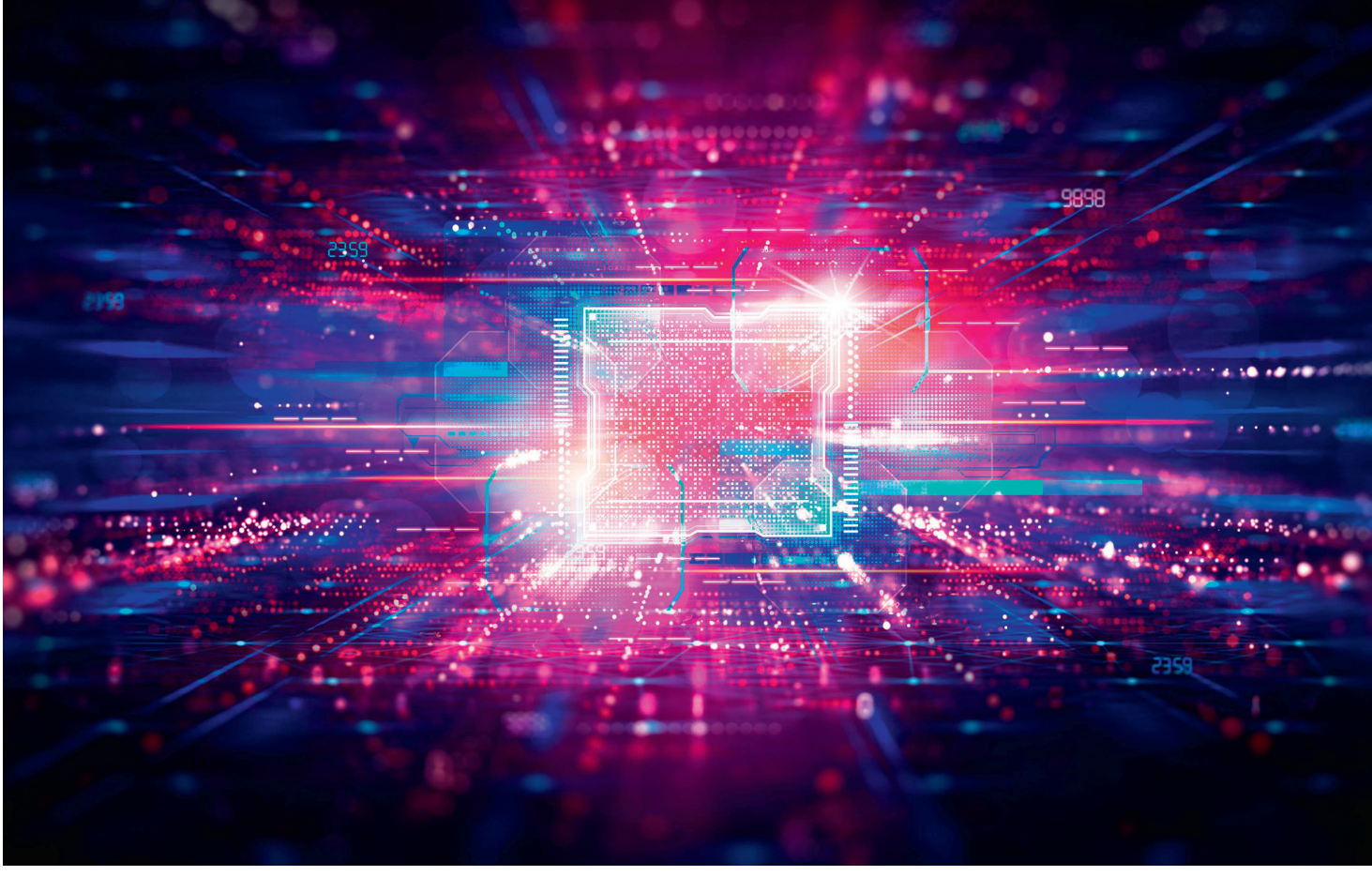

© ArtemisDiana | Shutterstock

Trainingsdaten für Fahrerassistenzsysteme

Synthetische Daten – ein Booster für den Einsatz von KI

Künstliche Intelligenz prägt schon heute automobile Entwicklungsprozesse. Besonders deutlich wird das beim Training von Umgebungswahrnehmungssystemen, das riesige Mengen hochwertiger Daten erfordert. Fachleute von ITK Engineering zeigen, dass synthetische Daten eine vielversprechende Lösung sind.

