Semantische Segmentierung der Umgebung auf Basis von 3D-Daten

Simon Kuhn

Wissenschaftliche Arbeit im Zuge des Fachwissenschaftlichen Seminares

Erstprüfer:	Prof. Dr. Christian Pfitzner
Betreuer:	Prof. Dr. Christian Pfitzner
Ausgabedatum:	23.03.2023
Abgabedatum:	31.08.2023

Inhaltsverzeichnis

1 Einleitung

1.1 Hintergrund und Motivation

In den letzten Jahren hat die Forschung im Bereich der autonomen Fahrzeuge und der Robotik enorme Fortschritte gemacht. Ein wichtiger Faktor für die Entwicklung dieser Technologien ist die Fähigkeit, die Umgebung ausreichend genau zu erkennen und zu verstehen. In diesem Zusammenhang hat die semantische Segmentierung der Umgebung auf Basis von 3D-Daten eine immer größere Bedeutung erlangt. Die semantische Segmentierung ist ein Verfahren zur automatischen Klassifizierung von Objekten und Strukturen in der Umgebung. Dabei werden jedem Pixel oder jedem Voxel in einem 3D-Modell eine bestimmte semantische Bedeutung zugeordnet, z.B. Straße, Gebäude, Bäume oder Fahrzeuge. Eine präzise und schnelle semantische Segmentierung ist eine wesentliche Voraussetzung für eine zuverlässige Navigation von mobilen Plattformen, wie autonomen Fahrzeugen oder Robotersystemen [1]. In dieser Arbeit wird die semantische Segmentierung der Umgebung auf Basis von 3D-Daten untersucht. Dabei sollen verschiedene Methoden und Ansätze für die semantischen Segmentierung, sowie bestehende Probleme dargestellt und bewertet werden.

1.2 Problemstellung und aktueller Stand

Trotz der Fortschritte im Bereich der Computer Vision gibt es noch immer einige Herausforderungen zu überwinden. Eines der Probleme besteht in der Komplexität der Umgebung. Ein 3D-Umfeld kann durch eine Vielzahl von verschiedenen Objekten und Strukturen, die miteinander interagieren und sich gegenseitig beeinflussen, besonders herausfordernd sein. Es ist schwierig, all diese Details genau zu erfassen und zu segmentieren, besonders wenn die Daten unvollständig oder fehlerhaft sind. Ein weiteres Problem ist die Notwendigkeit einer hohen Verarbeitungsge-

schwindigkeit. Die Verarbeitung von großen Datenmengen erfordert eine erhebliche Rechenleistung, um eine schnelle und präzise Segmentierung der Umgebung zu ermöglichen. Dies kann für viele Anwendungen, insbesondere für sich schnell bewegende mobile Geräte, eine Herausforderung darstellen [2].

Hinzu kommt die begrenzte Genauigkeit der Segmentierungsverfahren. Es gibt noch immer Schwierigkeiten bei der Unterscheidung zwischen ähnlichen Objekten, insbesondere wenn sie sich in Form oder Größe ähneln. Es ist schwierig, alle subtilen Unterschiede zu erfassen, die für eine präzise Segmentierung notwendig sind. Eine Vielzahl aktueller Entwicklungen beschäftigt sich mit der Verbesserung der Algorithmen. Dabei gelten besonders Deep-learning Methoden und Convolutional Neural Networks (CNNs) als vielversprechende Ansätze, um die Komplexität der Umgebung besser zu erfassen [3, 4].

2 Sensoren zur Erfassung von 3D-Daten

2.1 LiDAR-Sensoren

LiDAR-Sensoren, die auch unter dem Namen Light Detection and Ranging-Sensoren bekannt sind, stellen eine weit verbreitete Technologie zur Erfassung von 3D-Daten dar. Sie basieren auf dem Einsatz von Laserstrahlen, welche ausgesendet werden und von Objekten in der Umgebung reflektiert werden. Dabei kann zwischen Time of Flight (TOF) LiDAR und phasenbasiertem LiDAR unterschieden werden. Während TOF-LiDAR die Distanz über eine Messung der Laufzeit der lichtwelle bestimmt, erfolgt die Entfernungsmessung beim phasenbasierten LiDAR über die Auswertung der Phasenverschiebung der vom Objekt reflektierten Lichtwelle. Hierdurch können LiDAR-Sensoren hochgenaue Entfernungen zu den reflektierenden Objekten erfassen aus denen sich detaillierte 3D-Punktwolken erzeugen lassen, welche die Geometrie und räumliche Verteilung von Objekten in der Umgebung darstellen. Zusätzlich lassen sich LiDAR-Sensoren in Scanning-LiDAR und Non-Scanning-LiDAR untergliedern. Non-Scanning-LiDAR nutzt dabei einen statischen Laserstrahl, während Scanning-LiDAR einen sich bewegenden Laserstrahl nutzt. [5]

