KI-LOK

Prüfverfahren für KI-basierte
Komponenten im Eisenbahnbetrieb

Abschlussbericht

des Fraunhofer Instituts für Offene Kommunikationssysteme

für das Verbundprojekt des
Bundesministeriums für Wirtschaft und Klimaschutz

Finanziert von der Europäischen Union
NextGenerationEU

Gefördert durch:

Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

Autoren: Jürgen Großmann, Sami Kharma, Dorian Knoblauch, Henry Beiker, Johannes Viehmann

Inhaltsverzeichnis

1 Aufgabenstellung

Im Rahmen des Projekts KI-LOK wurden verschiedene Aufgabenstellungen formuliert, die zur Erreichung des Gesamtziels, nämlich der Entwicklung einer werkzeuggestützten Methodik zur Absicherung von KI-gestützten Bahntechnikkomponenten, definiert wurden. Diese Aufgaben lassen sich in drei Hauptbereiche unterteilen: die Entwicklung von V&V-Techniken und -Werkzeugen für lernende System, die Etablierung von Testverfahren und Methoden für KI-basierte Systeme sowie die Schaffung eines Zulassungsprozesses für sicherheitskritische KI-Anwendungen im Bahnbereich. Im Folgenden sind die zentralen Aufgabenstellungen des Projekts detailliert aufgeführt.

1. **Entwicklung von Methoden und Werkzeugen zur Prüfung von lernenden Systemen:** KI-LOK entwickelt Verfahren zur Verifikation und Validierung (V&V) von KI-basierten Bahnsystemen, die insbesondere die Besonderheiten maschinellen Lernens (ML) berücksichtigen. Dies umfasst die Validierung von extrafunktionalen Systemeigenschaften wie Sicherheit, Robustheit und Zuverlässigkeit, die Qualitätssicherung von Trainings- und Validierungsdaten und die Entwicklung von Methoden zur Steigerung der Transparenz und Nachvollziehbarkeit der Entscheidungen von ML-Modellen, um Sicherheitsanforderungen im Bahnbereich zu erfüllen.

   In diesem Kontext hat Fraunhofer FOKUS Verfahren zur Generierung von Testdaten aus formalen Spezifikationen relevanter Sicherheitseigenschaften entwickelt, mit dem Ziel, Testfälle für kritische Randfälle zu erzeugen. Diese Verfahren kombinieren dynamisches, suchbasiertes Testen mit formalen Methoden und statischer Analyse, um ML-Systeme zu prüfen. Zudem wurden klassische Verifikationsalgorithmen auf neuronale Netze angewendet, um zusätzliche Sicherheit zu gewährleisten.

2. **Entwicklung von automatisierten Testverfahren und Methoden für KI-Systeme:** Speziell für den Test von KI-Systemen ist eine hohe Anzahl von Testdaten notwendig, um der Komplexität und den Risiken von KI-System gerecht zu werden. Vor diesem Hintergrund entwickelt das KI-LOK Projekt Simulationsumgebungen zur Erzeugung realistischer Trainings- und Testdaten für KI-Systeme im Bahnbereich und erstellt Testautomatisierungswerkzeuge, die speziell die Unsicherheiten und die Komplexität von KI-basierten Lokomotivsystemen während ihrer Integration, Abnahme und Zulassung abdecken. Zusätzlich werden KI-Methoden verwendet, um Testfälle zu erzeugen, die realitätsnahe Umgebungsbedingungen simulieren und die Testabdeckung verbessern.

   In diesem Kontext hat Fraunhofer FOKUS einen risikobasierten Testansatz zur Priorisierung von Testfällen für Bahntechnik-Komponenten entwickelt. Dabei werden Feature-Vektoren und Kombinationen identifiziert, die aufgrund statistischer Unsicherheiten und domänenspezifischer Konsequenzen das Gesamtrisiko einer Anwendung erhöhen könnten. Parallel dazu wurde eine Testautomatisierungs- und Simulationsplattform erstellt, die es ermöglicht, sicherheitskritische Anforderungen zu prüfen.

3. **Prozessentwicklung für die Zulassung sicherheitskritischer KI-Systeme:** KI-LOK entwickelt einen Prozess zur sicheren Zulassung von KI-basierten Systemen für den Bahnbetrieb, der auf die speziellen Anforderungen sicherheitskritischer Software mit ML-Anteilen eingeht. Dieser

Prozess wird in enger Abstimmung mit Zulassungsbehörden wie dem Eisenbahnbundesamt gestaltet, um die Einhaltung aller relevanten Sicherheitsanforderungen zu gewährleisten.

Fraunhofer FOKUS integrierte in diesem Zusammenhang die im Projekt entwickelten Werkzeuge und Methoden mit bereits bestehenden Werkzeugen in einem Ende-zu-Ende-Testprozess, der speziell auf die Bedürfnisse der Bahntechnik zugeschnitten ist.

4. **Validierung und Demonstrator-Plattform:** Zur Validierung der entwickelten Techniken und Methoden werden zwei Fallstudien durchgeführt. Die erste Fallstudie befasst sich mit der Objekterkennung im Lichtraumprofil eines Zuges. Die zweite Fallstudie konzentriert sich auf die sichere Eigenlokation im Zug-Odometriesystem. Das System agiert rein Kamerabasiert hierbei werden die statischen Objekte zur Referenzierung herangezogen. Eine Erweiterung der Schnittstellen ist vorgesehen. Zu Evaluationszwecken wird eine Demonstrator-Plattform wird aufgebaut, mit der die Ergebnisse des Projekts praxisnah präsentiert und der Industrie zugänglich gemacht werden können, einschließlich der Integration der Methoden und Werkzeuge aus den Fallstudien zur Demonstration ihrer Effektivität im Entwicklungs- und Qualitätssicherungsprozess.

   In diesem Kontext integrierte Fraunhofer FOKUS die von Fraunhofer FOKUS im Projekt entwickelten Werkzeuge und Methoden in die Werkzeugkette des Demonstrators und erstellte einen experimentellen Laboraufbau für die Simulation der beiden Anwendungsfälle.

2 Voraussetzungen des Vorhabens

Das Projekt KI-LOK wurde in einem Kontext wachsender Anforderungen an die Sicherheit von KI-basierten Systemen in sicherheitskritischen Bereichen durchgeführt, insbesondere im Bahnwesen. Dies spiegelt sich in der zunehmenden Relevanz von Verifikations- und Validierungsmethoden (V&V) für KI-Systeme wider, die sowohl von Industrie als auch Regulierungsbehörden gefordert werden.

Mit dem steigenden Interesse an der Integration von maschinellem Lernen (ML) und KI in sicherheitskritische Systeme, etwa in der Bahntechnik, stieg auch der Bedarf an spezifischen V&V-Techniken, die den besonderen Herausforderungen solcher Technologien gerecht werden. Traditionelle Ansätze zur Softwareprüfung erwiesen sich für KI-basierte Systeme als unzureichend, da ML-Systeme sich potenziell kontinuierlich weiterentwickeln und im Gegensatz zu traditionellem Softwarecode zu komplex zur vollständigen Analyse sind. Daher stoßen klassische Verfahren zur Verifikation und Validierung schnell an ihre Grenzen.

Im Zusammenhang mit der fortschreitenden Digitalisierung und dem Interesse am Einsatz von KI in sicherheitskritischen Bereichen nahmen auch die regulatorischen Anforderungen zu. Für KI-Systeme im Bahnwesen, die in den hochregulierten Bereich sicherheitskritischer Infrastruktur fallen, müssen spezifische Zulassungsprozesse entwickelt und etabliert werden. Dies erfordert nicht nur technologische Fortschritte, sondern auch eine enge Zusammenarbeit mit den Regulierungsbehörden, wie beispielsweise dem Eisenbahnbundesamt.

Frühere Projekte und Forschungsarbeiten im Bereich der KI-Absicherung haben gezeigt, dass bestehende V&V-Methoden unzureichend sind, um die Herausforderungen von KI in sicherheitskritischen Bereichen vollständig zu adressieren. Insbesondere die fehlende Transparenz von Black-Box-Modellen und die Erklärbarkeit von KI-Entscheidungen wurden als kritische Faktoren identifiziert.

KI-LOK wurde in einer Umgebung durchgeführt, in der eine starke Nachfrage nach industrietauglichen Lösungen zur Sicherstellung der Qualität und Sicherheit von KI-Systemen bestand. Der Einsatz von Testautomatisierung und Simulationsumgebungen spielte eine zentrale Rolle, um realitätsnahe Bedingungen für die Testszenarien zu schaffen und die Anforderungen an sicherheitskritische Zulassungsverfahren zu erfüllen. Auch der Aufbau einer Demonstrator-Plattform zur praktischen Validierung der entwickelten Werkzeuge und Methoden war eine wichtige Voraussetzung für den Erfolg des Projekts.

3 Planung und Ablauf des Vorhabens

Im Rahmen des Projekts stehen drei wesentliche Forschungs- und Entwicklungsaspekte im Fokus, die jeweils zur Erreichung der übergeordneten Projektziele beitragen. Diese betreffen die Definition von Anforderungen aus der Bahntechnik (AP1), die Verifikation und Validierung von KI-Technologien (AP2), sowie die Entwicklung modellbasierter Techniken für den Test von KI-basierten Bahntechnikkomponenten (AP3). Zudem wird eine umfassende Methodik zur Absicherung (AP4) erarbeitet, um die Zuverlässigkeit und Sicherheit der entwickelten Techniken sicherzustellen. Es folgt eine grobe Aufschlüsselung der Hauptarbeitspakete.

3.1 Hauptarbeitspaket 1 (HAP1)

Das HAP1 konzentrierte sich auf die Definition von Fallstudien sowie die Erhebung und Integration der Anforderungen, die sowohl aus technischer als auch wissenschaftlicher Sicht an die im Projekt entwickelten Techniken und Methoden gestellt werden. Zudem enthielt dieses Arbeitspaket die Evaluation der Projektergebnisse anhand der Fallstudien.

3.2 Hauptarbeitspaket 2 (HAP2)

Im HAP2 wurde der Fokus auf die Entwicklung von Validierungstechniken für maschinelles Lernen (ML) gelegt. Hier wurden Methoden und Werkzeuge zur Prüfung der Modell- und Datenqualität sowie zur Anwendung geeigneter Testtechniken entwickelt, um die Qualität von ML-Komponenten sicherzustellen.

3.3 Hauptarbeitspaket 3 (HAP3)

Das HAP3 beschäftigte sich mit der Validierung autonomer Systeme (AS), die ML-Komponenten enthalten können. Im Mittelpunkt standen risikobasierte Testtechniken, die selbst KI-basierte Optimierungsverfahren nutzen, um die Abnahme, Zulassung und Zertifizierung dieser Systeme zu unterstützen.

3.4 Hauptarbeitspaket 4 (HAP4)

HAP4 verfolgte das Ziel, die in HAP2 und HAP3 entwickelten Ergebnisse durch eine umfassende Methodik für die Verifikation und Validierung zu integrieren. Eine Experimentierplattform wurde bereitgestellt, die sowohl die entwickelten Werkzeuge vereint als auch zur Schulung und Lehrzwecken genutzt werden kann.

3.5 Hauptarbeitspaket 5 (HAP5)

Schließlich wurden im HAP5 projektübergreifende Aktivitäten koordiniert, wie die Verbreitung und Verwertung der Projektergebnisse sowie der Wissenstransfer in Standardisierungs- und Normungsgremien. Damit ist sichergestellt, dass die Projekterkenntnisse nachhaltig in die Praxis und in technische Richtlinien überführt werden.

4 Wissenschaftlicher und technischer Ausgansstand

4.1 Genutzte bekannte Konstruktionen, Verfahren und Schutzrechte

Existierende Simulatoren oder Werkzeuge zum Bauen von Simulatoren, welche vom Fraunhofer FOKUS im Rahmen des Projekts genutzt wurden, sind unter anderem der Drohnen- und Autosimulator AirSim und die von Epic Games entwickelte Unreal Engine. Genutzte Datensätze sind u.a. PETA (pedestrian attribute dataset), Cityscapes, RailSem19 und das im Laufe der Projektlaufzeit veröffentlichte OSDaR23.

Ein bei der Demonstrator-Integration genutztes Werkzeug zur Visualisierung und Strukturierung von Zeitreihendaten jeglicher Art ist die in der Programmiersprache Rust entwickelte ReRun Applikation inklusive dazugehöriger Entwicklungskits.

Verfahren, welche im Rahmen des Projekts genutzt, evaluiert oder weiterentwickelt wurden und gegebenenfalls nicht bereits anderweitig hier erwähnt wurden, sind in den jeweiligen wissenschaftlichen Publikationen einsehbar und werden im Unterkapitel zum themenorientierten Ausgangszustand genauer aufgeführt.

4.2 Verwendete Fachliteratur und Informations- und Dokumentationsdienste

Konkrete Fachliteratur, welche oft auch genutzte Verfahren beschreibt, ist in den Kapiteln zu genutzten bekannten Konstruktionen und zum themenorientierten Ausgangszustand genauer aufgeführt. Das Fraunhofer FOKUS hat neben gängigen Quellen für wissenschaftliche Arbeiten keine gesonderten Informations- und Dokumentationsdienste genutzt.

4.3 Themenorientierter Ausgangszustand

4.3.1 Validierungs- und Verifikationstechniken für ML

Die Forschung an dedizierten Methoden zur Verifikation und Validierung von KI stand am Anfang. Klassische Verifizierungs- und Validierungsansätze wie Testen, Modellprüfung und Theorembeweisen sind bei der Verifikation unvollständiger Spezifikationen im Allgemeinen begrenzt. In den meisten Fällen ist es entweder unpraktisch oder unmöglich, ML mit bestehenden Techniken formal zu verifizieren. Derzeit wird an mehreren Stellen an der Anpassung von Werkzeugen zur Modellprüfung gearbeitet. Zum Einsatz kommen unter anderem Solver der Klassen Satisfiability Modulo Theory und Mixed Integer Programming (z.B. Katz, G. et al., *Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks*, CAV 2017). Einige Autoren kombinieren ML und Modellprüfung, so dass, falls die gewünschten logischen Eigenschaften durch ein trainiertes Modell nicht erfüllt werden, das Modell („Modellreparatur“) oder die Daten („Datenreparatur“), aus denen das Modell gelernt wird, systematisch modifiziert werden. Andere schlagen vor, die formale Verifikation mit verifizierter Laufzeitüberwachung zu kombinieren, sodass ein sicheres Lernen garantiert werden kann. Werkzeuge wie DeepXplore (Pei, K. et al., Commun. ACM 2019), DLFuzz (Guo, J. et al., ESEC/FSE 2018) und TensorFuzz (Odena, A. et al., ICML 2019) bieten verschiedene Metriken zur Quantifizierung der neuronalen Abdeckung und vereinfachen die Testautomatisierung. Für industrielle Anwendungen, insbesondere der Bahntechnik, war kein entsprechendes Werkzeug verfügbar.

