Bei diesem Beitrag handelt es sich um einen wissenschaftlich begutachteten und freigegebenen Fachaufsatz („reviewed paper“).

doi.org/10.37544/1436-4980-2024-11-12-43

Möglichkeiten der neuesten KI-Modelle in der Produktion

Multimodale Foundation-Modelle in der Produktion

H. Behnen, J.-H. Woltersmann, D. Wolfschläger, R. H. Schmitt

ZUSAMMENFASSUNG Aktuelle Herausforderungen in der Produktion, etwa der Fachkräftemangel, erhöhen die Notwendigkeit, Prozesse zu automatisieren und die Produktivität zu erhöhen. Multimodale Foundation-Modelle bieten diese Möglichkeit für eine Vielzahl an Anwendungen, indem sie aus heterogenen Informationsquellen Entscheidungen ableiten. Anwendungen um diese Technologie sind zurzeit jedoch rar. Dieser Beitrag gibt daher einen Überblick über die Potenziale und Herausforderungen dieser Modelle in der Produktion.

STICHWÖRTER

Produktionstechnik, Künstliche Intelligenz (KI), Industrie 4.0

Possibilities of the Latest AI Models in Production – Multi-Modal Foundation Models in Production

ABSTRACT Current challenges in production, such as shortage of skilled workers, increase the need to automate processes and increase productivity. Multi-modal foundation models address this automation demand for a variety of applications by deriving decisions based on heterogeneous information sources. However, applications around this technology are currently rare. This article therefore provides an overview of the potential and challenges of these models in production.

1 Einleitung

Unter dem Stichwort Industrie 4.0 wird seit den frühen 2010er Jahren in der Industrie zunehmend untersucht, durch den Einsatz von Internet- und Kommunikationstechnologie in Produktionsanlagen sowie die Integration diverser Sensoren sogenannte cyber-physische Systeme (CPS) zu bilden, um diese für die Optimierung von Produktionssystemen zu nutzen und so letztlich nachhaltiger, resilienter und wettbewerbsfähiger zu produzieren [1]. Dieses Potenzial basiert wesentlich darauf, dass große Mengen an Daten unterschiedlichen Formats und unterschiedlicher Messgrößen erfasst werden, diese im Digitalen Zwilling bereitgestellt werden und dort die Grundlage für intelligente Entscheidungsalgorithmen bilden. Diese intelligenten Verarbeitungssysteme müssen in der Lage sein, die gesammelten Daten zu prozessieren und Wissen aus ihnen abzuleiten. Seit einigen Jahren wird hierzu bereits der Einsatz Künstlicher Intelligenz (KI), meistens durch Modelle des Maschinellen Lernens (ML), untersucht. Typische Anwendungsfälle sind etwa der Einsatz von KI-Modellen zur Vorhersage Qualitäts-relevanter Merkmale (Predictive Quality) auf Basis prozessparallel erfasster Sensordaten, die ansonsten nur durch aufwendige nachgelagerte, zum Teil zerstörende, Prüfverfahren gemessen werden konnten [2]. Ein weiteres Beispiel ist der Einsatz von KI-Modellen zur Optimierung von Bewegungspfaden auf Basis von Reinforcement Learning, was zum Beispiel für Logistikanwendungen eingesetzt werden kann [3]. Obwohl der Nutzung von KI in der Produktion ein großes Potenzial zugesprochen wird, ist ihr Verbreitungsgrad im Produktivbereich, insbesondere bei kleinen und mittleren Unternehmen, noch sehr gering [2, 4]. Das ist aus folgenden Gründen der Fall:

- Geringe Generalisierungsqualität und Leistungsfähigkeit: Auch wenn KI-Modelle während des Trainings oder unter Simulationsbedingungen gute Genauigkeiten erzielen, ist ihre Performanz in realen, dynamischen, unstrukturierten Umgebungen oder bei Prozessänderungen oftmals stark limitiert [5].
- Begrenzte Transferfähigkeit und Aufgabenspezifität: Typischerweise werden KI-Modelle für eine spezifische Aufgabe, etwa die Detektion einer bestimmten Objektklasse, trainiert. Ändern sich die Anforderungen an die Aufgabe, muss ein neues Modell entwickelt werden. Dies ist gerade vor dem Hintergrund variantenreicher oder komplexer Produktportfolios mit erheblichem Aufwand verbunden [6].
- Verfügbarkeit von Datensätzen: Insbesondere im industriellen Kontext sind für die Modellierungsaufgaben passende Datensätze häufig nicht öffentlich verfügbar. Gründe sind die hohe Vielfalt möglicher Produkte und hohe Hürden beim Datenschutz [7]. Häufig müssen daher zu beginn von KI-Entwicklungsprozessen aufwendige Versuchsreihen aufgenommen werden, um eine vielfältige Datengrundlage für die Entwicklung von KI-Ansätzen zu generieren [6]. Dabei reicht es in der Regel nicht aus, lediglich Rohdaten zu erfassen und diese für das Training von KI-Modellen zu nutzen, sondern zusätzlich müssen diese Rohdaten mit Annotationen versehen werden, was meist mit großem manuellen Arbeitsaufwand einhergeht [8].
- Mangelnde Interpretierbarkeit: KI-Modelle sind meist Black-Box-Modelle, die eine Entscheidung treffen, ohne dass die Faktoren, die zu dieser Entscheidung führen, bekannt sind. Die Gründe für eine bestimmte Entscheidung sind damit nicht ohne Weiteres nachvollziehbar und können nicht direkt inter-