2.2 Tiefenkameras

Tiefenkameras basieren auf verschiedene Verfahren, um Entfernungen zu messen. Im Bereich der semantische Segmentierung kommen besonders Kamerasysteme, die auf Stereo-Vision, Time-of-Flight oder Structured Light basieren zum Einsatz. Kamerasysteme, die auf dem Prinzip der Stereo-Vision basieren, werden als Stereo Kameras bezeichnet. Bei diesen werden zwei räumlich getrennte Kameras verwendet, die gemeinsam Bilder von derselben Szene aus zwei leicht unterschiedlichen Perspektiven aufnehmen. Der dabei entstehende horizontale Versatz der beiden

Bilder wird als Disparität bezeichnet. Aus diesem lassen sich Tiefeninformationen des betrachteten Objektes berechnen [6]. Bei Time of Flight Kameras wird ein moduliertes Lichtsignal im Infrarotbereich ausgesendet und von Objekten in der Umgebung reflektiert. Über die Phasenverschiebung der Infrarotwelle lässt sich die Entfernung des Objektes zur Kamera berechnen [7]. 3D-Kameras auf Basis von Structured Light projizieren ein spezielles 2D-Muster auf das zu betrachtende Objekt. Aus der Verzerrung dessen, lassen sich Tiefeninformationen berechnen [8]. Die meisten Tiefenkameras stellen dabei die Tiefeninformationen in einem Bild aus Graustufen dar. Zusätzlich gibt es auch RGB-D Kameras, welche zusätzlich zu einer Structured Light oder TOF-Kamera über eine RGB-Kamera verfügen. Diese haben häufig eine höhere räumliche Auflösung und sind in der Lage zusätzlich Farbinformationen aufzunehmen, besitzen jedoch einen deutlich kleineren Arbeitsbereich [9].

2.3 Passive und aktive Sensoren

Grundsätzlich lassen sich Sensoren zur Gewinnnung von 3D-Daten in zwei Klassen unterscheiden. Aktive Sensoren wie LIDAR-Sensoren senden selbst Energie in Form von Laser- oder Lichtwellen aus, um Informationen über das Objekt zu sammeln. Die reflektierten Signale werden von der Sensor-Einheit aufgenommen und zur Berechnung von Tiefeninformationen verwendet. Im Gegensatz dazu erfordern passive Sensoren wie Stereokameras keine aktive Energiequelle, sondern nutzen das natürliche Licht, das von der Umgebung reflektiert wird. Der Vorteil von aktiven Sensoren besteht darin, dass sie unabhängig von der Umgebungshelligkeit arbeiten und auch bei Dunkelheit eingesetzt werden können. Passive Sensoren hingegen können bei schlechten Lichtverhältnissen Schwierigkeiten haben, genaue Tiefeninformationen zu liefern. Es ist jedoch anzumerken, dass passive Sensoren in der Regel kostengünstiger sind und eine höhere räumliche Auflösung bieten können.

2.4 Auswahl von Sensoren für die semantische Segmentierung

ie Wahl der geeigneten Sensoren für die semantische Segmentierung hängt von verschiedenen Faktoren ab, wie den Anforderungen der Anwendung, den Umgebungsbedingungen, dem Budget und den gewünschten Ergebnissen. Aspekte wie die benötigte Genauigkeit, räumliche Auflösung,

Reichweite, Echtzeitfähigkeit und Umgebungsbedingungen sollten bei der Auswahl von Sensoren berücksichtigt werden. Zum Beispiel benötigen Anwendungen im Bereich der autonomen Fahrzeuge möglicherweise Sensoren mit hoher Reichweite und Genauigkeit, während Anwendungen im Innenbereich möglicherweise Sensoren mit höherer räumlicher Auflösung und Echtzeitfähigkeit benötigen. Die Umgebungsbedingungen, wie schlechte Beleuchtungsbedingungen oder komplexe Geometrien, können ebenfalls die Leistung von Sensoren beeinflussen und die Wahl von geeigneten Sensoren beeinflussen. Das Budget ist ebenfalls ein wichtiger Faktor bei der Auswahl von Sensoren, da verschiedene Sensoren unterschiedliche Kosten haben können. Schließlich sollten auch die gewünschten Ergebnisse der semantischen Segmentierung berücksichtigt werden, da verschiedene Sensoren besser geeignet sein können, um bestimmte Objekte oder Strukturen in der Umgebung zu segmentieren. Zum Beispiel können Lidar-Sensoren aufgrund ihrer präzisen Tiefeninformationen und Reichweite gut geeignet sein, um Objekte wie Straßen, Gebäude oder Bäume zu segmentieren, während Kameras oder Tiefenkameras besser für die Segmentierung von Fußgängern oder Fahrzeugen geeignet sein können.

