KI Absicherung - Schlussbericht

Berichtszeitraum: 01.07.2019 - 30.06.2022

HCI, Heidelberg Collaboratory for Image Processing

Version 0.4

Laufzeit des Vorhabens: 01.07.2019 – 30.06.2022
Zuwendungsempfänger: HCI
Förderkennzeichen:19A19005S
Fälligkeitsdatum: 31.12.2022
Erstellungsdatum: 30.04.2025

Gefördert durch:

Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

Ein Projekt entwickelt von der
VDA Leitinitiative
autonomes und vernetztes Fahren

DOKUMENTENINFORMATION

KONTAKT
Dr. Christoph Langenhan & Andreas Blattmann
Computer Vision Group, Lehrstuhl Prof. Ommer
Ludwig-Maximilians-Universität München, Akademiestr. 7, 80799 München
für das
HCI, Heidelberg Collaboratory for Image Processing,
Universität Heidelberg, Postfach 10 57 60, 69047 Heidelberg
Tel.: +49 (0)89/2180-73452
E-Mail: c.langenhan@lmu.de

VERSIONSÜBERSICHT

Version	Datum	Kommentar	Autor
V0.1	21.12.2022	Dokument erstellt	Andreas Blattmann
V0.2	02.04.2025	Dokument erstellt	Christoph Langenhan
V0.3	11.04.2025	Dokument erstellt	Christoph Langenhan
V0.4	30.04.2025	Dokument erstellt	Christoph Langenhan

Inhaltsverzeichnis

1. Aufgabenstellung 4

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde 4

2.1 BMWi-Fachprogramm “Neue Fahrzeug- und Systemtechnologien“ 4

2.2 KI-Strategie der Bundesregierung 6

3. Planung und Ablauf des Vorhabens 7

4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde 13

4.1 Algorithmen zur Fußgängerkennung implementieren 13

4.2 Absicherbarkeit von KI-Modulen bewerten 13

5. Zusammenarbeit mit anderen Stellen 14

6. Eingehende Darstellung der Verwendung der Zuwendung 16

6.1 Multimodale Fusion von Pose- und Bounding-Box-Schätzern 22

6.2 Generative Analyse und unüberwachtes Lernen in der Posenschätzung 26

6.3 Datenbasierte Evaluation und urbane Szenensynthese 30

6.4 Robustifizierung durch Corner-Case-Augmentation 33

6.5 Interpretierbarkeit und Visualisierung tiefer Repräsentationen 37

6.6 Foundation Models für Sicherheit und Anpassungsfähigkeit 41

6.7 Offline-Validierung und Schwächendiagnose 42

6.8 Stable Diffusion als Baustein für Absicherung und Interpretierbarkeit generativer KI 43

6.9 Literatur 45

7. Wichtigsten Positionen des zahlenmäßigen Nachweises, ... 47

8. Notwendigkeit und Angemessenheit der geleisteten Arbeit 48

9. Voraussichtlichen Nutzens, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans 50

1. Aufgabenstellung

Ziel des Projekts KI-Absicherung ist die Entwicklung und Evaluation strukturierter Methoden und Maßnahmen zur sicherheitstechnischen Absicherung KI-basierter Einzelmodule im Kontext des automatisierten Fahrens. Aufbauend auf dem Stand der Technik soll ein in Fachkreisen anerkannter Ansatz zum Nachweis der Absicherbarkeit solcher Module erarbeitet werden. Ein besonderer Fokus liegt auf der Bestimmbarkeit verbleibender Restrisiken sogenannter „Blackbox"-Systeme. Die methodische Umsetzung erfolgt exemplarisch am sicherheitskritischen Use-Case der Fußgängererkennung, wobei bewusst auf eine Absichtserkennung verzichtet wird, um die Detektionsleistung klassischer KI-Modelle isoliert zu analysieren.

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Verbundprojekt KI Absicherung wurde unter Bezug zum BMWI/BMWK-Fachprogramm „Neue Fahrzeug- und Systemtechnologien" (NFST) sowie zur KI-Strategie der Bundesregierung durchgeführt.

2.1 BMWi-Fachprogramm "Neue Fahrzeug- und Systemtechnologien"

Das Fachprogramms formuliert u.A. den Bedarf für produkt- und anwendungsnahe Forschungsmaßnahmen in der Säule „Automatisiertes Fahren", um die „industrielle und wissenschaftliche Innovationskraft Deutschlands im Bereich des automatisierten Fahrens zu stärken". KI Absicherung weist einen konkreten Bezug zur Programmsäule „Automatisiertes Fahren" des BMWi- Fachprogramms auf. KI-Absicherung

- greift die im Themenfeld „Innovative Sensorik und Aktoriksysteme" definierten Herausforderungen auf, in dem es die Grundlage für neue (KI basierte) Wahrnehmungsfunktionen, die eine Umgebungserfassung in komplexen Szenarien für hoch- und vollautomatisches Fahren erst ermöglicht, schafft.

- erarbeitet Lösungen für die im Themenfeld „Angepasste Testverfahren und Validierung" dargestellte Herausforderung der Absicherung automatisierten Fahrfunktionen durch Entwicklung geeigneter Verfahren und Methoden, die die Grundlage für den Einsatz der Schlüsseltechnologie KI im automatisierten

Fahren schaffen.

Die konkreten Beiträge des Vorhabens KI-Absicherung zu den innerhalb des BMWi-Fachprogramms in den beiden benannten Themenfeldern formulierten Forschungszielen werden in nachfolgender Tabelle konkretisiert:

Forschungsziel	**Beitrag KI-Absicherung**
Themenfeld „Innovative Sensorik und Aktoriksysteme“	
Weiterentwicklung Sensorik- und Aktoriksysteme und dazu- gehöriger Architektur	Die Komplexität urbaner Szenarien lässt beim autonomen Fahren eine klassische, regelbasierte Programmierung von Algorithmen, besonders für das Szenenverstehen, nicht mehr zu. Im Vorhaben wurden die Potenziale KI basierter Wahrnehmungsfunktionen erschlossen.
	Hierfür bedarf es des Nachweises einer ausreichenden Robustheit bzw. Ausfallsicherheit derartiger Systeme. Entsprechend wurden im Vorhaben die Einflussgrößen auf die Sensorik selbst als auch die Auswirkungen auf das Ergebnis multisensorieller Wahrnehmung durch Erweiterung der Sensorik – z.B. um Tiefendaten und LiDAR-Sensorik) untersucht, qualifiziert und quantifiziert.
Themenfeld „Angepasste Testverfahren und Validierung“	
Neuartige Indikatoren, Testverfahren und Testmethoden / Entwicklung von standardisierbaren Tests und Testanforderungen	Den Kern des Vorhabens stellte die Definition von KPI zur Bestimmung der Wirksamkeit der KI-spezifischen Analyse-Methoden und -Maßnahmen dar.
	Diese Argumentation für ein methodisches Vorgehen zur

	Absicherung von KI-Funktionen wurde durch die Entwicklung und Nutzung geeigneter Testmethoden überprüft und mittels einer Teststrategie und eines Testplans generalisiert. Die entwickelte Teststrategie und der Testplan können als Ausgangspunkt für die Produktfreigabe KI-basierter Wahrnehmungsfunktionen dienen.

Tab. 1 Forschungsziele KI-Absicherung

2.2 KI-Strategie der Bundesregierung

Die Bundesregierung hat zwölf prioritäre Handlungsfelder formuliert, „um die Zukunft der KI und deren Nutzung in Deutschland und Europa mitgestalten zu können". Als ein zentrales Vehikel für die Umsetzung dieser Handlungsfelder wird die Forschungsförderung benannt. Das Vorhaben KI- Absicherung (im engeren und die KI-Projektfamilie im weiteren Sinn) hat unmittelbar zwei dieser Handlungsfelder adressiert:

- Forschung in Deutschland und Europa stärken, um Innovationstreiber zu sein

- Daten verfügbar machen und Nutzung erleichtern

Erklärbarkeit und Transparenz von KI sind für uns der Schlüssel für das Vertrauen in die KI. Für eine Vielzahl von Anwendungen der KI stellt sich die Technologie jedoch noch als sogenannte „Blackbox“ dar: Es ist für Nutzer und Betroffene eines KI-Systems oft weder nachvollziehbar noch transparent, wie das System zu Entscheidungen oder Ergebnissen gekommen ist; dies gilt schon für algorithmenbasierte Systeme zur Entscheidungsvorbereitung und erst recht für autonom arbeitende und entscheidende Systeme. Entscheidungen müssen sich nachvollziehen lassen, damit KI-Systeme als „trusted AI“ akzeptiert werden können und rechtlichen Anforderungen genügen.

Aufgrund der hohen Sicherheitsanforderungen im Automobilbereich kommt dort der Zuverlässigkeit von KI-Verfahren eine zentrale Rolle zu. Hierfür ist die Gewinnung ausreichender Trainingsdaten von besonderer Bedeutung. Zudem gilt es, auch mit Blick auf zukünftige Zulassungen neue Verfahren, Methoden und Gütekriterien zu entwickeln.

Vor diesem Hintergrund ist eine enge unternehmensübergreifende Zusammenarbeit der Automobilindustrie bei der Erzeugung/Aufnahme, Verwaltung und Auswertung von Fahr- und Sensordaten gefragt. Nur durch eine solche Kooperation zur Realisierung und Absicherung von KI- Systemen für das automatisierte Fahren wird sich die internationale Wettbewerbsfähigkeit künftig sicherstellen lassen.

3. Planung und Ablauf des Vorhabens

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

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

- AP1.1 „Technische Plattform“ diente der Bereitstellung der technischen Plattform, der Projektinfrastruktur für alle weiteren APs und somit sämtliche für das Projekt notwendigen Tools und Arbeitsprozesse.

- AP1.2 „Anforderungen (Performance-KPIs)“ widmete sich der systematischen Erfassung von quantitativen und qualitativen Anforderungen an eine synthetisierte Datengenerierung an TP2. Zudem wurden in AP1.2 auch die funktionalen Anforderungen und Gütekriterien für die Entwicklung der KI-Funktionalität definiert.

- AP1.3 „Implementieren von Algorithmen zur Fußgängerkennung“ stellte die Basisalgorithmen zur Fußgängererkennung aus monoskopischen Bildsequenzen bereit. Diese Algorithmen basieren auf DNNs und wurden im Wesentlichen mithilfe der von TP2 bereitgestellten Daten generiert Damit stellen sie die funktionale Basis für TP 3 und 4 bereit.

- AP1.4 „Erweiterung der Fußgängererkennung um Tiefendaten“ widmete sich der Entwicklung von Algorithmen zur Prädiktion und Tracking von 3D Bounding Boxen von Fußgängern in Weltkoordinaten. Zu diesem Zwecke wurden die Bilddaten mit Tiefendaten aus der Simulation angereichert.

- AP1.5 „Erweiterung um Posenschätzung“ entwickelte Algorithmen zur Fußgänger Posenschätzung, die im Gegensatz zur vergleichsweise groben Lokalisierung von Fußgängern mittels Bounding Boxen auf die Bestimmung bzw. Schätzung der aktuellen Körperhaltung abheben.

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

- AP2.1 „Toolkette für synthetische Datenerzeugung“ erarbeitete eine technische Verarbeitungskette entwickelt, die alle Verarbeitungsschritte umfasst, um von einer formalen Spezifikation einer Verkehrsszene zu den simulierten Sensordaten zu gelangen.

- AP2.2 „Corner Cases“ hatte zum Ziel Situationen zu identifizieren (Szene samt Kontext und dynamischen Objekten, in der die KI-Funktionalität ein nicht erwartetes und funktional nicht hinlängliches Ergebnis liefert, obwohl ein korrektes Verhalten erwartbar gewesen wäre. Des Weiteren fungierte AP2.2 als zentraler „Gate-Keeper“ für die Anforderungen an die Datengenerierung im gesamten Projekt. In dieser Rolle werden alle eingehenden Datenanforderungen konsolidiert und priorisiert an AP2.5 weitergeleitet.

- AP2.3 „Abstraktion von Sensorik“ beschäftigte sich mit der Frage der Übertragbarkeit von KI- Funktionen bezüglich ihrer Absicherbarkeit bei Änderung der Sensorik. Weiterhin wurde die Übertragbarkeit der KI-Funktionen bei Änderung der Sensorik durch Methoden des Transfer Learnings untersucht, sowie Konzepte zur Anpassung der Transfer

LearningMethoden erarbeitet.

- AP2.4 „Bewertung synthetischerDaten“ erarbeitete eine Methode zur vergleichenden Bewertung von Datensätzen unterschiedlicher Qualität.

- AP2.5 „Datengenerierung und Noisy Data“ generierte synthetische Trainings- und Validierungsdaten unter Einbezug von bewussten Verunreinigungen bzw. Störungen von Sensordaten.

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

- AP3.1 „Tracking State-of-Research“ übernahm die Aufgabe der Analyse und Überwachung der Forschungslandschaft sowie die Bewertung der entwickelten Mechanismen auf ihren Mehrwert hinsichtlich eines möglichen Beitrags zu einer stringenten Sicherheitsargumentation.

- AP3.2 „Wirksamkeits- und Sicherheits-KPIs für Machine Learning fokussierte auf die Identifikation und Belastbarkeit von Qualitätsmetriken.

- AP3.3 bis AP3.5 setzen sich expliziet mit Methoden und Maßnahmen auseinander, die entweder kein (AP3.5 „Blackbox Methoden und Maßnahmen“), ein beschränktes oder aber über vollständiges (AP3.4 „White-/Grexbox Methoden und Maßnahmen“) Wissen über die dahinterliegende KI Funktion haben, oder sogar funktionsverändernd wirken (AP3.3 „Funktional verändernde Methoden“).

- AP3.6 „Aggregierte Methoden und Maßnahmen“ setzte sich mit der Kombination von Methoden und Maßnahmen und deren Mehrwert hinsichtlich der Absicherbarkeit auseinander.

- Teilprojekt 4 (TP4) mit dem Kurztitel „Gesamtheitliche KI-Absicherungsstrategie“ hatte zum Ziel die Definition und exemplarische Umsetzung eines systematischen Vorgehens zur Formulierung einer

gesamtheitlichen Argumentation (Assurance Case) zur Absicherung einer KI-Funktion (Fußgängererkennung) vorzunehmen.