Der Ausganszustand in Hinblick auf die Transferierbarkeit von adversariellen Angriffen zeigt, dass zahlreiche Faktoren eine Rolle spielen. Frühere Arbeiten betonen, dass die Wahl des Angriffsalgorithmus die Übertragbarkeit adversarieller Beispiele stark beeinflussen kann (Liu, Y. et al., *Delving into Transferable Adversarial Examples and Black-Box Attacks*, 2017). So haben gradientenbasierte Methoden, wie der Fast Gradient Sign Method (FGSM) (Goodfellow, I.-J. et al., *Explaining and Harnessing Adversarial Examples*, ICLR (Poster) 2015), oft eine geringere

Transferierbarkeit im Vergleich zu suchbasierten Ansätzen. Weitere Forschungsergebnisse zeigen, dass die Modellarchitektur, Kapazität und Genauigkeit sowohl des Quell- als auch des Zielmodells eine entscheidende Rolle spielen (Wu, L. et al., *Understanding and Enhancing the Transferability of Adversarial Examples*, 2018). Besonders Modelle mit einer glatten Verlustfunktion neigen dazu, eine bessere Übereinstimmung der Gradienten zu ermöglichen, was die Transferierbarkeit von Angriffen verbessert. Während eine hohe Transferierbarkeit bei Modellen mit ähnlicher Genauigkeit festgestellt wurde, zeigt sich bei quantisierten Netzwerken eine deutlich geringere Übertragbarkeit von Angriffen, selbst wenn beide Netzwerke eine ähnliche Leistung aufweisen (Bernhard, R. et al., *Impact of Low-Bitwidth Quantization on the Adversarial Robustness for Embedded Neural Networks*, CW 2019). Der Transfer zwischen Vollpräzisionsmodellen und ihren quantisierten Versionen stellt dabei eine besondere Herausforderung dar. Obwohl in der Forschung bereits Algorithmen untersucht wurden, die die Transferierbarkeit von Angriffen auf quantisierte Netzwerke betreffen, liegt der Fokus oft auf Methoden, die Verlustgradienten nutzen oder schätzen (Matachana, A.-G. et al., *Robustness and Transferability of Universal Attacks on Compressed Models*, 2020 und Zhao, Z. et al., *To compress or not to compress: Understanding the Interactions between Adversarial Attacks and Neural Network Compression*, MLSys 2019). In realen Szenarien kann es sein, dass ein Angreifer keine genauen Informationen über die Quantisierungsstufe oder andere spezifische Eigenschaften des Zielmodells besitzt. Diese Lücken verdeutlichen den Bedarf an weiteren Untersuchungen in diesem Bereich.

Der Ausgangszustand im Bereich der Verifizierung und Validierung von Deep Neural Networks (DNN) wird durch die Black-Box-Eigenschaften dieser Modelle stark eingeschränkt. Diese Intransparenz macht es schwierig, die Entscheidungsprozesse der Modelle vollständig zu verstehen, was die Vertrauensbildung erschwert (Willers, O. et al., *Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks*, SAFECOMP 2020). Obwohl erklärbare Methoden keine vollständige Sicherheit garantieren, tragen sie dazu bei, Unsicherheiten in der Vorhersage zu reduzieren und Vertrauen in das Modellverhalten zu schaffen. Einige Methoden bieten quantitative Erklärungen, die Einblicke in den Entscheidungsprozess geben und helfen, irrelevante Merkmale zu entfernen, die das Ergebnis nicht beeinflussen (Marco Tulio, R. et al., *Explaining the Predictions of Any Classifier*, ACM SIGKDD 2016). Verschiedene Ansätze zur Erklärbarkeit, insbesondere modell-agnostische Methoden, analysieren maschinelle Lernmodelle, indem sie Eingabedaten manipulieren und die resultierenden Vorhersagen untersuchen (Christoph, M. et al., *Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges*, ECML PKKD 2020). Saliency-Methoden, die oft für die Bildklassifikation verwendet werden, visualisieren die Wichtigkeit einzelner Bildpixel, sind jedoch in ihrer Fähigkeit, tiefergehende Konzepte zu interpretieren, begrenzt (Karen, S. et al., *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*, ICLR 2013). Post-hoc-Methoden bieten eine weitere Möglichkeit, indem sie die Beziehung zwischen den Eingabefaktoren und den Modellparametern nachträglich analysieren, um einfache Erklärungen zu liefern (Poursabzi-Sangdeh, F. et al., *Manipulating and Measuring Model Interpretability*, CHI 2021). Ein vielversprechender Ansatz der Erklärbarkeit sind Concept Activation Vectors (CAV) (Kim, B. et al., *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors*, ICML 2017). Hierbei werden Modelle auf der Ebene von Konzepten statt auf der Pixel-Ebene interpretiert.

Zur Überprüfung von Deep Neural Networks (DNNs) wurden bereits viele Methoden und Metriken entwickelt, die sich vor allem auf Testadäquanzkriterien stützen. Zu den frühesten Vorschlägen gehört das Neuron-Coverage-Kriterium von Pei et al. (2017), das die Abdeckung der Aktivierungen von Neuronen misst. Auf dieser Grundlage entwickelten Ma, Juefei-Xu, F. Zhang et al. (2018) zusätzliche Metriken, wie die Neuron Boundary Coverage, die Aktivierungen außerhalb festgelegter Grenzen erfasst. Weitere Ansätze umfassen die T-Way Combination Sparse Coverage (Ma, Juefei-Xu, Xue et al.,

2019), die die Kombinationen aktivierter Neuronen analysiert. Sun et al. (2018) führten Metriken wie Sign-Sign und Value-Sign Coverage ein, die auf der Idee basieren, Signifikanzänderungen in neuronalen Aktivierungen zwischen Schichten zu erfassen. Kim et al. (2019) schlugen Surprise Adequacy (LSA und DSA) vor, die basierend auf der Abweichung von Aktivierungsmustern im Vergleich zu Trainingsdaten arbeiten. Odena et al. (2019) und Ma, F. Zhang et al. (2018) führten zusätzlich die Konzepte Simplified-DSA und Mutation Score ein, um Abweichungen in der Modellleistung bzw. durch eingeführte Fehler zu messen. Trotz dieser Vielfalt gibt es Kritik an der Anwendbarkeit und Skalierbarkeit dieser Methoden, insbesondere in realen DNNs mit komplexen Strukturen.

Auch die Bewertung der Datenqualität ist entscheidend für die Zuverlässigkeit von maschinellen Lernmodellen, insbesondere zur Erkennung von "Out-of-Distribution"-Daten, also Daten, die außerhalb der Trainingsverteilung liegen. Frühere Ansätze zur Verbesserung der Datenqualität umfassen die Synthese neuer Daten (Simard, P.-Y. et al., *Best practices for convolutional neural networks applied to visual document analysis*, ICDAR 2003), oft mittels Generative Adversarial Networks (GANs), die jedoch Probleme wie mode-collapse aufweisen und nicht alle Datenverteilungen abbilden können. Alternativ werden Daten aus öffentlichen Archiven oder dem Internet gesammelt, was jedoch häufig zu ungenauer oder verfälschter Datenqualität führt. Zur Lösung solcher Probleme wird vermehrt auf die Erkennung von "Out-of-Distribution"-Daten gesetzt, z. B. um Veränderungen der Datenverteilung (Data Drift) zu identifizieren (Yang, J. et al., *Generalized Out-of-Distribution Detection: A Survey*, CoRR 2021). Hierbei wird überprüft, ob ein Datenpunkt zur Verteilung der Trainingsdaten passt, indem Wahrscheinlichkeiten von Klassifikatoren genutzt werden. Dies hat Implikationen und Wert zur Verbesserung von vollautomatisierten Testprozessen, insbesondere in der Phase der Datensammlung und Datenaufbereitung.

4.3.2 Modellbasiertes Testen von KI-basierten Bahntechnikkomponenten

Das systematische Testen von Software ist eine der bekanntesten und effektivsten Verifikations- und Validierungsmethoden für softwarebasierte Systeme. Die ständig zunehmende Komplexität der Testaufgaben in Kombination mit dem Mangel an Flexibilität, Wartbarkeit und Nutzbarkeit bei der Erstellung und Wartung der erforderlichen Modellartefakte im Rahmen des modellbasierten Testens ist immer noch eine Herausforderung, die es zu lösen gilt. Speziell für sicherheitskritische Systeme bietet eine systematische Kombination von Risikobewertung und Testen eine Möglichkeit, die bewerteten Risiken des Softwareprodukts als Leitfaktor zur Steuerung aller Phasen eines Testprozesses zu verwenden. In den letzten Jahrzehnten hat die Forschung industrietaugliche Techniken zur Steigerung der Qualität, Effizienz und Zuverlässigkeit von Testprozessen entwickelt, die auch im Bahnbereich Einzug gehalten haben. Selbst moderne, modellbasierte und suchbasierte Testansätze sind jedoch nicht flexibel und leistungsfähig genug, um die für den Test von KI-basierten Bahntechnikkomponenten erforderlichen Anforderungen abzudecken.

Während viele Simulatoren im Automobilbereich (z.B. Carla) existieren, existierten keine Simulatoren für den Bahnbereich, welche den von uns identifizierten Anforderungen zum systematischen Test von Perzeptionssystemen, welche KI-Komponenten enthalten, gerecht werden. Existierende Bahnsimulatoren (z.B. Train Simulator, RailSim) sind als Spiel konzipiert und wurden daher mit signifikant anderem Fokus entwickelt.

Die Entwicklung einer Simulationsumgebung kann ressourcenintensiv sein, insbesondere für Teams mit begrenztem Budget. In solchen Fällen ist es notwendig, auf bestehende Simulatoren zurückzugreifen, auch wenn diese Einschränkungen aufweisen. Sie können dennoch wertvolle Einblicke in die Tests autonomer Systeme bieten. So wurde z.B. der Rennsimulator TORCS für

Reinforcement Learning zur Spurhaltung verwendet (Sallab, A.-E. et al., *End-to-End Deep Reinforcement Learning for Lane Keeping Assist*, CoRR 2016), und "Grand Theft Auto V" genutzt, um Objekt-Erkennungsmodelle zu testen (Wang, D. et al., *Deep object-centric policies for autonomous driving*, ICRA 2019). Während diese Ansätze kostengünstig sind, gelten fortschrittliche Simulatoren wie CARLA und LGSVL in der Automobilindustrie als führend. Im Gegensatz dazu gibt es in der Eisenbahnindustrie nur wenige Simulationsmöglichkeiten. TrainSim (D'Amico, G. et al., *Trainsim: A railway simulation framework for lidar and camera dataset generation*, IEEE TITS 2023) ist ein bemerkenswerter Simulator, der speziell für autonome Zugsysteme entwickelt wurde und Funktionen wie simulierte Kameras und LIDAR bietet. Kommerzielle Spiele wie "Train Simulator Classic" könnten zwar angepasst werden, verfügen jedoch nicht über spezialisierte Tools für autonome Systeme. CARLA, entwickelt mit der Unreal Engine 4, ist ein Open-Source-Simulator für autonome Autos und bietet städtische Umgebungen mit vordefinierten Verkehrsregeln. Es unterstützt drei Haupttestmethoden: Modular Pipeline (Umweltwahrnehmung, Routenplanung und Fahrzeugsteuerung), Imitationslernen (von menschlichen Fahrern) und Reinforcement Learning (belohnungsbasierte Optimierung). Die Flexibilität von CARLA liegt in seiner Open-Source-Natur, die Anpassungen und die Integration von simulierten Sensordaten wie Kameras und GPS ermöglicht. LGSVL, entwickelt auf der Unity-Engine, lässt sich in andere Tools integrieren und ermöglicht die Erstellung digitaler Zwillinge von realen Umgebungen sowie anpassbare Testszenarien mit steuerbaren Verkehrsbedingungen. Es unterstützt auch die Implementierung eigener Sensoren, was Flexibilität bei der Prüfung sensorgestützter Algorithmen bietet. Während es viele Simulatoren für autonome Autos gibt, ist TrainSim die Hauptoption für autonome Züge. Entwickelt mit der Unreal Engine 4, bietet es simulierte Sensoren und eine API für den Bau individueller Gleise sowie ein GEO-World-System zur Simulation unterschiedlicher Landschaften. Allerdings sind die vollständigen Implementierungsdetails von TrainSim nicht leicht zugänglich.

