Munk und S. Burton, Structuring Validation Targets of a Machine Learning Function Applied to Automated Driving. In: Gallina B., Skavhaug A., Bitsch F. (eds) Computer Safety, Reliability, and Security. SAFECOMP . Lecture Notes in Computer Science, vol 11093, Springer, 2018.

Berichtsblatt

1. ISBN oder ISSN -	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht
3. Titel Individueller Schlussbericht zum Verbundbericht „KI-Absicherung“ Verbundprojekt: KI-Absicherung - Methoden und Maßnahmen zur Absicherung von KI basierten Wahrnehmungsfunktionen für das automatisierte Fahren; Teilvorhaben: Fußgängererkennung, bei der die Fußgänger mit der minimal umgebenden 3D Box umschlossen werden	
4. Autor(en) [Name(n), Vorname(n)] Nagel, Alexander Sämann, Timo	5. Abschlussdatum des Vorhabens 30.06.2022 6. Veröffentlichungsdatum 15.12.2022 7. Form der Publikation Online + Gebundener Bericht
8. Durchführende Institution(en) (Name, Adresse) Valeo Schalter und Sensoren GmbH (Verbundpartner) Laiernstrasse 12 74321 Bietigheim-Bissingen AUDI Aktiengesellschaft Opel Automobile GmbH Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung eingetragener Verein Intel Deutschland GmbH Universität Heidelberg Robert Bosch Gesellschaft mit beschränkter Haftung Hella Aglaia Mobile Vision GmbH Merantix Labs GmbH Continental Automotive Technologies GmbH ZF Friedrichshafen AG Technische Universität München Luxoft GmbH VAIVA GmbH Argo AI GmbH Accenture Song Content Germany GmbH umlaut systems GmbH Bergische Universität Wuppertal Visteon Electronics Germany GmbH FZI Forschungszentrum Informatik Deutsches Forschungszentrum für Künstliche Intelligenz GmbH QualityMinds GmbH Deutsches Zentrum für Luft- und Raumfahrt e.V. Bayerische Motoren Werke Aktiengesellschaft e:fs TechHub GmbH Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung eingetragener Verein Projektkoordinator: VOLKSWAGEN AG	9. Ber. Nr. Durchführende Institution Valeo Schalter und Sensoren GmbH 10. Förderkennzeichen: 19A19005H 11. Seitenzahl 169
12. Fördernde Institution (Name, Adresse) Bundesministerium für Wirtschaft und Klimaschutz (BMWK) Scharnhorststr. 34-37 10115 Berlin	13. Literaturangaben 62 14. Tabellen 14 15. Abbildungen 67
16. Zusätzliche Angaben -	
17. Vorgelegt bei (Titel, Ort, Datum) Technische Informationsbibliothek (TIB), Deutsche Forschungsberichte, Postfach 6080, D-30060 Hannover, 22.11.2022	

18. Kurzfassung

KI Absicherung ist ein Projekt der KI Familie der Leitinitiative. Es wurde aus der VDA Leitinitiative autonomes und vernetztes Fahren initiiert und entwickelt und wurde vom Bundesministerium für Wirtschaft und Klimaschutz gefördert. Der hier vorliegende Schlussbericht beschreibt die Arbeiten und Ergebnisse des Projektpartners Valeo Schalter und Sensoren GmbH, welche in allen 5 Teilprojekten beteiligt war. Schwerpunkte im ersten Teilprojekt setzte Valeo in der 3D Bounding Box Fußgängererkennung. Diese Erkennung wurde im AP1.3 (Co-Lead-Rolle) auf Basis einer monokularen Kamera erzeugt und in AP1.4 (Lead-Rolle) auf Basis einer Fusion von Kamera und Lidar Sensoren. Im Teilprojekt 2 lag der Fokus auf dem Einsatz von Transfer-Learning Methoden. In AP2.3 (Co-Lead-Rolle) wurden diese Techniken eingesetzt, um die KI-Funktion auf ein geändertes Sensor-Setup anzupassen. In AP2.4 (Co-Lead-Rolle) untersuchte Valeo die Anpassung der KI-Funktion bezüglich der Qualität der synthetischen Trainingsdaten. Im dritten Teilprojekt konzentrierte sich Valeo auf die Entwicklung von introspektiven Methoden und Maßnahmen mittels Plausibilisierungstechniken (AP3.4, Lead-Rolle). Im Teilprojekt 4 lag der inhaltliche Schwerpunkt von Valeo in der Strukturierung und Formalisierung des Eingaberaumes (AP4.1, Lead-Rolle). Das fünfte Teilprojekt diente dem Projektmanagement, der Ergebnisverbreitung, sowie der Kommunikation mit Normungsgremien.