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

- AP4.2 „Gesamtstruktur der Argumentation“ hatte sich zum Ziel gesetzt, die Sicherheitsziele der eingesetzten Funktion (z.B. jeder relevante Fußgänger wird rechtzeitig erkannt, so dass ausgewichen oder gebremst werden kann), die geforderten Zielgrößen (zulässigen Wertebereiche), der KI-Gesamtfunktionskontext als auch die übergeordnete Systemarchitektur – soweit zum Nachweis der Absicherbarkeit benötigt – zu definieren.

- AP4.3 „Argumentation für eine abgesicherte KI-Funktion“ nahm die Arbeiten unter der Prämisse an, dass eine KI-Funktion aufgrund der inherenten Komplexität und Datenabhängigkeit schwierig als alleinstehende Funktion ohne zusätzliche Maßnahmen als sicher argumentiert werden kann. Entsprechend war es das Ziel des Arbeitspaketes ein Vorgehen zur systematischen Herleitung einer Sicherheitsargumentation auf Basis von Evidenzen basierend auf den Methoden und Maßnahmen aus TP3 und AP4.4 exemplarisch zu erarbeiten. Dies geschah ohne den Anspruch einen vollständigen Sicherheitsnachweis für eine konkrete Umsetzung einer KI basierte Fußgängererkennung erarbeiten zu können.

- AP4.4 „Testmethoden und Wirksamkeitsuntersuchung“ hat sich zum Ziel gesetzt Testmethoden zur Anwendung zu bringen, mit denen es möglich ist die in AP4.1 durchgeführte Domänenanalyse und die in AP4.2 definierten Sicherheitsziele zum einen eine Abdeckung des Eingaberaums und zum anderen die Wirksamkeit der Einzelmaßnahmen zu untersuchen.

- AP4.5 „KI-Teststrategie und KI-Testplan“ widmete sich dem allgemeinen Vorgehen zum Testen einer KI-Funktion. Welche Teststrategie vor dem Hintergrund der Absicherbarkeit von KI- Funktionen am besten geeignet scheint, wurde hierbei unabhängig von der konkreten Funktionsausprägung betrachtet. Darauf aufbauend war das Ziel für die KI-Funktion Fußgängererkennung auf Basis des konkreten Assurance Case unter Verwendung der KI- Teststrategie auszuarbeiten, welcher KI-Testplan geeignet erscheint.

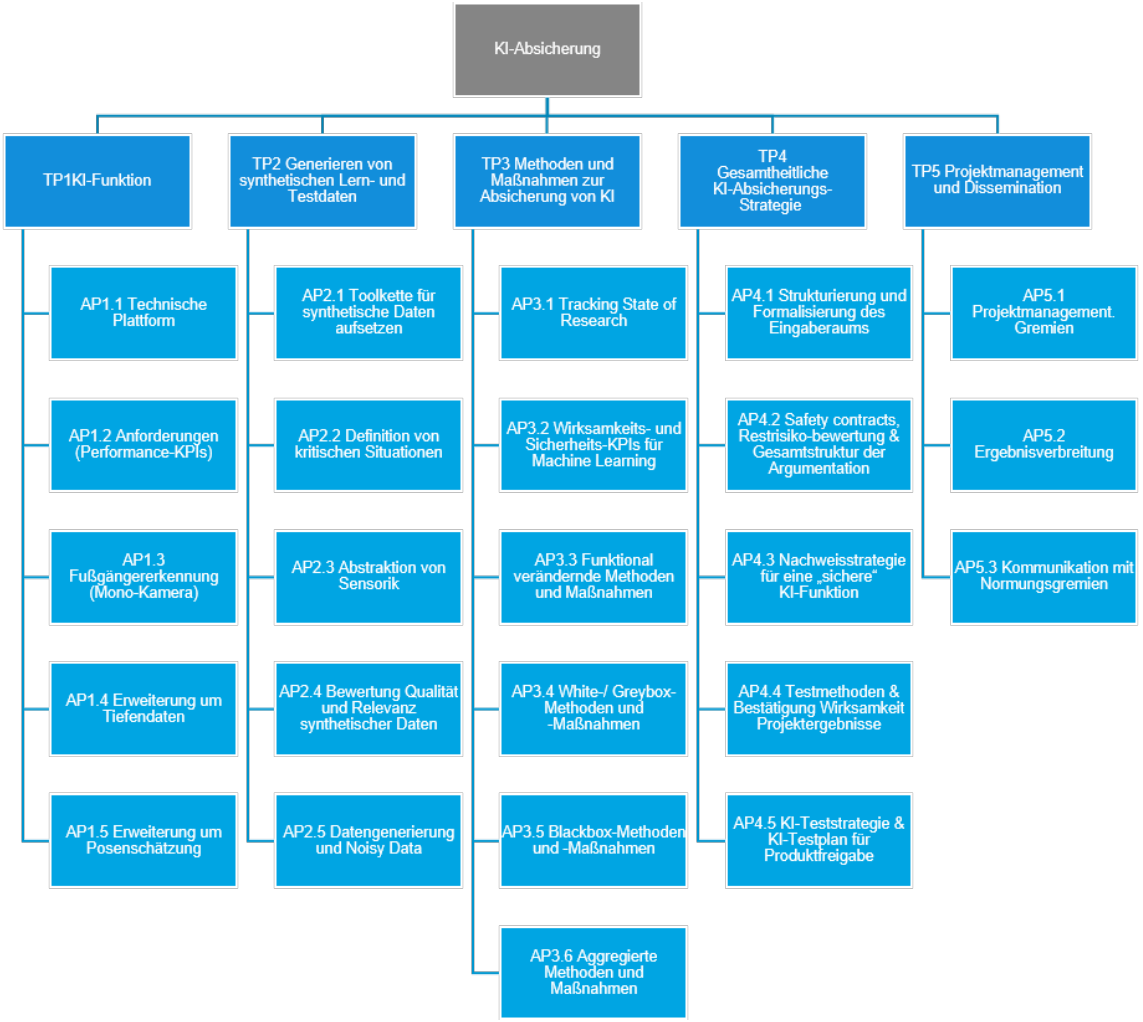

Abb. 1: Projektstruktur KI-Absicherung

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

- AP5.1 „Projektmanagement“ hatte die übergeordnete Aufgabe der technischen Gesamtkoordination mit der inhaltlichen und zeitlichen Organisation der technischen Arbeiten sowie die Sicherstellung der Qualität der Ergebnisse. Hierzu wurden die Arbeiten der einzelnen Arbeitspakete inhaltlich abgestimmt und nachverfolgt sowie der Arbeitsfortschritt hinsichtlich der definierten Meilensteine bestimmt. Bei inhaltlichen Konfliktsituationen und Planabweichungen wurden Vorschläge für adäquate Gegenmaßnahmen erarbeitet.

- AP5.2 „Ergebnisverbreitung“ hatte die Aufgabe die Kommunikation des Vorhabens und dessen Ergebnissen nach außen voranzutreiben. Eine Hauptaufgabe bestand in der zielgruppengerechten Aufbereitung und Zur-Verfügung-Stellung von relevanten Projektinformationen durch geeignete Kommunikationsmittel. Zum anderen wurden Veranstaltungen (Halbzeit- und Abschlusspräsentation) konzipiert, vorbereitet und umgesetzt.

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

4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Hier wird nur auf die Teilprojekte 1 und 3 eingegangen, da das HCI nur in diese aktiv involviert war. Eine ausführliche Gesamtübersicht über alle Teilprojekte ist im Gesamtschlussberichts des Projekts enthalten.

4.1 Algorithmen zur Fußgängerkennung implementieren

Die Fußgängererkennung stellt seit den 1990er-Jahren ein zentrales Forschungsfeld im Automobilbereich dar. Erste prototypische Systeme wurden aufgrund des hohen Anteils von Fußgängern an den Verkehrstoten entwickelt. Neben Zweiradfahrern zählen sie zu den am stärksten gefährdeten Verkehrsteilnehmenden. Systeme zur automatischen Fußgängererkennung leisten daher einen essenziellen Beitrag zur Kollisionsvermeidung.

In heutigen Fahrerassistenzsystemen erfolgt dies u. a. durch Warnsignale bei detektierten Gefährdungssituationen. In autonom fahrenden Fahrzeugen steigen die Anforderungen an die Erkennung deutlich. Während frühe bildbasierte Ansätze auf manuell definierten Merkmalen basierten, haben sich in den letzten Jahren tiefenlernbasierte Verfahren durchgesetzt. Insbesondere Algorithmen auf Basis neuronaler Netze ermöglichen eine signifikant höhere Erkennungsgenauigkeit.

Für den Automotive-Bereich ist neben der Detektion auch die Analyse des Fußgängerverhaltens entscheidend, um adäquate Fahrzeugreaktionen zu ermöglichen. Hierzu reicht eine Lokalisierung mittels Bounding Boxen nicht aus; erforderlich sind präzise Posenanalysen sowie die Modellierung ihrer zeitlichen Dynamik. Diese bilden die Grundlage für die Klassifikation aktueller sowie die Prädiktion zukünftiger Verhaltensweisen von Fußgängern.

4.2 Absicherbarkeit von KI-Modulen bewerten

Das alleinige Testen künstlicher Intelligenz (KI) reicht zur Absicherung automatisierter Fahrsysteme nicht aus. Gründe hierfür sind:

- 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).

- Die begrenzte Kenntnis über Generalisierungsfähigkeit und Stabilität trainierter Modelle.

- Die fehlende Interpretierbarkeit der aus Trainingsdaten abgeleiteten Parameter – insbesondere bei tiefen neuronalen Netzen – erschwert eine gezielte Testung auf Basis nachvollziehbarer Merkmalsrepräsentationen.

Daraus ergibt sich die Notwendigkeit neuer Bewertungsmechanismen für KI-Systeme. Diese umfassen:

- Methoden: Technologien und Software zur Analyse der Absicherbarkeit in den Phasen Training, Test und Inferenz.

- Maßnahmen: Daraus abgeleitete Mechanismen zur Verbesserung oder Absicherung, etwa durch Laufzeitüberwachung oder aktive Ergebnisinterpretation.

Obwohl entsprechende Ansätze in der Forschung existieren, fehlt vielfach eine belastbare Bewertung ihrer Wirksamkeit in sicherheitskritischen Kontexten. Häufig werden in wissenschaftlichen Studien vereinfachte Aufgabenstellungen und Datenformate genutzt (z. B. einfache Klassifikation auf Bildern statt komplexer Prädiktionen mit multimodalen Sensordaten), was deren Übertragbarkeit einschränkt. Für den industriellen Einsatz – etwa im automatisierten Fahren – bedarf es jedoch validierter Verfahren, die unter realitätsnahen Bedingungen getestet wurden. Die bisherige Forschung orientiert sich eher an kontrollierten Experimenten als an ingenieurwissenschaftlich fundierten Prüfverfahren, wie sie für eine spätere Produktzulassung erforderlich wären.

Zudem mangelt es an domänenspezifischen Vergleichsstudien, die eine differenzierte Bewertung und Auswahl geeigneter Methoden ermöglichen. Auch fehlt eine systematische Überführung theoretischer Erkenntnisse in praktisch anwendbare Maßnahmen zur funktionalen Absicherung von KI-Systemen.

5. Zusammenarbeit mit anderen Stellen

Während der Projektlaufzeit arbeitete KI Absicherung kontinuierlich daran, ausgereifte Konzepte und Ergebnisse in geeignete Gremien und Fachprojekte zur

Absicherung KI-basierter Module einzubringen. So floss das im Projekt entwickelte Konzept einer evidenzbasierten Sicherheitsargumentation in die ISO PAS 8800 (Safety and Artificial Intelligence) sowie in die Fortschreibung der DIN Normungsroadmap KI 2.0 (Schwerpunkt Mobilität) ein.

Zudem wurden Kooperationen mit weiteren Initiativen aufgebaut, u.a. mit dem BMWK-Projekt Safetrain und dem vom Land NRW geförderten Projekt Zertifizierte KI. Auch das BSI wurde über zentrale Projektergebnisse informiert und wirkte – ebenso wie DIN und ISO – an der Abschlussveranstaltung mit.

Die im Projekt entwickelte Ontologie wurde mit den Projekten VV Methoden und KI Data Tooling abgestimmt. Solche Harmonisierungsschritte sollten auch künftig, z. B. im Rahmen der ASAM-Initiativen OpenOntology und OpenODD, fortgeführt werden. Ebenso könnten definierte Label- und Metainformationen aus der gemeinsamen Spezifikation (E1.2.3 / E4.1.4b) als Input für andere Projekte oder die OpenLabel-Initiative dienen.

Die Zusammenarbeit mit externen Stellen erfolgte primär durch Projektpartner, die auch in verwandten Projekten oder relevanten Gremien aktiv sind. Dies stellt sicher, dass die im Projekt entwickelten Konzepte über die Projektlaufzeit hinaus wirksam bleiben und nachhaltig implementiert werden.

6. Eingehende Darstellung der Verwendung der Zuwendung

Die zunehmende Integration künstlicher Intelligenz (KI) in sicherheitskritische Systeme, insbesondere im Bereich des automatisierten Fahrens, stellt neue Anforderungen an die Zuverlässigkeit, Nachvollziehbarkeit und Sicherheit dieser Technologien. Während KI-basierte Verfahren bereits beeindruckende Leistungen in der Objekterkennung und Szeneninterpretation erzielen, bleiben zentrale Herausforderungen wie Robustheit gegenüber unbekannten Situationen, Generalisierbarkeit auf neue Umgebungen und die Interpretierbarkeit von Entscheidungen weitgehend ungelöst.

Gerade im urbanen Straßenverkehr ist die Fähigkeit zur präzisen Erkennung und Interpretation menschlichen Verhaltens essenziell, um potenzielle Gefährdungen frühzeitig zu erkennen und entsprechende Reaktionen einzuleiten. Die Erkennung von Fußgängern, ihre Haltung, Bewegungsrichtung und Absichten zählt zu den komplexesten Aufgaben für autonome Systeme, da sie mit einer Vielzahl an Variablen wie Kleidung, Körperform, Umweltbedingungen oder verdeckten Sichtlinien konfrontiert sind. Entsprechend hoch ist der Bedarf an Verfahren, die über herkömmliche Objektdetektion hinausgehen und ein tieferes Verständnis menschlicher Bewegung ermöglichen.

