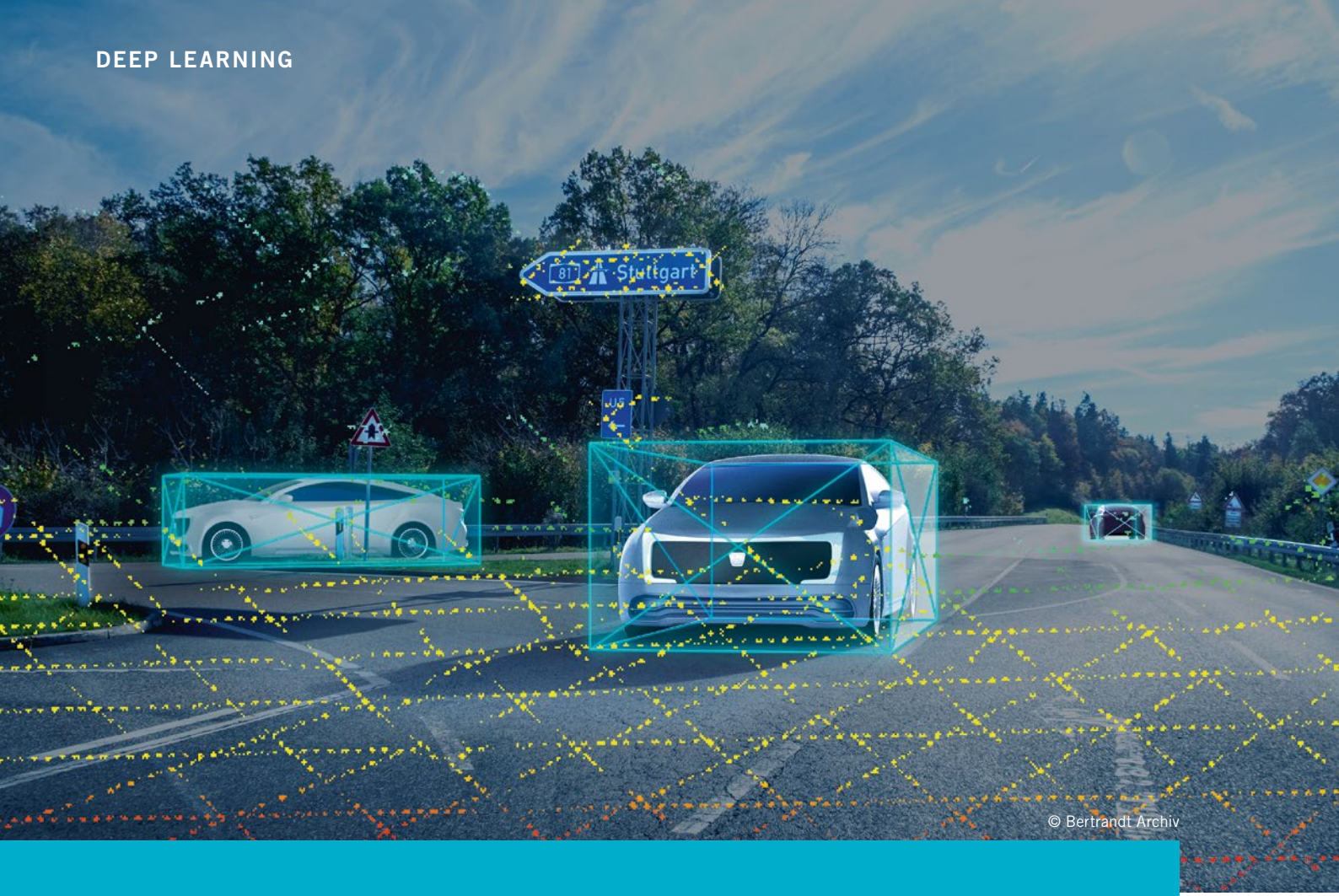

© Bertrandt Archiv

Deep-Learning-basierte 3-D-Objekterkennung – Daten, Training und Absicherung

Deep-Learning-Algorithmen und neuronale Netze sind wichtige Bausteine von automatisierten Fahrfunktionen zukünftiger autonomer Fahrzeuge. Bertrandt hat im Rahmen des Forschungsprojektes „3-D-CamLiFusion“ eine 3-D-Objekterkennung per Deep-Learning-basierter Fusion von Kamera- und Lidardaten erarbeitet. Das Projekt deckt den gesamten Deep-Learning-Workflow ab.