19. Schlagwörter

Automatisiertes Fahren, Absicherung, KI-Absicherung Transfer-Learning, Bounding-Boxen, Fusion, Neuronale Netze, KI, Künstliche Intelligenz, Leitinitiative, Normung, Absicherungsstrategie

20. Verlag	21. Preis
-	-

Document Control Sheet

1. ISBN or ISSN -	2. type of document (e.g. report, publication) Final Report
3. title Individueller Schlussbericht zum Verbundbericht „KI-Absicherung“ Joint project: KI-Absicherung - Methoden und Maßnahmen zur Absicherung von KI basierten Wahrnehmungsfunktionen für das automatisierte Fahren; Sub-project: Fußgängererkennung, bei der die Fußgänger mit der minimal umgebenden 3D Box umschlossen werden	
4. author(s) (family name, first name(s)) Nagel, Alexander Sämann, Timo	5. end of project 30.06.2022
	6. publication date 15.12.2022
	7. form of publication Online + written report
8. performing organization(s) (name, address) Valeo Schalter und Sensoren GmbH (Verbundpartner) Laiernstrasse 12 74321 Bietigheim-Bissingen AUDI Aktiengesellschaft Opel Automobile GmbH Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung eingetragener Verein Intel Deutschland GmbH Universität Heidelberg Robert Bosch Gesellschaft mit beschränkter Haftung Hella Aglaia Mobile Vision GmbH Merantix Labs GmbH Continental Automotive Technologies GmbH ZF Friedrichshafen AG Technische Universität München Luxoft GmbH VAIVA GmbH Argo AI GmbH Accenture Song Content Germany GmbH umlaut systems GmbH Bergische Universität Wuppertal Visteon Electronics Germany GmbH FZI Forschungszentrum Informatik Deutsches Forschungszentrum für Künstliche Intelligenz GmbH QualityMinds GmbH Deutsches Zentrum für Luft- und Raumfahrt e.V. Bayerische Motoren Werke Aktiengesellschaft e:fs TechHub GmbH Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung eingetragener Verein Project coordinator: VOLKSWAGEN AG	9. originator’s report no. Valeo Schalter und Sensoren GmbH
	10. reference no. 19A19005H
	11. no. of pages 169
12. sponsoring agency (name, address) Federal Ministry for Economic Affairs and Climate Action Scharnhorststr. 34-37 10115 Berlin	13. no. of references 62
	14. no. of tables 14
	15. no. of figures 67
16. supplementary notes -	
17. presented at (title, place, date) Technische Informationsbibliothek (TIB), Deutsche Forschungsberichte, Postfach 6080, D-30060 Hannover, 22.11.2022	

18. abstract

KI-Absicherung is a project of the AI family of the flagship initiative (Leitinitiative). It was initiated and developed from the VDA flagship initiative (Leitinitiative) for autonomous and networked driving and was funded by the Federal Ministry for Economic Affairs and Climate Action. This final report describes the work and results of the project partner Valeo Schalter und Sensoren GmbH, which was involved in all 5 sub-projects. In the first sub-project, Valeo focused on the 3D bounding box pedestrian detection. This detection was generated in AP1.3 (co-lead role) based on a monocular camera and in AP1.4 (lead role) based on a fusion of camera and lidar sensors. In sub-project 2, the focus was on the use of transfer learning methods. In AP2.3 (co-lead role) these techniques were used to adapt the AI function to a different sensor setup. In WP2.4 (co-lead role), Valeo investigated the adaptation of the AI function regarding the quality of the synthetic training data. In the third sub-project, Valeo concentrated on the development of introspective methods and measures using plausibility techniques (AP3.4, lead role). In sub-project 4, Valeo's main focus was on the structuring and formalization of the input domain (AP4.1, lead role). The fifth sub-project was used for project management, dissemination of results and communication with standardization committees.

19. keywords

Automated driving, validation, AI validation, transfer learning, bounding boxes, fusion, neural networks, AI, safe-AI, artificial intelligence, flagship initiative, standardization, validation strategy

20. publisher	21. price
-	-