3 Datengrundlage und Vorverarbeitung

3.1 3D-Datenformate und Datentypen

Bei der semantischen Segmentierung von 3D-Daten spielen die zugrunde liegenden 3D-Datenformate eine elementare Rolle. Diese bilden die Grundlage für die Erfassung, Speicherung und Verarbeitung von 3D-Daten, die für die semantische Segmentierung verwendet werden. In diesem Kapitel werden die beiden verbreitetsten 3D-Datenformate und Datentypen untersucht, die in der Forschung und Praxis eingesetz werden.

Ein wichtiges 3D-Datenformat ist das Punktewolkenformat, das häufig von LiDAR-Sensoren erzeugt wird. Punktewolken sind Sammlungen von 3D-Punkten, die die Oberfläche von Objekten in der Umgebung darstellen. Sie können in verschiedenen Dateiformaten gespeichert werden, wie beispielsweise dem ASCII-Format oder dem binären LAS-Format (LASer File Format), das speziell für LiDAR-Daten entwickelt wurde. Diese Formate ermöglichen die Speicherung von großen Mengen an Punkten mit 3D-Koordinaten, Intensitätsinformationen und weiteren Attributen, die zur semantischen Segmentierung verwendet werden können. Im Vergleich zu Bildern weißen Punktwolken eine stark variable Punktdichte auf. Dies lässt sich durch Faktoren, wie eine ungleichmäßige Abtastung des Raumes, der Verdeckung von Objekten und der relativen Ausrichtung des Objektes zum Sensor begründen [10]. Das Punktwolkenformat erzeugt dabei große Datenmengen, was zu einer rechenintensiven Verarbeitung der Daten führen kann.

Neben Punktewolken werden auch 3D-Gitter oder Voxel-Daten oft für die semantische Segmentierung verwendet. Voxel sind volumetrische Elemente, die den Raum in einem dreidimensionalen Gitter unterteilen. Jedes Voxel enthält dabei Informationen über Materialeigenschaften, Farbe oder Textur des Objektes an diesem Punkt. Voxel-Daten können in verschiedenen Formaten gespeichert werden, wie zum Beispiel das binäre OctoMap-Format oder dem ASCII-Format. Sie bieten dabei eine effektive Möglichkeit, komplexe dreidimensionale Strukturen darzustellen und

weiter zu analysieren. Die Voxeldichte im Raum kann dabei frei gewählt werden und bestimmt so die Auflösung und den Speicherbedarf des Datensatzes.

3.2 Vorverarbeitungsschritte wie Filterung, Normalenberechnung und Downsampling

Die Vorverarbeitung von 3D-Daten ist ein wichtiger Schritt in der semantischen Segmentierung, um die Qualität und Genauigkeit der Segmentierungsergebnisse zu verbessern. In diesem Kapitel werden verschiedene Vorverarbeitungsschritte wie Filterung, Normalenberechnung und Downsampling betrachtet, die oft in der Praxis angewendet werden.

Zu Beginn wird häufig eine Filterung des Datensatzes durchgeführt. Dadurch wird unerwünschtes Rauschen oder Ausreißer im Datensatz unterdrückt und dessen Qualität verbessert. Dies kann durch verschiedene Filtertechniken erfolgen, wie zum Beispiel durch Medianfilter, Gaußsche Filter [11] oder region growing und Bilateralfilter [12]. Diese Filter können angewendet werden, um die Qualität des Datensatzes und somit auch das Ergebnis der Semantischen Segmentierung zu verbessern.