Die 3-D-Objekterkennung ist ein zentraler Bestandteil der Umfeldwahrnehmung von autonomen Fahrzeugen und Fahrerassistenzsystemen, der per Deep Learning (DL) umgesetzt wird. Insbesondere kosteneffiziente, aber dafür niedrigauflösende Lidarsensoren mit 4 bis 32 Ebenen werden in den nächsten Jahren für fortgeschrittene Fahrerassistenzsysteme (SAE-Level 2 und 3) in Privatfahrzeugen einsetzbar sein. Durch die geringe Auflösung der Lidardaten ist ihre Fusion mit detailreicheren Kameradaten besonders relevant und vielversprechend. Ansätze aus der aktuellen Forschung setzen hierbei auf tiefe neuronale Netze, die nicht nur isoliert lernen, in den Kamera- und Lidardaten die Merkmale der Objekte zu detektieren, sondern zusätzlich auch lernen, diese Datenströme und ihre Merkmale miteinander zu fusionieren.

ERSTELLUNG SYNTHETISCHER KAMERA- UND LIDARDATEN

Synthetische und simulierte Daten für Training und Absicherung von neuronalen Netzen bieten den Vorteil, dass der hohe Aufwand zum Labeln der Daten

AUTOREN

Dr. Cay-Christian Kalmbach
ist Lead Expert für KI/Deep Learning mit Fokus auf automatisiertes Fahren bei Bertrandt Technologie GmbH in Tappenbeck/Wolfsburg.

Dr. Xinxing Wang
ist Teamleiter bei Bertrandt Ingenieurbüro GmbH in Gaimersheim/Ingolstadt.

Jochen Schwenninger
ist Lead Expert im Bereich ADAS und Autonomes Fahren bei Bertrandt Technologie GmbH in Regensburg.

Christian Freystein
ist Entwicklungsingenieur im Bereich Infotainmentsysteme bei Bertrandt Ingenieurbüro GmbH in Tappenbeck/Wolfsburg.

entfällt. Besonders das Training von kamerabasierten KI-/DL-Funktionen benötigt allerdings Bilddaten mit einem hohen Grad an Fotorealismus, um den enormen Merkmalsreichtum der realen Welt ausreichend genau abzubilden. Simulierte Daten, die lediglich mithilfe von 3-D-Engines computergeneriert wurden, weisen häufig einen nicht ausreichenden Grad an Fotorealismus für das Training und die Absicherung neuronaler Netze auf [1].

In der aktuellen Forschung gibt es daher einen Trend zur Erstellung von synthetischen Daten, die reale Daten als Grundlage nehmen [2, 3]. Dies verknüpft den Vorteil des geringen Labeling-Aufwands mit der großen Merkmalsvielfalt von realen Daten. Darauf aufbauend wurde im Projekt „3-D-CamLiFusion" ein Ansatz zur Erstellung konsistenter Kamera- und Lidardaten entwickelt, der reale Daten als Basis verwendet und einen hohen Grad an Fotorealismus und Detailreichtum erzielt. Hierbei werden Straßenszenen eines realen, vollständig annotierten 3-D-Objekterkennungsdatensatzes (hier: der A2D2-Datensatz [4]) durch zusätzliche, synthetisch erstellte Objekte erweitert.

Um eine möglichst große Variation der Kamerainformationen der synthetischen Objekte zu erreichen, wird eine spezielle DL-Methode, ein generatives neuronales Netz (Generative Adversarial Network, GAN), eingesetzt. Hierfür werden im Forschungsprojekt die Ansätze StyleGAN2 [5] und SeFa [6] miteinander kombiniert. Die Idee lässt sich am Beispiel von Autos als zu erkennenden Objekten erklären: Als Input für das GAN dient das Bild eines roten Autos der Marke A. Das GAN ist nun in der Lage, daraus das Kamerabild eines gelben Autos mit neuen Designmerkmalen zu erstellen, **BILD 1**. Verschiedene optische Merkmale können gezielt manipuliert werden, etwa die Farbe, Konturen, Reflexionen, Lichtverhältnisse. So können nahezu beliebig viele Objektvariationen erstellt werden. Andere Merkmale, wie die Kameraperspektive, die Form oder die Größe des Objekts, können beibehalten werden, sodass die für die 3-D-Objekterkennung relevanten und von Menschen erstellten Labels zu Größe, Orientierung und Segmentierung weiterhin gelten und kein erneuter manueller Aufwand notwendig ist.