In den letzten Jahren hat die Forschung erhebliche Fortschritte bei der Nutzung generativer KI zur Verbesserung synthetischer Bilddaten erzielt, um fotorealistische Ergebnisse zu erzeugen. Ein zentrales Thema ist die Überwindung der "Domain Gap" zwischen realen und synthetischen Daten. (Wang R. et al., *Improving the Effectiveness of Deep Generative Data*, CVPR) zeigten, dass Modelle wie StyleGAN-XL zwar schnelle Konvergenz bieten, aber schlechter generalisieren, da synthetische Bilder oft nur häufige Merkmale darstellen. (Zhu J.-Y. et al., *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks* (*CycleGAN*), ICCV) stellten das Tool CycleGAN vor, das durch unüberwachtes Lernen Bildübersetzungen ohne gepaarte Trainingsdaten ermöglicht. Diese Methode verbessert den Realitätsgrad von Bildern und kann ein Bild auch im Stil verändern, hat jedoch Schwierigkeiten bei großen geometrischen Veränderungen. (Guo X. et al., *Generative AI for Synthetic Data Generation: Methods, Challenges and the Future*, arXiv) betonten die Bedeutung von fortschrittlichem Prompt Engineering und Few-Shot-Learning, um synthetische Daten vielfältiger und aufgabenspezifischer zu machen, wodurch die Datenqualität und Relevanz für spezifische Aufgaben verbessert wird. Diese Arbeiten verdeutlichen die Herausforderungen und Fortschritte in der synthetischen Bildgenerierung mittels generativer KI. Insbesondere in der Autodomäne wurden auch Methoden zur Verbesserung synthetischer Daten mit KI entwickelt. In der Arbeit von (Richter S.-R. et al., *Enhancing Photorealism Enhancement*, IEEE Transactions on Pattern Analysis and Machine Intelligence) wird eine Methode zur Verbesserung der Fotorealität synthetischer Bilder vorgestellt. Die Autoren nutzen ein neuronales Netz, das Zwischenrepräsentationen von klassischen Rendering-Pipelines verwendet, und trainieren es mit einem neuartigen adversarialen Ziel, um starke Überwachung auf verschiedenen perzeptuellen Ebenen zu gewährleisten. Diese Arbeit zeigt, dass die Unterschiede in der Szenenlayout-Verteilung zwischen gängigen Datensätzen eine Ursache für

Artefakte in früheren Methoden sein könnten, und schlägt Verbesserungen in der Netzarchitektur und Patch-Sampling-Strategie vor.

4.3.3 Absicherungs- und Abnahmeprozesse für KI-basierten Bahntechnikkomponenten

Zwei der größten technischen Hürden bei der Begutachtung und Zulassung von lernenden Systemen nach EN50128 sind (a) der Mangel an Transparenz (Black-Box-Modell) und (b) die fehlende Fähigkeit einer KI zur Erklärung einer Entscheidung. Um das Transparenzproblem (a) zu lösen, wurden verschiedene Ansätze zur Modellinterpretation, zur Modell-/Entscheidungserklärung, zur Modellzerlegung, sowie zur Extraktion von Entscheidungsbäumen aus neuronalen Netzen entwickelt. Zur Lösung des Erklärbarkeitsproblems (b) wurden in den letzten Jahrzehnten verschiedene Ansätze als Kombination von maschinellem Lernen und symbolischem Lernen realisiert. Die Erklärung einer Entscheidung wird in erster Linie mit der Fähigkeit zur Repräsentation des Wissens über mehrere symbolischen Systeme sowie mit Verfahren zur Robustheit des Lernens realisiert. Für die Bahntechnik muss als Erklärung die Beherrschung der tolerierbaren Gefährdungsrate (Tolerable Hazard Rate, THR) nachgewiesen werden, was beim Stand der Technik allerdings noch nicht gelungen ist.

5 Zusammenarbeit mit anderen Stellen

Fraunhofer FOKUS hat im Rahmen des Projekts Kontakt zu nationalen und internationalen Standardisierungsgremien (DIN/DKE NA 043-01-42, ETSI MTS) aufgebaut. Darüber hinaus hat sich Fraunhofer FOKUS an der Ausarbeitung der zweiten Ausgabe der "Normungsroadmap KI" durch die Aktivitäten der AG "Qualität, Konformität und Zertifizierung" beteiligt und dort Anforderungen, Erfahrungen und Erkenntnisse u.a. aus dem KI-LOK Projekt eingebracht. Weiterhin hat Fraunhofer FOKUS Ergebnisse aus dem KI-LOK Projekt als initiale Version eines neuen Work Items bei ETSI MTS eingereicht (ETSI DTR/MTS-103910 - MTS AI Testing Test Methodology and Test Specification for AI-enabled Systems). Das Dokument befasst sich mit dem Testen von KI-fähigen Systemen zum Zweck der Standardisierung und erläutert Testmethodologien und Methoden für die Testspezifikation. Es identifiziert Anforderungen für das Testen und macht Vorschläge, wie die technischen Aspekte der Zertifizierung der Vertrauenswürdigkeit von KI in Normungskontexten angegangen werden können. Zusätzlich hat Fraunhofer FOKUS im letzten Jahr an der IEEE P3407 Standardisierungsinitiative „Standard for End-to-End Software Testing Automation Tools“ mitgewirkt. Diese Arbeit stärkt die Standardisierungsbemühungen in der Testautomatisierung für KI-gestützte Systeme und schafft eine zusätzliche Basis für die Projektergebnisse.

Schlussendlich wurde im Rahmen des Projekts mehrfach Treffen mit dem Deutschen Zentrum für Schienenverkehrsforschung (DZSF) abgehalten, um Ergebnisse des Projekts zu teilen und Möglichkeiten zur Standardisierung im Bahnbereich auszuloten.

6 Erzielte Ergebnisse des Fraunhofer FOKUS

Im Folgenden werden die Projektergebnisse, an denen das Fraunhofer FOKUS mitgewirkt hat, oder welche das Fraunhofer FOKUS selbst voll erzielte, im Einzelnen aufgeführt. Die Hauptergebnisse sind:

1. **Testautomatisierungs- und Simulationsplattform Perception-Lab:** Im Rahmen des Projekts wurde die Testautomatisierungs- und Simulationsplattform Perception-Lab entwickelt und entlang der Projektfallstudien evaluiert. Perception-Lab ermöglicht es, interaktive veränderbare visuelle 3D Welten zu erstellen, in denen virtuelle Zugfahrten durchgeführt werden können. Im Verlauf dieser virtuellen Zugfahrten werden fotorealistische Bilddaten erzeugt, die für Tests von maschinellen Lernsystemen im Bahnbereich verwendet werden.

Durch die Beschreibung von Szenarien in einer eigens entwickelten Beschreibungssprache konnten relevante Umgebungen realistisch simuliert und die Szenarioerstellung automatisiert werden. Schnittstellen zum Einlesen und Verändern der Szenarien unterstützen interaktive und suchbasierte Testansätze und die Integration mit den Werkzeugen der Projektpartner. Durch die Realisierung virtualisierter Zugfahrten ermöglicht Perception-Lab die Generierung von Testdaten für ausgewählte Rand- und Sonderfälle, die in der Realität nicht einfach realisierbar sind, und behebt somit das Problem fehlender Testdaten für den Test von automatisierten Zugsystemen. Das Perception-Lab erweitert die bereits bestehenden Fraunhofer FOKUS Testwerkzeuge um eine Plattform, mit der KI-basierte Perzeptionssysteme systematisch getestet werden können.

2. **Entwicklung einer probabilistischen und risikoorientierten Methode zur Testauswahl und Gewichtung (ODD-TA):** Fraunhofer FOKUS entwickelte eine risikobasierte Teststrategie, die sich auf probabilistische Ontologien stützte. Grundlage für diesen Ansatz ist eine Ontologie, die die Operational Design Domain (ODD) eines automatisierten Zuges beschreibt. Diese Ontologie ermöglicht es, die verschiedenen Umgebungseigenschaften, wie etwa Wetterbedingungen, Tageszeiten oder spezifische Szenarien, systematisch zu modellieren. Auf dieser Basis wurde eine probabilistische und risikoorientierte Methode zur Testauswahl entwickelt, die es ermöglicht, kritische Testszenarien gezielt zu priorisieren. Die Ontologie diente nicht nur zur systematischen Darstellung der Umgebung, sondern auch zur Identifizierung von Feature-Vektoren, die das Gesamtrisiko signifikant erhöhen könnten. Die Darstellung der Testergebnisse erfolgt in einem Dashboard, das die relevanten Informationen zur Testabdeckung und den Testergebnissen visualisiert. Die Arbeiten zur probabilistischen und risikoorientierten Testauswahl wurden gemeinsam mit dem Projektpartner IT-Power Solutions durchgeführt und im Rahmen des Demonstrators evaluiert.

3. **Methoden zur Validierung und Verifikation von ML-Modellen:** Ein zentraler Bestandteil des Projekts war die Entwicklung von Methoden zur Validierung und Verifikation von ML-Modellen. Hierbei kamen spezielle Testverfahren zum Einsatz, die dynamische, suchbasierte Tests mit formalen Methoden und statischer Analyse kombinieren. Ein zentrales Element der Validierungs- und Verifikationsmethoden ist die Einführung des **Pixel-wise Testing**. Diese Methode ermöglicht es, die relevanten Merkmale, die die Vorhersagen für jedes einzelne Pixel eines Bildes ausmachen, detailliert zu analysieren. Dadurch konnten gezielt neue Testeingaben definiert werden, um sicherheitskritische Fehlklassifikationen zu identifizieren. Die Methode erlaubt es zudem, semantische Segmentierungsmodelle auf einer granularen Ebene zu testen, was besonders für den Einsatz in sicherheitskritischen Anwendungen wie dem Bahnbetrieb von hoher Bedeutung ist. Das Pixel-wise Testing erweitert die von Fraunhofer FOKUS entwickelte Fuzz-Testing Plattform Fuzzino und ermöglicht eine präzise Überprüfung der Vorhersagequalität von ML-Systemen in komplexen Szenarien. Zusätzlich wurden Techniken wie das "randomized smoothing" eingesetzt, um die Robustheit der Modelle zu überprüfen sowie eine Methode entwickelt, mit der sich auf Basis deskriptiver Techniken “out-of-domain"-Daten, also Ausreißer in Datensätzen, erkennen lassen.

Im Folgenden werden diese Ergebnisse weiter ausgeführt und detailliert beschrieben.

6.1 Testautomatisierungs- und Simulationsplattform Perception-Lab

Um die Qualität von KI-System zur Objekterkennung im Bahnbereich zu prüfen, ist es notwendig, solche Perzeptionssysteme auch Situationen auszusetzen, welche normalerweise nicht auftreten. Hierzu bedarf es speziell für das maschinelle Lernen aufgearbeitetes Bildmaterial, welches im

Bahnbereich nicht in hinreichenden Mengen existiert und in der Regel nicht öffentlich zugänglich ist. Die Untersuchung diverser existierender Bahn-Simulatoren sowie anderer Simulatoren wie AirSim und allgemeiner 3D-Engines durch Fraunhofer FOKUS ergab, dass bestehende Lösungen nicht die spezifischen Anforderungen für den systematischen Test von KI-Perzeptionssystemen im Bahnbereich erfüllen. Im Gegensatz zum Automobilbereich fehlen im Bahnbereich geeignete Simulatoren, die den identifizierten Anforderungen für den systematischen Test von KI-basierten Perzeptionssystemen entsprechen.

Das Perception-Lab vereint erfolgreich die Ergebnisse der Hauptarbeitspakete 1 bis 4 und stellt damit eine vollständige Werkzeugkette bereit, die den Projektanforderungen gerecht wird. Das Perception-Lab ermöglichte eine Vielzahl von Experimenten, die wertvolle Einblicke in die Leistungsfähigkeit von KI-basierten Perzeptionssystemen im Bahnbereich lieferten.

6.1.1 Allgemeines zum Perception-Lab

Die Simulationskette des Perception-Labs generiert automatisiert relevante Testdaten, die präzise auf die Anforderungen von KI-basierten Perzeptionssystemen abgestimmt sind und als Test-Stimuli bereitgestellt werden können. Hierbei werden Szenarien mittels einer von Fraunhofer FOKUS entwickelten Szenario-Beschreibungssprache ausgedrückt. Die einzelnen Instanzen eines Szenarios werden in einer 3D Umgebung generiert, aus der dann das benötigte Bildmaterial entnommen wird. Im letzten Schritt der Kette wird das KI-System den Bildern ausgesetzt und die Reaktion mit der Erwartung verglichen. Mithilfe der von uns entwickelten probabilistischen und risikobasierten Methoden, welche in einem folgenden Kapitel näher erläutert werden, lässt sich so eine Operational Design Domain (ODD) spezifizieren und vollautomatisiert risikobasiert testen. Das Fraunhofer FOKUS hat die technische Machbarkeit dieser Methodik in einem Prototyp demonstriert und dies in der wissenschaftlichen Arbeit *Test and Training Data Generation for Object Recognition in the Railway Domain* (SEFM 2022, Jürgen Großmann et al.). Technisch ist dies mittels der Unreal Engine 5 gelöst, welche als Grundlage zur Implementierung der 3D-Visualisierung dient. Diese ist in Teilen über den Unreal Engine 5 Editor entwickelt, aber enthält auch in C++ implementierte Komponenten wie den TCP-Server, welcher die API zur Übermittlung von Testfällen bereitstellt. Diese API wird über einen in Python entwickelte Software angesprochen, welche wiederum die Nutzung des Simulators auf flexible Weise nach außen bereitstellt. Hiermit werden kompliziertere Implementationsdetails und der dem Simulator zugrundeliegende Technologie-Stack vollständig vom Endnutzer wegabstrahiert und durch eine einfach zu nutzende Python-Schnittstelle verfügbar gemacht.

Da generierte Daten synthetischer Natur sind, wurde eine Experimentreihe zur Bestimmung von Güteanforderungen an generierte Bilddaten durchgeführt. Die dadurch identifizierten Gütekriterien und dessen Einhaltung durch den Simulator dienen als Nachweis dafür, dass die von der Simulationsumgebung generierten Bilder als Teststimulus geeignet sind.

6.1.2 Simulator API und Testfallspezifikation

Die eigens entwickelte Szenenbeschreibungssprache ermöglicht die automatisierte Erstellung von Bilddaten mit präziser Kontrolle über Szeneneigenschaften wie Wetterbedingungen und Uhrzeit. Dies ermöglicht es, detaillierte und realitätsnahe Testfälle zu generieren. Zusätzliche Bilddaten mit kritischen Fällen, wie etwa Bäumen auf dem Schienenweg, können automatisch generiert werden. Dadurch können bestehende Datensätze zum Testen von ML-Systemen gezielt in besonders risikorelevanten Bereichen verstärkt werden. Durch eine Modifikation der maschinenlesbaren Spezifikation ist es möglich, die Datensätze gezielt an die Erfordernisse anzupassen. Aufgrund der vollständigen Kontrolle über das Simulationssystems können pixelgenaue und garantiert korrekte

semantische Segmentierungen (u.a. auch Instanzensegmentierungen) und Tiefendaten zusammen mit den Bilddaten erzeugt werden.