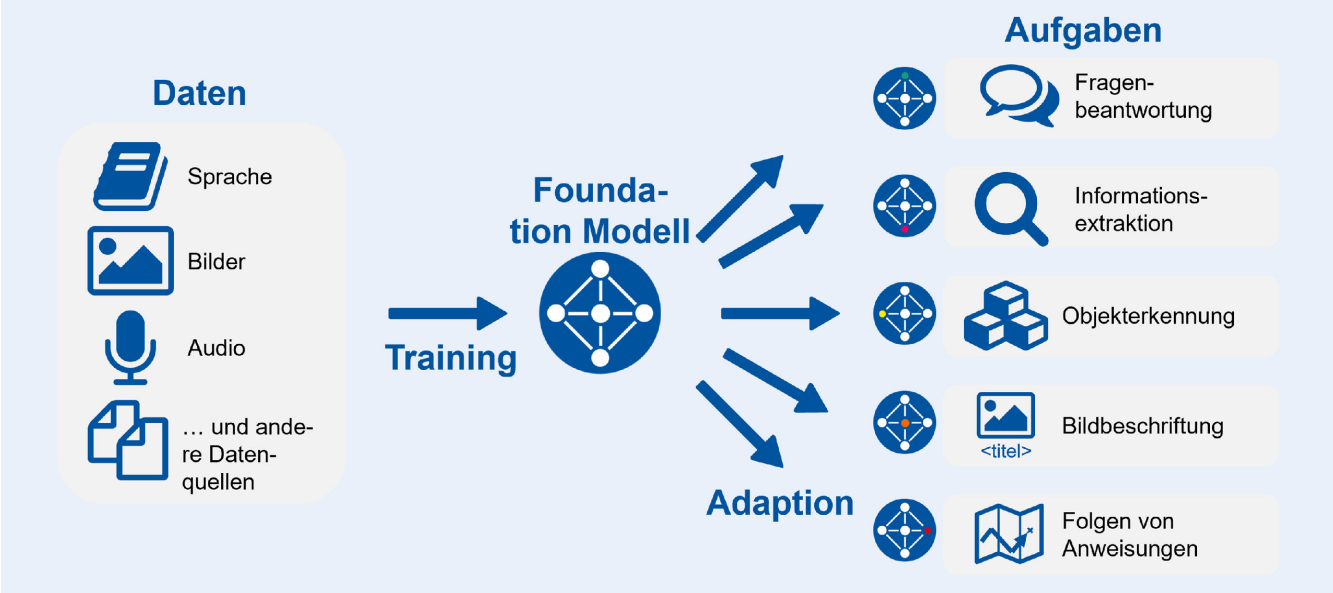

Bild 1. Prinzipbild multimodaler Foundation-Modelle. *Grafik: WZL | RWTH Aachen University in Anlehnung an [12]*

pretiert und damit im Sinne einer Optimierung des Produktionssystems angepasst werden [9].
- Unimodalität: KI-Modelle werden typischerweise lediglich auf einem Typ von Daten, etwa Bilddaten, trainiert [10]. Andere zur Verfügung stehende Datenmodalitäten werden für die Modellierung nicht einbezogen, sodass in vielen Szenarien nur ein Teil der zur Verfügung stehenden Information über das Produktionssystem und seine Umgebung genutzt wird.

Mit dem Auftreten großer Sprachmodelle (LLMs) hat sich spätestens seit der Entwicklung von GPT-3 [11] ein Paradigmenwechsel im Bereich KI ergeben. Auf Basis moderner Modellarchitekturen wurde bei der Entwicklung festgestellt, dass Modelle mit steigender Menge an Trainingsdaten, Rechenzeit und Modellparametern emergente Eigenschaften zeigen, das heißt., Fähigkeiten für die Durchführung von Aufgaben entstanden sind, die außerhalb der Trainingsverteilung liegen [12]. Dies hat zum großen KI-Hype in den letzten Jahren und der Entwicklung immer größerer und leistungsstärkerer Modelle geführt. Diese Modelle, die auf solch großen Datenmengen trainiert wurden und sich durch geringe Anpassungen zur Erfüllung unterschiedlicher Aufgaben nutzen lassen, werden Foundation-Modelle genannt. Sie besitzen das Potenzial die obengenannten Mängel klassischer KI-Verfahren zu adressieren. Als ein aktuell stark wachsendes Forschungsfeld bietet insbesondere der Einsatz multimodaler Foundation-Modelle, d.h., Modelle, die verschiedene Datentypen aus unterschiedlichen Quellen verarbeiten und gemeinsame Repräsentationen dieser Daten erstellen können, die Möglichkeit, die heterogenen Daten, die im Zuge von Industrie 4.0 in Unternehmen erhoben werden, für verschiedene Zwecke zu verarbeiten und neue Erkenntnisse abzuleiten. Diese Vorteile werden auch bereits in Industrieunternehmen erkannt, wie etwa eine vom KI-Bundesverband im Rahmen des LEAM-Projekts durchgeführte Umfrage zeigt, wonach 71,4 % der befragten kleinen und mittleren Unternehmen Sprachmodelle und bereits 37,5 % multimodale Modelle als eine relevante Technologie für ihre Prozesse sehen [13]. Insbesondere in der Produktion gibt es aber aktuell noch wenige Anwendungsfälle für diese Technologie [14].

Ziel dieses Beitrags ist es daher, einen Überblick über die Möglichkeiten multimodaler Foundation-Modelle zu liefern, bestehende Ansätze in der Produktion vorzustellen und zu erläutern, welchen Mehrwert diese Modelle für bestimmte Industriezweige und Anwendungsfälle bieten können. In diesem Kontext wird das vom Bundesministerium für Bildung und Forschung als eines von vier deutschen KI-Servicezentren geförderte Projekt „WestAI“ vorgestellt und ausgearbeitet, welche Rolle dies beim Erschließen dieser Potenziale bieten kann [15]. Abschließend werden neue Herausforderungen dieser Modelle sowie mögliche Lösungsmaßnahmen dargestellt.

2 Multimodale Foundation-Modelle

Foundation-Modelle lassen sich als Untermenge des Deep Learnings interpretieren, also als Untermenge tiefer, künstlicher neuronaler Netze. Diese neuronalen Netze werden seit vielen Jahren für Modellierungsaufgaben untersucht. Verglichen mit traditionellen Deep-Learning-Modellen unterscheiden sich Foundation-Modelle im Wesentlichen durch vier Faktoren: ihre Größe, ihre Fähigkeit des In-Context-Learnings, das bedeutet, sich über Beschreibungen auf neue Aufgaben übertragen zu lassen, ihre Architektur basierend auf Transformer-Modulen und ihre Trainingsmethode des self-supervised Learnings. Diese vier Merkmale kombiniert mit dem Vortraining auf großen, multimodalen Datensätzen führen zu Modellen, die in der Lage sind, verschiedene Datentypen zu verarbeiten und dieses verarbeitete Wissen auf eine Vielzahl an nachgelagerten Aufgaben zu überführen [12], **Bild 1**.