Das vom Bundesministerium für Wirtschaft und Klimaschutz (BMWK) geförderte Projekt „KI-Absicherung“ zielt darauf ab, neuartige methodische und technologische Konzepte zur Sicherung von KI-gestützten Wahrnehmungssystemen im Fahrzeugumfeld zu entwickeln. Zentraler Bestandteil ist dabei die Entwicklung, Analyse und Validierung von Verfahren, die eine zuverlässige, erklärbare und anpassungsfähige Objekterkennung ermöglichen – unter realistischen und auch außergewöhnlichen Bedingungen (Corner Cases).

Innerhalb dieses Gesamtprojekts beteiligte sich das Heidelberg Collaboratory for Image Processing (HCI) mit zwei forschungsstarken Gruppen: der Computer Vision Group (CVG) unter Leitung von Prof. Ommer und dem Visual Learning Lab (VLL) unter Leitung von Prof. Rother. Während sich die CVG auf tiefenlernbasierte Verfahren zur Erkennung und Analyse menschlicher Posen konzentrierte [1][2], entwickelte das VLL generative Modelle zur robusten Darstellung, Synthese und Erweiterung von Szenarien.

Ziel des vorliegenden Projektberichts ist es, die erarbeiteten Konzepte, Modelle und Evaluationen systematisch darzustellen und deren Beitrag zur Absicherung von KI-

Systemen im Kontext autonomer Mobilität zu analysieren. Dabei wird insbesondere die Rolle von generativer Modellierung, multimodaler Fusion, unüberwachtem Lernen und Visualisierungstechniken hervorgehoben. Auf Basis der ursprünglichen Projektziele wird zudem eine Gegenüberstellung von Planung und Ergebnis vorgenommen.

Die HCI Computer Vision Gruppe beteiligte sich im Teilprojekt TP1 (AP1.3, 1.5) mit Arbeiten zur präzisen Schätzung von Fußgängerposen und -bewegungen. Aufbauend auf Forschungsarbeiten zur Verhaltens- und Absichtserkennung stand die detaillierte Bewegungsanalyse im Fokus, um menschliches Verhalten robuster zu erfassen und vorherzusagen. Eine kontinuierliche Evaluierung der Robustheit begleitet die Projektlaufzeit.

In TP3 (AP3.3, 3.4) entwickelte das HCI Visual Learning Lab tiefe generative Modelle zur Analyse und erklärbaren Repräsentation menschlicher Bewegung. Ziel war ein Modell, das Einflussfaktoren trennt, um Detektionsergebnisse interpretierbar und abgesichert zu machen. Das HCI Visual Learning Lab untersuchte, wie simulierte Trainingsdaten durch neue Bildgebungssysteme (z. B. Lichtfeldkameras) und Bildverbesserungsmethoden wie GANs optimiert werden können, um eine hohe Übertragbarkeit auf reale Szenarien zu gewährleisten. Ein zentrales Ergebnis ist die Entwicklung von Stable Diffusion, ein effizientes zweistufiges Diffusionsmodell im latenten Raum, das breite Anwendung in Wissenschaft und Industrie findet und die Basis für das aktuell vom BMWK geförderte Projekt „nxtAIM“ bildet.

Die beiden beteiligten Gruppen des HCI Visual Learning Labs, das Visual Learning Lab (VLL) von Prof. Rother und die Computer Vision Group (CVG) von Prof. Ommer zeichneten hierbei für die im Folgenden aufgelisteten Ergebnisse verantwortlich:

Erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele für die Arbeitspakete von HCI im Umfang von 116PM		
	Vorgegebene Ziele	**Erzielte Ergebnisse**
TP1: KI-Funktion – 36PM / **AP1.3 - 12PM – CVG**	E1.3.3 - Implementierung der funktionalen Algorithmen unter Einbeziehung von Personenmodellen	
	Ein online trainiertes Modell passt sich anhand vorheriger Frames und geschätzter Fußgängerposen (AP1.5) an die Umgebung an. Durch Vergleich der aktuellen Bounding Box mit der modellbasierten Vorhersage soll eine Korrektur zur Verbesserung der Personendetektion erzeugt werden.	Es wurde ein Modell entwickelt, das Posenschätzung und Bounding-Box-Detektion fusioniert, um Robustheit der Fußgängerdetektion zu erhöhen. Ergänzend erzeugt ein generatives Modell poseunabhängige Erscheinungsrepräsentationen, die mit beliebigen Posen kombinierbar sind. Die daraus generierten synthetischen Bilder verbessern die Bounding-Box-Schätzungen. Quantitative Auswertungen zeigen, dass die kombinierte Methode die Detektionsgenauigkeit gegenüber Einzellösungen steigert.
AP1.5 - 24PM - CVG	E1.5.1 - Angaben zum Stand der Wissenschaft	
	Kontinuierliche Literatur-Recherche zu Algorithmen und Datensätzen	Die Literaturrecherche identifizierte zentrale Ansätze zur Absicherung KI-basierter Systeme im Fahrzeugkontext. Diese Erkenntnisse bilden die Grundlage für die Ausrichtung des Projekts und sind in den "Related Work"-Abschnitt der Veröffentlichungen eingeflossen.
	E1.5.4 - Algorithmus zur 2D-/3D-Posenschätzung aus reinen RGB-Daten	
	Es sollte ein selbstüberwachter Ansatz zur 2D/3D-Posenschätzung aus nicht annotierten RGB-D-Daten entwickelt werden, der Bewegungen berücksichtigt, vortrainierte Posenrepräsentationen adaptiert und durch Online-Lernen an veränderte Umgebungen anpasst.	Ein generatives Framework zur präzisen Analyse von Posenveränderungen wurde entwickelt, das mittels invertierbarer neuronaler Netzwerke (INNs) dynamische Bewegungen modelliert und interaktive Werkzeuge wie iPOKE ermöglicht. Ergänzend wurde ein hybrider Ansatz zur Posenschätzung etabliert, das überwachte und unüberwachte Modelle kombiniert, um die Robustheit durch höhere Diversität zu steigern. Das Modell lernt dabei implizite, nicht-joint-basierte Repräsentationen menschlicher Körperteile zur Posenrekonstruktion.

TP3: Methoden und Maßnahmen zur Absicherung von KI – 66PM + 14PM = 80 PM		E1.5.7 - Evaluation der Ergebnisse anhand der KPIs	
		Evaluation der Ergebnisse anhand der KPIs aus AP1.2	Zur Evaluierung der synthetischen Datengenerierung wurde das Posenschätzungsmodell MargiPose (vgl. AP1.5.2, TUM) eingesetzt. Es wurde zunächst auf einem Originaldatensatz trainiert und anschließend auf den augmentierten Datensatz feinjustiert. Die Performanceverbesserung durch die synthetischen Daten diente als zentrales Kriterium zur Bewertung der Datenqualität anhand der definierten KPIs.
	AP3.3 - 66PM – CVG - Prof. Ommer + VLL – Prof. Rother	E3.3.1 - Datenauswahl und Evaluierung - VLL - Prof. Rother	
		Zur Absicherung und Generalisierung neuronaler Netzwerke sollte der Augmentationsraum gezielt charakterisiert, gesampelt und analysiert werden. Die Pipeline nutzt synthetische und reale Daten, um kritische Situationen abzudecken. White-Box-Analysen, Heatmap-Korrelationen und Virtual Active Learning sollten eine optimierte Auswahl und Nutzung relevanter Trainings- und Testsamples ermöglichen.	Die Evaluation erfolgte über FID-/KID-Metriken und Segmentierungsleistung. Während einfache Objekteinbettung sichtbare Artefakte verursachte, konnten die entwickelten Anpassungen diese Unterschiede deutlich verringern. Mask R-CNN profitierte in Form signifikanter AP-Steigerungen über alle Instanzklassen hinweg. Bei Panoptic DeepLab blieben die Effekte uneinheitlich.

E3.3.2 - Datensatz-Veränderung - Augmentations Pipeline - VLL - Prof. Rother	
Zur Verbesserung der Generalisierungsfähigkeit sollte eine differenzierbare Datenaugmentierung eingesetzt werden, bei der realistisch gerenderte Personenmodelle in bestehende Bilddaten eingefügt werden. Neben einfachen, nichtdifferenzierbaren Verfahren umfasst die erweiterte Methode Geometrieanpassung, Verdeckungsbehandlung und Image Matting. Eine differenzierbare Render-Pipeline sollte eine End-to-End-Optimierung und das Lernen von Materialien, Texturen und Beleuchtung erlauben. Zusätzlich sollte eine Domain Adaption durch Style Transfer zwischen synthetischen und realen Lichtfeld-Daten realisiert werden, um den Einfluss des Realismus auf die Modellleistung zu untersuchen.	Es wurde eine Pipeline zur kosteneffizienten Erweiterung urbaner Datensätze durch automatisches Labeling mit synthetischen Objekten entwickelt. Ziel war die Reduktion manueller Annotationen und die Verbesserung der Modellrobustheit. Die Methode ergänzt reale Szenen durch realistisch gerenderte Fahrzeuge und Fußgänger unter Berücksichtigung von Bild-, Segmentierungs- und Kameradaten. Visuelle Inkonsistenzen wurden durch domänenangepasste Beleuchtung und kameratypische Bildfilter minimiert.
E3.3.4 - Funktions-Robustifizierung – Posendetektion - CVG - Prof. Ommer	
Zur Absicherung von Personen- und Posendetektion sollte ein generativer Ansatz entwickelt werden, der Pose und Erscheinungsbild explizit trennt. Durch die Kombination von VAE, GAN und Disentangling Loss wird Appearance Transfer ermöglicht, um Detektionen zu überprüfen und zu korrigieren. Ein transitives Vergleichsmodell sollte die Absicherung gegenüber früheren Modulen erlauben und eine effiziente Forward Propagation ermöglicht automatische Korrektur.	Zur Erhöhung der Robustheit von Posenschätzern gegenüber sicherheitskritischen Corner Cases wurde ein Verfahren zur gezielten Datenaugmentation seltener Körperposen im Straßenverkehr entwickelt. Ein generatives Modell erzeugt realistische Bilddaten auf Basis zuvor identifizierter, unterrepräsentierter Posen. Das Fine-Tuning des Posenschätzers MargiPose auf diesen augmentierten Daten verbesserte die Erkennungsleistung signifikant. Ergänzend wurden adaptive Mechanismen mittels Retrieval-Augmented Generation sowie ein hybrider Syntheseansatz mit Transformer- und Diffusionsmodellen integriert. Daraus entstand u.a. das ressourceneffiziente Modell Stable

		Die Arbeit baut auf CVPR'18-Vorarbeiten auf.	Diffusion, das breite Anwendung findet.
	AP3.4 - 14PM - CVG	E3.4.1 – Plausibilisierung - Gelernte Repräsentation	
		Basierend auf früheren Arbeiten des HCI (CVPR'18, ECCV'18) sollte ein generatives Modell zur Visualisierung von Personen- und Posenrepräsentationen in statischen Bildern entwickelt und anschließend auf Bewegungsinformationen ausgeweitet werden.	Zur Erhöhung der Transparenz tiefen neuronaler Netze wurde ein Visualisierungsmodell entwickelt, das Fokusbereiche und ignorierte Bildinhalte sichtbar macht. Es dient der Validierung gelernter Repräsentationen und der Analyse von Modellinvarianzen. Der Ansatz wurde erfolgreich auf den Posenschätzer MargiPose angewendet, auf der ECCV 2020 präsentiert und in ein Foundation Model (Stable Diffusion) integriert. Ergänzend ermöglicht ein invertierbares Interpretationsnetzwerk eine semantische Analyse ohne Leistungsverlust – ein Fortschritt für die KI-Absicherung.
		E3.4.5 - Offline-Validierung - Bewertung von Schwächen einer gelernten Repräsentation	
		Entwicklung einer Schnittstelle zu bestehenden Repräsentationen sowie zur Testumgebung, um durch eine effiziente Rekonstruktion von RGB-Bildern die Reaktion auf kritische Testfälle – wie Corner Cases, simulierte Sensorfehler oder adversarielle Beispiele – analysieren zu können.	Die Schnittstelle wurde zu Beginn des Projektes im Rahmen der Forschung zu E3.4.1c entwickelt und bereitgestellt. Alle im Bericht gezeigten qualitativen Ergebnisse wurde mittels der Schnittstelle erzeugt.

Tab. 2: Quantitative Ergebnisse des vorgeschlagenen Ansatzes zur Robustifizierung

Ziel ist es, neben der reinen Leistungsfähigkeit auch Aspekte wie Robustheit, Generalisierbarkeit, Interpretierbarkeit und Redundanz zu gewährleisten. Im Rahmen des Projekts wurden daher verschiedene funktionale Algorithmen entwickelt, die durch Fusion von Methoden, generative Modellierung und tiefe Repräsentationslernen zu einer belastbaren KI-Absicherung beitragen.

6.1 Multimodale Fusion von Pose- und Bounding-Box-Schätzern

In E1.3.3 wurde ein Modell entwickelt, das Fußgängerrepräsentationen aus den Posenschätzungsmodellen von AP1.5 nutzt, um die Detektionen der Bounding-Box-Schätzer aus AP1.3 zu korrigieren [1][2]. Ziel war es, durch die Fusion beider Methoden zusätzliche Robustheit und Redundanz zu erzielen – analog zur etablierten Sensorfusion in der Datenerfassung. Dabei wird die unterschiedliche Wirkweise der beiden Ansätze gezielt genutzt: Im Fall einer Fehldetektion des Bounding-Box-Modells kann die vorhergesagte Pose des Posenschätzers zur Korrektur herangezogen werden. Das entwickelte Modell fungiert hierbei als Bindeglied und ermöglicht eine effiziente Kombination beider Signale.

Im Teilprojekt E1.3.3e wurde dazu ein tiefes neuronales Netz implementiert, das die visuelle Erscheinung einer Person – wie Kleidung, Körperform und andere stilistische Merkmale – unabhängig von deren Körperpose darstellt. Die CVG entwickelte ein Modell, das diese Erscheinung in einer kompakten, niedrigdimensionalen Repräsentation erfasst. Diese Darstellung enthält keine Poseinformation, kann aber gezielt mit beliebigen Körperposen kombiniert werden, um neue, synthetische Bilder der Person zu generieren. Damit wird das in E1.3.3e formulierte Ziel erreicht, das Modell auf eine bestimmte Pose zu konditionieren. Das Modell erfüllt folgende Kernfunktionen:

- Trennung von Erscheinung und Pose sowie kompakte Darstellung der Erscheinung,
- Kombination dieser Erscheinung mit beliebigen Posen,
- Generierung entsprechender Bilder durch das neuronale Netz,
- Lernen einer Wahrscheinlichkeitsverteilung über Erscheinungen, sodass auch ohne konkrete Vorlage plausible Personendarstellungen erzeugt werden können.