Wichtig ist, dass die Lidar-Punktwolke des synthetisch erzeugten Objekts sowohl zum Kamerabild als auch zur Lidarsensor-Konfiguration passt. Dafür wird ein Modul des Ansatzes MonoPSR [7] verwendet, das im Trainingsprozess gelernt hat, nur aus Bilddaten eine hochaufgelöste Punktwolke vorherzusagen. Die Punktwolke des synthetischen Objekts wird anschließend an die vorgesehene Stelle im 3-D-Raum transferiert und an die spezielle Charakteristik der Lidar-Sensorkonfiguration angepasst.

Mit dem entwickelten Ansatz kann, aufbauend auf realen Daten, eine große Menge an hochqualitativen Kamera- und Lidardaten automatisiert erstellt werden, die sowohl zum Training als auch zur Absicherung dienen können. Der Ansatz ist grundsätzlich für jede Art von Objekten geeignet.

GELERNTE FUSION VON KAMERA- UND LIDARDATEN

Der entwickelte Ansatz ist eine Kombination von zwei Ansätzen aus der aktuellen Forschung: Frustum PointNets [8] und MonoPSR [7]. Diese bauen auf einer 2-D-Objekterkennung im Kamerabild auf, **BILD 2**, da diese Funktion heutzutage sehr ausgereift ist. Bei der gewählten Sensorkonfiguration ist es sehr unwahrscheinlich, dass ein Objekt in den Lidardaten erkannt wird, nicht aber im Kamerabild. Daher bietet die 2-D-Objekterkennung im Kamerabild die höchste Wahrscheinlichkeit, ein Objekt initial zu detektieren.

Anschließend an die 2-D-Objekterkennung folgt die 3-D-Objekterkennung mit Fusion der Kamera- und Lidardaten. Das neuronale Netz lernt aus großen Datenmengen die Fusion auf Merkmalsebene. Durch diese Fusionsart werden wesentlich mehr Rohdaten mit einbezogen als bei einer reinen Fusion auf Objektebene, wie sie in aktuellen Serienfahrzeugen zu finden ist. Je nachdem, ob keine, wenige oder viele gemessene Lidar-Punkte für das Objekt vorliegen, lernt das neuronale Netz, die verfügbaren Informationen zu gewichten und zu fusionieren.

Für einen Menschen reichen die Bildinformationen des Objekts in Kombination mit einem einzelnen gemessenen Lidar-Punkt auf dem Objekt bereits aus, um mithilfe der eigenen dreidimensionalen Vorstellung des Objekts eine präzise 3-D-Bounding-Box am richtigen Ort zu kennzeichnen. In der entwickelten Architektur wird hierfür die gemessene niedrigauflösende Lidar-Punktwolke mit einer aus dem Bild vorhergesagten hochauflösenden Objekt-Punktwolke ergänzt/angereichert, **BILD 2** (Pfad 1). Ein Modul des neuronalen Netzes lernt, diese detaillierte Lidar-Punktwolke allein aus Bildinformationen vorherzusagen. Dieses Modul wird mit echten Lidardaten aus einem hochauflösenden Sensor trainiert. Bei der Anwendung im Fahrzeug (der sogenannten „Inferenz") verwendet es jedoch nur Bildinformationen, um die hochauflö-