Der wesentliche Befähiger von Foundation-Modellen ist ihre Größe. Die Größe beziehungsweise die Skala bezieht sich dabei auf die drei in der Einleitung beschriebenen Merkmale der Anzahl der Modellparameter, der Größe der Trainingsdatensätze sowie der Menge an Rechenleistung, mit der diese Modelle trainiert werden. Die Skala eines Modells ist zurzeit der wesentliche Treiber für die Weiterentwicklungen großer KI-Modelle. Hintergrund sind Zusammenhänge, die zwischen Modellqualitäten und ihrer Größe identifiziert werden können. Ab gewissen Modell-

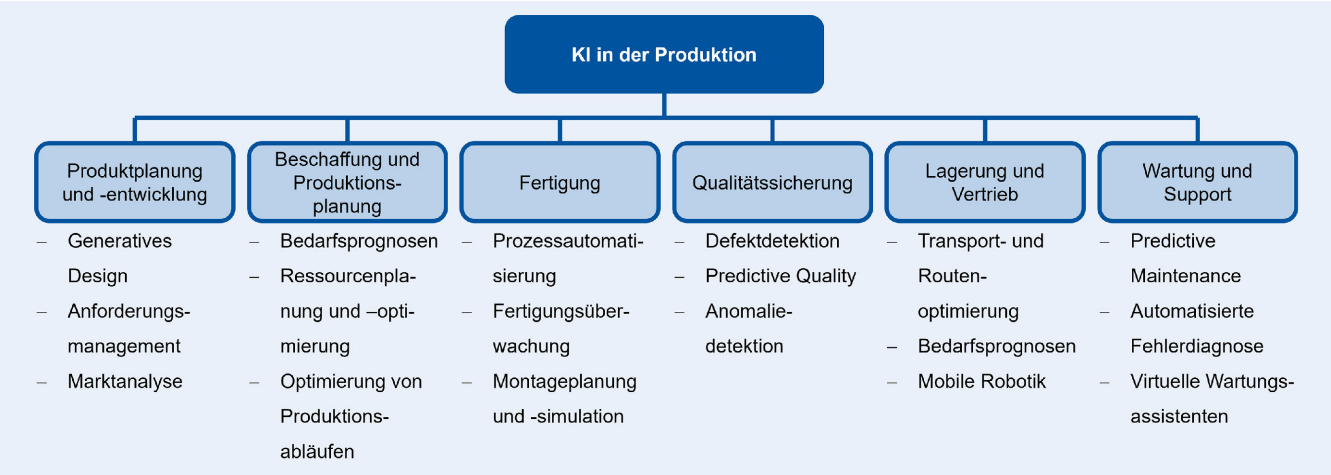

Bild 2. Anwendungsbereiche künstlicher Intelligenz in der Produktion. *Grafik: WZL | RWTH Aachen University*

größen wird über eine Reihe an Modellierungsaufgaben das in der Einleitung beschriebene emergente Verhalten festgestellt, welches sich mit einer Vergrößerung einer der drei genannten Faktoren kontinuierlich verbessert und über sogenannte Skalengesetze vorhergesagt werden kann [16]. Das emergente Verhalten drückt sich dabei in der Regel dadurch aus, dass sich die Modelle ohne weiteres Training (Zero-Shot) oder durch den Einsatz einer geringen Menge an Beispieldaten (Few-Shot) auf neue Aufgaben anwenden lassen. Diese Eigenschaft macht Foundation-Modelle für viele Anwendungen interessant, denn häufig ist das Bereitstellen einer hinreichend großen Datenmenge für das Training eines Modells der limitierende Faktor bei KI-Entwicklungsprojekten. Bei LLMs ist es beispielsweise möglich, das Modell durch eine textuelle Beschreibung der Aufgabe, auch Prompt genannt, auf eine neue Aufgabe zu adaptieren. Diese Eigenschaft beschreibt einen zweiten Befähiger der Foundation-Modelle, das In-Context-Learning. In-Context-Learning beschreibt die Flexibilität eines Foundation-Modells. Im Gegensatz zu traditionellen KI-Modellen lassen sich diese Modelle durch Beschreibungen, für gewöhnlich in textueller Form, auf eine Vielzahl von Aufgaben übertragen. Für Modelle, die keine unmittelbare Anwendbarkeit zeigen, besteht darüber hinaus die Möglichkeit, im sogenannten Fine-Tuning die vortrainierten Foundation-Basismodelle mit eigenen Trainingsdaten für spezifische Aufgaben zu optimieren [13]. Um die Rechenaufwände dabei zu verringern, können Methoden des parametereffizienten Fine-Tuning wie LoRA eingesetzt werden [17].

Dritter Befähiger von Foundation-Modellen ist ihre Modellarchitektur, die heutzutage nahezu ausschließlich auf Transformer-Modulen basiert. Vor diesem Hintergrund wird häufig von der Homogenisierung der Modellarchitekturen gesprochen, denn _waren es vor einigen Jahren noch häufig Änderungen an den Modellarchitekturen, die zu signifikanten Leistungssteigerungen von KI-Modellen geführt haben, wurde die Transformer-Architektur seit ihrer Einführung im Jahr 2017 nicht wesentlich weiterentwickelt [12]. Ein Transformer-Modell überführt die Bedeutung von Informationseinheiten, den „Token“, zum Beispiel Buchstabengruppen in der Sprachverarbeitung, über Embedding-Modelle zu numerischen Repräsentationen, die dann über den sogenannten Aufmerksamkeits- beziehungsweise Attention-Mechanismus im Kontext der gesamten Eingangsgrößen interpretiert werden. Dieser Attention-Mechanismus bildet den wesentlichen Bestandteil eines Transformer-Modells und ermöglicht so, auch komplexe Beziehungen zwischen Eingangsdaten zu modellieren [18]. Obwohl dieses Vorgehen für die Sprachverarbeitung entwickelt wurde, hat sich gezeigt, dass es auch sehr effektiv für Bilder, Audio und weitere Modalitäten adaptiert werden kann.

Trotz aller dieser drei bisher genannten Befähiger wäre das Training vieler heutiger Foundation-Modelle nicht ohne ihre Trainingsmethode, das self-supervised-Learning möglich. Hintergrund ist, dass KI-Modelle typischerweise im Rahmen eines Optimierungsprozesses, dem Training, auf Ein- und Ausgangsdaten optimiert werden und so den funktionalen Zusammenhang über das Bestimmen von Kosten- beziehungsweise Maximierungsfunktionen der Ein- und Ausgangsgrößen approximieren. Das Erzeugen dieser Ausgangsgrößen beziehungsweise Label oder Annotationen ist in der Praxis jedoch häufig mit großen Aufwänden verbunden. An dieser Stelle bietet das self-supervised Learning eine Lösungsansatz, denn es basiert darauf, dass aus ungelabelten Daten über bestimmte Verarbeitungsschritte automatisiert Annotationen erzeugt werden. Dies geschieht im Bereich der Sprachverarbeitung etwa durch das Maskieren bestimmter Satzteile, die das Modell während des Vortrainings prädizieren muss oder die Aufgabe, zu bestimmen, ob ein bestimmter Satz vom Kontext her auf einen vorherigen folgen kann. Self-supervised Learning macht es so möglich, Modelle auf großen, nicht annotierten Datenmengen zu trainieren [12].