In der praktischen Anwendung wird ausgehend von einer Bounding Box eines Fußgängers (aus E1.3.3a) sowie einer geschätzten Pose (aus AP1.5) eine alternative Repräsentation der Person generiert. Diese kann erneut durch den Bounding-Box-Detektor verarbeitet werden. Die daraus resultierende alternative Detektion wird mit der ursprünglichen Bounding Box verglichen. Bei Abweichungen

kann ein Korrektursignal abgeleitet werden, was die Robustheit gegenüber Fehldetektionen deutlich erhöht.

Besonders relevant ist hierbei die Fähigkeit des Modells, sich an verändernde Umgebungsbedingungen anzupassen. Während des Trainings wird eine parametrische Wahrscheinlichkeitsverteilung über Erscheinungen gelernt, was nicht nur Sampling ohne konkretes Eingabebild erlaubt, sondern auch eine kontinuierliche Anpassung der Verteilung während des Einsatzes ermöglicht. Damit ist das Modell in der Lage, sich dynamisch an neue visuelle Bedingungen anzupassen – eine zentrale Voraussetzung für den Einsatz im autonomen Fahren.

Die Evaluation des Modells erfolgte zunächst auf öffentlichen Datensätzen wie DeepFashion, Market1501 und Human3.6M. Hierbei wurden die generierten Bilder unter anderem mittels der Metriken Frechet Inception Distance (FID) und Structural Similarity (SSIM) bewertet, die eine hohe Übereinstimmung mit den realen Daten belegten [3]. Nach erfolgter Datenlieferung aus TP2 konnte das Modell schließlich auch auf den projektspezifischen Daten angewendet werden, wodurch die Korrekturmechanismen für reale Anwendungsfälle weiter validiert wurden.

Zusätzlich wurde ein selbstüberwachter Algorithmus zur 2D/3D-Posenschätzung entwickelt (E1.5.4), der unüberwacht Körperhaltungen erkennt und eine diskriminative Merkmalsraumdarstellung lernt. Die Kombination beider Modelle ermöglicht es, Posen zu korrigieren und gleichzeitig durch generierte Darstellungen alternative Sichten auf dieselbe Szene zu liefern – eine Eigenschaft, die in AP3.3.4 für die automatische Funktionsüberprüfung von Posendetektoren verwendet wird. Das System kann so erkennen, in welchen Situationen Unsicherheiten bestehen und gezielt synthetische Daten zur Nachschulung erzeugen, wodurch sich die Gesamtzuverlässigkeit der Detektion weiter steigern lässt [4].

Abb. 2: Beispielhafte Generierungen des Modells zur Korrektur der Detektionsnetze. Die obere Reihe zeigt das Ursprungsbild. Die untere Reihe zeigt das durch Verwendung der Posenschätzung erhaltene generierte Bild unseres Modells. In der ersten Spalte ist z.B. zu sehen, dass so verdeckte Fußgänger in der Bildvordergrund geholt werden können, was offensichtlich für einen Personendetektor von Vorteil ist

Abbildung 2 zeigt exemplarische qualitative Generierungen des Modells, die als Eingabe für das Bounding-Box-Modell dienten, um alternative Schätzungen zu erzeugen. Quantitative Auswertungen belegen, dass die Kombination aus ursprünglicher und alternativer Schätzung bessere Ergebnisse erzielt als die ursprüngliche Schätzung allein. Das zugrunde liegende generative Modell (Details siehe Zwischenbericht M1–M6) wurde im Rahmen des Releases auf den verfügbaren öffentlichen Datensätzen umfassend qualitativ und quantitativ evaluiert – mit durchweg positiven Ergebnissen. Dabei kamen Metriken wie der Frechet Inception Distance (FID) und die Structural Similarity (SSIM) zum Einsatz, die insbesondere auf DeepFashion (FID: 28.38; SSIM: 0.79), Human3.6M (FID: 5.88; SSIM: 0.89) und Market1501 (FID: 9.34; SSIM: 0.44) eine hohe Übereinstimmung der generierten Bilder mit den Originaldaten nachwiesen [1].

Die Arbeiten zur Entwicklung effektiver Repräsentationen und Visualisierungen konzentrierten sich zunächst auf generative Methoden zur Trennung komplexer visueller Merkmale wie Körperhaltung und Erscheinung. Frühere Arbeiten nutzten getrennte Kodierungen zur isolierten Darstellung dieser Merkmale. Eine Erweiterung im Rahmen von E3.3.4 ermöglichte die gezielte Synthese von Personenbildern durch systematische Kombination beliebiger Erscheinungen mit beliebigen Körperposen. Dies bildet die Grundlage für eine gezielte Nachschulung des Posendetektors bei

systematisch fehlerhaften Erkennungen bestimmter Posen oder Erscheinungen.

Darüber hinaus wurde in E3.4.1c ein Modell entwickelt, das die von einem Detektionsnetzwerk als relevant erachteten Bildbereiche sichtbar macht. Die Idee der „Plausibilisierung gelernter Repräsentationen“ beruht dabei auf der Visualisierung jener Aspekte eines Bildes, die vom Detektionsmodell genutzt werden. Durch die Rekonstruktion des Originalbildes mit Fokus auf detektionsrelevante Regionen kann geprüft werden, ob das Modell tatsächlich auf sinnvolle Merkmale zugreift. Abweichungen zwischen rekonstruiertem Bild und Eingangsbild erlauben zudem eine automatisierte Online-Erkennung von Fehldetektionen. Im Abgleich mit einer geeigneten Pixeldifferenzmetrik ergeben sich so robuste Verfahren zur Identifikation unsicherer Vorhersagen – ein essenzieller Schritt zur funktionalen Absicherung von KI-Systemen [2].

Parallel dazu wurden in E3.4.5e Schnittstellen zur Testumgebung entwickelt, um die Reaktion des Systems auf kritische Beispiele visuell darzustellen. Insbesondere bei Modellen, die unter Realbedingungen operieren, ermöglicht diese Kombination aus Visualisierung und quantitativer Analyse ein tiefes Verständnis über Unsicherheiten und Schwachstellen – sowohl während des Trainings als auch zur Laufzeit [3].

6.2 Generative Analyse und unüberwachtes Lernen in der Posenschätzung

Wir haben ein neuartiges generatives Framework zur Analyse von Haltungsvariationen entwickelt [3][4]. Dieses ermöglicht präzise zeitliche Vergleiche von Haltungsveränderungen sowie die Analyse und Visualisierung subtiler Verhaltensabweichungen. Ergänzend wurden neue Datensätze zur detaillierten Haltungsanalyse erstellt. Zur Modellierung dynamischer Bewegungsabläufe wurden invertierbare neuronale Netzwerke (INNs) eingesetzt, die eine verhaltensabhängige Synthese menschlicher Dynamiken ermöglichen. Daraus gingen interaktive Werkzeuge wie iPOKE [9] hervor, die eine fein abgestufte Benutzungssteuerung erlauben und später auf dynamische Texturen erweitert wurden.

Analog zur Bounding-Box-Erkennung ist auch bei der Posenschätzung eine redundante Modellarchitektur von Vorteil. Eine gängige Strategie besteht in der Verwendung von Ensembles, bei denen mehrere überwacht trainierte Modelle kombiniert und deren Vorhersagen aggregiert werden. Da alle Modelle auf identischen Annotationen basieren, zeigen sie jedoch oft ähnliche Fehlerprofile, was die Redundanz einschränkt. Um diesem Problem zu begegnen, wurde in E1.5.4 durch die CVG die Eignung eines unüberwachten Posenschätzungsverfahrens als Ergänzung zu überwachten Methoden untersucht.

Das entwickelte unüberwachte Modell basiert auf der Idee, Körperhaltungen durch eine generische Repräsentation zu erfassen, ohne auf explizite Posenannotationen zurückgreifen zu müssen. Es nutzt Invarianz- und Äquivarianzeigenschaften zwischen Form und Erscheinung, um während des Trainings für jedes Körperteil spezifische Merkmalsrepräsentationen zu erlernen. Aus diesen Repräsentationen werden zur Testzeit die Schlüsselstellen (Keypoints) extrahiert, welche die Körperhaltung beschreiben.

Zur Validierung des Ansatzes wurden qualitative und quantitative Experimente auf öffentlich verfügbaren Datensätzen wie DeepFashion, Human3.6M und iPER durchgeführt. Der Vergleich mit einem zuvor entwickelten Verfahren zeigte eine insgesamt bessere Performance des neuen Modells, insbesondere bei der Erkennung fein differenzierter Körperhaltungen. Diese Ergebnisse führten dazu, dass das neuere Modell zur weiteren Anwendung auf Projektdaten priorisiert wurde, sobald Posenannotationen verfügbar sind.

Ein besonderer Vorteil ergibt sich durch die Kombination des unüberwachten Verfahrens mit bestehenden überwachten Posenschätzern aus AP1.5. Die strukturelle Unterschiedlichkeit der Lernverfahren führt zu divergenten Fehlerbildern, was die Ensemble-Robustheit erhöht. Dies wurde bereits in der Literatur vermutet und im Projektkontext durch das Verhalten der Modelle auf synthetischen und realen Daten weiter untermauert.

Ergänzt wurde der Ansatz durch Konzepte aus E3.3.4, in denen generative Modelle zur systematischen Erkennung von Unsicherheiten in der Posenschätzung eingesetzt werden. Mittels Clustering ähnlicher Posen und einer nachfolgenden Analyse von Detektionsunsicherheiten können „schwierige“ Beispiele identifiziert werden. Diese dienen dann als Grundlage für ein gezieltes Nachtrainieren der Modelle und verbessern so deren Gesamtrobustheit.

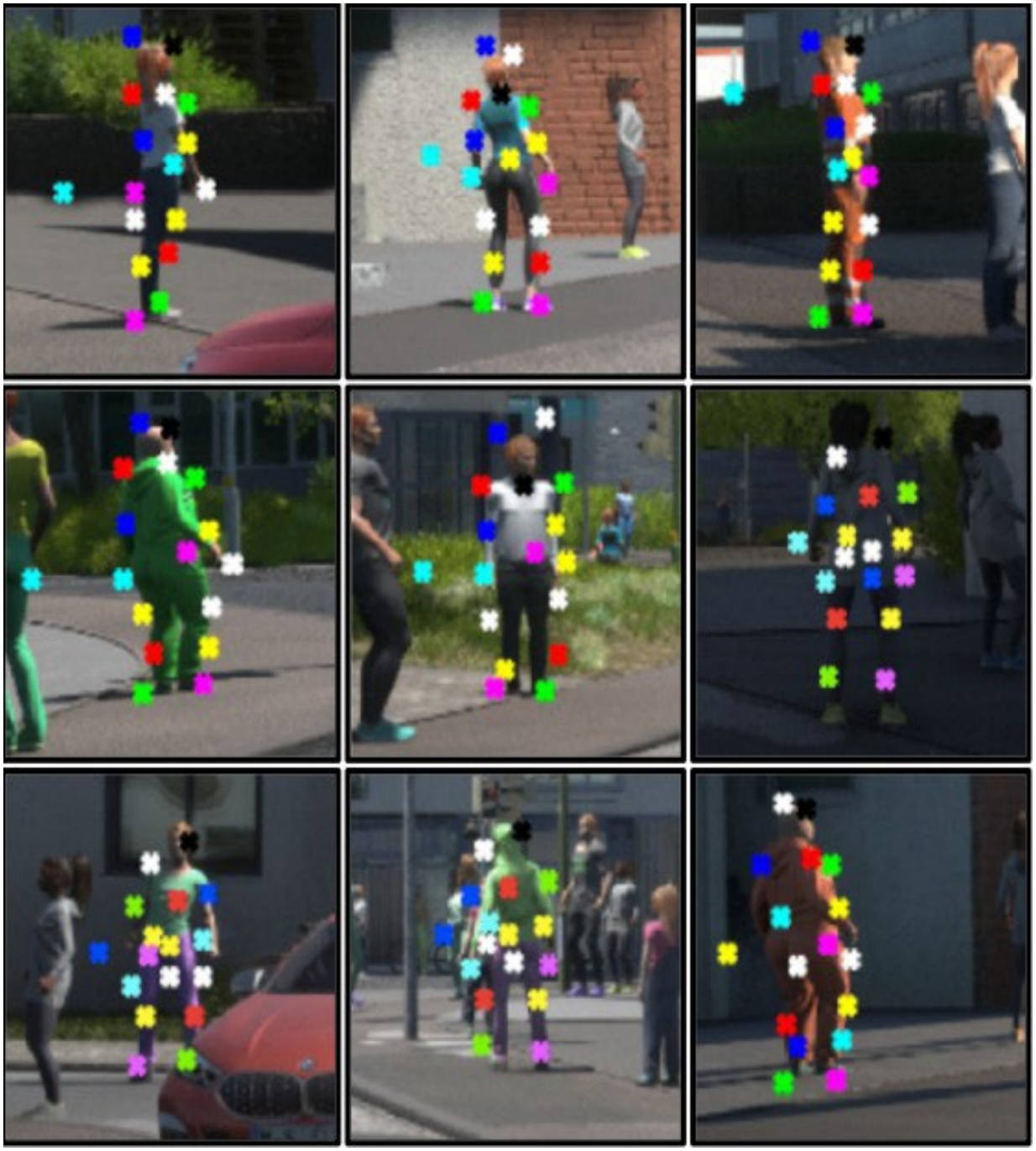

Abb. 3: Ergebnisse des in E1.5.4 entwickelten unüberwachten Posenschätzungsalgorithmus

Diese Netzwerke ermöglichen es, latente Repräsentationen verlustfrei zu entpacken und somit sowohl Erscheinung als auch Formfaktoren visuell interpretierbar zu machen. Abbildung 3 zeigt, dass die durch das Modell erkannten Ankerpunkte nicht explizit mit menschlichen Gelenken übereinstimmen, jedoch eine kohärente Darstellung des Körpers ermöglichen. Diese Eigenschaft resultiert direkt aus dem unüberwachten Lernansatz, wie er im Teilprojekt E1.5.4 verfolgt wurde: Anstatt auf vorgegebene Gelenkdefinitionen zurückzugreifen, lernt das Netzwerk in einer vollständig selbstüberwachten Weise Schwerpunkte für Körperteile, die sich durch

rekonstruktive Konsistenz in den Bilddaten auszeichnen.