Johannes Schmidt, Peter Allard

Maschinelles Lernen kommt in der Automobilbranche zunehmend zum Einsatz. Bereits heute nutzen die Wahrnehmungssysteme moderner Fahrzeuge vielfach Modelle auf Basis Künstlicher Intelligenz (KI). Eine zentrale Herausforderung: die Beschaffung großer Mengen qualitativ hochwertiger Daten. So sind zum Training des Wahrnehmungssystems eines Fahrzeugs mehrere hundert Terrabyte an Daten notwendig [1], und zeitgleich ist ihre Vielfalt entscheidend, speziell die Abdeckung seltener und kritischer Szenarien. Der Aufbau von Pipelines zur datenschutzkonformen Aufnahme und Verarbeitung von Realdaten stellt einen erheblichen Aufwand dar.

Vorteile synthetischer Daten

Synthetische Daten dagegen bieten entscheidende Vorteile. Sie lassen sich in beliebig großen Mengen erzeugen und zielgerichtet auf die Darstellung spezifischer Szenarien zuschneiden. Die Zahl an weltweiten Veröffentlichungen in Forschung und Industrie zu diesem Thema wächst in den letzten Jahren stark [2]. Interessant macht sie auch der niedrigere Kosten- und Zeitaufwand im Vergleich zur Aufnahme von Realdaten. Dieser wird durch Einschränkungen in der Verarbeitung und Nutzung durch die Datenschutz-Grundverordnung (DSGVO) noch erhöht.

Synthetische Daten können orts-, zeit- und wetterunabhängig erzeugt werden (**Bild 1**). Auch Edge Cases, die in der Realität nur selten auftreten, lassen sich synthetisch in großem Umfang bereitstellen.

Zudem lässt sich der in Realdaten oft auftretende Bias – eine Überrepräsentation gewisser Faktoren – durch

Bild 1: Synthetisch erzeugtes Szenario mit japanischen Fahrbahnmarkierungen bei unterschiedlichen Wetterbedingungen. © ITK Engineering

die Beimischung synthetischer Daten verringern. Nutzt ein deutsches Unternehmen z. B. hauptsächlich reale Daten im lokalen Umfeld, wird das trainierte System auf heimische Umgebungen fokussiert und im Ausland schlechter arbeiten.

Auch im Bereich des Labelings punkten synthetische Daten durch Zeit- und Kosteneinsparungen sowie verlässliche Qualität gegenüber realen Daten [3], da der Prozess zur Generierung der Labels vollständig automatisiert werden kann.

Ihr volles Potenzial entfalten synthetisch erzeugte Daten, wenn ihre Qualität sichergestellt ist. Diese wird anhand der Ähnlichkeit zwischen synthetischen und realen Daten bemessen. Den Grad des Realismus synthetischer Daten zu erhöhen, ist Gegenstand der Forschung, wobei aktuell vielversprechende Verfahren auf die Nutzung von Modellen aus dem Bereich der generativen KI setzen [4].

Vorteile und Limitationen Generativer Neuronaler Netze

Synthetische Daten, die beispielsweise aus einer Simulation stammen, können mittels solcher Modelle nachverarbeitet werden. So wird auch diskutiert, ob synthetische Daten künftig den Großteil der Trainingsdaten bilden können [5]. Wissenschaftliche Erkenntnisse unterstützen diese Einschätzung [6].

ITK Engineering evaluiert passende Methoden und erarbeitet Toolings zur Erzeugung synthetischer Daten. Da generative neuronale Netze hier besonders leistungsfähig sind, liegt es nahe, sie zur Datenerzeugung fotorealistischer Bilder zu nutzen. Hürden sind die entstehenden Artefakte in den generierten Bildern sowie die bisher unzuverlässige Bereitstellung von Ground-Truth-Informationen zu den generierten Szenen [7].