Multimodale Foundation-Modelle basieren auf den gleichen Prinzipien, wenden diese jedoch entsprechend auf verschiedene Datentypen an, wie Bild 1 zeigt. Daten verschiedener Modalitäten werden in der Regel über verschiedene Embedding-Modelle in einen gemeinsamen Repräsentationsraum überführt, der es dann ermöglicht, Beziehungen zwischen verschiedenen Datentypen leicht zu berücksichtigen. Auf Basis dieser gemeinsamen Repräsentationen lassen sich so für viele Modellierungsaufgaben höhere Genauigkeiten erzielen [19].

3 Identifikation von Anwendungsfällen und Potenzialen

Grundsätzlich gibt es für den Einsatz von KI in der Produktion typische Anwendungsbereiche, die sich über die gesamte Produktion erstrecken, **Bild 2**. In der Produktplanung und -entwicklung werden KI-Modelle etwa in der Sprachverarbeitung für das

Anforderungsmanagement eingesetzt, um den Kontext von Anforderungsdokumenten auf gewisse Schlüsselbegriffe zu kontrahieren, die dann nachgelagert verarbeitet werden können [20]. Ein weiteres Beispiel ist der Bereich des generativen Designs, bei welchem generative KI-Modelle eingesetzt werden, um zum Beispiel eine Vielzahl innovativer 3D-Konzeptentwürfe in sehr kurzer Zeitspanne zu erzeugen und damit etwa mechanische Eigenschaften optimieren [21]. In der Beschaffung und Produktionsplanung werden KI-Ansätze zum Beispiel für die Bedarfsprognose eingesetzt. [22] analysieren etwa verschiedene KI-Modelle, um den Bedarf nach bestimmten Rohmaterialien auf Basis historischer Daten vorherzusagen, so passgenaue Bestellungen durchzuführen und Kosten einzusparen [22]. Im Bereich Fertigung existiert eine Vielzahl an Anwendungen zur Montageplanung und -simulation. Reinforcement-Learning-basierte KI-Ansätze können hier unter anderem verwendet werden, um das Scheduling von Montagesequenzen hinsichtlich verschiedener Kriterien, zum Beispiel Durchlaufzeiten, zu optimieren [23]. Eines der wohl bekanntesten Anwendungsfelder von künstlicher Intelligenz in der Produktion lässt sich mit der Defektdetektion der Qualitätssicherung zuweisen. Hier ermöglichen etwa bildbasierte Ansätze die prozessparallele Überwachung von Qualitätsmerkmalen und substituieren so teure nachgelagerte, zum Teil zerstörende, Messverfahren. Ein Beispiel hierzu ist das bei der Firma Trumpf untersuchte System zur Schätzung von Qualitätsmerkmalen wie der Oberflächenrauheit beim Laserschneiden durch den Einsatz von Convolutional Neural Networks [24]. Auch in den Fertigung und Qualitätssicherung nachgelagerten Schritten wie Lagerung und Vertrieb oder Wartung und Support sind KI-Ansätze etwa für Logistik oder die vorausschauende Wartung (Predictive Maintenance) bekannt [25, 26].

Auch wenn diese Auflistung zeigt, dass es für KI-Anwendungen bereits eine Vielzahl möglicher Anwendungen im Produktionskontext gibt, ist der Bedarf nach weiterführenden und besseren Ansätzen gegeben. Dieser Bedarf entsteht aus den aktuellen Anforderungen und Herausforderungen in der Produktionstechnik, etwa durch Faktoren wie die zunehmende Individualisierung von Produkten und den damit einhergehenden erhöhten Flexibilitätsanforderungen [27], kürzeren beziehungsweise zirkulären Produktlebenszyklen, die durch das erhöhte Nachhaltigkeitsbewusstsein immer präsenter werden [28], dem sich anbahnenden Fachkräftemangel in Industrienationen wie Deutschland [28] oder den geopolitischen Krisen der letzten Jahre und den damit verbundenen Schwierigkeiten, bestehende Lieferketten aufrechtzuerhalten [28]. In Kombination mit den in der Einleitung herausgestellten Defiziten traditioneller KI-Modelle und den Möglichkeiten multimodaler Foundation-Modelle, die im vorhergehenden Kapitel erläutert wurden, zeigt sich, dass letztere einen großen Beitrag für die Zukunft der Produktion liefern können.

Eine primäre Charakteristik multimodaler Foundation-Modelle, die hier eine wichtige Rolle spielt, ist die einfache Übertragbarkeit dieser Modelle auf verschiedene Aufgaben. Während klassische KI-Modelle für jede neue Aufgabe auf Basis von Daten, die häufig in aufwendigen Versuchen aufgenommen werden, neu trainiert werden müssen, lassen sich multimodale Foundation-Modelle deutlich einfacher anpassen, da sie als Basismodell bereits auf einer so großen Datenmenge trainiert wurden, dass sie ein implizites Verständnis einer Reihe an Objekten und Assoziationen gelernt haben [12]. Diese Charakteristik wird durch das sogenannte Grounding verstärkt, das heißt dem Lernen von Assoziationen von Objekten zur physischen Welt, welches über verschiedene Modalitäten hinweg gute Ergebnisse erzielt [19]. Dies wird etwa für Perzeptionsmodelle für Roboter-Greifaufgaben in diversen Arbeiten untersucht [29, 30]. Die grundsätzliche Idee in diesen Anwendungen ist es, die Repräsentationen von reinen Bildsequenzen um textuelle Sprache zu erweitern, um dem Roboter so mehr Kontext und ein besseres Verständnis zu vermitteln. Ein Beispiel für ein solches Modell ist das von der Open X-Embodiment Collaboration vorgeschlagene RT-2-X-Modell, welches Instruktionen in natürlicher Sprache in Kombination mit Bilddaten verarbeitet und Handlungsempfehlungen für die Robotersteuerung in natürlicher Sprache ausgibt [31, 32]. Durch das Training dieses Modells auf Bild-Sprache-Paaren wird das Szenenverständnis maßgeblich verbessert, sodass der Roboter in der Lage ist, auch komplexe Anweisungen zu interpretieren und korrekt zu verarbeiten, **Bild 3**. Das Modell zeigt dabei eine starke Übertragbarkeit auf neue Anwendungen und kann diese einfach durch neue Instruktionen mit neuen textuellen Eingangsgrößen verarbeiten [31]. Die Möglichkeit, ein Basismodell ohne jegliches Training und damit Datenaufwände auf eigene individuelle Anwendungen allein durch das Setzen von textuellen Randbedingungen zu übertragen, stellt einen wesentlichen Vorteil multimodaler Foundation-Modelle dar [19]. Dies kann auch für Anomaliedetektionen im Bereich der Qualitätssicherung genutzt werden. Der Einsatz verschiedener Prompt-Templates für Bild-Sprache-Modelle wurde so etwa in [33] auf einer Reihe von Benchmark-Datensätzen, etwa dem MVTec AD-Datensatz [34], einem Datensatz speziell für die industrielle Qualitätssicherung ausgewertet. Dort wurde festgestellt, dass Anomalien über gewisse Prompt-Templates bereits hochgenau detektiert werden können, unter der Voraussetzung, dass ein Referenzbild eines Gutteils ebenfalls in das Modell eingegeben wird [33], Bild 3.