Die Szenenbeschreibungssprache nutzt das maschinenlesbare Format JSON und repräsentiert Szenenelemente lokaler und globale Natur auf deskriptive Weise. Der Simulator, die zentrale Komponente des Perception-Labs, verarbeitet Testfälle in der Szenenbeschreibungssprache und realisiert diese.

Die in der Sprache beschreibbaren globalen Eigenschaften oder Modifikationen der Szene sind:

- **Uhrzeit**
  Dies wirkt sich auf die Lichtverhältnisse aus, u.a. die Sonnenposition, atmosphärischen Eigenschaften, welche je nach Winkel das Sonnenlicht beeinflussen, oder aber auch die Reflektion des Monds, etc.
- **Wetterbedingungen**
  Die Wetterbedingungen lassen sich spezifizieren basierend auf den folgenden Parametern:
  - Wolkendichte
  - Regenstärke
  - Schneestärke
  - Blitzauftrittsdichte
  - Windstärke
  - Nebeldichte
  - Staubdichte

Das Zusammenspiel solcher Szeneneigenschaften wird mitsimuliert. Das bedeutet, dass z.B. eine hohe Windstärke sich auch auf die visuelle Darstellung von Regen auswirkt, und eine hohe Wolkendichte Konsequenzen für die Belichtung der Szene hat (wie z.B. die Schattenwirkung von Wolken).

Die Eigenschaften haben auch sekundäre Konsequenzen für die Szene wie z.B. die Bedeckung von Schnee, das Tropfen und Herabfließen von Regenwasser an Objekten, oder die Bildung von Pfützen. Diese Wirkungen sind im Simulator realisiert und benötigen keine gesonderte Spezifikation.

Neben den globalen Szeneneigenschaften lässt sich auch die Szene selbst spezifizieren. Dies enthält den Verlauf von Schienen, den Verlauf von Strommasten, und die Position von Objekten auf oder entlang der Gleise.

Neben **Schienen** und **Strommasten** sind folgende Objekte spezifizierbar:

- **Laubbäume** (verschiedene Varianten)
- **Nadelbäume** (verschiedene Varianten)
- **Menschen** (verschiedene Menschen mit verschiedenen Eigenschaften)
- **PKWs** (Verschiedene Autos mit verschiedenen Eigenschaften)
- **Signale**

Diese Menge an Objekten ist einfach erweiterbar. Aufgrund der Notwendigkeit von 3D Assets zur Darstellung von Objekten in der Simulationsumgebung wurden insbesondere domänenspezifische Objekte wie Signale unter der Nutzung von echten Aufnahmen in 3D Modelle und Texturen überführt.

Die durch die Sprache spezifizierte Szene kann, mit einer Wahl der zu befahrende Strecke (eine Szene kann natürlich ggf. mehrere Schienenstrecken enthalten), über eine Netzwerkschnittstelle an den Simulator übergeben werden. Konkret betreibt jede Instanz des Simulators einen eigenen TCP-Server,

welcher eine API bereitstellt, welche wiederum Testfallspezifikationen im oben beschriebenen Format entgegennimmt. Dies erlaubt es, den gesamten Perception-Lab Workflow auch auf verteilten Systemen einzusetzen, was zu signifikanten Vorteilen in der Skalierbarkeit auf serieller und paralleler Ebene führt. Aufgrund der Nutzung des Unreal Engine 5 Technologie Stacks ist es möglich, den Simulator auf allen gängigen Plattformen auszuführen.

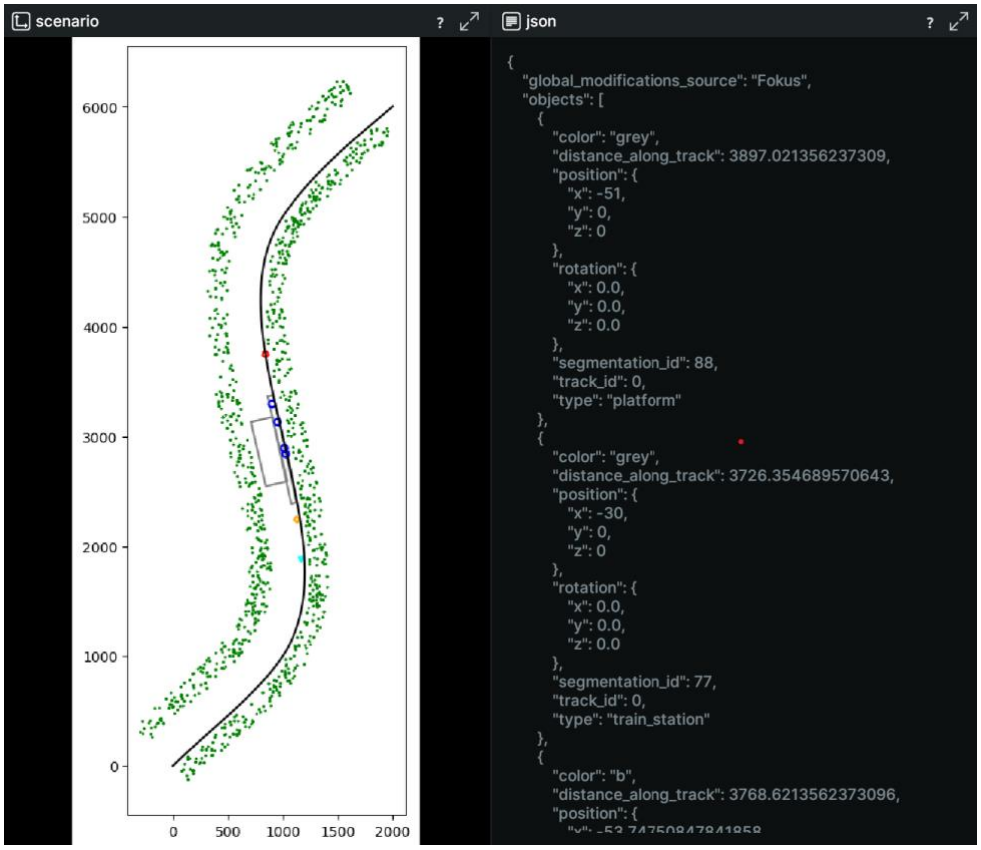

Abbildung 1 – Rechts: Ein Ausschnitt der JSON-Datei welche eine Szene spezifiziert. Links: Eine 2D-Visualisierung der durch die JSON spezifizierten Szene.

6.1.3 Nachverarbeitung von Simulatorausgaben mit generativer KI (GenKI)

Die Testresultate zeigen, dass die synthetischen Bilder von real trainierten Modellen erfolgreich verarbeitet werden, was die hohe Qualität der simulierten Testdaten bestätigt. Hierzu wurden mehrere Objekterkennungs- und Bildsegmentierungsmodelle trainiert und als Systeme unter Test (SUTs) eingesetzt. Dies sind u.a. sowohl ein U-Net als auch ein YOLO-basiertes Modell. Diese Modelle variieren hinsichtlich der Anzahl der Parameter und der erkennbaren Merkmale. Hiermit ist gezeigt, dass die synthetischen Bilder von Systemen verarbeitet werden können, welche auf real Daten trainiert wurden. In das Training der SUT-Modelle wurden auch die Daten des während der Projektlaufzeit veröffentlichten OSDAR23-Datensatzes aufgenommen.

Unter Nutzung der Simulatorausgaben, insbesondere der neben den normalen Bildern mitgelieferten Tiefen- und Segmentierungsmaps, kann durch das Verwenden von generativer KI optional der Realismusgrad der Bilder weiter erhöht werden. Hierzu werden konkret latente Diffusionsmodelle, zum Beispiel Stable Diffusion, verwendet. Diese Modelle können auch durch natürliche Sprache und der Nutzung von LoRA/LyCORIS weiter konditioniert werden.

Abbildung 2 – Oben: Das durch den Simulator generierte Bild. Unten: Vier durch GenKI nachverarbeitete Bilder. Allen vier Bildern liegen die gleichen Eingabebilder zu Grunde (Simulatorbild, Segmentierungsmap, Tiefenmap), haben jedoch unterschiedliche Textbeschreibungen zur Stilveränderung.

Die Anwendung von Methoden wie CycleGAN ist möglich und einfach in den Prozess zu integrieren, aber führt im Falle des Perception-Labs zu vernachlässigbarem Mehrwert. Durch CycleGAN können existierende Szenenbilder z.B. in andere Jahreszeiten oder Lichtverhältnisse mittels generativer Netzwerke überführt werden. Dies ist allerdings bereits im Simulator selbst implementiert, was durch die sich dadurch ergebenden perfekten Labels deutlich besser ist.

6.1.4 6.1.4. Demonstrator und Evaluation

Um vielfältigere Tests durchführen zu können, wurde eine Szene aus der ODD (Operational Design Domain), passend zu den Anforderungen unseres Industriepartners Hitachi nachgebildet. Diese Szene erweitert die Möglichkeiten, spezifische Testdaten zu generieren:

1. Signalobjekte wurden hinzugefügt, um Objekterkennung von Signalen und Signalzuständen testen zu können.
2. Mehrere Oberspannungsleitungsobjekte wurden eingeführt, um den Realismusgrad entlang des Gleises zu erhöhen.

3. Eine Objektkonstellation eines Bahnhofes wurde hinzugefügt, um sicherheitskritische Einfahrten in einem Personenbahnhof simulieren zu können.

Das Perception-Lab ist sowohl mit vollautomatisiert-generierten Testfällen als auch manuell nachgebildeten Szenen evaluiert worden. Hierzu wurden verschiedene Modelle, u.a. das vom Projektpartner Hitachi entwickelte Perzeptionssystem, aber auch von Fraunhofer FOKUS angepasste Varianten des weit verbreiteten YOLO-Objekterkennungsmodell als System-under-Test verwendet. Das Perception-Lab identifizierte sicherheitskritische Defekte, wie die Fehlinterpretation von Signalen oder den Zusammenbruch von Objekterkennung und Segmentierung im Allgemeinen unter besonderen Lichtverhältnissen, was die Effektivität des Systems zur Verbesserung der Modellrobustheit unterstreicht.

Abbildung 3 – Oben: Ein fälschlicherweise als vertikale Stange erkannte Region wird vom Simulator und Evaluationssystem als Fehler erkannt. Unten: Analog mit nicht-existenten Lampen.

6.1.5 ODD-Ontologie-getriebener Perception-Lab Prozess

Der komplette Prozess, welcher das Perception-Lab mit ODD- (Operational Design Domain-) Spezifikationen durch eine probabilistisch annotierte Ontologie vereint, ist demnach wie folgt.

1. Zunächst werden aus der probabilistisch annotierten Ontologie abstrakte Testfälle generiert. Die Ontologie bietet hierbei Aussagen zu allgemeinen Umgebungsbedingungen wie Licht und Wetter, zur Geometrie und Topologie der Schienenabschnitte, aber auch zu lokalen Pertubationen und Szenenelementen. Die daraus hergeleiteten Testfälle können in Bezug auf die ODD mit Blick auf die Auftrittswahrscheinlichkeit im Ganzen oder in Teilen untersucht werden. Diese Testfälle werden daraufhin konkretisiert und in das vordefinierte maschinenlesbare Format überführt, welches der Simulator annimmt.

2. Der Simulator erzeugt automatisiert realitätsnahe Szenendaten, die aus der Perspektive eines fahrenden Zuges gewonnen werden und eine detaillierte Grundlage für präzise Tests von KI-Systemen bilden. Aufgrund der vollen Kontrolle über die Umgebung können hiermit neben den Bildern selbst auch pixelgenaue Segmentierungsmaps (sowohl semantisch als auch instanzenegmentiert) und Tiefenmaps generiert werden. Diese Bilddaten können mithilfe generativer KI weiter verbessert werden, um den Realismusgrad zu erhöhen und damit eine noch realistischere Umgebung für die Testprozesse zu schaffen. Die vom Simulator nun generierten Daten können dann dem zu testenden Perzeptionssystem als Eingabe übermittelt werden. Aufgrund der vom Simulator mitgenerierten korrekten pixelgenauen Grundwahrheit können anschließend die Ausgaben des Perzeptionssystems sicher korrekt evaluiert werden.
3. Die Resultate der Evaluation können ggf. verwendet werden, um wiederum Szenenmodifikationen durchzuführen, um iterativ Regionen im Raum der Eingaben zu finden, in welchen sich das Perzeptionssystem als ggf. unzureichend herausstellt.

Die folgende Abbildung stellt den Prozess genauer dar. Hierzu wird (1) die Szenenbeschreibung über die Schnittstelle an den Simulator übergeben, (2) diese Szenenbeschreibung durch die Schnittstelle verarbeitet und Bilddaten durch den Simulator generiert, (3) der generierte Bilddatensatz vom Simulator exportiert und bereitgestellt, (4) optional dieser Datensatz durch GenKI nachverarbeitet, (5) ein Teil der Daten dem System-under-Test als Eingabe geliefert und die Ausgabe mit anderen vom Simulator bereitgestellte Daten über einen Vergleich evaluiert, und (6) Einblick durch die Evaluation geboten, welche Szenenmodifikationen sich wie auf das zu testende System auswirken und so weiter gesteuert werden können.