Ergänzt wird dieser Ansatz durch die in E3.4.1c entwickelten Visualisierungstechniken, bei denen die durch das Detektionsnetzwerk als relevant erachteten Bildbereiche durch ein Decoder-Netzwerk rekonstruiert werden. Dabei entsteht eine differenzierte Darstellung jener Aspekte eines Bildes, auf die sich das Modell tatsächlich stützt. Weichen rekonstruiertes und ursprüngliches Bild signifikant voneinander ab, deutet dies auf Unsicherheiten oder fehlerhafte Fokusregionen hin. Diese Information kann genutzt werden, um nicht nur technische Schwächen aufzudecken, sondern auch die Nachvollziehbarkeit der Modellentscheidungen zu erhöhen – ein wichtiger Aspekt im Kontext sicherheitskritischer KI-Anwendungen.

Durch die Nutzung unterschiedlicher Modellarchitekturen mit divergierenden Lernansätzen ergibt sich ein Ensemble mit höherer Fehlertoleranz. Insbesondere die Kombination aus überwachten und unüberwachten Verfahren bietet strukturell komplementäre Fehlerprofile, wodurch Fehldetektionen einzelner Modelle seltener das Gesamtsystem beeinflussen. Die Evaluation erfolgte durch rekonstruktive Fehlermaßzahlen sowie durch qualitative Visualisierungen der abgeleiteten Posen auf verschiedenen Datensätzen (z. B. DeepFashion, Human3.6M), wie sie im Rahmen der CVG-Arbeiten durchgeführt wurden.

Darüber hinaus wurde im Projektabschnitt E3.4.5e eine Schnittstelle entwickelt, die diese visuellen Rückmeldungen auch in die Testumgebung integriert. So können Entwickler und Tester systematisch beobachten, wie das Netzwerk auf kritische Eingabebeispiele reagiert – nicht nur als Metrik, sondern auch in Form von nachvollziehbaren Visualisierungen. Dieses Werkzeug unterstützt nicht nur die Absicherung im Entwicklungsprozess, sondern kann perspektivisch auch zur Laufzeit eingesetzt werden, um bei kritischen Situationen im Fahrzeugkontext den menschlichen Fahrer zu informieren oder automatisiert einzugreifen.

6.3 Datenbasierte Evaluation und urbane Szenensynthese

Zur Verbesserung der Trainingsgrundlage wurde in E3.3.1f ein automatisiertes Labeling-Verfahren für urbane Szenen entwickelt. Ziel war es, den hohen Aufwand manueller Annotation zu reduzieren und gleichzeitig qualitativ hochwertige Datensätze bereitzustellen, insbesondere für sicherheitskritische Anwendungen im urbanen Raum. Aufbauend darauf wurden in E3.3.2c verschiedene Augmentations- und Synthesestrategien umgesetzt, mit denen realistische Mixed-Reality-Datensätze erzeugt werden konnten.

Die zentral entwickelte Pipeline basiert auf dem Cityscapes-Datensatz als realer Basis und erweitert diesen systematisch durch die Platzierung fotorealistischer synthetischer Objekte (z. B. Fußgänger, Fahrzeuge). Dafür wurde ein umfangreicher 3D-Modellbestand mit ca. 1000 Personenmodellen und 500 Fahrzeugmodellen generiert und durch Variationen in Kleidung und Farbe diversifiziert.

Um eine realitätsnahe Integration dieser Objekte zu ermöglichen, wurden mehrere technische Maßnahmen ergriffen:

- Beleuchtungsschätzung: Ein eigenes Datenset (Heidelberg360) wurde erstellt, um aufgenommene Szenen mit der jeweils korrekten Sonnenposition zu annotieren. Diese Daten dienten als Grundlage für die präzise Umgebungslichtanpassung synthetischer Objekte.

- Geometrische Platzierung: Durch Segmentierungsinformationen, Tiefenprognosen und die Vogelperspektivendarstellung realer Szenen wurden Kollisionen mit bestehenden Objekten vermieden und realistische Positionierungen gewährleistet.

- Domänenabgleich: Die generierten Szenen wurden in einem Post-Processing-Schritt stilistisch an reale Daten angepasst, u. a. durch Helligkeits- und Farbkorrektur (→ "Learning-Based Post-Processing Pipeline").

Die Segmentierungsmethoden Mask R-CNN, Panoptic DeepLab v3 und PointRend wurden auf diesen angereicherten Datensätzen hinsichtlich ihrer Robustheit gegenüber domänenspezifischen Variationen untersucht [5][6]. Dabei zeigte sich, dass Mask R-CNN die größten Performancegewinne verzeichnete – insbesondere in Kombination mit den neuen Mixed-Reality-Daten. Dieses Ergebnis belegt nicht nur

die Wirksamkeit der Szenensynthese, sondern auch den positiven Einfluss strukturierter Datenaugmentation auf die Segmentierungsgenauigkeit bei urbanen Szenen.

Die aus diesen Arbeiten hervorgehenden Datensätze ermöglichen nicht nur eine gezielte Analyse domänenspezifischer Schwächen, sondern dienen im Projektkontext auch als Grundlage für die robuste Evaluierung weiterer KI-Komponenten, z. B. der Posendetektion in E3.3.4. Diese Kopplung von Synthese, automatischer Annotation und robuster Evaluation stellt ein zentrales Element der KI-Absicherungsstrategie dar.

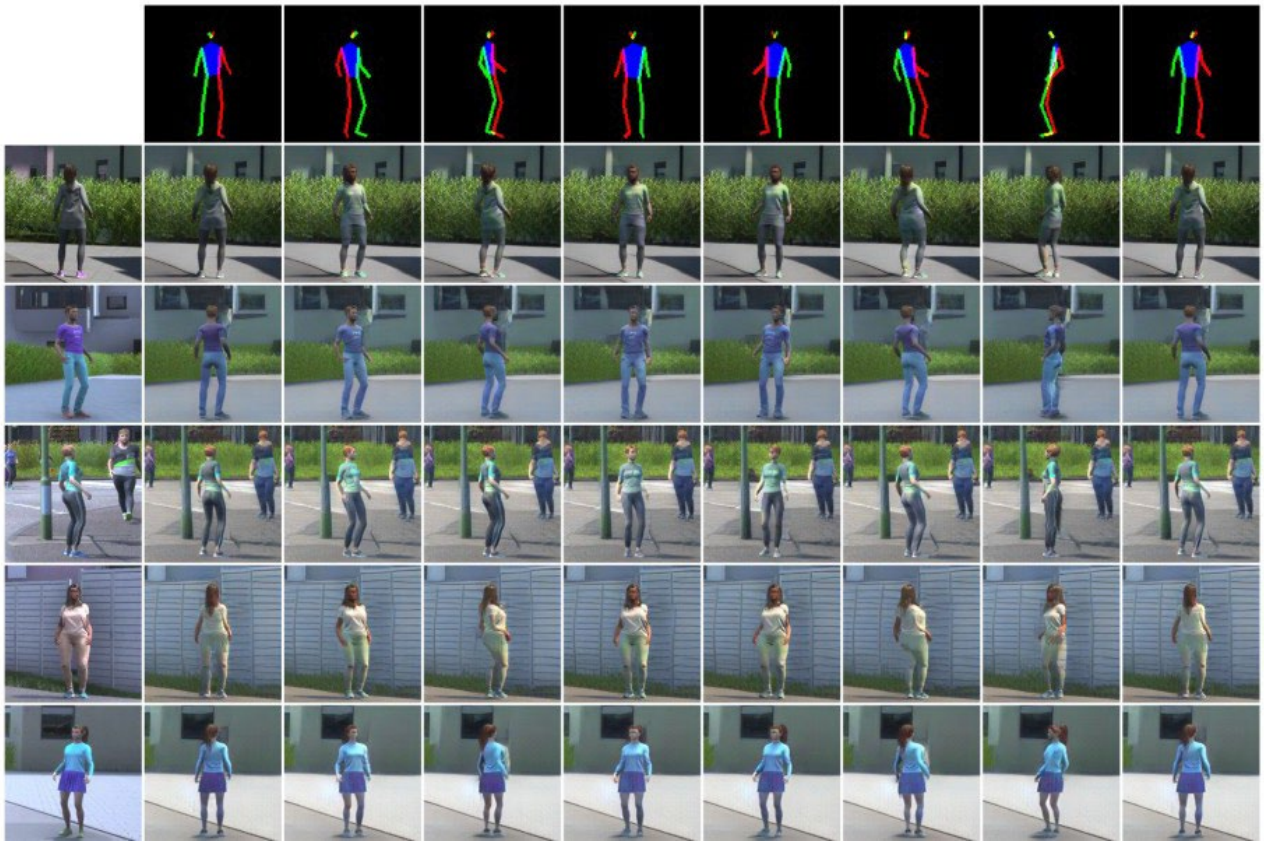

Abb. 4: Ergebnisse des in E3.3.4 erforschten und entwickelten Modells zur Vervielfältigung von Posen. Dieses Beispiel zeigt die Kombination einer Menge von Posen (erste Reihe) mit einer Menge von Erscheinungen (erste Spalte). Hierdurch können die oben abgebildeten Posen, für die es je nur genau ein Ground Truth Bild gibt, künstlich vervielfältigt werden.

E3.3.2 ergänzte diese Arbeit durch eine Augmentation-Pipeline, die reale urbane Szenen um synthetische Fußgänger, Fahrzeuge und andere Verkehrsteilnehmer erweitert. Abbildung 4 demonstriert die qualitativ hochwertigen Integrationsergebnisse. Die Pipeline basiert auf einer Kombination traditioneller

Rendering-Methoden mit lernbasierten Post-Processing-Schritten und wurde im Rahmen der sogenannten "Learning-Based Post-Processing Pipeline (6PM)" entwickelt.

Ein zentrales Merkmal dieser Pipeline ist die präzise visuelle Kohärenz zwischen synthetischen und realen Objekten. Dies wurde insbesondere durch die Anwendung domänenspezifischer Filter für Kameraeffekte, Tiefenunschärfe, Bewegungsunschärfe und Farbcharakteristika erreicht. Ein besonderes Augenmerk lag auf der Beleuchtungssituation: Mithilfe eines eigens erstellten HDR-Datensatzes (Heidelberg360) konnten reale Lichtverhältnisse in Außenbereichen rekonstruiert und für die korrekte Schatten- und Helligkeitsmodellierung synthetischer Objekte genutzt werden. Die so gewonnenen HDR-360°-Lichtinformationen trugen maßgeblich zur fotorealistischen Darstellung bei.

Für die Platzierung der Objekte wurden Tiefeninformationen, Segmentierungsdaten und eine Analyse der urbanen Raumverteilung (basierend auf dem Cityscapes-Datensatz) genutzt. Dadurch konnten synthetische Fußgänger realistisch auf Gehwegen oder am Straßenrand platziert werden, ohne mit vorhandenen Objekten wie Fahrzeugen oder Laternenmasten zu kollidieren. Dies wurde durch eine projektionstechnische Analyse in Vogelperspektive realisiert, bei der Bounding-Box-Kollisionen in der Szene automatisiert vermieden wurden.

Die Qualität der erzeugten Szenen ermöglichte den Einsatz in der robusten Trainings- und Testdatengenerierung für mehrere KI-Komponenten. So wurden die augmentierten Datensätze beispielsweise verwendet, um Instanz-Segmentierungsverfahren wie Mask R-CNN und Panoptic DeepLab zu trainieren und im Anschluss auf originalen Cityscapes-Validierungsdaten zu evaluieren. Die Experimente zeigten, dass Modelle, die mit den augmentierten Daten trainiert wurden, signifikante Performanceverbesserungen im Hinblick auf Generalisierbarkeit und Domänenrobustheit erzielten.

Darüber hinaus wurde der Datensatz perspektivisch zur Evaluierung neuronaler Detektoren aus dem Projekt KI-Absicherung vorbereitet. Insbesondere für E3.3.4 wird das synthetische Material genutzt, um generative Modelle auf „kritische Situationen“ zu testen, wie sie im urbanen Verkehr real auftreten – aber in realen Trainingsdaten selten oder gar nicht vertreten sind. Die Kombination aus Datenvielfalt, visueller Realitätsnähe und automatisierter Annotation bildet somit ein zentrales Werkzeug zur funktionalen Absicherung von KI-Systemen.

6.4 Robustifizierung durch Corner-Case-Augmentation

Im Rahmen von E3.3.4 wurde ein Verfahren zur künstlichen Vervielfältigung seltener Corner Cases entwickelt, um deren Repräsentation im Lerndatensatz eines Posendetektionsmodells gezielt zu verbessern. Der Fokus lag hierbei auf ungewöhnlichen oder seltenen Körperposen, wie sie im realen Straßenverkehr etwa durch spontane Bewegungen, ungewöhnliche Perspektiven oder Verdeckungen entstehen. Diese Posen stellen in der Regel nur einen geringen Teil klassischer Trainingsdaten dar, sind aber für sicherheitskritische Anwendungen von zentraler Bedeutung.

Das Verfahren basiert auf einem generativen Modell, das ausgehend von einer gegebenen Pose synthetische Bilddaten erzeugt. Diese Bilder zeigen Personen, die exakt die spezifizierte Körperhaltung einnehmen, wobei Erscheinung und Pose unabhängig modelliert und flexibel kombinierbar sind. Die Umsetzung erfolgte durch die CVG als Erweiterung eines Variational Autoencoders, wie er bereits in früheren Teilprojekten (z. B. E1.3.3e) eingesetzt wurde.

Zur Identifikation seltener, augmentationswürdiger Posen wurde ein Clusteringverfahren im 2D-Bildkoordinatenraum der annotierten Keypoints durchgeführt. Dieser Ansatz gruppiert ähnliche Posen basierend auf geometrischen Kriterien und erlaubt die Quantifizierung der Verteilung im Datensatz. Auf diese Weise konnten unterrepräsentierte Cluster als "kritisch unterversorgt" erkannt und systematisch durch das generative Modell angereichert werden. Abbildung 5 zeigt exemplarisch die Ergebnisse dieser Clusteranalyse, die bereits im Vorfeld auf öffentlich verfügbaren Datensätzen wie Human3.6M und iPER erprobt wurde.