Normalenberechnung: Die Berechnung von Normalen ist ein optionaler Schritt, um die geometrische Information der 3D-Daten zu erfassen. Dabei wird für jeden Punkt oder Voxel des Datensatzes ein Normalenvektor bestimmt. Normalen sind Vektoren, die senkrecht zur Oberfläche von Objekten in der Umgebung stehen und gibt so Aufschluss über die Orientierung der Oberflächen der Szene. SCHÄTZVERAHReN

Downsampling: Das Downsampling von 3D-Daten beschreibt die Reduzierung der Datenpunkte eines Datensatzes, um die Verarbeitungsgeschwindigkeit und den Speicherbedarf zu reduzieren. Dies kann durch verschiedene Techniken erfolgen. Eine sehr schnelle und recheneffiziente Methode ist hierbei das Random-Downsampling. Dabei wird eine absolute oder relative Punktanzahl eines Datensatzes aus diesem entfernt. Die entfernten Punkte werden dabei zufällig ausgewählt. Nachteil ist hierbei, dass sich ursprünglich ungleichmäßig abgetastete Bereich des Datensatzes noch vergrößern können, und so Lücken in den Punktwolken entstehen können. Deshalb wird hierfür häufig auf das Voxel-Grid-Downsampling zurückgegriffen. Dabei wird der Raum, in der sich die Punktwolke befindet, in gleichmäßige Voxel unterteilt, wobei für jedes Voxel der Schwerpunkt der darin enthaltenen Punkte berechnet wird. Danach wird eine bestimmte

Anzahl an Punkten, häufig durch ihre nähe zum Schwerpunkt, ausgewählt und in das Voxel-Grid übernommen. Ziel des Downsamplings ist es, die Datenmenge zu reduzieren, während wichtige strukturelle und semantische Informationen erhalten bleiben.

3.3 Datenannotation und Ground Truth-Erstellung

Das Erstellen annotierter Daten und eines Ground Truths sind entscheidende Schritte bei der semantischen Segmentierung von 3D-Daten. Um neuronale Netze für die Semantische Segmentierung zu trainieren, werden annotierte Datensätze benötigt. Hierbei muss für einen Teil der Rohdaten eine manuelle Klassifikation der Bilder erfolgen, indem semantische Labels oder Klasseninformationen manuell oder automatisch den 3D-Daten zugeordnet werden. Die Qualität und Genauigkeit der Datenannotation sind entscheidend für die Leistungsfähigkeit von semantischen Segmentierungsalgorithmen. Die Erstellung des Ground Truths umfasst die Erstellung von referenzbasierten Segmentierungsergebnissen, die als Grundlage für das Training und die Evaluation von Segmentierungsmodellen dienen. Die Ground Truth kann manuell oder automatisch erstellt werden, um die Zuverlässigkeit und Vergleichbarkeit von Segmentierungsergebnissen sicherzustellen und die Qualität von trainierten Modellen zu überprüfen. Die Verwendung von Ground Truths ist jedoch nicht nur auf das Training beschränkt, sondern kann auch für die Validierung und Bewertung von Modellen verwendet werden.

4 Grundlegende Verfahren der semantischen Segmentierung

Ziel der Semantischen Segmentierung von 3D-Daten ist es, jedem Punkt im Raum einer bestimmten Kategorie zuzuordnen und dadurch Bereiche des Bildes in klassifizierte Objekte zu unterteilen. Hierfür gibt es verschiedene Ansätze, die auf erweiterten neuronalen Netzen basieren.

4.0.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) sind eine Art von künstlichen neuronalen Netzen, die speziell für die Verarbeitung von Bildern entwickelt wurden. Sie bestehen aus mehreren Schichten, darunter Convolutional Layers, Pooling Layers und Fully Connected Layers, die miteinander verbunden sind. Die Architektur der CNNs ist dabei nicht vorgegeben, folgt aber in der Praxis immer einer ähnlichen Vorlage. Ein Input-Layer beinhaltet die Pixelwerte des Bildes. Darauf folgen in der Regel ein oder zwei Convolutional-Layers, woraufhin sich ein Pooling-Layer anreiht. Die Kombination aus Convolutional- und Pooling-Layer kann dabei je nach Komplexität beliebig oft im Netz vorkommen. Am Ende folgt ein Fully-Connected-Layer an den der Output anknüpft.[GRAFIK?] Die Convolutional Layers können vereinfacht als Filter-Layer betrachtet werden und extrahieren Merkmale aus den Eingabebildern. Dabei wird jeder Bereich des Bildes mit Kernels (Filtern) gefalten und erzeugen eine 2D-Aktivierungskarte. Die kernels besitzen dabei häufig kleine räumliche Dimensionalitäten, erstrecken sich aber über die Gesamte Tiefe des Eingangsbildes. Pooling Layers dienen dazu, die Dimensionen der Ausgabe des Convolutional-Layers zu reduzieren und somit die Rechnenkomplexität des Modells zu verringern. Die Fully Connected Layers am Ende des Netzes verarbeiten schließlich die extrahierten Aktivierungen und versuchen daraus Klassifizierungsergebnisse zu gewinnen. Der Hauptanwendungsbereich von CNNs liegt

dabei in der Klassifikation von Bildern in vorbestimmte Kategorien. [13].