Anhand des letzten Anwendungsfalls lässt sich eines der weiteren Potenziale multimodaler Foundation-Modelle erkennen. Die Anomalien werden in dem gegebenen Setup nicht nur detektiert, sondern das Modell liefert Erklärungen, warum das gegebene Modell einen Defekt enthält oder nicht. Diese Charakteristik der Erklärbarkeit der Modellausgänge kann für viele Anwendungsfälle relevant sein. Am Werkzeugmaschinenlabor der RWTH Aachen wird daran geforscht, den Einsatz solcher Bild-Sprache-Modelle für die Interpretation von Computertomographie-(CT-)Bilddaten heranzuziehen. Aufgrund zahlreicher Einflussfaktoren auf die CT-Bildgebung ist die Detektion in diesen Bildern oftmals herausfordernd und die Differentiation von CT-Artefakten und tatsächlichen Defekten für KI-Modelle und menschliche Entscheider schwierig. Textuelle Modellausgänge, die als Unterscheidungsstützung für menschliche Operatoren dienen, können hier einen großen Mehrwert bieten und die Entscheidungsqualität bei der CT-Prüfung erheblich verbessern.

Auch im Bereich des betrieblichen Wissensmanagements bieten die Bild-Sprache-Kopplung beziehungsweise multimodale Repräsentationen im Allgemeinen Potenziale. Hier wird bereits in vielen Unternehmen der Einsatz von LLM-Chatbot-Lösungen untersucht, die an Datenbanken angebunden sind, welche in vektorisierter Form Repräsentationen des betriebsinternen, textuellen Wissens enthalten [35]. Dadurch werden Chatbot-Systeme in die Lage gebracht, fundierte Antworten zu unternehmensinternen Fragestellungen zu beantworten, auch wenn die Sprachmodelle selbst nicht auf dieser Datengrundlage trainiert wurden, indem stattdessen semantische Ähnlichkeiten zwischen Nutzerfragen

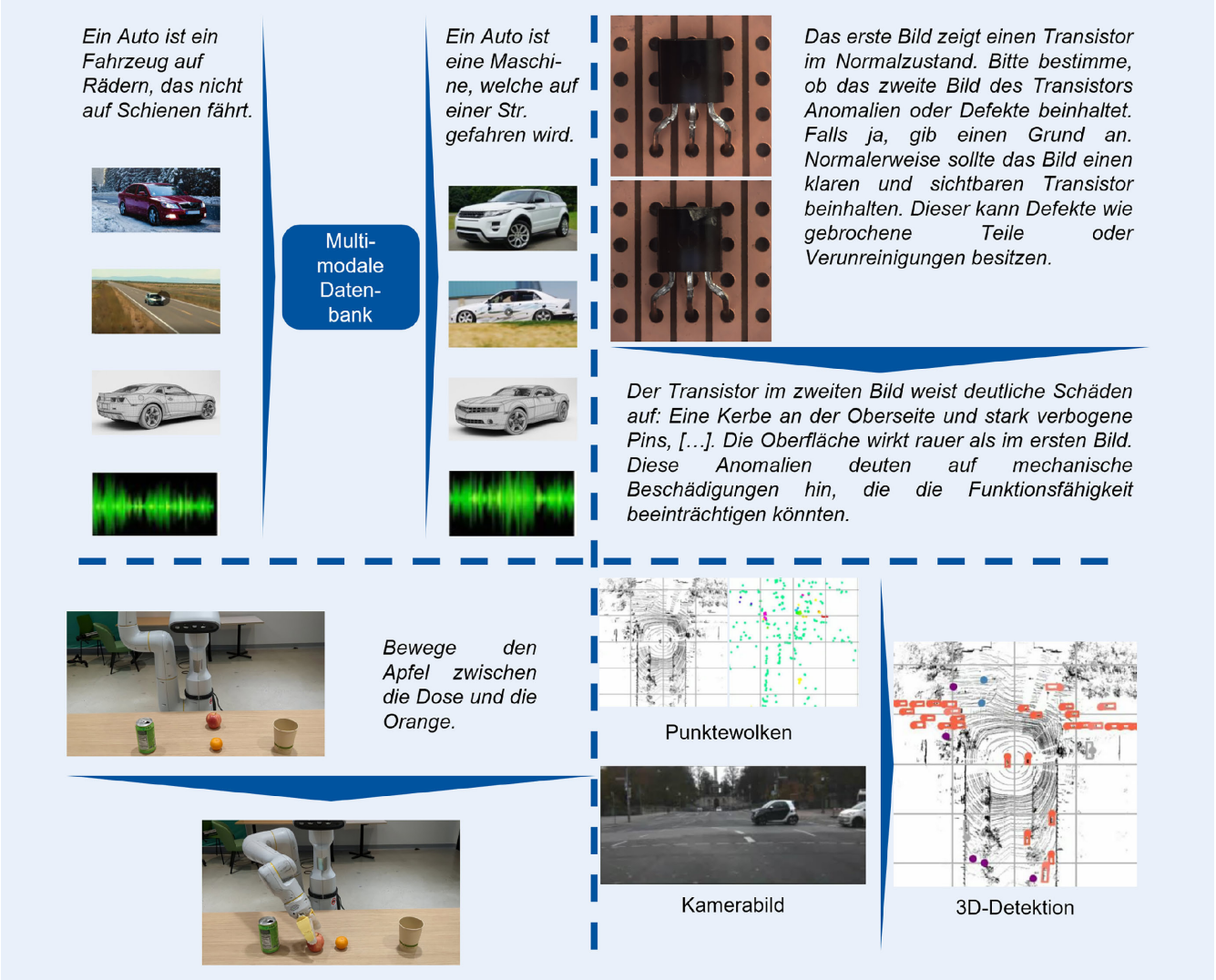

Bild 3. Übersicht verschiedener Anwendungsmöglichkeiten multimodaler Foundation-Modelle in der Produktion. Links oben: RT-2-X-Modell zur automatisierten Robotersteuerung bei komplexen Handlungsinstruktionen. Oben rechts: Anomaliedetektion über Bild-Sprache-Modelle durch Prompttemplates. Unten links: Cross-Modal-Retrieval zum innerbetrieblichen Wissensmanagement. Unten rechts: DeepFusion-Ansatz zur 3D-Objektdetektion von Objekten im Kontext des autonomen Fahrens. *Grafik: WZL | RWTH Aachen University auf Basis von [31, 33, 34, 37, 38]*