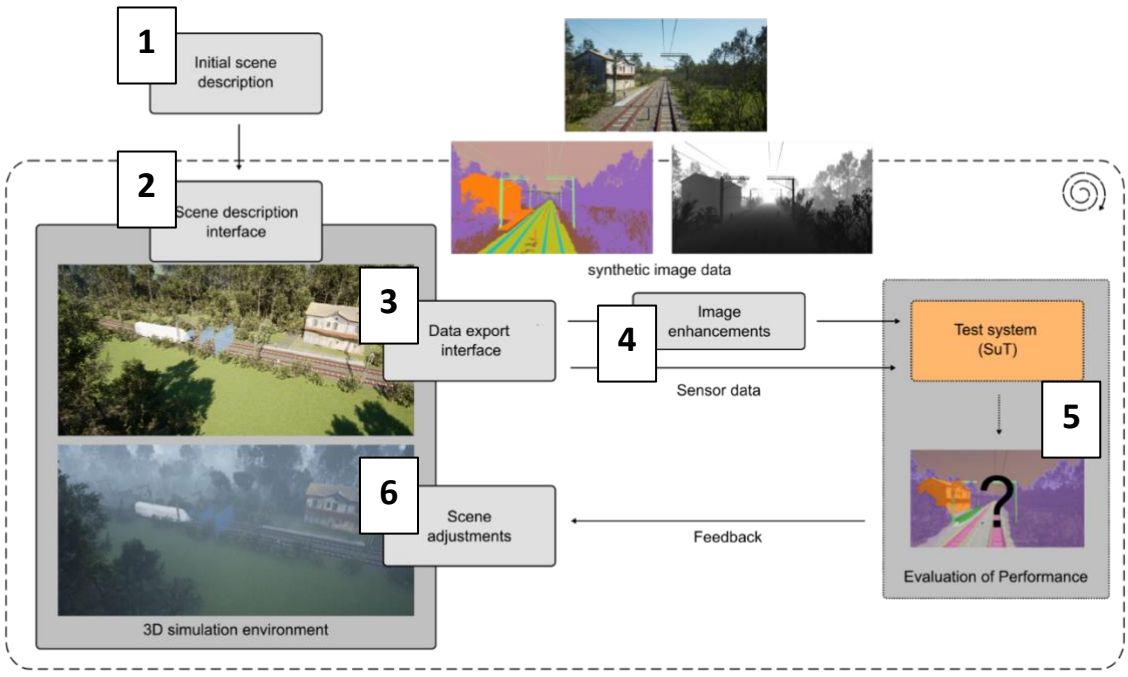

Abbildung 2 - Der integrierte Testprozess unter Nutzung des Simulators.

6.1.6 Dashboard

Mithilfe des ReRun Visualisierungswerkzeugs können vom Simulator generierte Daten z.B. wie in der nachstehenden Abbildung dargestellt werden. Dieses Werkzeug fungiert als flexibles und erweiterbares Dashboard für das Perception-Lab. Damit wurde es ermöglicht, die Testläufe effizienter zu gestalten und die generierten Daten systematisch auszuwerten und zu visualisieren.

Dargestellt sind (1) das vom Simulator generierte Bild, (2) die vom Simulator generierte Segmentierungsmap, (3) die vom Simulator generierte Tiefenmap, (4) die Tiefenmap als 3D-Punktwolke dargestellt, (5) die durch GenKI verarbeiteten Bilder, (6) die System-under-Test (SuT) Vorhersagen auf dem Simulatorbild und (7) die SuT Vorhersagen auf den GenKI Bildern.

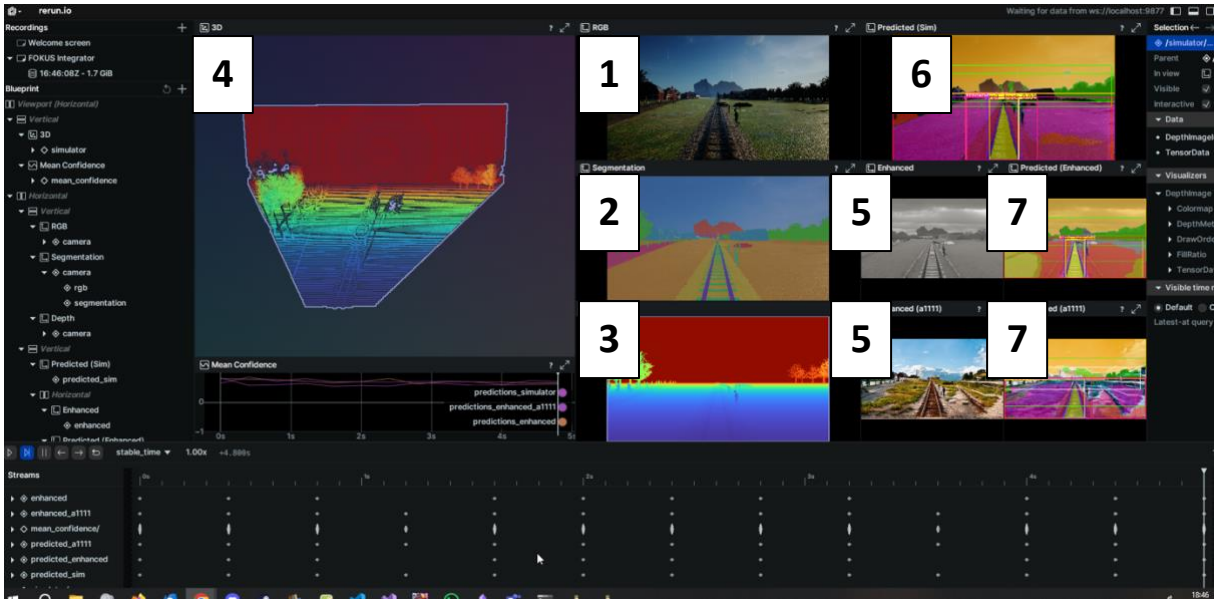

Abbildung 3 - Visualisierung von Simulatordaten und Teilen der Perzeptionssystemevaluation in ReRun.

Das Dashboard ist sowohl manuell als auch programmatisch anpassbar und lässt sich einfach zur Visualisierung und Analyse von neuen Datenströmen erweitern. Das erlaubt die einfache Entwicklung von Plugins für das Perception-Lab, welche die Funktionalität der Werkzeugkette über den jetzigen Stand hinaus erweitern. Sowohl die Darstellung von Visualisierungen als auch die Zusammenführung und Nachverarbeitung von Daten ist durch plug-and-play Python Scripts anpassbar.

6.2 Probabilistische und risikoorientierte Methode zur Testauswahl und Gewichtung

Fraunhofer FOKUS hat im Rahmen der Hauptarbeitspakete 2 und 3 mit der ODD-TA-Methode eine probabilistische und risikoorientierte Methode zur Testauswahl implementiert, die eine präzise und effiziente Ermittlung relevanter Testszenarien ermöglichen. Eine initial von Fraunhofer FOKUS entwickelte Ontologie ermöglicht eine umfassende Beschreibung der relevanten Eigenschaften der Operational Design Domain (ODD) eines automatisierten Zuges. Die ODD basiert auf den Erkenntnissen der Fallstudie Rangierfahrt und gewährleistet eine detaillierte Modellierung des Zugbetriebs in verschiedenen Szenarien. In Anlehnung an eine in der Automobilindustrie etablierte Aufteilung beschreibt die Klassenhierarchie unserer Ontologie auf der obersten Klassenebene ein Schichtenmodell mit Klassen den Klassen *RailroadSystem*, *Actor*, *Scenery*, *EnvironmentalConditions* sowie *TechnicalSystem*.

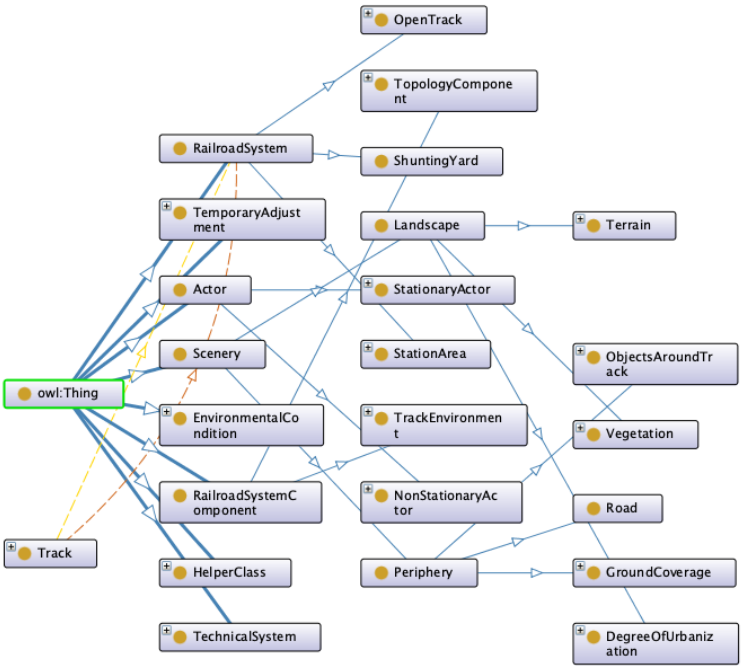

Abbildung 4 - Darstellung der ersten Klassenebenen der KI-LOK Ontologie.

Unter diesen Basisklassen wird dann weiter differenziert. So können Akteure stationär (z.B. ein Licht- oder Formsignal) und nicht stationär (z.B. Lebewesen oder Fahrzeuge) sein. Die Umweltbedingungen enthalten die Jahres- sowie Tageszeit und das Wetter. Dadurch, dass die Ontologie maschinenlesbar ist, können abhängig vom ausgewählten Szenario eine beliebige Anzahl Objekte mit gewünschten Eigenschaften programmatisch ausgewählt werden. Fraunhofer FOKUS hat die Ontologie im Laufe des Projektes regelmäßig erweitert, sodass insbesondere die Eigenschaften von Fahrzeugen, Personen und Personengruppen (auf und neben dem Gleis) ausführlich spezifiziert werden können. Das Fahrzeug und die Person sind Akteure, haben also eine Position, Ausrichtung und Geschwindigkeit. Das Aussehen ist wählbar. Das Fahrzeug kann ein LKW oder ein PKW verschiedener Modelle sein. Die Kleidung der Person oder die Karosserie des Fahrzeugs können einfarbig oder gemustert sein (geblümt, kariert, getarnt usw.); oder sie können zu speziellen Diensten gehören (z. B. Polizei oder Bahnwartung), was ihr Aussehen bestimmt. An den Seiten des Fahrzeugs kann Werbung angebracht werden. Eine Visualisierung eines Ausschnitts der Ontologie folgt.

Die Ontologie kann insbesondere durch die Integration mit dem Perception-Lab als Grundlage für eine vollautomatisierte Testmethode von Perzeptionssystemen im Bahnbereich genutzt werden.

Fraunhofer FOKUS hat dann mit der ODD-TA-Methode ein risikobasiertes, kombinatorische Verfahren zu Generierung von Testfällen basierend auf der Ontologie erarbeitet. Hierbei werden sowohl die semantischen Einschränkungen, die durch die Ontologie definiert sind, beachtet und mit dem sogenannten N-wise Testing ein Ansatz gewählt, der die Anzahl der Testfälle auch für die Komplexität der KI-LOK Ontologie beherrschbar macht. Das Verfahren wurde implementiert und evaluiert. Diese Arbeiten liefen in enger Abstimmung und Koordination mit dem Partner IT-Power Solutions.

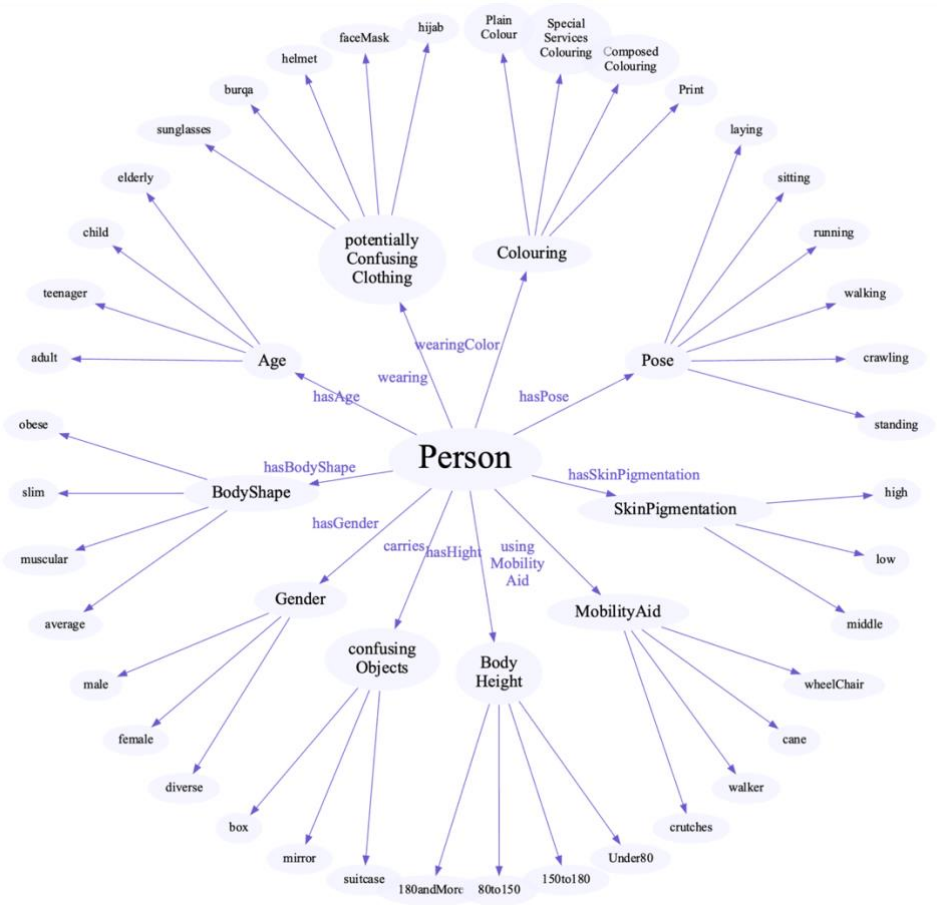

Abbildung 7 - Der Ausschnitt "Person" der Ontologie.

Die gemeinsam mit dem Partner IT-Power Solutions erfolgte Erweiterung der Ontologie um Parameter zur Einbindung empirischer Informationen ermöglicht präzisere Vorhersagen über die Auftrittswahrscheinlichkeit von Objekten und Szenarien. Diese Annotationen erlauben es, statistische Vorhersagen über die Auftrittswahrscheinlichkeit verschiedener Objekte und Szenarien zu modellieren und bei der Testgenerierung berücksichtigen zu können. Dies wird als Basis für eine risikobasierte Testgenerierung genutzt.