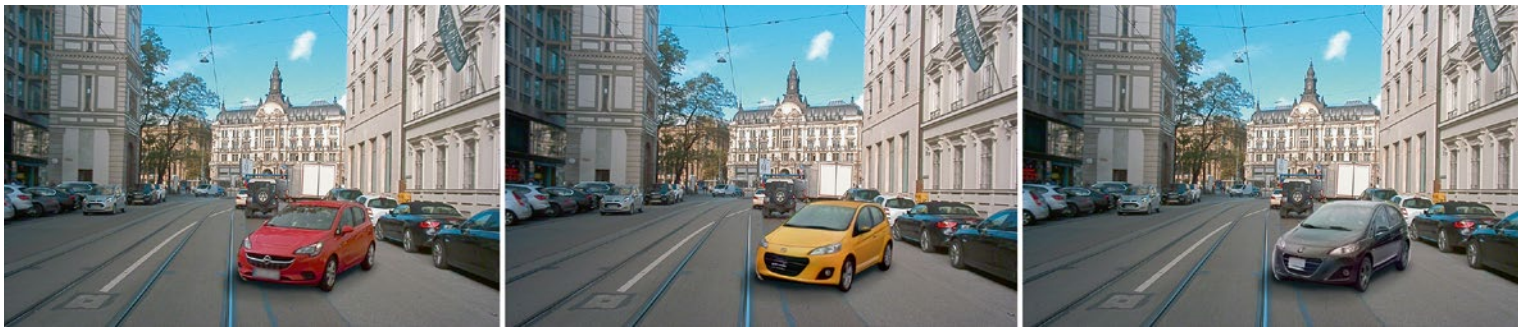

BILD 1 Synthetisch eingefügtes Auto in reale Hintergrundszene des A2D2-Datensatzes [4]; Auto ohne GAN-Manipulation (links); GAN-generiertes Auto mit gezielter Änderung der Autofarbe (Mitte, rechts) (© Bertrandt Archiv)

sende Punktwolke des Objekts zu generieren. Schließlich wird die vorhergesagte Lidar-Punktwolke mit den gemessenen Lidardaten fusioniert und ergänzt letztere. So stehen für die zu erkennenden Objekte detailreiche Punktwolken-Informationen zur Verfügung, im Fahrzeug wird daher lediglich ein niedrigauflösender Lidar-Sensor benötigt.

Aus dem Kamerabild werden Position, Größe und Abmessungen des zu erkennenden Objekts abgeleitet und mit den gemessenen Lidardaten in einem KI-Modul fusioniert, das die finale 3-D-Bounding-Box abschätzt, **BILD 2** (Pfad 2). Eine 3-D-Bounding-Box kann daher sowohl mit gemessenen Lidar-Punkten als auch nur mit Kamera-Informationen vorhergesagt werden.

Durch den modularen Aufbau können gezielt Zwischenergebnisse gelernt werden, die an anderen Stellen der Architektur fusioniert werden. Diese Zwischenergebnisse der einzelnen Module sind vom Menschen interpretierbar und helfen dabei, die Funktionsweise des neuronalen Netzes nachvollziehbar und im Fehlerfall analysierbar zu machen, was bei sicherheitskritischen Anwendungen von großer Bedeutung ist.

Die beschriebene Architektur der 3-D-Objekterkennung wird im Projekt auf der Pegasus-Plattform von Nvidia implementiert, um Echtzeit-Lauffähigkeit zu erreichen. Das neuronale Netz wird mithilfe von Nvidia TensorRT optimiert. Neben grundlegenden Optimierungen und eigenen TensorRT-Operationen in den Programmiersprachen C++ und CUDA wird die Genauigkeit der Gewichte des neuronalen Netzes zugunsten der Laufzeit reduziert (von 32-Bit Float auf 8-Bit Integer; int8-Inferenz), ohne die Erkennungsgenauigkeit signifikant zu beeinflussen. Verschiedene Teile der Architektur werden zur optimalen Ausnutzung der vorhandenen Ressourcen des Steuergeräts parallelisiert.

ABSICHERUNG NEURONALER NETZE

Die Absicherung von tiefen neuronalen Netzen ist essenziell für sicherheitskritische Anwendungen. Neuronale Netze funktionieren grundsätzlich anders als klassische regelbasierte Algorithmen, was sich in ihren teilweise Black-Box-ähnli-