und Dokumenten gesucht werden und diese dem Sprachmodell für die Antwortgeneration bereitgestellt werden (Retrieval Augmented Generation) [36]. Die Erweiterung auf semantische Suchen in multimodalen Datenquellen, das sogenannte Cross-Modal Retrieval, kann die Potenziale dieser Technologie nochmals verbessern, denn häufig ist betriebsinternes Wissen nicht nur in Texten natürlicher Sprache enthalten, sondern auch in Bildern oder Bildsequenzen [37]. Das Suchen nach passenden Informationen in multimodalen Datenquellen kann so etwa das Training und die Einarbeitung für neues Personal verbessern, den Kundensupport entlasten, indem Kunden eine Reihe von Problemen bereits selbstständig diagnostizieren können oder die Planung von Prüfversuchen erleichtern, indem automatisiert in Datenbanken nach Parametereinstellungen für bereits verprobte Bauteile etwa anhand von Ähnlichkeiten von CAD-Modellen gesucht wird, Bild 3.

Die bisher vorgestellten Anwendungsbereiche beziehen sich primär auf die Kopplung von Bild und Sprache. Im Kontext von Industrie 4.0 sind jedoch auch vor allem Modelle interessant, die neben diesen beiden Modalitäten Sensordaten mit in die Modellierung aufnehmen. Ein Beispiel ist die von Google und der Johns Hopkins University vorgestellte „DeepFusion"-Modellserie, die die Fusion von Kameradaten und Lidar-Sensoren zur 3D-Objektdetektion im Bereich des autonomen Fahrens untersucht [38]. Die Nutzung multimodaler Eingangsgrößen führt zu erhöhten Genauigkeiten und erhöhter Robustheit der Ansätze. Dies wird so auch im Bereich der mobilen Robotik beziehungsweise Industrierobotik untersucht und soll die Detektionsmöglichkeiten von sogenannten Open-Set-Modellen, das heißt, Modellen, die Instanzen beliebiger Objektklassen detektieren können, verbessern [39, 40], Bild 3.

Die Übersicht zeigt, dass im Bereich multimodaler Foundation-Modelle primär der Einsatz von Bild-Sprache-Modellen untersucht wird, sowie Anwendungen mit Ausnahme der Robotikdomäne rar sind. Vorteile multimodaler Foundation-Modelle werden vor allem in den Bereichen einfache Übertragbarkeit, etwa dem Ermöglichen von Open-Set-Detektionen ohne die aufwendige Aufnahme von Trainingsdaten, der erhöhten Robustheit und Generalisierbarkeit sowie der verbesserten Erklärbarkeit gesehen.

Diese Charakteristiken können prinzipiell für eine Vielzahl der in Bild 2 dargestellten Anwendungsbereiche genutzt werden und dort zu verbesserten Prozessen führen. Aktuell mangelt es hier aber noch an konkreten Anwendungsfällen. Um eine weitere Verbreitung zu erreichen und Anwendungsfälle zu identifizieren, bieten die vom BMBF geförderten KI-Servicezentren kostenlose Informationen und Erstgespräche an und stellen unter bestimmten Voraussetzungen entsprechende Rechenressourcen oder vertiefende Ausarbeitungen von Ideen über Forschungskooperationen zur Verfügung.

4 Herausforderungen für den Einsatz multimodaler Foundation-Modelle

Obwohl multimodale Foundation-Modelle großes Potenzial für verschiedene Aufgaben versprechen, nutzen, wie die Übersicht aus dem vorangegangenen Abschnitt zeigt, produzierende Unternehmen diese Technologie aktuell kaum. Dies ist primär auf die folgenden Herausforderungen zurückzuführen [6, 12, 13]:

- Domänenlücke zwischen Trainingsdaten bestehender multimodaler Foundation-Modelle und industriellen Datensätzen
- Geringe Verfügbarkeit öffentlicher industrieller Trainings- und Benchmarkdatensätze
- Unterrepräsentiertheit bestimmter Datenmodalitäten
- Verfügbarkeit von Rechen- und Dateninfrastruktur
- Geringe oder nicht vorhandene KI-Kompetenz
- Bedenken zur Sicherheit und Verlässlichkeit

Wesentliches Defizit für den Einsatz multimodaler Foundation-Modelle in der Produktion sind Daten. Insbesondere stellt die Verfügbarkeit großer, branchenübergreifender, offener Datensätze ein Problem dar. Dies hat zur Folge, dass nahezu alle großen Modelle auf Datensätzen vortrainiert werden, die über eine ausgesprochene Domänenlücke zu produktionsnahen Daten verfügen. Diese vortrainierten Modelle lassen sich so häufig nur mit großen Leistungseinbußen auf die meist speziellen industriellen Probleme, etwa spezielles industrielles Vokabular, überführen oder müssen durch Fine-Tuning auf die spezifischen Problemstellungen angepasst werden [41]. Erste Initiativen wie die Open X-Embodiment Collaboration stellen solche Datensätze zur Verfügung, allerdings fehlt es im Allgemeinen weiterhin an branchenübergreifenden Datensätzen [32]. Herausforderungen im Kontext der Daten beziehen sich darüber hinaus auch auf die Unterrepräsentiertheit bestimmter Modalitäten, die in Produktionsumgebungen besonders relevant sind. Für Zeitreihen- oder 3D-Punktewolkendaten von Sensoren etwa sind kaum Modelle verfügbar und ihre Leistungsfähigkeit zusätzlich eingeschränkt [12]. Die Verfügbarkeit von Datensätzen bezieht sich dabei nicht nur auf Datensätze, die zur Entwicklung derartiger Modelle vorhanden sein müssen, sondern auch auf Benchmarks, die einen fairen Vergleich verschiedener Modelle ermöglichen und Entwickler bei der Auswahl geeigneter Basismodelle unterstützen sollen [42].