Das umfassende Verfahren zum Testen von ML-basierten Black-Box-Systemen gewährleistet eine systematische und gründliche Überprüfung der Systeme auf Basis probabilistischer Methoden. Auf Basis des Konzepts einer probabilistisch erweiterten Ontologie wurde ein Endkriterium für das Testen ML-basierter Systeme definiert. Dieses Konzept wurde zu einem systematischen Testprozess zusammengefasst, der unabhängig von der Entwicklung nach der Black-Box-Methode durchgeführt werden kann. Analog zur traditionellen Systementwicklung werden die Tests unabhängig von den Design- und Entwicklungsaktivitäten spezifiziert und am zu testenden System (SUT) durchgeführt, um die korrekte Implementierung zu überprüfen. Statistische Kriterien für das Testende geben an, wie gründlich die einzelnen Aspekte der Systemumgebung (ODD) getestet werden und bestimmen, wann der Testprozess abgeschlossen werden kann. Dies wurde in der Arbeit *Outline of an Independent Systematic Blackbox Test for ML-based Systems* (IEEE AITest 2024, Hans-Werner Wiesbrock et al.) publiziert. Weiterführende Experimente wurden in der Arbeit *On a systematic Test of ML-based systems: Experiments on test statistics* (IEEE AITest 2024, Nicolas Grube et al.) veröffentlicht. Die ODD-

TA-Methode wurde darüber hinaus mit RACOMAT, einem von Fraunhofer FOKUS entwickelten Werkzeug zur Risikoermittlung, integriert.

Darüber hinaus hat Fraunhofer FOKUS Experimente zur synthetischen Erzeugung von Testdaten mittels metamorpher Transformationen durchgeführt. In Zusammenarbeit mit den Projektpartnern hat Fraunhofer FOKUS als Leiter des Arbeitspakets 3 das Dokument „Objectives and strategies for automated testing of AI-based perception systems in railroad engineering“ verfasst, in welchen gemeinsam mit den Projektpartnern die Anforderungen und Möglichkeiten des Testens von KI-basierten Bahn-Systemen beschrieben sind. Methoden zur Validierung und Verifikation von ML-Modellen

6.3 Methoden zur Validierung und Verifikation von ML-Modellen

Im Rahmen des Hauptarbeitspakets 2 wurde ein Ansatz entwickelt, der es ermöglicht, die Sicherheitsnachweisführung gegenüber Regulatoren im industriellen Einsatz präzise und nachvollziehbar zu gestalten. Motiviert ist dies durch Probleme wie: Was bedeutet es, wenn ein ML-System sich zu 62% sicher ist, einen Menschen nah am Lichtraumprofil der Bahnstrecke erkannt zu haben? Wie hoch müssen die Erkennungswerte sein, damit das Restrisiko der darauf basierenden Entscheidungen akzeptiert werden kann? Und wie lässt sich bezüglich unerwarteter Daten argumentieren – bezüglich der Erkennung von Objekten, für die das Perzeption-System weder trainiert noch getestet wurde? Aus diesen Fragen wurde unter anderem der Ansatz entwickelt, die Leistung menschlicher Lokführer als Maßstab zu nehmen – eine KI gesteuerte LOK soll in kritischen Situationen mindestens so gut reagieren können, wie ein menschlicher Lokführer das im regulären Betrieb zustande bringt. Um bezüglich der Zulassung bzw. allgemein über die Sicherheit eines KI-basiertes Systems für den autonomen Zugbetrieb gegenüber Regulatoren argumentieren zu können, muss die neue Technologie als mindestens genauso sicher wie der aktuelle Stand der Technik, also menschliche Lokführer, nachgewiesen werden. Dazu müssen belastbare Metriken gefunden werden, mit deren Hilfe sich die Fähigkeit KI-basierter Systeme im Vergleich zu für die Zugsteuerung autorisierten menschlichen Lokführern messen und beurteilen lassen. Fraunhofer FOKUS hat untersucht, welche Metriken nach dem aktuellen Forschungsstand bekannt sind, um Menschen und Maschinen miteinander zu vergleichen. Es wurde analysiert, inwiefern diese Metriken die tatsächliche Kompetenz erfassen können. Die Unterschiede zwischen menschlicher und künstlicher Intelligenz machen die Erstellung, Durchführung und Auswertung von Experimenten zum Mensch-Maschine-Vergleich zu einer Herausforderung; hier besteht noch grundlegender Forschungsbedarf.

6.3.1 OOD-Pruning

Da die Qualität von KI-basierten Systemen stark von der Qualität der zugrunde liegenden Daten abhängt, hat sich Fraunhofer FOKUS u.a. auch mit dem Thema der Datenqualität befasst. Das von Fraunhofer FOKUS hierfür entwickelte Verfahren OOD-Pruning (Out-of-Distribution-Pruning) erkennt Verteilungsabweichungen in Bilddatensätzen und bereinigt diese basierend darauf. Dieses bewertet die Datenqualität eines Datensatzes in Bezug auf eine gewünschte Verteilung oder einen semantischen Bereich von Daten. Fraunhofer FOKUS hat hiermit einen einfachen und leistungsfähigen Ansatz entwickelt, der die Bereinigung von Daten auf der Grundlage einer deskriptiven Definition von gewünschten und unerwünschten Daten ermöglicht. Die von Fraunhofer entwickelte Methode macht sich die Fähigkeiten des Cross-Language Image Pre-training (CLIP) Modells zunutze, sowohl textuelle als auch visuelle Eingaben in einen gemeinsamen Merkmalsraum einzubetten. In diesem Raum stellt die Kosinus-Ähnlichkeit die semantische Ähnlichkeit zwischen zwei beliebigen Merkmalsvektoren dar und erlaubt es so, Bilddaten und Bilddatenbeschreibungen miteinander ins Verhältnis zu setzen. Ausgehend von einer geeigneten textuellen Beschreibung der gewünschten Merkmale lassen sich über

das CLIP-Verfahren hochwertigen Bilddaten einem Merkmalsvektor zuordnen, der der Codierung des Bildes selbst entspricht. Anstelle von Beispielen genügt daher eine Beschreibung der gewünschten Szene, um geeignete Bilddaten auswählen zu können. OOD-Pruning ist zwar nur anteilig übertragbar in den Bahnbereich. Allerdings kann die Methode als Komponente einer komplexeren Datenaufbereitungspipeline eingesetzt werden, um Datenqualität abzusichern. Die OOD-Pruning Methode wurde in der wissenschaftlichen Arbeit *Image Dataset Quality Assessment Through Descriptive Out-of-Distribution Detection* (KI 2024, Sami Kharma et al.) publiziert.

6.3.2 Randomized Smoothing

Neben den zugrundeliegenden Daten hat Fraunhofer FOKUS auch verschiedene Ansätze zur Verifikation neuronaler Netze im Bereich der Bilderkennung auf ihre Eignung für das KI-LOK-Projekt untersucht, und einen dieser Ansätze zur Anwendung gebracht. Dabei handelt es sich um den Verifikationsansatz durch „randomized smoothing“ – eine Technik, bei der mithilfe von zufälligen kleinen Bildveränderungen eine stochastisch beweisbare Robustheit erzeugt werden kann. Dabei wird eine auf semantische Segmentierung spezialisierte Variante dieser Technik auf KI-LOK übertragen und diese an neuronalen Netzen erprobt, welche auf die Erkennung für das Zugsystem relevanter Strukturen trainiert sind. Die Arbeit, welche die Methode anhand von Segmentierungsproblemen aus dem Bahnbereich evaluiert, umfasste:

1. eine ausführliche Einführung in Anforderungen aus dem Bahnbereich sowie die theoretische Einführung in die Konzepte Verifikation, zertifizierte Robustheit im Allgemeine und das Verfahren des „Randomized Smoothing“,
2. die Anwendung des „Randomized Smoothing“ für Segmentierungsaufgaben aus dem Bereich auf Basis zweier Datensätze mit unterschiedlichen Eigenschaften,
3. ein Versuchsaufbau zur Prüfung der Skalierungsfähigkeit des „Randomized Smoothing“ im Hinblick auf die Verifikationsanforderungen aus dem Bahnbereich, und
4. eine systematische Evaluation der Versuchsergebnisse sowie der Vergleich mit Ergebnissen aus anderen Domänen (Citiscapes Datensatz)

Die Implementierung des Ansatzes und die Evaluation der Ergebnisse entlang zweier Datensätze aus dem Bahnbereich zeigen die prinzipielle Anwendbarkeit des „Randomized Smoothing“ für Segmentierungsaufgaben im Bahnbereich.

In Abbildung 5 werden die Ergebnisse der Experimente zusammengefasst. Im Vergleich der Ergebnistabellen der beiden Korrekturarten „Bonferroni“ und „Holm“ zeigen sich – insbesondere bei geringen Werten für τ und σ – kaum bis gar keine Unterschiede.

- Zeit: Die Zeiten ändern sich innerhalb eines Szenarios bei unterschiedlicher Wahl von τ und σ nicht wesentlich, korrelieren jedoch stark mit der Bildgröße der verarbeiteten Eingabebilder. Geringfügige Schwankungen sind auf eine ungleichmäßige Auslastung der CPU durch parallele Nutzung des Gerätes zurückzuführen.
- Genauigkeit: Es ist zu beobachten, dass die absolute Genauigkeit erwartungsgemäß bei steigenden τ- und σ-Werten sinkt. Im Comic-Train-Szenario erfolgt bereits bei sehr geringen τ-Werten ein starker Abfall der absoluten Genauigkeit auf unter 20%. Im Railsem19-Szenario zeigt sich ein deutlich geringerer Abfall der absoluten Genauigkeit. Im Vergleich zur allgemeinen Genauigkeit liegt die Genauigkeit der Schienenerkennung signifikant höher als die allgemeine, sowohl absolut als auch relativ gesehen. Die relative Genauigkeit sinkt bei steigenden σ-Werten nur geringfügig und steigt teilweise sogar ein wenig.

- Enthaltungen: Im Vergleich zum Cityscapes-Referenzmodell kommt es in den untersuchten Szenarien zu deutlich mehr Enthaltungen. Bereits für geringe sigma-Werte ist ein höherer Anteil von Enthaltungen zu beobachten. Während dieser Anstieg im RailSem19-Szenario noch moderat verläuft, steigt im Comic-Train-Szenario der Anteil von Enthaltungen sprunghaft auf nahezu 100%.

Modell	τ	σ	Genauigk. gesamt	Genauigk. Schienen	Enthaltung gesamt	Enthaltung Schienen	Zeit
Cityscapes	0.5	0.25	0.91	NaN	0.05	NaN	261.04
Cityscapes	0.75	0.25	0.87	NaN	0.12	NaN	263.05
Comic Train	0.5	0.05	0.94	0.84	0.04	0.11	36.86
Comic Train	0.5	0.1	0.74	0.52	0.24	0.45	36.55
Comic Train	0.5	0.15	0.29	0.17	0.66	0.81	38.42
Comic Train	0.5	0.2	0.15	0.05	0.53	0.94	42.33
Comic Train	0.5	0.25	0.12	0.03	0.51	0.96	45.91
Comic Train	0.5	0.33	0.1	0.03	0.51	0.97	47.78
Comic Train	0.5	0.5	0.07	0.02	0.55	0.98	46.3
Comic Train	0.75	0.05	0.9	0.69	0.09	0.3	42.52
Comic Train	0.75	0.1	0.31	0.13	0.69	0.86	38.52
Comic Train	0.75	0.15	0.1	0.03	0.9	0.97	38.04
Comic Train	0.75	0.2	0.06	0.03	0.93	0.97	37.33
Comic Train	0.75	0.25	0.04	0.02	0.94	0.98	44.62
Comic Train	0.75	0.33	0.02	0.02	0.96	0.98	53.06
Comic Train	0.75	0.5	0.0	0.0	1.0	1.0	48.34
RailSem19	0.5	0.05	0.78	0.79	0.06	0.04	19.48
RailSem19	0.5	0.1	0.75	0.77	0.09	0.08	16.94
RailSem19	0.5	0.15	0.71	0.75	0.12	0.11	16.36
RailSem19	0.5	0.2	0.65	0.72	0.16	0.14	16.1
RailSem19	0.5	0.25	0.57	0.69	0.21	0.17	15.97
RailSem19	0.5	0.33	0.45	0.61	0.26	0.23	16.06
RailSem19	0.5	0.5	0.3	0.45	0.29	0.31	16.24
RailSem19	0.75	0.05	0.75	0.76	0.12	0.1	16.49
RailSem19	0.75	0.1	0.71	0.73	0.18	0.16	16.41
RailSem19	0.75	0.15	0.65	0.69	0.25	0.22	16.65
RailSem19	0.75	0.2	0.55	0.64	0.34	0.28	26.18
RailSem19	0.75	0.25	0.46	0.6	0.41	0.33	21.32
RailSem19	0.75	0.33	0.35	0.51	0.49	0.4	21.25
RailSem19	0.75	0.5	0.22	0.35	0.54	0.52	19.47

(a) Bonferroni-Korrektur

Modell	τ	σ	Genauigk. gesamt	Genauigk. Schienen	Enthaltung gesamt	Enthaltung Schienen	Zeit
Cityscapes	0.5	0.25	0.92	NaN	0.05	NaN	261.85
Cityscapes	0.75	0.25	0.88	NaN	0.11	NaN	264.23
Comic Train	0.5	0.05	0.94	0.85	0.03	0.1	36.97
Comic Train	0.5	0.1	0.76	0.54	0.22	0.44	36.76
Comic Train	0.5	0.15	0.29	0.18	0.66	0.81	38.55
Comic Train	0.5	0.2	0.15	0.05	0.52	0.93	42.42
Comic Train	0.5	0.25	0.12	0.03	0.5	0.96	45.9
Comic Train	0.5	0.33	0.11	0.03	0.5	0.97	48.01
Comic Train	0.5	0.5	0.08	0.02	0.54	0.98	46.29
Comic Train	0.75	0.05	0.9	0.71	0.08	0.27	42.85
Comic Train	0.75	0.1	0.31	0.13	0.69	0.86	38.85
Comic Train	0.75	0.15	0.1	0.03	0.9	0.97	38.2
Comic Train	0.75	0.2	0.06	0.03	0.93	0.97	37.65
Comic Train	0.75	0.25	0.04	0.02	0.94	0.98	44.96
Comic Train	0.75	0.33	0.02	0.02	0.96	0.98	52.56
Comic Train	0.75	0.5	0.0	0.0	1.0	1.0	48.56
RailSem19	0.5	0.05	0.78	0.79	0.05	0.04	19.52
RailSem19	0.5	0.1	0.76	0.78	0.08	0.07	16.95
RailSem19	0.5	0.15	0.72	0.75	0.11	0.1	16.53
RailSem19	0.5	0.2	0.65	0.73	0.15	0.13	16.22
RailSem19	0.5	0.25	0.57	0.69	0.2	0.16	16.15
RailSem19	0.5	0.33	0.46	0.61	0.25	0.22	16.11
RailSem19	0.5	0.5	0.3	0.45	0.28	0.3	16.35
RailSem19	0.75	0.05	0.76	0.77	0.11	0.09	16.6
RailSem19	0.75	0.1	0.72	0.74	0.17	0.15	16.59
RailSem19	0.75	0.15	0.66	0.7	0.23	0.21	16.82
RailSem19	0.75	0.2	0.57	0.65	0.32	0.26	26.22
RailSem19	0.75	0.25	0.48	0.61	0.39	0.31	21.81
RailSem19	0.75	0.33	0.36	0.53	0.47	0.39	21.61
RailSem19	0.75	0.5	0.22	0.36	0.52	0.5	19.72

(b) Holm-Korrektur

Abbildung 5 - Zusammenfassung der Versuchsergebnisse.