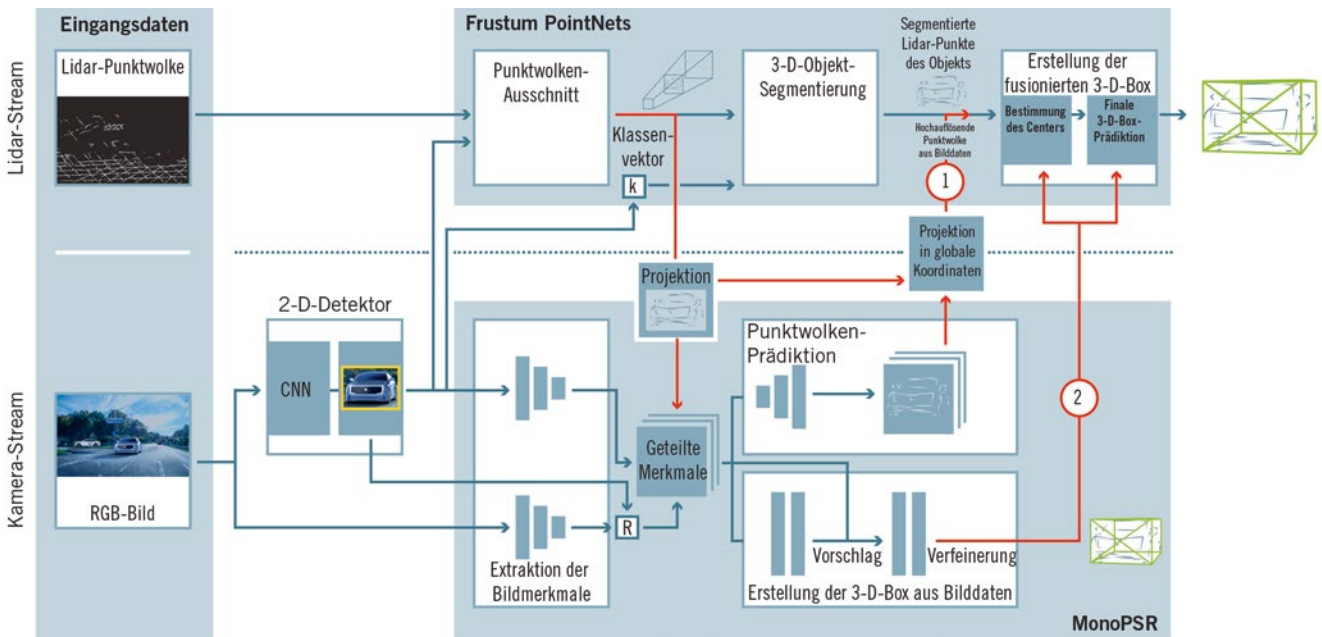

BILD 2 Schematische Darstellung zur Architektur der 3-D-Objekterkennung; Darstellung inspiriert durch [7] und [8] (© Bertrandt Archiv)

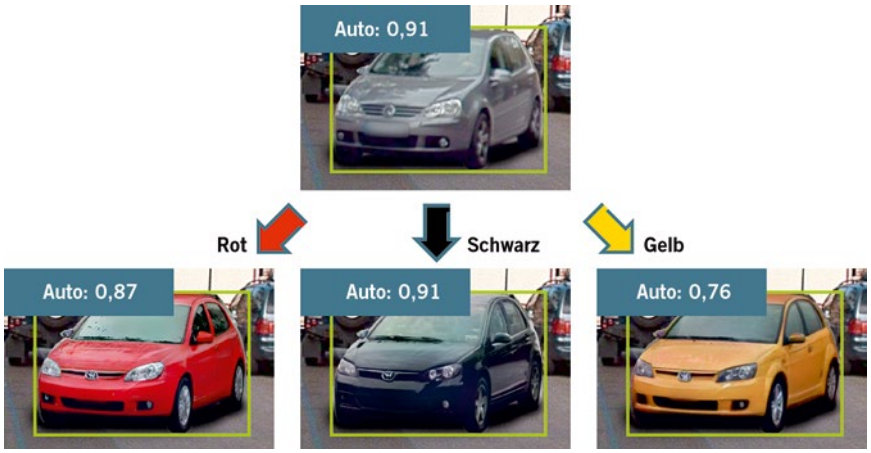

BILD 3 Durch die mittels GAN-Manipulation veränderte Farbe kann die Robustheit des neuronalen Netzes in Bezug auf diese Objekteigenschaft untersucht werden (© Bertrandt Archiv)

chen Eigenschaften und ihrer schwierigen Erklärbarkeit widerspiegelt. Folglich werden neuartige, für neuronale Netze angepasste Methoden der Absicherung benötigt. Bei der Absicherung von kamerabasierten Systemen wird eine enorme Menge an vielfältigen, fotorealistischen und mit Labels gekennzeichneten Daten benötigt. Vor allem für selten auftretende oder sicherheitskritische Szenarien (der sogenannte „long tail") ist es teuer und aufwendig, diese Daten zu gewinnen. Das Projekt „3-D-CamLiFusion" beschäftigt sich daher mit zwei zentralen Fragen: der Methodik zur Absicherung von neuronalen Netzen über aussagekräftige Metriken sowie der Verwendbarkeit von GAN-generierten synthetischen Daten für die KI-Absicherung.