Neben Herausforderungen im Kontext Daten erschweren insbesondere auch die notwendigen Anforderungen an Daten- und Recheninfrastruktur die Nutzung großer, multimodaler KI-Modelle [13]. Dies ist mehr noch als bei traditionellen Modellen für große, multimodale KI-Modelle eine Herausforderung. Zum einen müssen für multimodale Modelle Daten aus verschiedenen Quellen fusioniert werden, was ebenfalls mit Mehraufwänden in Entwicklung und Betrieb verbunden ist [43]. Zum anderen übersteigen die erforderlichen Rechenressourcen beim Hosting von Foundation-Modellen die der klassischen KI-Modelle um ein Vielfaches [13]. Gleiches gilt für das Fine-Tuning dieser Modelle. Für die Nutzung dieser Technologie wird somit aufgrund der Homogenisierung der Modellarchitekturen neben der KI-Expertise mehr und mehr die Expertise im Bereich des Hochleistungsrechnens relevant. Cloudanbieter und andere kommerzielle Dienste bieten die benötigte Hardware und ermöglichen so, die hohen Rechenanforderungen zu bewältigen. Hier bestehen jedoch häufig bei Industrieunternehmen Bedenken zur Datensicherheit und Verlässlichkeit, weswegen sich die Auswahlmenge an Anbietern häufig stark verringert [13].

Viele der genannten Herausforderungen sind aktuell stark fokussierte Forschungsbereiche. Insbesondere die Verfügbarkeit an Daten und Modellen für bestimmte Domänen wird aktuell in vielen, auch europäischen oder deutschen, Forschungsinitiativen wie „LAION“ untersucht [42]. WestAI bietet an dieser Stelle _Unternehmen und Forschenden Hilfe, in diesem dynamischen Umfeld den Überblick durch das Angebot von Beratungs- und Schulungsleistungen zugeschnitten auf die Bedürfnisse bei der Nutzung großer KI-Modelle zu bekommen. Zusätzlich werden im Rahmen des Projekts kostenfrei Rechenressourcen von bis zu 10 000 GPU-Stunden auf moderner Hochleistungsrecheninfrastruktur bereitgestellt. WestAI kann so dazu beitragen, Herausforderungen beim Technologietransfer multimodaler Foundation-Modelle aufzuheben.

5 Fazit und Ausblick

Traditionelle KI-Ansätze stoßen vor dem Hintergrund aktueller Anforderungen an Flexibilität und Robustheit in der Produktion an ihre Grenzen. Der Einsatz multimodaler Foundation-Modelle verspricht, diese Limitierungen aufzuheben und für eine Vielzahl möglicher Anwendungen eingesetzt zu werden. Primäre Vorteile dieser Technologie liegen in der besseren Übertragbarkeit auf verschiedene Aufgaben, der verbesserten Genauigkeit und Robustheit dieser Modelle sowie in der Nutzung textueller Ausgangsgrößen zur Erklärung von Modellentscheidungen. Diese Eigenschaften bieten das Potenzial auch traditionell schwierig zu automatisierende Aufgaben in der Produktion, etwa die Montage, zu automatisieren. Aktuelle Anwendungen im Produktionsbereich sind jedoch noch die große Ausnahme. Lediglich in der Robotikdomäne werden diese Modelle bereits in Anwendungen eingesetzt, zum Beispiel zur automatisierten Generierung von Handlungsanweisungen für die Robotersteuerung bei Greifaufgaben. Eine Herausforderung im Bereich der multimodalen Foundation-Modelle ist die Übertragung vortrainierter Basismodelle auf die industrielle Domäne, die in klassischen Datensätzen, mit denen diese Modelle vortrainiert sind, kaum enthalten ist, sodass Methoden des Modelltransfers, etwa Fine-Tuning, angewendet werden müssen. Für diesen Modelltransfer beziehungsweise die Nutzung großer KI-Modelle ist zusätzlich eine leistungsfähige Recheninfrastruktur notwendig, die insbesondere für kleine und mittlere Unternehmen nicht leicht verfügbar ist. Das Forschungsprojekt WestAI bietet interessierten Unternehmen und Forschenden an dieser Stelle die Möglichkeit, diese Rechenressourcen in einem gewissen Rahmen für Forschungsfragestellungen kostenfrei zu nutzen sowie konkrete industrielle Anwendungsfälle zu besprechen und gemeinsam im Rahmen von Forschungskooperationen Lösungen zu entwickeln.

Weiterer Forschungsbedarf im Bereich multimodaler Foundation-Modelle liegt unter anderem in der Erweiterung offener Datensätze auf zusätzliche Modalitäten sowie dem Benchmarking von Modellen auf industriellen Datensätzen. Darüber hinaus sind insbesondere für industrielle Anwendungen Projekte mit Leuchtturmcharakter erforderlich, die den Nutzen und die Transferierbarkeit dieser Modelle an realen Anwendungen zeigen.

FÖRDERHINWEIS

Gefördert durch das Bundesministerium für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01IS22094D WEST-AI.