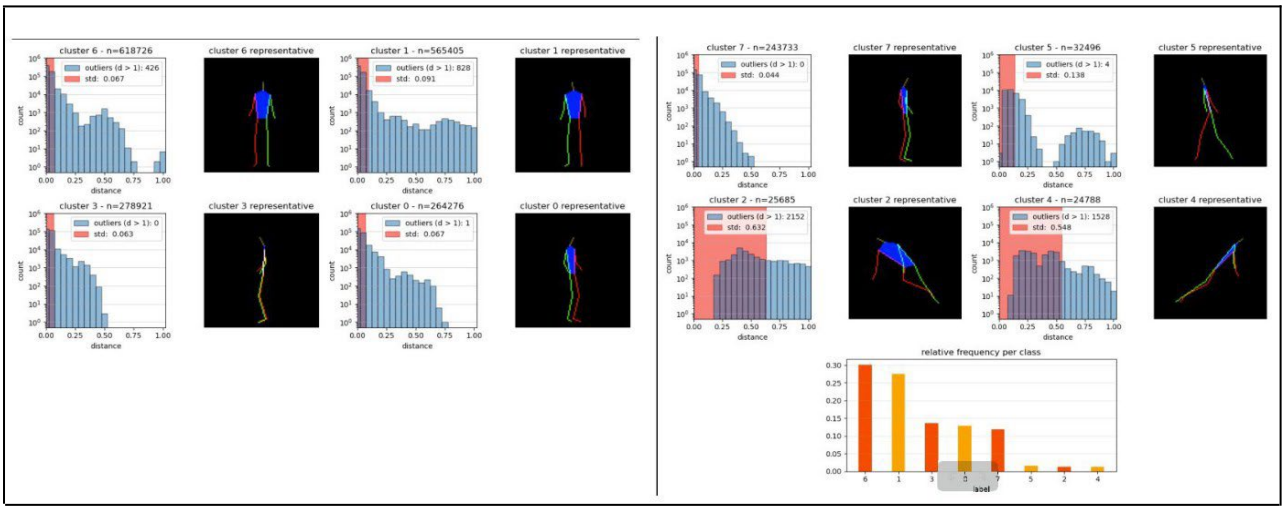

Abb. 5: Ergebnisse des Posenclusterings auf einer Datentranche für n=8 Cluster. Die dargestellten Posen zeigen jeweils die Durchschnittspose der entsprechenden Cluster auf. Die Balkendarstellung

rechts unten gibt für jedes Cluster Aufschluss über die Häufigkeit von Posen des entsprechenden Clusters in den untersuchten Daten. Für diesen Fall sind z.B. die Cluster 2,4 und 5 stark unterrepräsentiert.

Das Clustering erlaubt darüber hinaus die Bestimmung sogenannter "schwerer Beispiele", d. h. Kombinationen aus Pose und Erscheinung, bei denen bestehende Modelle systematisch zu Fehlern neigen. Die Detektionsunsicherheit wurde dabei durch Mittelwerte über die erkannten Gütemaße innerhalb jedes Clusters quantifiziert. Diese Methode diente nicht nur der gezielten Nachschulung des Posendetektors in E3.3.4, sondern beeinflusste auch die Datengenerierung in TP2, wo synthetische Daten durch diese Analyse gezielter gewichtet und erweitert wurden. Die hier entwickelten Techniken unterstützen die übergeordneten Ziele der KI-Absicherung, indem sie erlauben, potenziell fehleranfällige Situationen aktiv zu identifizieren und mit synthetischen Beispielen zu verstärken. In Verbindung mit E1.5.7, das eine quantitative Evaluation anhand projektspezifischer Key Performance Indicators (KPIs) vorsieht, bietet dieses Verfahren eine belastbare Grundlage zur Validierung der Robustheit gegen seltene, aber sicherheitsrelevante Situationen im Straßenverkehr.

Nach dem Clustering identifizierte unterrepräsentierte Posen wurden mithilfe des beschriebenen generativen Verfahrens augmentiert. Diese gezielt synthetisierten Beispiele erweiterten den Trainingsdatensatz insbesondere in jenen Bereichen, in denen Standarddatensätze eine unzureichende Dichte aufweisen – etwa bei seltenen Bewegungsabläufen, extremen Körperhaltungen oder durch Verdeckungen erschwerten Situationen.

Zur Evaluierung diente das Posenschätzungsmodell MargiPose (vgl. AP1.5.2, TUM), das zunächst auf einem Originaldatensatz trainiert und anschließend auf dem augmentierten Datensatz feinjustiert wurde. Diese Feinabstimmung (Finetuning) zielte darauf ab, das Modell robuster gegenüber den zuvor identifizierten „Corner-Case“-Beispielen zu machen. Die Methode greift dabei auf Ergebnisse aus E3.3.4 zurück, in dem nicht nur die Generierung, sondern auch die gezielte Auswahl und Integration schwerer Trainingsbeispiele systematisiert wurde.

Das Finetuning führte insbesondere bei seltenen Posen zu einer deutlich verbesserten Modellleistung, was die Wirksamkeit der vorgeschlagenen Robustifizierungsstrategie empirisch bestätigt. Zur quantitativen Bewertung kamen zwei etablierte Metriken zum Einsatz:

- Mean Per Joint Position Error (MPJPE): misst den durchschnittlichen euklidischen Fehler zwischen vorhergesagten und Ground-Truth-Körpergelenken (je niedriger, desto besser),

- mean Distance (mD): bewertet die mittlere Distanz der geschätzten Gelenkpositionen im Verhältnis zur idealen Posenkonfiguration.

Tabelle 3 zeigt die Ergebnisse für beide Metriken im Vergleich zwischen dem Originalmodell und dem feinjustierten Modell ("MargiPose augmented"). Letzteres zeigt insbesondere in jenen Clustern, die zuvor als unterrepräsentiert identifiziert wurden, signifikante Verbesserungen. Diese Resultate stimmen mit den Zielsetzungen von E1.5.7 überein, in dem die modellbasierten Verbesserungen entlang definierter KPIs systematisch überprüft werden sollen – sobald entsprechende Posenannotationen in TP2 vollständig vorliegen.

Method	MPJPE	mD
MargiPose	0.046	0.0204
MargiPose augmented	0.042	0.0197

Tab. 3: Quantitative Ergebnisse des vorgeschlagenen Ansatzes zur Robustifizierung

Darüber hinaus bestätigt die durchgeführte Evaluation eine zentrale Hypothese aus dem Arbeitspaket: Die Kombination von generativer Corner-Case-Synthese mit zielgerichtetem Finetuning führt zu einem belastbareren Modellverhalten in sicherheitsrelevanten Extremsituationen. Damit leistet die Methode nicht nur einen Beitrag zur Modellgenauigkeit, sondern auch zur funktionalen Absicherung von KI-Komponenten im automatisierten Straßenverkehr.

Zudem entwickelten wir interaktive Anpassungsmechanismen auf Basis von Retrieval-Augmented Generation [11], die es kleineren Modellen ermöglichen, sich flexibel und ohne kostspieliges Fine-Tuning an neue Aufgabenstellungen anzupassen. Dies erlaubt den dynamischen Zugriff auf externe Wissensquellen zur Kontextanreicherung und stellt eine ressourceneffiziente Alternative zum klassischen Transferlernen dar.

Zur Optimierung der Bildsynthese wurde ein hybrider Ansatz verfolgt, bei dem Transformer-Architekturen mit klassischen Verfahren kombiniert wurden. Eine nachgelagerte Verfeinerung mittels eines diffusionsbasierten Modells [11] führte zu einer signifikanten Steigerung der Bildrealitätsnähe. Auf dieser Grundlage entstand Stable Diffusion [2][1], ein ressourcenschonendes, zweistufiges Diffusionsmodell im latenten Raum, das mit über 300 Millionen Downloads breite Anwendung in Forschung und Industrie fand. Stable Diffusion bildet zudem die technologische Basis des vom BMWK geförderten Projekts „nxtAIM“.

So wurde im Teilprojekt E1.3.3e ein Modell entwickelt, das Erscheinung und Pose einer Person trennt, um im komprimierten latenten Raum realistische Bilder zu generieren. Durch die separate Modellierung von Erscheinungsmerkmalen (z. B. Kleidung, Körperbau) und Posen konnten neue Personendarstellungen erzeugt werden, ohne das gesamte Modell neu trainieren zu müssen. Diese Latentrepräsentationen dienen analog zur Verwendung bei Diffusionsmodellen der Reduktion des Rechenaufwands bei gleichzeitiger Erhaltung der Bildqualität.

Darüber hinaus wurde in E3.3.4 ein generatives Modell zur systematischen Kombination von Erscheinung und Pose eingesetzt, welches insbesondere in sicherheitsrelevanten Anwendungen (z. B. bei seltenen Körperhaltungen im Straßenverkehr) eingesetzt wird. Die so erzeugten Bilder konnten gezielt zur Nachschulung von Posendetektoren genutzt werden, um die Modellrobustheit gegenüber Corner Cases zu verbessern – ein Vorgehen, das in seiner Wirkung stark an promptbasierte Steuerung und rekursive Verfeinerungsschritte in modernen Diffusionsmodellen erinnert.

In E3.4.1c wurden schließlich rekonstruktive Verfahren entwickelt, um die vom Modell als relevant erachteten Merkmalsregionen sichtbar zu machen. Auch hier wird – ähnlich wie in Stable Diffusion – im Latent Space operiert, was eine rekonstruktive Bewertung von Vorhersagen auf Pixelebene ermöglicht. Dieses Prinzip erlaubt es, Unsicherheiten nicht nur numerisch, sondern visuell zu interpretieren, und stärkt damit die Erklärbarkeit und Absicherung generativer KI-Modelle im Einsatzkontext.

6.5 Interpretierbarkeit und Visualisierung tiefer Repräsentationen

Ein häufig und berechtigterweise geäußerter Kritikpunkt an tiefen neuronalen Netzen ist ihr sogenanntes „Black-Box-Verhalten“. Es bleibt in der Regel unklar, auf welcher Grundlage eine Entscheidung getroffen wird. Wird etwa ein Fußgänger erkannt, ist nicht ersichtlich, ob die Erkennung auf tatsächlichen Merkmalen eines Fußgängers oder auf irrelevanten Korrelationen basiert. Gerade im Kontext der Absicherung solcher Systeme ist Transparenz über die Entscheidungsgrundlage essenziell, um mögliche Schwächen zu identifizieren.

Zur Adressierung dieser Problematik wurde in E3.4.1c ein Visualisierungsmodell entwickelt, das die relevanten Bildbereiche, auf die sich das Detektionsmodell fokussiert, als RGB-Bild darstellt. Es visualisiert dabei die von TP1 gelernten Repräsentationen von Posen und Personen in statischen Frames. Die verwendete Methode basiert auf einer Kombination aus Autoencoder-Strukturen und einem Decoder-Netzwerk, das ausschließlich auf jene latenten Merkmale zurückgreift, die vom Modell als detektionsrelevant eingestuft wurden. Dadurch entsteht eine Rekonstruktion des Bildes, die gezielt nur jene Bildbereiche enthält, auf die das Modell tatsächlich „achtet“.

Zusätzlich ergänzt das Modell nicht beachtete Bildelemente mittels einer generativen Komponente [9][10], sodass im Idealfall eine realistische Darstellung der erkannten Pose bzw. Person entsteht. Dabei wird eine parametrische Wahrscheinlichkeitsverteilung über Erscheinungsmerkmale gelernt, die es ermöglicht, ignorierte Bildbereiche plausibel zu ergänzen. Diese Ergänzung liefert einen doppelt interpretierbaren Output: einerseits ein rekonstruierbares visuelles Feedback, andererseits quantitative Hinweise auf Modellunsicherheiten.

Das Modell erfüllt zwei zentrale Zwecke:

- Validierung der Repräsentation: Kann die Körperpose einer Person nicht korrekt rekonstruiert werden, deutet dies darauf hin, dass das Detektionsmodell diese Pose nicht adäquat erfasst hat – ein kritischer Hinweis insbesondere bei Posendetektoren. Diese Eigenschaft wurde im Projekt gezielt genutzt, um systematische Schwächen im Posenschätzer MargiPose (vgl. AP1.5.2) zu identifizieren und zu adressieren.

- Visualisierung von Invarianzen: Aspekte des Bildes, die vom Detektionsmodell ignoriert werden, werden über ein generatives Verfahren ergänzt. Diese "Differenzbilder" geben Rückschlüsse auf die internen Repräsentationen und ermöglichen dadurch Einblicke in die Entscheidungslogik des Netzes. Damit ist eine visuelle Form der Explainable AI (XAI) realisiert worden, die über rein numerische Unsicherheiten hinausgeht.

Eine entsprechende Visualisierung der Invarianzen ist in Abbildung 6 dargestellt. Dort wird sichtbar, wie gut das Modell in der Lage ist, Posen zu rekonstruieren – und wo rekonstruierte Strukturen unscharf oder nicht plausibel erscheinen. Die Abweichung zwischen Original und Rekonstruktion dient in diesem Zusammenhang als qualitativer Unsicherheitsindikator. Im Projektkontext wurde dieses Verfahren auch genutzt, um visuelle Rückmeldungen in der Testumgebung darzustellen (E3.4.5e), was zusätzlich zur Systemdiagnose und späteren Laufzeitkontrolle beiträgt.

Abb. 6: Beispielhafte Visualisierungen der Invarianzen des Posenschätzers MargiPose. Die oberste Reihe zeigt das Eingabebild. Die folgenden Reihen zeigen Samples des Visualisierungsmodells. Man sieh, dass das Posenmodell invariant gegenüber der Erscheinung und Textur der Person im Eingabebild ist und sich nur auf die Poseninformation konzentriert.

Da die Ergebnisse sehr überzeugend waren, wurde der Ansatz zudem verallgemeinert und auf der European Conference on Computer Vision (ECCV) 2020 vorgestellt [4]. Des Weiteren stellt er eines der Kapitel des im Rahmen von KI-Absicherung herausgegebenen Fachbuchs Deep Neural Networks and Data for Automated Driving [16] dar (siehe auch Punkt 2.3). Diese Sichtbarkeit verdeutlicht die wissenschaftliche Relevanz und Anwendungsnähe des entwickelten Ansatzes.