4.0.2 Fully Convolutional Networks (FCNs)

Fully Convolutional Networks (FCNs) sind eine Weiterentwicklung von Convolutional Neural Networks (CNNs), die speziell für die Aufgabe der semantischen Segmentierung von Bildern entwickelt wurden. Im Gegensatz zu herkömmlichen CNNs, die für die Klassifizierung und Erkennung von Objekten in Bildern ausgelegt sind, können FCNs jedes Pixel eines Eingabebildes klassifizieren und somit die räumliche Information beibehalten. FCNs verwenden dabei ausschließlich Convolutional-Layers und Pooling-Layers, nicht aber Fully-Connected-Layers. Dies ermöglicht es eine Merkmalskarte des Eingabebildes zu erzeugen, auf der jedes Pixel einer bestimmten Klasse zugeordnet wird. Durch die Verwendung von FCNs können somit komplizierte Zusammenhänge innerhalb von Bildern auf der Ebene der Pixel identifiziert werden, was für Anwendungen wie die autonome Navigation oder Objekterkennung von großer Bedeutung ist. [14] Bekannte und erfolgreiche Verfahren wie das U-Net, verwenden dabei eine Encode-Decoder-Architektur, um das Klassifizierungsergebnis in der Dimensionalität des Eingangsbildes darzustellen.

4.0.3 Encoder-Decoder-Architekturen

Encoder-Decoder-Architekturen stellen eine spezielle Architektur von neuronalen Netzen dar. Der Name leitet sich aus deren Aufbau ab, welcher aus zwei Hauptkomponenten, einem Encoder und einem Decoder, besteht. Der Encoder verwendet typischerweise CNNs, um das Eingabebild schrittweise in eine kompakte, abstrakte Repräsentation zu komprimieren, die die Merkmale des Bildes stark reduziert enthält. Dabei werden die Positionen der maximalen Aktivierungen der Ebene während des Max-Pooling Prozesses gespeichert. Die gespeicherten Daten werden als Pooling-Indizes bezeichnet. Der Decoder verwendet häufig Deconvolutional-Neuronale-Netzwerke, um das Ergebnis des Encoder-Netzwerkes wieder in die Dimensionalitäten des Eingangsbildes zurückzuführen. Das Upsampling erfolgt dabei nicht linear, sondern verwendet die Pooling-Indizes des zugehörigen Encoding-Schrittes, um die Aktivierungen an der richtigen Position des Bildes wiederherzustellen.

Die Semantische Segmentierung erfolgt dabei am Ende des Decoding Netzwerkes. [14] SOFT-

Abbildung 4.1: Beschreibung des Bildes

MAX

4.0.4 Conditional Random Fields (CRFs)

Conditional Random Fields (CRFs) sind probabilistische Modelle, die zur Modellierung von sequentiellen Daten eingesetzt werden. Im Gegensatz zu Markov Random Fields (MRFs) ermöglichen CRFs eine Modellierung von Abhängigkeiten zwischen den Ausgaben der einzelnen Knoten, um somit ein besseres Ergebnis bei der Inferenz zu erzielen. In einem CRF-Modell wird jeder Knoten durch eine Funktion repräsentiert, die seine Zustände modelliert. Die Funktionen können sowohl globale Merkmale der Daten als auch lokale Merkmale des Knoten und seiner Nachbarn berücksichtigen. CRFs zielen darauf ab, die bedingte Wahrscheinlichkeit einer Ausgabe für eine gegebene Eingabe zu modellieren, indem sie die Abhängigkeiten zwischen den Zuständen der Knoten im Modell berücksichtigen. In der Praxis werden CRFs oft in Kombination mit CNNs eingesetzt, um semantische Segmentierungsaufgaben auf Bildern durchzuführen. Dabei kann das CNN zur Extraktion von Merkmalen und das CRF zur Modellierung von Abhängigkeiten zwischen den Ausgaben der Knoten eingesetzt werden.