Literatur

[1] Vogel-Heuser, B.: Handbuch Industrie 4.0. Heidelberg: Springer Vieweg 2017
[2] Tercan, H.; Meisen, T.: Machine learning and deep learning based predictive quality in manufacturing: a systematic review. Journal of Intelligent Manufacturing 33 (2022) 7, S. 1879–1905
[3] Yan, Y.; Chow, A. H.; Ho, C. P. et al.: Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities. Transportation Research Part E: Logistics and Transportation Review 162 (2022), S. 102712
[4] Wiggers, K.: IDC: For 1 in 4 companies, half of all AI projects fail. Internet: https://venturebeat.com/ai/idc-for-1-in-4-companies-half-of-all-ai-projects-fail/. Zugriff am 04.08.2024
[5] Freiesleben, T.; Grote, T.: Beyond generalization: a theory of robustness in machine learning. Synthese 202 (2023) 4, S. 109
[6] Hu, Y.; Quanting Xie; Jain, V. et al.: Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis. 2023
[7] Azmoodeh, A.; Dehghantanha, A.: Big Data and Privacy: Challenges and Opportunities. In: Choo, K.-K. R.; Dehghantanha, A. (Hrsg.): Handbook of Big Data Privacy. Cham: Springer International Publishing 2020, S. 1–5
[8] Peres, R. S.; Jia, X.; Lee, J. et al.: Industrial Artificial Intelligence in Industry 4.0 – Systematic Review, Challenges and Outlook. IEEE Access 8 (2020), S. 220121–220139
[9] Paleyes, A.; Urma, R.-G.; Lawrence, N. D.: Challenges in Deploying Machine Learning: A Survey of Case Studies. ACM Computing Surveys 55 (2022) 6
[10] Munikoti, S.; Stewart, I.; Horawalavithana, S. et al.: Generalist Multimodal AI: A Review of Architectures, Challenges and Opportunities. 2024
[11] Brown, T. B.; Mann, B.; Ryder, N. et al.: Language Models are Few-Shot Learners. 2020
[12] Bommasani, R.; Hudson, D. A.; Adeli, E. et al.: On the Opportunities and Risks of Foundation Models. 2022
[13] Bienert, J.; Broch, R.; Bunk, P. et al.: Große KI-Modelle für Deutschland. 2023
[14] Singla, A.; Sukharesvsky, A.; Yee, L. et al.: The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. Internet: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai. Zugriff am 31.08.2024
[15] WestAI – KI-Servicezentrum: KI-Services aus NRW für Deutschland. Internet: https://westai.de/. Zugriff am 12.09.2024
[16] Kaplan, J.; McCandlish, S.; Henighan, T. et al.: Scaling Laws for Neural Language Models. 2020
[17] Edward J. Hu; Yelong Shen; Phillip Wallis et al.: LoRA: Low-Rank Adaptation of Large Language Models. 2021
[18] Vaswani, A.; Shazeer, N.; Parmar, N. et al.: Attention Is All You Need. 2023
[19] Li, C.; Gan, Z.; Yang, Z. et al.: Multimodal Foundation Models: From Specialists to General-Purpose Assistants. 2023
[20] Zhao, L.; Alhoshan, W.; Ferrari, A. et al.: Natural Language Processing (NLP) for Requirements Engineering: A Systematic Mapping Study. 2020
[21] Yüksel, N.; Börklü, H. R.: A Generative Deep Learning Approach for Improving the Mechanical Performance of Structural Components. Applied Sciences 14 (2024) 9
[22] Aktepe, A.; Yanık, E.; Ersöz, S.: Demand forecasting application with regression and artificial intelligence methods in a construction machinery company. Journal of Intelligent Manufacturing 32 (2021) 6, S. 1587–1604
[23] Göppert, A. M. R.: Artificial intelligence in online scheduling of dynamically interconnected assembly systems, RWTH Aachen University, 2022
[24] Tatzel, L.; Puente León, F.: Image-based roughness estimation of laser cut edges with a convolutional neural network. Procedia CIRP 94 (2020), S. 469–473
[25] Woschank, M.; Rauch, E.; Zsifkovits, H.: A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics. Sustainability (Switzerland) 12 (2020) 9
[26] Carvalho, T. P.; Soares, Fabrízzio A. A. M. N.; Vita, R. et al.: A systematic literature review of machine learning methods applied to predictive maintenance. Computers and Industrial Engineering 137 (2019)
[27] Göppert, A.; Hüttemann, A.; Jung, S. et al.: Frei verkettete Montagesysteme. Ein Ausblick. Zeitschrift für wirtschaftlichen Fabrikbetrieb 113 (2018) 3, S. 151–155
[28] Demary, V.; Matthes, J.; Plünnecke, A. et al.: Gleichzeitig: Wie vier Disruptionen die deutsche Wirtschaft verändern. Herausforderungen und Lösungen. Köln 2021
[29] Wake, N.; Kanehira, A.; Sasabuchi, K. et al.: GPT-4V(ision) for Robotics: Multimodal Task Planning from Human Demonstration. 2024
[30] Guo, D.; Xiang, Y.; Zhao, S. et al.: PhyGrasp: Generalizing Robotic Grasping with Physics-informed Large Multimodal Models. 2024
[31] Brohan, A.; Brown, N.; Carbajal, J. et al.: RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. 2023
[32] Open X-Embodiment Collaboration; O'Neill, A.; Rehman, A. et al.: Open X-Embodiment: Robotic Learning Datasets and RT-X Models. 2024
[33] Xu, X.; Cao, Y.; Chen, Y. et al.: Customizing Visual-Language Foundation Models for Multi-modal Anomaly Detection and Reasoning. 2024
[34] Bergmann, P.; Batzner, K.; Fauser, M. et al.: The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. International Journal of Computer Vision 129 (2021) 4, S. 1038–1059
[35] Xu, Z.; Cruz, M. J.; Guevara, M. et al.: Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering, 2024, S. 2905–2909
[36] Yunfan Gao; Yun Xiong; Xinyu Gao et al.: Retrieval-Augmented Generation for Large Language Models: A Survey. 2024
[37] Zhou, K.; Hassan, F. H.; Hoon, G. K.: The State of the Art for Cross-Modal Retrieval: A Survey. IEEE Access 11 (2023), S. 138568–138589
[38] Yingwei Li; Adams Wei Yu; Tianjian Meng et al.: DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection. 2022
[39] Jang, D.-S.; Cho, D.-H.; Lee, W.-C. et al.: Unlocking Robotic Autonomy: A Survey on the Applications of Foundation Models. International Journal of Control, Automation and Systems 22 (2024) 8, S. 2341–2384
[40] Le, X.; Yu, N.; Zhang, S. et al.: ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding. 2024
[41] Ji, W.; Li, J.; Bi, Q. et al.: Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications. Machine Intelligence Research 21 (2024) 4
[42] Gadre, S. Y.; Ilharco, G.; Fang, A. et al.: DataComp: In search of the next generation of multimodal datasets. 2023
[43] Riddiough, J.: Challenges and Solutions in Multimodal Models. Internet: https://aimodels.org/multimodal-artificial-intelligence/challenges-and-solutions-multimodal-models/. Zugriff am 24.10.2024

Hannes Behnen, M. Sc.
hannes.behnen@wzl-iqs.rwth-aachen.de
Tel. +49 241 80-25836
Foto: WZL | RWTH Aachen University

Jan-Henrik Woltersmann, M. Sc.

Dominik Wolfschläger, M. Sc.

Prof. Dr.-Ing. Robert H. Schmitt

WZL | RWTH Aachen University
Intelligence in Quality Sensing (IQS)
Lehrstuhl für Informations-, Qualitäts- und Sensorsysteme in der Produktion
Campus-Boulevard 30, 52074 Aachen
https://wzl.rwth-aachen.de