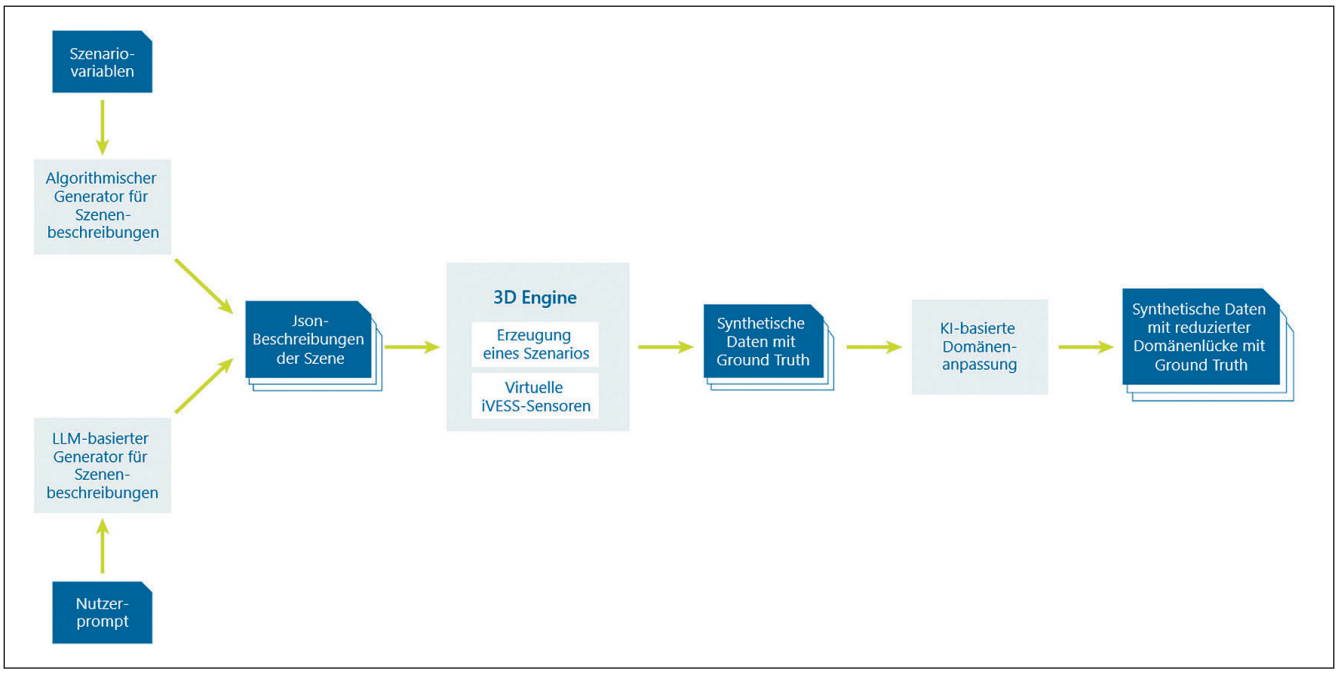

Bild 2: Datenerzeugungspipeline für gelabelte synthetische Sensordaten © ITK Engineering

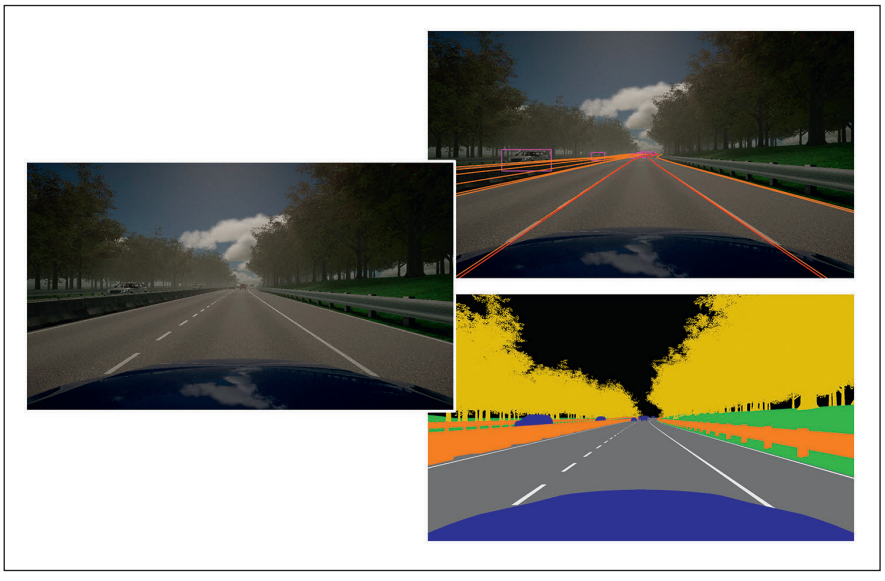

Bild 3: Beispiel eines mit der Datenpipeline erzeugten Szenarios (links) inklusive Labels für Fahrbahnmarkierungen und Begrenzungsrahmen für Fahrzeuge (rechts oben) sowie Segmentierungsmaske (rechts unten) ©TK Engineering

Game Engines als Alternative

Neben generativen KI-Modellen sind Game Engines wie Unity, Unreal oder Carla zur Datengenerierung geeignet. Sie bieten umfangreiche Möglichkeiten zur Gestaltung von virtuellen 3D-Umgebungen, in denen realitätsnahe Daten erzeugt werden können. Der Vorteil: Die Ground-Truth-Informationen sind inhärent verfügbar, da die Positionen aller Objekte in der Szene zu jedem Zeitpunkt bekannt sind. Die Herausforderung bei der Datengenerierung besteht allerdings im hohen zeitlichen Aufwand für den manuellen Aufbau der Szenarios.