Um eine detaillierte Steuerung und Interpretierbarkeit bei der Bildrepräsentation und -synthese zu erreichen, untersuchten wir das Disentanglement visueller Merkmale – insbesondere über Erscheinung, Pose und Struktur hinweg. Aufbauend auf dieser Merkmalstrennung wurde ein generatives Framework konzipiert, das die Basis des Foundation-Modells Stable Diffusion [11,13] ist. Dabei kommt eine Kreuz-Attention-Mechanik in einer Encoder-Decoder-Architektur zum Einsatz, die unterschiedliche Benutzereingaben wie Text, Tiefenkarten und Segmentierungsmasken gleichzeitig verarbeiten kann. Konzeptuell ähnelt dies der modularen Entkopplung von Pose und Erscheinung, wie sie in E1.3.3e durch ein niedrigdimensionales Erscheinungsmodell für Fußgänger realisiert wurde.

Für die nachträgliche Interpretierbarkeit wurde in E3.4.1c ein invertierbares Interpretationsnetzwerk eingeführt [6], das semantische Verständnis über die internen Merkmalsrepräsentationen eines trainierten Modells ermöglicht – ohne dessen Leistung im Inferenzprozess zu beeinträchtigen. Diese Netzwerke erlauben es, modellinterne Invarianzen zu rekonstruieren, indem nicht beachtete Bildteile generativ ergänzt und mit den tatsächlichen Aktivierungen im Netzwerk verglichen werden. Die Methode nutzt ein rekonstruktives Vergleichsverfahren zur Visualisierung von Unsicherheiten in Detektionsmodellen – etwa bei Posenerkennung oder Objektdetektion – und liefert somit visuell interpretierbare Hinweise auf mögliche Schwächen oder Überanpassungen.

Diese Methodik wurde nicht nur in Standalone-Tests validiert, sondern auch in eine prototypische Testumgebung (siehe E3.4.5e) integriert, um visuelle Rückmeldungen zur Systementscheidung während der Anwendung zu ermöglichen. Das erlaubt eine erweiterte Funktionsüberprüfung in der Absicherung generativer und diskriminativer KI-Systeme, insbesondere bei sicherheitskritischen Entscheidungen im Straßenverkehr.

Insgesamt verbesserten diese Fortschritte die Nutzbarkeit, Transparenz und Sicherheitsbewertung generativer Modelle signifikant – ein entscheidender Beitrag zur Zielstellung der KI-Absicherung und zur Operationalisierung moderner Foundation-Modelle im Bereich des automatisierten Fahrens.

6.6 Foundation Models für Sicherheit und Anpassungsfähigkeit

Foundation Models wie Stable Diffusion nutzen Cross-Attention, um multimodale Eingaben (z. B. Text, Segmentierungskarten, Tiefeninformationen) zu verarbeiten und in realistische Bildinhalte zu übersetzen [11], [12]. Diese Fähigkeit zur Kombination verschiedenartiger Modalitäten ist besonders relevant im Kontext automatisierten Fahrens, wo Sensorfusion und kontextuelle Deutung zentral sind.

Das Modell wurde gezielt auf Sicherheitsaspekte trainiert und ermöglicht eine kontrollierte Generierung kritischer Szenarien – beispielsweise für Corner-Case-Tests oder Edge-Case-Augmentationen. Entsprechende Zielstellungen wurden im Projekt KI-Absicherung u. a. in E3.3.4 adressiert, wo generative Modelle dazu eingesetzt wurden, seltene Posen und Verkehrssituationen synthetisch zu erzeugen, um die Trainingsbasis von Erkennungsmodellen gezielt zu erweitern und deren Verhalten in kritischen Kontexten zu untersuchen.

Retrieval-Augmented Generation (RAG) unterstützt dabei die dynamische Anpassung an neue Aufgaben, indem externe Wissensquellen einbezogen werden – ein Konzept, das in E1.5.4 mit der Kombination von überwachten und unüberwachten Posenschätzern eine Parallele findet: Durch die strukturell unterschiedlichen Lernmethoden werden fehleranfällige Situationen robuster erfasst, ohne dass eine vollständige Modellneuschulung nötig ist.

Ergänzend wurde ein invertierbares Interpretationsnetzwerk (INN) integriert, das Repräsentationen entpackt und modifiziert, ohne die Modellleistung zu beeinträchtigen [8], [13], [14]. Vergleichbare Konzepte kamen in E3.4.1c zur Anwendung, wo latente Merkmalsräume rekonstruiert und visuell aufbereitet wurden. Durch den Rekonstruktionsfehler konnte erkannt werden, ob das Modell ein Objekt oder eine Pose tatsächlich verstanden hatte – oder lediglich auf korrelierende Randmerkmale reagierte. Dies erlaubt eine direkte Funktionsüberprüfung im Sinne der Absicherung.

In der Anwendung wurde das Verfahren darüber hinaus mit dem Posenschätzer MargiPose (AP1.5.2) kombiniert und in einer interaktiven Testumgebung (E3.4.5e) visualisiert. Die Kombination von erklärbaren Modellkomponenten mit einer latenten Steuerbarkeit bietet eine zukunftsweisende Grundlage für sichere, adaptive KI-Systeme im Verkehrskontext.

6.7 Offline-Validierung und Schwächendiagnose

Die in E3.4.1 entwickelte Visualisierungsmethode wurde im Rahmen von E3.4.5 in eine automatisierte Testumgebung überführt. Ziel war es, eine visuelle Reaktion auf sicherheitskritische Eingaben sichtbar zu machen und strukturelle Schwächen des zugrundeliegenden Modells zu identifizieren [15], [16]. Die visuelle Analyse erlaubt nicht nur eine Bewertung der Entscheidung des Modells, sondern auch Rückschlüsse auf dessen interne Repräsentation – insbesondere dann, wenn klassische Metriken keine differenzierte Diagnose ermöglichen.

Die integrierte Schnittstelle erlaubt die Darstellung neuronaler Fokusbereiche direkt im Bild und basiert auf einer projektspezifischen Erweiterung eines rekonstruktiven Interpretationsnetzwerks (vgl. E3.4.1c). Dabei wird das Bild ausschließlich aus jenen latenten Merkmalsbereichen rekonstruiert, auf die das Modell tatsächlich zugreift. Durch diesen selektiven Rekonstruktionsansatz werden Bildbereiche, die ignoriert oder falsch interpretiert wurden, automatisch visuell hervorgehoben. Dies macht inkonsistente oder fehlende Detektionen nachvollziehbar und erklärbar – ein entscheidender Schritt zur Absicherung sicherheitskritischer KI-Komponenten im Bereich automatisierter Fahrfunktionen.

Ein besonderer Fokus der Testumgebung lag auf der Offline-Validierung generativer sowie klassischer Detektionsmodelle (z. B. MargiPose aus AP1.5.2). In simulierten Umgebungsszenarien konnten gezielt Testfälle erzeugt und durch das Visualisierungssystem analysiert werden. Dabei zeigte sich, dass bestimmte Posen, Verdeckungen oder untypische Erscheinungen systematisch zu Fehlklassifikationen führten. Diese Erkenntnisse wurden in E3.4.5 genutzt, um auf Basis der Visualisierungen Cluster ähnlicher Fehlerbilder zu bilden und so zielgerichtete Nachbesserungen der Modelle zu ermöglichen.

Alle im Bericht gezeigten Visualisierungen wurden über diese Schnittstelle erzeugt und dokumentieren anschaulich die Funktionsweise und den Diagnosewert der Methode. Durch die klare visuelle Rückmeldung lassen sich nicht nur Schwächen identifizieren, sondern auch potenzielle Korrekturstrategien für die Modellarchitektur und das Training ableiten.

6.8 Stable Diffusion als Baustein für Absicherung und Interpretierbarkeit generativer KI

Im Rahmen des Projekts KI-Absicherung wurde der Einsatz moderner Foundation Models wie Stable Diffusion systematisch untersucht und auf sicherheitsrelevante Herausforderungen im Kontext des automatisierten Fahrens übertragen. Stable Diffusion kombiniert eine zweistufige, latenzbasierte Bildsynthese mit einer leistungsfähigen Cross-Attention-Mechanik, die multimodale Eingaben – darunter Textbeschreibungen, Segmentierungskarten oder Tiefeninformationen – simultan verarbeiten kann. Diese Eigenschaft eröffnet neue Möglichkeiten für die gezielte Generierung komplexer urbaner Szenen, wie sie z. B. in E3.3.2 durch die Augmentation realer Verkehrsdaten mit synthetischen Fußgängern und Fahrzeugen praktiziert wurde.

Ein zentrales Anwendungsfeld der Technologie liegt in der Corner-Case-Augmentation: In E3.3.4 wurde ein generatives Framework entwickelt, das seltene oder schwierige Körperhaltungen (z. B. durch Verdeckung oder ungewöhnliche Perspektiven) gezielt vervielfältigt. Aufbauend auf einem Clustering realer Pose-Daten wurden unterrepräsentierte Konfigurationen erkannt und durch Stable-Diffusion-ähnliche Modelle synthetisch ergänzt. Die generierten Szenen verbesserten nachweislich die Modellleistung bei seltenen Fällen, insbesondere beim Posenschätzer MargiPose. Das Fine-Tuning mit den augmentierten Daten führte zu einer signifikanten Reduktion des Mean Per Joint Position Error (MPJPE) und der mittleren Distanz (mD) – ein Beleg für die Wirksamkeit datengestützter Robustifizierungsstrategien.

Zur weiteren Absicherung wurde ein invertierbares Interpretationsnetzwerk (INN) entwickelt (E3.4.1c), das intern gelernte Merkmalsrepräsentationen entpackt und visualisiert, ohne die Inferenzleistung zu beeinträchtigen. In Anlehnung an die Prinzipien latenzbasierter Generierung bei Stable Diffusion, rekonstruiert dieses Netzwerk gezielt jene Bildbereiche, die das Modell bei seiner Entscheidung tatsächlich verwendet. Unterschiede zwischen Originalbild und Rekonstruktion offenbaren verdeckte Unsicherheiten oder inkonsistente Fokusbereiche – z. B. bei falscher Posenerkennung oder durch irrelevante Korrelationen. Diese Funktion wurde in E3.4.5 in eine automatisierte Testumgebung überführt, in der sicherheitskritische Eingaben visualisiert und modellinterne Schwächen identifiziert werden können.

Stable Diffusion wurde darüber hinaus als technologische Grundlage für das vom BMWK geförderte Projekt „nxtAIM“ hervorgehoben. Es demonstriert, wie durch die Entkopplung von Erscheinung, Pose und Szene eine kontrollierte Szenengenerierung realisiert werden kann, die sowohl für das Training als auch für die Diagnose und Validierung von KI-Modellen relevant ist. Die Kombination mit Retrieval-Augmented Generation ermöglicht eine dynamische Aufgabenanpassung – ohne kostenintensives Re-Training.

Die übergreifende Stärke von Stable-Diffusion-basierten Methoden liegt in ihrer Fähigkeit zur Interpretierbarkeit, Repräsentationssteuerung und Reproduzierbarkeit. Ihre Integration in die Systemarchitektur des KI-Absicherungsprojekts unterstreicht, wie moderne generative Modelle nicht nur als Content-Erzeuger, sondern als Werkzeuge für Transparenz und Sicherheit fungieren können.

6.9 Literatur

[1] Rombach, Robin; Blattmann, Andreas; Lorenz, Dominik; Esser, Patrick; Ommer, Björn; High- Resolution Image Synthesis with Latent Diffusion Models; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[2] Esser, Patrick; Rombach, Robin; Blattmann, Andreas; Ommer, Björn; ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis; Neural Information Processing Systems (NeurIPS), 2021.

[3] Rombach, Robin; Esser, Patrick; Ommer, Björn; Making Sense of CNNs: Interpreting Deep Representations & Their Invariances with INNs; IEEE European Conference on Computer Vision (ECCV), 2020.

[4] Rombach, Robin; Esser, Patrick; Ommer, Björn; Network-to-Network Translation with Conditional Invertible Neural Networks; Neural Information Processing Systems (NeurIPS) (Oral), 2020.

[5] Roth, Karsten; Milbich, Timo; Ommer, Björn; PADS: Policy-Adapted Sampling for Visual Similarity Learning; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2020.

[6] Milbich, Timo; Roth, Karsten; Bharadhwaj, Homanga; Sinha, Samarth; Bengio, Yoshua; Ommer, Björn; Cohen, Joseph Paul; DiVA: Diverse Visual Feature Aggregation for Deep Metric Learning; IEEE European Conference on Computer Vision (ECCV), 2020.

[7] Milbich, Timo; Roth, Karsten; Brattoli, Biagio; Ommer, Björn; Sharing Matters for Generalization in Deep Metric Learning Journal; IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022.

[8] Rombach, Robin; Esser, Patrick; Blattmann, Andreas; Ommer, Björn; Invertible neural networks for understanding semantics of invariances of CNN representations; In: Deep Neural Networks and Data for Automated Driving, 2022.

[9] Dorkenwald, Michael; Büchler, Uta; Ommer, Björn; Unsupervised Magnification of Posture Deviations Across Subjects; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[10] Esser, Patrick; Rombach, Robin; Ommer, Björn; A Disentangling Invertible Interpretation Network for Explaining Latent Representations; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[11] Blattmann, Andreas; Rombach, Robin; Oktay, Kaan; Ommer, Björn; Semi-Parametric Neural Image Synthesis; Neural Information Processing Systems

(NeurIPS), 2022.

[12] Braun, Sandro; Esser, Patrick; Ommer, Björn; Unsupervised Part Discovery by Unsupervised Disentanglement; Proceedings of the German Conference on Pattern Recognition (GCPR) (Oral), Tübingen, 2020.

[13] Blattmann, Andreas; Milbich, Timo; Dorkenwald, Michael; Ommer, Björn; Understanding Object Dynamics for Interactive Image-to-Video Synthesis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[14] Blattmann, Andreas; Milbich, Timo; Dorkenwald, Michael; Ommer, Björn; iPOKE: Poking a Still Image for Controlled Stochastic Video Synthesis; Proceedings of the International Conference on Computer Vision (ICCV), 2021.