Das Railsem19-Modell zeigt insgesamt ein stabileres Verhalten als das Comic-Train-Modell und erst bei höheren σ-Werten treten gehäuft Enthaltungen und Fehlklassifizierungen auf. Dabei ist anzumerken, dass die Schienenumgebung – insbesondere die eigene Fahrstrecke – weiterhin überdurchschnittlich zuverlässig erkannt wird, auch wenn es noch weiteren Forschungsbedarf gibt im Hinblick auf die Performanz des Ansatzes sowie die Befürchtung, dass die Zertifizierungsradien zu klein sind bzw. zu viele Enthaltungen vorkommen (d.h. die Fälle, für die keine Zusicherung gemacht werden kann). Grundsätzlich sollte nicht vergessen werden, dass es sich bei der Verifikation um eine punktweise Robustheitsprüfung handelt, also dass die Robustheit des Modells nur auf bestimmten Bildern getestet wird. Die tatsächliche Robustheit des Modells auf der Gesamtmenge aller vorkommenden Bilder kann daher nur angenähert werden. Sie basiert auf der Annahme, dass die im Testdatensatz vorliegenden

Bilder eine ausreichend breite Abdeckung der natürlich vorkommenden Bilder bietet. Somit hängt die Aussagekraft über die Robustheit des Modells maßgeblich von der Güte des ausgewählten Testdatensatzes ab. Details zu dem Ansatz finden sich in (Gerlach 2024)

6.3.3 Adversarielle Risiken

Neben der Leistung von Deep Neural Networks (DNNs) ist es wichtig, auch potenzielle Sicherheitsrisiken zu berücksichtigen. DNNs sind dafür bekannt, anfällig für adversariale Angriffe zu sein. Darüber hinaus hat sich gezeigt, dass diese adversarialen Beispiele von dem Ursprungsnetzwerk, in dem sie erstellt wurden, auf ein Black-Box-Zielnetzwerk übertragbar sind. Da der Einsatz von Deep Learning auf eingebetteten Geräten zunehmend an Bedeutung gewinnt, wird es relevant, die Übertragungseigenschaften adversarieller Beispiele zwischen komprimierten Netzwerken zu untersuchen. Fraunhofer FOKUS hat, unter Betrachtung von Quantisierung als eine Technik zur Netzwerkkompression, die Leistung von übertragungsbasierten Angriffen bewertet. Insbesondere untersucht wurde der Fall, in dem das Ursprungs- und das Zielnetzwerk mit unterschiedlichen Bitbreiten quantisiert sind. Wir haben untersucht, wie algorithmusspezifische Eigenschaften die Übertragbarkeit beeinflussen, indem wir verschiedene Algorithmen zur Generierung adversarieller Beispiele berücksichtigen. Darüber hinaus wurde die Übertragbarkeit in einem realistischeren Szenario, in dem sich das Ursprungs- und das Zielnetzwerk nicht nur in der Bitbreite, sondern auch in anderen modellspezifischen Eigenschaften wie Kapazität und Architektur unterscheiden, analysiert. Unsere Ergebnisse zeigen, dass die Quantisierung zwar die Übertragbarkeit verringert, jedoch bestimmte Angriffstypen diese verbessern können. Zudem lässt sich die durchschnittliche Übertragbarkeit adversarieller Beispiele zwischen quantisierten Versionen eines Netzwerks nutzen, um die Übertragbarkeit auf quantisierte Zielnetzwerke mit unterschiedlicher Kapazität und Architektur zu schätzen. Details zu den Ergebnissen finden sich in *(Shresta 22)* und (Shrestha et al. 24).

6.3.4 Vertrauensschaffung und Erklärbarkeit

Um Vertrauen in KI-Komponenten zu schaffen, ist auch die Erklärbarkeit der Komponente essenziell. Bei Fraunhofer FOKUS wurde hierzu im Rahmen einer Masterarbeit untersucht, wie das Testen mit sogenannten “concept activation vectors” angewendet werden kann. Testing mit Concept Activation Vectors (TCAV) ist eine leicht verständliche und einfach anzuwendende Methode zur Interpretierbarkeit von ML-Modellen. TCAV erweist sich als nützliches Werkzeug zur nachträglichen Analyse und kann die Modellergebnisse bereits in der Entwicklungsphase optimieren. In dieser Arbeit wurden Sicherheitsbedenken im Zusammenhang mit neuronalen Netzen (DNNs) vorgestellt und TCAV als ein Ansatz zur Lösung dieser Probleme vorgeschlagen. Mithilfe von Testfällen wurde versucht, verschiedene Sicherheitsbedenken zu adressieren und zu klären, ob die Methode als Sicherheitsargument zur Erkennung von zugrunde liegenden Problemen in DNNs eingesetzt werden kann. Details zu dem Ergebnis finden sich in (Kulkarni 22).

6.3.5 Pixel-Wise-Testing

Das Pixel-Wise-Testing stellt einen neuartigen Ansatz zur Bewertung KI-basierter Perzeptionssysteme dar, bei dem die Auswirkungen von Manipulationen auf einzelnen Pixeln bzw. Pixelgruppen und nicht die aggregierten Änderungen im Gesamtbild im Mittelpunkt stehen, wodurch eine detailliertere uns systeamtische Analyse der Modellrobustheit ermöglicht wird. Diese Methode ist besonders im Eisenbahnbereich von Bedeutung, wo semantische Segmentierungsmodelle komplexe und sicherheitskritische Szenarien wie unterschiedliche Wetterbedingungen, Lichtverhältnisse und ungewöhnliche Gleiskonfigurationen zuverlässig verarbeiten müssen. Durch das Aufdecken von Schwachstellen auf granularer Ebene erhöht das Pixel-Wise-Testing die Robustheit und Vertrauenswürdigkeit von KI-Systemen und gewährleistet so einen sichereren Einsatz in

Eisenbahnanwendungen. Pixel-Wise-Testing erlaubt es, die relevanten Merkmale, die die Vorhersage für ein einzelnes Pixel ausmachen, über die Merkmalshierarchie in semantischen Segmentierungsmodellen zu lokalisieren. Hierdurch können Testeingaben neu definiert werden und es kann identifiziert werden, was ein geeignetes Adäquatheitskriterium für diese Testeingaben ist. Das Pixel-Wise-Testing wurde in Form eines Frameworks implementiert, um einen Konzeptnachweis für die Durchführung des Testens von DNNs für semantische Segmentierung zu schaffen. Ein zentrales Problem ist, dass die korrekte Funktionsweise dieser Systeme eine umfassende Test- und Validierungsstrategie erfordert. Traditionelle Testmethoden, die sich lediglich auf ungesehene Eingaben stützen, haben sich als ineffektiv erwiesen, insbesondere wenn es um ungewöhnliche oder seltene Eingaben geht, die potenziell zu fehlerhaften Ausgaben führen können. Diese sogenannten "corner cases" sind besonders kritisch, da sie in der realen Welt oft nicht ausreichend berücksichtigt werden. Zudem erschwert die Black-Box-Natur von DNNs das Auffinden solcher Eingaben, was die Testbarkeit weiter einschränkt. Vorhandene Testmethoden konzentrieren sich häufig auf Modelle zur Bildklassifikation und sind daher nicht direkt auf semantische Segmentierungsmodelle anwendbar. Es zeigt sich bei der Analyse von existierenden Methoden schnell, dass bestehende Metriken nicht als geeignete Testkriterien für semantische Segmentierungsmodelle fungieren können und dass neue Metriken entwickelt, werden müssen, die die spezifische Entscheidungslogik dieser Modelle berücksichtigen. Ein wesentlicher Bestandteil der "Pixel-Wise-Testing"-Methode ist die Definition von "bestimmenden Pixelregionen", welche den Einfluss auf die Vorhersage eines Pixels darstellen. Diese Regionen werden durch das Rückverfolgen der Merkmale in den verschiedenen Schichten des DNNs bis hin zum Eingabebild identifiziert. Hierzu werden sogenannte Pixel-Bildausschnitte erstellt. Diese Ausschnitte umfassen einen definierten Bereich um einen bestimmten Pixel und dienen als Testfälle für individuelle Pixel-Vorhersagen. Die Methodik umfasst auch die Entwicklung einer neuen Testangemessenheitsmetrik, die quantifiziert, wie wahrscheinlich es ist, dass ein bestimmter Pixel-Bildausschnitt fehlerhafte Vorhersagen auslöst. Hierbei wird ein Ansatz verfolgt, der sich an den internen Aktivierungswerten des DNN orientiert. Um diese Werte zu approximieren, werden sogenannte Pixel-Tiefenvektoren erstellt. Diese Vektoren enthalten Aktivierungswerte aus den verschiedenen Aktivierungskarten des Modells und ermöglichen eine differenzierte Analyse zwischen korrekt und falsch klassifizierten Pixeln. Die gesamte Testing-Prozedur wird durch einen strukturierten Prozess ergänzt, der eine systematische Generierung von Testfällen für ungewöhnliche natürliche Eingaben umfasst. Zunächst wird die Verteilung falsch klassifizierter Pixel analysiert, um festzustellen, welche Klassen am häufigsten betroffen sind. Anschließend werden ähnliche Cluster falsch klassifizierter Pixel ausgewählt und analysiert. Details zu dem Ansatz finden sich in (Petersen 22)

7 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Das Projekt KI-LOK hat sowohl auf europäischer als auch nationaler Ebene erfolgreich als Wegbereiter für nachhaltige Partnerschaften zwischen Unternehmen und Organisationen zur Erforschung und Entwicklung industriegerechter Lösungen für den Test von sicherheitskritischen Systemen mit KI-Komponenten fungiert. Die innovativen Forschungs- und Entwicklungsergebnisse des Projekts stärken sowohl Europa als auch den Standort Deutschland, indem sie Test- und Prüflösungen bereitstellen, die effektive Maßnahmen zur Qualitätssteigerung von KI-Komponenten ermöglichen. Darüber hinaus befähigen die im Rahmen von KI-LOK entwickelten Verfahren die beteiligten Unternehmen, eine führende Rolle bei der Bereitstellung innovativer Testlösungen für sicherheitskritische Systeme zu übernehmen.

Die durchgeführten Arbeiten sowie die eingesetzten Ressourcen waren erforderlich und angemessen, da sie der im Projektantrag detaillierten Planung entsprachen und alle im Arbeitsplan festgelegten Aufgaben erfolgreich bearbeitet und mit innovativen Ergebnissen abgeschlossen wurden.

8 Fortschritte bei anderen Stellen während des Vorhabens

Es sind Fraunhofer FOKUS keine relevanten neuen Forschungsergebnisse im Bereich KI-Systeme für die Bahn von dritter Seite bekannt, welche die durchgeführte Forschungsarbeit im KI-LOK Projekt vorweggenommen oder überflüssig gemacht hätten.

Es wurde Kontakt aufgenommen zum ITEA3 Project 17032 CyberFactory#1, BMBF FKZ 01IS18061D, HTW Berlin, Prof. Dr. Carsten Thomas, der sich in diesem Projekt mit dem Thema Prozess-FMEA (Failure Mode and Effects Analysis) für ML-basierte Bahnsysteme beschäftigt.

9 Veröffentlichungen

Im Rahmen des KI-LOK Projekts wurde von Fraunhofer FOKUS eine Vielzahl von Veröffentlichungen publiziert. Das sind sowohl Arbeiten, welche auf wissenschaftlichen Konferenzen veröffentlicht wurden, als auch Abschlussarbeiten. Diese werden im Folgenden genauer aufgeführt.

9.1 Wissenschaftliche Veröffentlichungen

9.1.1 Veröffentlichungen auf Konferenzen und in Zeitschriften

- Wiesbrock, H.-W., & Großmann, J. (2022). Test- und Trainingsdatengenerierung für die Objekterkennung im Bahnbereich. In *Berlin Workshop on Artificial Intelligence for Engineering Applications (AI4EA 2022)*.
- Großmann, J., Grube, N., Kharma, S., Knoblauch, D., Krajewski, R., Kucheiko, M., Wiesbrock, H.-W. (2022). Test and Training Data Generation for Object Recognition in the Railway Domain. In *Lecture Notes in Computer Science: Software Engineering and Formal Methods, SEFM 2022* (Vol. 13765). Springer.
- Shrestha, A., Großmann, J. (2022). Properties that Allow or Prohibit Transferability of Adversarial Attacks among Quantized Networks. In *Proceedings of the 5th International Conference on Automation and Software Testing*.
- Wiesbrock, H.-W., Großmann, J. (2024). Outline of an Independent Systematic Blackbox Test for ML-based Systems. In *Proceedings of IEEE AITest 2024*. IEEE Xplore. Zusammenarbeit mit IT-Power Solutions.
- Grube, N., Massah, M., Tebbe, M., Wancura, P., Wiesbrock, H.-W., Großmann, J., Kharma, S. (2024). On a Systematic Test of ML-based Systems: Experiments on Test Statistics. In *Proceedings of IEEE AITest 2024*. IEEE Xplore.
- Kharma, S., Großmann, J. (2024). Image Dataset Quality Assessment Through Descriptive Out-of-Distribution Detection. In *Proceedings of the 47th German Conference on Artificial Intelligence, KI 2024*. Springer.