Bei der Methodik zur Absicherung werden nach Cheng et al. [9] vier verschiedene Aspekte der Validation von neuronalen Netzen betrachtet: Robustheit, Interpretierbarkeit, Korrektheit (Correctness) und (Daten-)Vollständigkeit (Completeness) – die sogenannten RICC-Kriterien. Für jedes dieser Kriterien werden verschiedene Metriken, wie etwa Szenarien-Abdeckung, szenarienbasierte Performanz-Minderung, störungsbedingte Konfidenz-Minderung, **BILD 3**, sowie verschiedene Heatmap-basierte Metriken, an die Objekterkennungsaufgabe angepasst und für die Absicherung verwendet. Über die Berechnung der störungsbasierten Konfidenz-Minderung beispielsweise kann untersucht werden, ob etwa die Objektfarbe die Detektions-Konfidenz beeinflusst. So wurde das gelbe Auto, **BILD 3**, mit signifikant verringerter Konfidenz erkannt, was auf eine geringe Robustheit des neuronalen Netzes in Bezug auf die untersuchte Objekt-Eigenschaft hindeutet. Um das neuronale Netz in Bezug auf die RICC-Kriterien abzusichern, werden die im Projekt erstellten synthetischen Daten genutzt.

ZUSAMMENFASSUNG

Im 3-D-CamLiFusion-Projekt wurde eine Methode zur Deep-Learning-basierten Fusion von Kamera- und Lidar-Daten für 3-D-Objekterkennung entwickelt. Mithilfe von synthetischen Daten kann für Training und Validation eine große Menge unterschiedlicher Objekte und Szenarien automatisiert erstellt werden. Durch die Kombination von realen Hintergrundszenen mit GAN-generierten Objekten ist es dabei möglich, spezifische Objekteigenschaften (Farbe, Fahrzeugstil und -form, Umgebungslichtverhältnisse, Entfernung, Orientierung, Verdeckung oder andere Objektdetails) gezielt zu kontrollieren beziehungsweise zu manipulieren und trotzdem den für kamerabasierte neuronale Netze notwendigen Grad an Fotorealismus zu erreichen.

Aus dem Wechselspiel von Generierung synthetischer Daten, Training des neuronalen Netzes und Identifikation von Fehlerfällen mit Hilfe der Absicherungsmetriken entsteht ein iterativer Absicherungsprozess, der die Performanz und Robustheit des neuronalen Netzes stetig verbessert.

LITERATURHINWEISE

[1] Wrenninge, M; Unger, J.: Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. Online: https://arxiv.org/abs/1810.08705, aufgerufen: 2018
[2] Li, W.; et al.: AADS: Augmented Autonomous Driving Simulation using Data-driven Algorithms. Online: https://arxiv.org/abs/1901.07849v1, aufgerufen: 2019
[3] Yang, Z.; et al.: SurfelGAN: Synthesizing Realistic Sensor Data for Autonomous Driving. Conference on Computer Vision and Pattern Recognition, virtual, 2020
[4] Geyer, J.; et al.: A2D2: Audi Autonomous Driving Dataset. Online: https://arxiv.org/abs/2004.06320, aufgerufen: 2020
[5] Viazovetskyi, Y.; et al.: StyleGAN2 Distillation for Feed-forward Image Manipulation. In: Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J.-M.: Proceedings Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020
[6] Shen, Y.; et al.: Closed-Form Factorization of Latent Semantics in GANs. Online: https://arxiv.org/abs/2007.06600v1, aufgerufen: 2020
[7] Ku, J.; et al.: Monocular 3-D Object Detection Leveraging Accurate Proposals and Shape Reconstruction. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019
[8] Qi, C. R.; et al.: Frustum PointNets for 3-D Object Detection from RGB-D Data. Online: https://arxiv.org/abs/1711.08488v1, aufgerufen: 2017
[9] Cheng. C.-H.; et al.: Towards Dependability Metrics for Neural Networks. 16th ACM/IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE), Beijing, 2018

DANKE

Das Projekt „3D-CamLiFus" wird durch die Initiative „Neustart Niedersachsen Innovation" teilweise durch Mittel des Landes Niedersachsen gefördert. Dank geht zudem an die aktuellen und ehemaligen Kollegen Lennart Arendt, Mohammad Al Zoubi, Houssem Braham, Torsten Hafer, Daniel Hafner, Walid Haoues, Heiner Kowarsch, Niyantkumar Mehta, Abdelmawla Rizk und Mayuran Surendran.

DIESER BEITRAG IST IM E-MAGAZIN VERFÜGBAR UNTER:
www.emag.springerprofessional.de/atz