[15] Blattmann, Andreas; Milbich, Timo; Dorkenwald, Michael; Ommer, Björn; Understanding Object Dynamics for Interactive Image-to-Video Synthesis; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[16] Blattmann, Andreas; Milbich, Timo; Dorkenwald, Michael; Ommer, Björn; Behavior-Driven Synthesis of Human Dynamics; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

7. Wichtigsten Positionen des zahlenmäßigen Nachweises,

Die Position 0812 – Personal stellt die wesentliche Grundlage für die Durchführung der im Projekt KI-Absicherung erzielten Arbeiten dar. Das hier eingesetzte Personal war zwingend erforderlich, um die Konzeption, Umsetzung und Evaluierung zentraler Technologien im Bereich generativer KI – insbesondere im Zusammenhang mit Foundation Models wie Stable Diffusion – durchzuführen.
Konkret wurden die Personalressourcen für folgende Aufgabenbereiche eingesetzt:

- Forschung und Entwicklung generativer Verfahren: Die Entwicklung und Adaption latenzbasierter Diffusionsmodelle erforderte fundierte Expertise in neuronalen Netzarchitekturen, Bildverarbeitung und Deep Learning, insbesondere zur Trennung und Synthese von Pose- und Erscheinungsmerkmalen (vgl. E1.3.3e, E3.3.4).

- Corner-Case-Augmentation und Datengenerierung: Für die systematische Identifikation, Modellierung und Vervielfältigung sicherheitskritischer Posen und Szenarien war ein hoher manuell-wissenschaftlicher Aufwand notwendig, einschließlich Clustering-Methoden, Evaluation von Unsicherheiten und Modell-Feinabstimmung (siehe E3.3.4).

- Implementierung interpretierbarer KI: Die Entwicklung des invertierbaren Interpretationsnetzwerks (INN) sowie dessen Integration in eine Testumgebung (E3.4.1c, E3.4.5e) erforderte umfangreiche Programmier- und Analysearbeiten, um die Erklärbarkeit der Modelle sicherzustellen und ihre Absicherbarkeit im Systemkontext zu ermöglichen.

- Evaluierung und Validierung: Personal wurde zudem für die Konzeption und Durchführung von Experimenten benötigt, die die Leistungsverbesserung durch augmentierte Daten sowie die Funktionsweise generativer Module quantitativ nachweisen (z. B. MPJPE, SSIM, FID).

Die in Position 0812 veranschlagten Personalmittel bildeten somit die operative und inhaltliche Grundlage für nahezu alle zentralen Ergebnisse des Projekts – insbesondere im Hinblick auf die Entwicklung und Absicherung generativer KI-Methoden für sicherheitsrelevante Anwendungen im automatisierten Straßenverkehr.

8. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die im Projekt KI-Absicherung geleistete Arbeit war sowohl aus wissenschaftlich-technischer als auch aus sicherheitsstrategischer Sicht notwendig und in vollem Umfang angemessen. Angesichts der zunehmenden Verbreitung von KI-Systemen in sicherheitskritischen Bereichen wie dem automatisierten Fahren entsteht ein hoher Bedarf nach Verfahren, die nicht nur leistungsfähig, sondern vor allem zuverlässig, nachvollziehbar und absicherbar sind. Insbesondere das Verhalten solcher Systeme in Randfällen, auch bekannt als Corner Cases, stellt eine zentrale Herausforderung dar, da diese in realen Datensätzen nur unzureichend abgebildet sind, gleichzeitig aber potenziell sicherheitsrelevante Fehler provozieren können. Die durch das Projekt entwickelten generativen Methoden, unter anderem Stable Diffusion, wurden genau zu diesem Zweck eingesetzt: Um jene seltenen, aber kritischen Szenarien systematisch und kontrolliert zu erzeugen, die in der realen Welt kaum beobachtbar, aber für die Modellvalidierung essenziell sind.

Die Auswahl, Entwicklung und Weiterentwicklung eines Foundation Models mit latenter Steuerbarkeit war in diesem Zusammenhang nicht nur technologisch ambitioniert, sondern aus Sicht der Projektziele zwingend erforderlich. Die spezifische Architektur von Stable Diffusion – insbesondere die Trennung von Informationsverarbeitung im latenten Raum und die modulare Steuerung über multimodale Inputs wie Text, Segmentierung oder Tiefeninformationen – erlaubte die gezielte Erzeugung relevanter Szenarien, ohne aufwendig reale Daten erheben oder annotieren zu müssen. Dies reduzierte nicht nur den Ressourceneinsatz erheblich, sondern ermöglichte zugleich eine inhaltlich fundierte Erweiterung bestehender Datengrundlagen. Die Arbeit war auch deshalb angemessen, weil sie auf mehreren Ebenen einen klaren Mehrwert für die funktionale Sicherheit generativer KI bietet: durch verbesserte Robustheit in der Vorhersage, durch gezielte Trainingsdatenanreicherung und durch die Integration transparenter Mechanismen zur Entscheidungsnachvollziehbarkeit.

Die Visualisierungskomponenten, die in den Teilprojekten E3.4.1c und E3.4.5e realisiert wurden, sind in diesem Zusammenhang besonders hervorzuheben. Sie erlauben nicht nur die semantische Rekonstruktion latenter Repräsentationen, sondern machen über rekonstruktive Bildanalysen auch sichtbar, auf welchen Merkmalen eine Modellentscheidung tatsächlich basiert – eine Funktion, die für die

Absicherung von KI-Systemen im Straßenverkehr unerlässlich ist. Gleichzeitig wurde ein innovatives, invertierbares Interpretationsnetzwerk (INN) entwickelt, das diese Visualisierungen ohne Leistungsverlust ermöglicht. Die Integration dieser Technik in eine automatische Testumgebung zeigt, dass die geleistete Arbeit nicht nur experimentell, sondern auch anwendungsnah und praktisch verwertbar war. Sie eröffnet perspektivisch auch eine Laufzeitanalyse sicherheitskritischer Systeme und damit eine wichtige Grundlage für zukünftige regulatorische Standards.

Ein weiterer Aspekt der Angemessenheit ergibt sich aus dem hohen interdisziplinären Anspruch der Aufgabenstellung: Die Verbindung von Deep Learning, generativer Modellierung, Absicherungslogik und interpretierbarer KI erforderte hochqualifiziertes wissenschaftliches Personal, das sowohl theoretische als auch anwendungsnahe Expertise einbringen konnte. Die in der Position 0812 – Personal geleisteten Beiträge waren daher nicht nur notwendig für das Erreichen der Projektziele, sondern bildeten die operative Kernleistung, ohne die die technologischen Fortschritte – insbesondere im Bereich der Corner-Case-Augmentation, Szenensynthese, Fehlerdiagnose und Visualisierung – nicht möglich gewesen wären.

Insgesamt zeigt sich, dass die durchgeführten Arbeiten methodisch sauber begründet, technisch zukunftsweisend und sicherheitsstrategisch dringend erforderlich waren. Sie stellen einen wesentlichen Beitrag zur vertrauenswürdigen Nutzung von KI im automatisierten Fahren dar und bilden eine belastbare Grundlage für den sicheren Einsatz generativer Systeme in realen Anwendungskontexten.

9. Voraussichtlichen Nutzens, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Die im Rahmen des Projekts KI-Absicherung entwickelten Methoden im Bereich generativer KI – insbesondere unter Einsatz latenzbasierter Foundation Models wie Stable Diffusion – weisen ein hohes Maß an Verwertbarkeit auf, sowohl im industriellen als auch im wissenschaftlichen Kontext. Stable Diffusion wurde im Projekt nicht nur als Werkzeug zur Bildsynthese betrachtet, sondern vielmehr als technologischer Kern für eine neue Klasse kontrollierbarer, absicherbarer KI-Systeme. Die Fähigkeit des Modells, multimodale Eingaben wie Textbeschreibungen, Segmentierungskarten oder Tiefeninformationen über Cross-Attention zu verarbeiten, wurde gezielt genutzt, um realistische urbane Szenarien synthetisch zu generieren. Diese Szenen dienen nicht nur der Datenanreicherung, sondern auch der gezielten Robustifizierung von Posendetektoren durch Corner-Case-Augmentation. Besonders hervorzuheben ist hierbei die Fähigkeit, seltene oder sicherheitskritische Posen künstlich zu erzeugen und dadurch gezielt in den Trainingsprozess zu integrieren. Die resultierenden Leistungssteigerungen – gemessen u. a. durch eine Reduktion des Mean Per Joint Position Error (MPJPE) und der mittleren Distanz (mD) – belegen die Wirksamkeit der Methode. Das generative Verfahren wurde im Rahmen von KI-Absicherung auch als Bestandteil einer automatisierten Testumgebung genutzt, in der synthetisch erzeugte kritische Szenarien visuell analysiert und modellinterne Unsicherheiten sichtbar gemacht werden konnten. Dies wurde durch die Integration eines invertierbaren Interpretationsnetzwerks (INN) realisiert, dass es erlaubt, latente Repräsentationen ohne Leistungsverlust zu entpacken, zu visualisieren und in Teilen zu modifizieren. Auf Basis dieser Architektur konnten Entscheidungsgrundlagen sichtbar gemacht und damit die sogenannte Black-Box-Problematik tiefen neuronalen Netzen adressiert werden. Die Visualisierung neuronaler Fokusbereiche, wie sie in E3.4.1c und E3.4.5e entwickelt und erprobt wurde, ermöglicht eine nachvollziehbare Schwächendiagnose, die über rein numerische Evaluationsmetriken hinausgeht. Die Ergebnisse dieses Forschungsstrangs wurden bereits auf internationalen Fachkonferenzen wie der ECCV vorgestellt und als Kapitel im Sammelband Deep Neural Networks and Data for Automated Driving publiziert, was die wissenschaftliche Anschlussfähigkeit und Relevanz zusätzlich unterstreicht.

Aus Sicht des fortgeschriebenen Verwertungsplans ergeben sich daraus vielfältige

Anwendungsmöglichkeiten. Die entwickelten Verfahren eignen sich insbesondere für die industrielle Integration in die Validierungspipeline automatisierter Fahrfunktionen, etwa zur Erzeugung seltener Testszenarien, zur gezielten Nachschulung sicherheitskritischer KI-Komponenten oder zur Entwicklung neuer Absicherungsstrategien. Darüber hinaus lassen sich die Visualisierungs- und Analyseverfahren als Werkzeuge für Entwickler:innen einsetzen, um Detektionsmodelle effizient und nachvollziehbar zu prüfen und weiterzuentwickeln. Die zugrunde liegenden Technologien sind darüber hinaus nicht auf den Bereich Mobilität beschränkt, sondern prinzipiell auf andere Domänen übertragbar – etwa in der industriellen Bildverarbeitung, im Bereich medizinischer Diagnostik oder in sicherheitskritischen Anwendungen der Robotik. Nicht zuletzt bieten die Ergebnisse des Projekts eine methodische Grundlage für zukünftige regulatorische Anforderungen an transparente, absicherbare KI-Systeme. Stable Diffusion steht damit exemplarisch für eine neue Generation generativer KI, die nicht nur leistungsfähig, sondern auch vertrauenswürdig, steuerbar und erklärbar ist – und damit zentrale Voraussetzungen für den realweltlichen Einsatz erfüllt.

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht

3. Titel
KI-Absicherung, Teilprojekte des HCI

4. Autor(en) [Name(n), Vorname(n)]	5. Abschlussdatum des Vorhabens 30.06.2022
Langenhan, Christoph Blattmann, Andreas Rother, Carsten Ommer, Björn	6. Veröffentlichungsdatum 31.05.2025
	7. Form der Publikation Bericht

8. Durchführende Institution(en) (Name, Adresse)	9. Ber. Nr. Durchführende Institution
Heidelberg Collaboratory for Image Processing, Universität Heidelberg Berliner Str. 43 D-69120 Heidelberg	10. Förderkennzeichen 19A19005S
	11. Seitenzahl 51

12. Fördernde Institution (Name, Adresse)	13. Literaturangaben 16
Bundesministerium für Wirtschaft und Klimaschutz (BMWK) 53107 Bonn	14. Tabellen 3
	15. Abbildungen 6

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)

18. Kurzfassung

Im Rahmen des Verbundsprojektes KI Absicherung wurden Methoden und Maßnahmen für die Absicherung KI-basierter Funktionen für autonome Fahrzeuge entwickelt und deren Eignung hinsichtlich einer stringenten Sicherheitsargumentation untersucht. Gegeben des Standes der Technik wurde ein umsetzbarerer und im Expertenkreis anerkannter Weg einer strukturierten Herangehensweise zum Nachweis der Absicherbarkeit von KI-Modulen aufgezeigt und, zusammen mit allen anderen Inhalten des Vorhabens, an einem definierten und abgrenzbaren Use-Case methodisch erarbeitet.

Die vom HCI durchgeführten Arbeiten fokussierten sich hierbei zum einen auf die Entwicklung und Implementierung von State-of-the-Art Detektionsmethoden zur Fußgängererkennung. Der Hauptfokus lag jedoch auf der Erfoschung von Methoden zur Absicherung dieser Detektionsmethoden, insbesondere in den Bereichen Data-Augmentation, Funktionsrobustifizierung und White-Box-Methodenentwickung.

19. Schlagwörter
Automatisiertes Fahren, Künstliche Intelligenz, tiefe neuronale Netze, Interpretierbarkeit, Sicherheit für Automatisiertes Fahren

20. Verlag	21. Preis

Document Control Sheet

1. ISBN or ISSN	2. type of document (e.g. report, publication)
3. title KI-Absicherung, contributions of HCI	
4. author(s) (family name, first name(s)) Langenhan, Christoph Blattmann, Andreas Rother, Carsten Ommer, Björn	5. end of project 30.06.2022 6. publication date 31.05.2025 7. form of publication report
8. performing organization(s) (name, address) Heidelberg Collaboratory for Image Processing, Universität Heidelberg Berliner Str. 43 D-69120 Heidelberg	9. originator's report no. 10. reference no. 19A19005S 11. no. of pages 51
12. sponsoring agency (name, address) Bundesministerium für Wirtschaft und Klimaschutz (BMWK) 53107 Bonn	13. no. of references 16 14. no. of tables 3 15. no. of figures 6
16. supplementary notes	
17. presented at (title, place, date)	
18. abstract The joint project 'KI.Absicherung' developed means and methods for safeguarding AI-based functions for automated cars and investigated their suitability with regard to a stringent safety argumentation. Based on current state of the art methods a feasible way for a structured approach to prove the securability of AI modules was shown and, together with all other contents of the project, methodically elaborated on a defined use case. The works of the HCI Heidelberg within this project focused on investigating and developing state-of-the-art methods for AI-based pedestrian detection as well as exploring approaches for safeguarding these detection methods. The main focus here was put on approaches for data augmentation, function robustification and white-box methods.	
19. keywords Automated Driving, Artificial Intelligence, Deep Neuronal Networks, Interpretability, Safe Automated Driving	
20. publisher	21. price