9.1.2 Abschlussarbeiten

- Shrestha, A. (2022) *Properties that Allow or Prohibit Transferability of Adversarial Attacks among Quantized Networks*. Masterarbeit, Monomaster Informatik, Technische Universität Berlin. Vertiefung der Ergebnisse der gleichnamigen veröffentlichten Arbeit.

- Kulkarni, O. P. (2022) *Interpreting Safety Relevant Concepts in Deep Neural Networks*. Masterarbeit, Monomaster Informatik, Berliner Hochschule für Technik (BHT).
- Kucheiko, M. (2022) *Generation of Abstract Scene Descriptions for Testing of Visual Perception of ML-based Autonomous Locomotives*. Bachelorarbeit, Monobachelor Informatik, Technische Universität Berlin.
- Petersen, L. (2022) *Methods and Metrics for Testing DNN Models in Test Cases with Unusual Natural Inputs in the Application of Semantic Segmentation*. Masterarbeit, Monomaster Informatik, Technische Universität Berlin.
- Gerlach, L. (2024) *Zertifizierung von Bilderkennungssystemen im Bahnbereich unter Verwendung von „Randomized Smoothing“*. Bachelorarbeit, Monobachelor Informatik, Humboldt-Universität zu Berlin.
- Kharma, S. (2024) *Dataset Quality Assessment Through Descriptive Out-of-Distribution Detection*. Masterarbeit, Monomaster Informatik, Humboldt-Universität zu Berlin. Vertiefung der Ergebnisse der gleichnamigen veröffentlichten Arbeit.
- Beiker, H. *Improving Object Detection Models by Simulating False Negative Prone Scenarios*. Masterarbeit in Arbeit, Monomaster Informatik, Technische Universität Berlin. Voraussichtlicher Abschluss Ende 2024.

9.2 Sonstige nennenswerte Veröffentlichungen und Präsentationen

- Hemzal, G., Strobel, T., Großmann, J., Schlingloff, B.-H., Leuschel, M., Sadeghipour, S., Firnkorn, J. (2021) *KI-LOK – Ein Verbundprojekt über Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb*. Signal+Draht, Oktober 2021. In Kooperation mit Projektpartnern, Darstellung bisheriger Projektergebnisse.
- Großmann, J., Grube, N., Kharma, S., Knoblauch, D., Krajewski, R., Kucheiko, M., Wiesbrock, H.-W. (2023) *KI-LOK – Ein Verbundprojekt über Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb*. Signal+Draht, April 2023. In Kooperation mit Projektpartnern, Darstellung bisheriger Projektergebnisse
- Knoblauch, D. (2023) Virtual Testing of AI Object Detection Systems in UAVs and Trains. Vortrag auf dem 32. Safe TRANS Industrial Day. Präsentation zur Absicherung von KI-Objekterkennungssystemen in Drohnen und Zügen unter Nutzung von virtuellen Testumgebungen durch Paratrust.ai zur Simulation vielfältiger Szenarien und Optimierung des Testprozesses.

Literaturverzeichnis

Katz, G. et al. (2017), *Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks*, CAV 2017

Pei, K. et al. (2019), DeepXplore, Commun. ACM 2019

Guo, J. et al. (2018), *DLFuzz*, ESEC/FSE 2018

Odena, A. et al., *TensorFuzz (2019)*, ICML 2019

Liu, Y. et al. (2017), *Delving into Transferable Adversarial Examples and Black-Box Attacks*, 2017

Goodfellow, I.-J. et al. (2015), *Explaining and Harnessing Adversarial Examples*, ICLR (Poster) 2015

Wu, L. et al. (2018), *Understanding and Enhancing the Transferability of Adversarial Examples*, 2018

Bernhard, R. et al. (2019), *Impact of Low-Bitwidth Quantization on the Adversarial Robustness for Embedded Neural Networks*, CW 2019

Matachana, A.-G. et al. (2020), *Robustness and Transferability of Universal Attacks on Compressed Models*, 2020

Zhao, Z. et al. (2019), *To compress or not to compress: Understanding the Interactions between Adversarial Attacks and Neural Network Compression*, MLSys 2019

Willers, O. et al. (2020), *Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks*, SAFECOMP 2020

Marco Tulio, R. et al. (2016), *Explaining the Predictions of Any Classifier*, ACM SIGKDD 2016

Christoph, M. et al. (2020), *Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges*, ECML PKKD 2020

Karen, S. et al. (2013), *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*, ICLR 2013

Poursabzi-Sangdeh (2021), F. et al., *Manipulating and Measuring Model Interpretability*, CHI 2021

Kim, B. et al., *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors*, ICML 2017

Simard, P.-Y. et al. (2003), *Best practices for convolutional neural networks applied to visual document analysis*, ICDAR 2003

Yang, J. et al. (2021), *Generalized Out-of-Distribution Detection: A Survey*, CoRR 2021

Sallab, A.-E. et al. (2016), *End-to-End Deep Reinforcement Learning for Lane Keeping Assist*, CoRR 2016

Wang, D. et al. (2019), *Deep object-centric policies for autonomous driving*, ICRA 2019

D'Amico, G. et al. (2023), *Trainsim: A railway simulation framework for lidar and camera dataset generation*, IEEE TITS 2023

Wang R. et al. , *Improving the Effectiveness of Deep Generative Data*, CVPR

Zhu J.-Y. et al., *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks* (*CycleGAN*), ICCV

Guo X. et al., *Generative AI for Synthetic Data Generation: Methods, Challenges and the Future*, arXiv

Richter S.-R. et al., *Enhancing Photorealism Enhancement*, IEEE Transactions on Pattern Analysis and Machine Intelligence

Wiesbrock, H.-W., & Großmann, J. (2022). Test- und Trainingsdatengenerierung für die Objekterkennung im Bahnbereich. In Berlin Workshop on Artificial Intelligence for Engineering Applications (AI4EA 2022).

Großmann, J., Grube, N., Kharma, S., Knoblauch, D., Krajewski, R., Kucheiko, M., Wiesbrock, H.-W. (2022). Test and Training Data Generation for Object Recognition in the Railway Domain. In Lecture Notes in Computer Science: Software Engineering and Formal Methods, SEFM 2022 (Vol. 13765). Springer.

Shrestha, A., Großmann, J. (2022). Properties that Allow or Prohibit Transferability of Adversarial Attacks among Quantized Networks. In Proceedings of the 5th International Conference on Automation and Software Testing.

Wiesbrock, H.-W., Großmann, J. (2024). Outline of an Independent Systematic Blackbox Test for ML-based Systems. In Proceedings of IEEE AITest 2024. IEEE Xplore. Zusammenarbeit mit IT-Power Solutions.

Grube, N., Massah, M., Tebbe, M., Wancura, P., Wiesbrock, H.-W., Großmann, J., Kharma, S. (2024). On a Systematic Test of ML-based Systems: Experiments on Test Statistics. In Proceedings of IEEE AITest 2024. IEEE Xplore.

Kharma, S., Großmann, J. (2024). Image Dataset Quality Assessment Through Descriptive Out-of-Distribution Detection. In Proceedings of the 47th German Conference on Artificial Intelligence, KI 2024. Springer.

Shrestha, A. (2022) *Properties that Allow or Prohibit Transferability of Adversarial Attacks among Quantized Networks*. Masterarbeit, Monomaster Informatik, Technische Universität Berlin. Vertiefung der Ergebnisse der gleichnamigen veröffentlichten Arbeit.

Kulkarni, O. P. (2022) *Interpreting Safety Relevant Concepts in Deep Neural Networks*. Masterarbeit, Monomaster Informatik, Berliner Hochschule für Technik (BHT).

Kucheiko, M. (2022) *Generation of Abstract Scene Descriptions for Testing of Visual Perception of ML-based Autonomous Locomotives*. Bachelorarbeit, Monobachelor Informatik, Technische Universität Berlin.

Petersen, L. (2022) *Methods and Metrics for Testing DNN Models in Test Cases with Unusual Natural Inputs in the Application of Semantic Segmentation*. Masterarbeit, Monomaster Informatik, Technische Universität Berlin.

Gerlach, L. (2024) *Zertifizierung von Bilderkennungssystemen im Bahnbereich unter Verwendung von „Randomized Smoothing“*. Bachelorarbeit, Monobachelor Informatik, Humboldt-Universität zu Berlin.

Kharma, S. (2024) *Dataset Quality Assessment Through Descriptive Out-of-Distribution Detection*. Masterarbeit, Monomaster Informatik, Humboldt-Universität zu Berlin. Vertiefung der Ergebnisse der gleichnamigen veröffentlichten Arbeit.

Beiker, H. *Improving Object Detection Models by Simulating False Negative Prone Scenarios*. Masterarbeit in Arbeit, Monomaster Informatik, Technische Universität Berlin. Voraussichtlicher Abschluss Ende 2024.

ISBN oder ISSN	Berichtsart (Schlussbericht oder Veröffentlichung)
DOI 10.24406/publica-3918	Schlussbericht

Titel
KI-LOK – Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb, Abschlussbericht des Fraunhofer Instituts für Offene Kommunikationssysteme

Autor(en) [Name(n), Vorname(n)]
Großmann, Jürgen
Kharma, Sami
Knoblauch, Dorian
Henry, Beiker,
Johannes Viehmann

Abschlussdatum des Vorhabens
September, 2024

Veröffentlichungsdatum
09.12.2024

Form der Publikation
Elektronische Veröffentl.

Durchführende Institution(en) (Name, Adresse)
Fraunhofer FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin

Ber. Nr. Durchführende Institution

Förderkennzeichen
19I21007D

Seitenzahl
30

Fördernde Institution (Name, Adresse)
Bundesministerium für Wirtschaft und Klimaschutz (BMWK)
10115 Berlin

Literaturangaben
38

Tabellen
0

Abbildungen
8

Zusätzliche Angaben

Vorgelegt bei (Titel, Ort, Datum)

Kurzfassung
Bestehende Verifikations- und Validierungsmethoden für KI-Systeme stoßen insbesondere bei sicherheitskritischen Anwendungen an ihre Grenzen, da sie die spezifischen Herausforderungen von Black-Box-Modellen und kontinuierlich lernenden Systemen unzureichend adressieren. Ziel des Projekts KI-LOK war die Entwicklung innovativer Testmethoden und Werkzeuge, um die Sicherheit und Zuverlässigkeit KI-basierter Bahntechniksysteme zu gewährleisten und den Zulassungsprozess durch robuste Prüfverfahren zu unterstützen. Das Projekt entwickelte mit der Perception-Lab Plattform ein System zur automatisierten Testdatengenerierung, das risikoorientierte und probabilistische Testansätze mit generativer KI und Ontologien kombiniert, um realitätsnahe Szenarien zu simulieren und Randfälle zu testen. Perception-Lab ermöglicht den automatisierten Test von kamerabasierten Perzeptionsdaten durch die simulative Erzeugung fotorealistischer, synthetischer Daten. Methoden wie das Pixel-wise Testing, der Einsatz probabilistischer Ontologien und die Nutzung generative KI erlauben eine systematische Testauswahl und zeigen Wege auf, den bestehenden Sim2Real-Gap, d.h. die systematischen Unterschiede zwischen Simulationsdaten und realen Daten, zu überwinden. Die Ergebnisse zeigen eine hohe Relevanz für die Praxis, einschließlich der Anwendbarkeit in sicherheitskritischen Industrien und der Unterstützung bei der Standardisierung und Zulassung KI-basierter Systeme.
Die entwickelten Methoden und Werkzeuge zeigen einen vielversprechenden Ansatz für die Absicherung sicherheitskritischer KI-Systeme und sind auf andere Industrien wie Automobil oder Luftfahrt übertragbar.

Schlagwörter
KI-LOK, KI, Künstliche Intelligenz, Bahntechnik, Prüfverfahren, Sicherheitskritische Systeme

Verlag	Preis
Fraunhofer PUBLICA	

1. ISBN or ISSN

DOI 10.24406/publica-3918

2. type of document (e.g. report, publication)

Final report

3. title

KI-LOK - Test procedure for AI-based components in railway operations, Final Report of the Fraunhofer Institute for Open Communication Systems

4. author(s) (family name, first name(s))

Großmann, Jürgen
Kharma, Sami
Knoblauch, Dorian
Henry, Beiker,
Johannes Viehmann

5. end of project

September 2024

6. publication date

December 9th, 2024

7. form of publication

electronic

8. performing organization(s) (name, address)

Fraunhofer FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin

9. originator's report no.

10. reference no.

19I21007D

11. no. of pages

30

12. sponsoring agency (name, address)

Bundesministerium für Wirtschaft und Klimaschutz (BMWK)
10115 Berlin

13. no. of references

38

14. no. of tables

0

15. no. of figures

8

16. supplementary notes

17. presented at (title, place, date)

18. abstract

Existing verification and validation methods for AI systems reach their limits particularly in safety-critical applications, as they inadequately address the specific challenges of black-box models and continuously learning systems. The aim of the KI-LOK project was to develop innovative test methods and tools to ensure the safety and reliability of AI-based railway technology systems and to support the approval process with robust test procedures. The project developed the Perception-Lab platform, a system for automated test data generation that combines risk-oriented and probabilistic test approaches with generative AI and ontologies to simulate realistic scenarios and test edge cases. Perception-Lab enables the automated testing of camera-based perception data by simulating the generation of photorealistic, synthetic data. Methods such as pixel-wise testing, the use of probabilistic ontologies and generative AI allow for systematic test selection and show ways to overcome the existing sim2real gap, i.e. the systematic differences between simulation data and real data. The results show a high practical relevance, including applicability in safety-critical industries and support for standardisation and approval of AI-based systems. The developed methods and tools show a promising approach for the validation of safety-critical AI systems and are transferable to other industries such as automotive or aerospace.

19. keywords

artificial intelligence, railway technology, test procedure, safety-critical systems

20. publisher

Fraunhofer PUBLICA

21. price