Praxisbeispiel Datenerzeugungspipeline

Eine von ITK Engineering erarbeitete Generierungspipeline für synthetische Daten kombiniert traditionelle Methoden basierend auf Game Engines mit generativen Ansätzen (**Bild 2**). Durch gezielte Automation werden so die Stärken beider Vorgehensweisen kombiniert. Die virtuellen Umgebungen werden in standardisierter maschinenlesbarer Form beschrieben, die von einem Generator in der Game Engine geparst und in 3D-Szenen übersetzt werden. Der Generator übernimmt dann die prozedurale Generierung der 3D-Szene, die Platzierung von Sensoren sowie die Extraktion der damit erzeugten Rohdaten und Labelinformationen. Für die Datenerzeugung lassen sich Sensorsimulationsframeworks nutzen, wie das von ITK Engineering entwickelte Framework iVESS (individual Virtual Environment and Sensor Simulation). Es ermöglicht die Simulation eines breiten Spektrums an Sensoren wie Radar, Lidar, Ultraschall oder von Kameramodellen. Dieses Vorgehen setzt auf die Vorteile der herkömmlichen Verfahren, die deterministische und vollständig reproduzierbare Ergebnisse ermöglichen und Artefakte in den Daten vermeiden.

Insbesondere zwei Bereiche bieten vielversprechende Einsatzmöglichkeiten für generative Methoden. Entsprechend gepromptete Large Language Models (LLMs) lassen sich zu Beginn der Datenerzeugungspipeline einsetzen, um Szenenbeschreibungen mittels natürlicher Sprache zu erstellen. Diese Beschreibungen fungieren als Schnittstelle zur deterministischen Datenerzeugung. Hier ist auch eine Domain Randomization leicht umsetzbar, die per Prompt eine Randomisierung des abgebildeten Szenarios ermöglicht. Der zweite Einsatzbereich liegt in der Nachbearbeitung des Sensordatenoutputs. Verfahren der Domänenanpassung erlauben die Nachbearbeitung von Kamerabildern, wodurch sich die Ähnlichkeit zu realen Daten erhöht. Beispielsweise bietet Nvidia mit Cosmos Transfer ein KI-Modell für diesen Verarbeitungsschritt an [8]. So werden synthetische Daten realen Daten ähnlicher, die Qualität als Trainingsdaten steigt.

Zusammenfassung

Der Bedarf an qualitativ hochwertigen Trainingsdaten ist eine der zentralen Herausforderungen für die erfolgreiche Entwicklung KI-basierter Wahrnehmungssysteme für automatisierte Fahrfunktionen. Synthetisch erzeugte Daten können bei Training und Testing dieser Systeme entscheidend helfen. Die iVESS-Datengenerierungspipeline von ITK Engineering integriert traditionelle und generative Methoden und bietet einen Lösungsansatz zur effizienten Erzeugung qualitativ hochwertiger synthetischer Daten. ■ (sh)

www.itk-engineering.de

Quellenverzeichnis

[1] Nvidia Developer (HRSG.): Training AI for selfdriving vehicles: the challenge of scale. Online: https://developer.nvidia.com/blog/training-self-driving-vehicles-challenge-scale/, aufgerufen: 16. Februar 2025

[2] Tools for Innovation Monitoring (TIM) (Hrsg.): EDPS Weak Signals 2022–2023. Online: https://www.timanalytics.eu/TimTechPublic/dashboard/index.jsp#/space/s_1836?ds=224919, aufgerufen: 16. Oktober 2025

[3] CVAT (Hrsg.): Calculating the cost image annotation for AI projects: Annotating solo. Online: https://www.cvat.ai/blog/calculating-the-cost-of-solo-image-annotation-for-ai-projects, aufgerufen: 16. Oktober 2025

[4] Zhao et al.: Exploring Generative AI for Sim2Real in Driving Data Synthesis. Online: https://arxiv.org/abs/2404.09111, aufgerufen: 16. Oktober 2025

[5] Gartner (Hrsg.): Is synthetic data the future of AI? Online: https://www.gartner.com/en/newsroom/press-releases/2022–06–22-is-synthetic-data-the-future-of-ai, aufgerufen: 16. Oktober 2025

[6] Burdorf et al.: Reducing the Amount of Real World Data for Object Detector Training with Synthetic Data. Online: https://arxiv.org/abs/2202.00632, aufgerufen: 16. Oktober 2025

[7] Ma, W. et al.: Generating Images with 3D Annotations Using Diffusion Models. Online: https://arxiv.org/abs/2306.08103, aufgerufen: 16. Oktober 2025

[8] Abu Alhaija et al.: Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control. Online: https://arxiv.org/abs/2503.14492

Johannes Schmidt ist Entwicklungsingenieur Simulation bei ITK Engineering, Frankfurt am Main.
© ITK Engineering

Peter Allard ist Entwicklungsingenieur Visual Computing bei ITK Engineering, Frankfurt am Main.
© ITK Engineering