4.0.5 Region-based Convolutional Neural Networks (R-CNNs)

Region-based Convolutional Neural Networks (R-CNNs) sind eine Weiterentwicklung von Convolutional Neural Networks, die speziell für die Aufgabe der Objekterkennung in Bildern entwickelt wurden. Im Gegensatz zu herkömmlichen CNNs, die eine feste Größe der Eingabebilder erfordern, verwenden R-CNNs eine Region Proposal Technik, um Regions of Interest (ROI) innerhalb des Bildes zu detektieren. Anschließend wird nur auf die ausgewählten Bereiche ein FCN angewendet, um diese auf Pixelebene zu klassifizieren. R-CNNs erzielen eine höhere Genauigkeit als herkömmliche CNNs bei der Erkennung von Objekten in Bildern und werden daher häufig in der Robotik und im autonomen Fahren eingesetzt. Dies liegt daran, dass sie weniger Rechenleistung benötigen und gleichzeitig Störeinflüsse, die außerhalb der ROI liegen, keinen Einfluss auf die Klassifizierung nehmen können. [15]

4.1 Evaluierung von Verfahren zur semantischen Segmentierung

Die Evaluierung von Verfahren zur semantischen Segmentierung erfolgt in der Regel anhand von Metriken wie der İntersection over Union"(IoU), auch "Jaccard Index"genannt. Dieser Wert gibt an, wie viel Prozent der vorhergesagten Pixel tatsächlich richtig klassifiziert wurden im Verhältnis zu den tatsächlich vorhandenen Pixeln. Weitere Metriken sind die "Pixelgenauigkeit"(Pixel Accuracy), die "Klassen-Genauigkeit"(Class Accuracy) und die "Mittlere-Klassen-Genauigkeit"(Mean Class Accuracy). Für die Evaluierung wird in der Regel ein Testdatensatz verwendet, der sowohl Bilder als auch Ground-Truth-Masken enthält. Anhand dieser Daten wird das Verfahren trainiert und anschließend auf dem Testdatensatz ausgewertet. Die Bewertung der Ergebnisse ermöglicht die Beurteilung der Leistung des Verfahrens und die Vergleichbarkeit mit anderen Ansätzen.

5 Anwendungszenarien der semantischen Segmentierung

5.1 Autonomes Fahren

Im Bereich der autonomen Fahrzeuge spielt die semantische Segmentierung eine entscheidende Rolle, um eine präzise und zuverlässige Wahrnehmung der Umgebung zu ermöglichen. Semantische Segmentierung bezieht sich auf die Fähigkeit, ein Echtzeitbild in verschiedene Klassen oder Kategorien zu segmentieren, wobei jedem Pixel eine bestimmte Bedeutung zugewiesen wird.

Die semantische Segmentierung ermöglicht es autonomen Fahrzeugen, ihre Umgebung genau zu analysieren und wichtige Informationen über Straßenverhältnisse, Verkehrszeichen, Fußgänger, Fahrzeuge und andere Hindernisse zu extrahieren. Durch die präzise Klassifizierung jedes Pixels im Bild kann das Fahrzeug Hindernisse erkennen und entsprechend darauf reagieren, indem es beispielsweise seine Geschwindigkeit anpasst oder Hindernisse umfährt.

Ein entscheidender Vorteil der semantischen Segmentierung besteht darin, dass sie eine detailliertere und kontextbezogene Wahrnehmung der Umgebung ermöglicht. Durch die genaue Zuordnung von Klassen zu den erkannten Objekten kann das Fahrzeug komplexe Szenarien besser verstehen und angemessene Entscheidungen treffen. Zum Beispiel kann es zwischen verschiedenen Arten von Fahrzeugen unterscheiden und Prioritäten entsprechend den Verkehrsregeln setzen.

5.2 Robotik in der Industrie

Die semantische Segmentierung spielt auch im Bereich der Industrierobotik eine bedeutende Rolle. Industrieroboter werden häufig in anspruchsvollen Produktionsumgebungen eingesetzt, in

denen eine präzise Wahrnehmung und Interpretation der Umgebung von entscheidender Bedeutung ist. Durch die semantische Segmentierung können Roboter die visuelle Erfassung von Objekten verbessern und deren Kategorisierung ermöglichen.

Die semantische Segmentierung ermöglicht es Industrierobotern, einzelne Objekte oder Regionen in einem Bild oder einer Szene zu identifizieren und zu isolieren. Dies ist besonders wichtig, wenn es darum geht, spezifische Objekte oder Teile in einer komplexen Umgebung zu erkennen und zu handhaben. Durch die präzise Segmentierung von Objekten können Roboter zielgerichtete und genaue Manipulationen durchführen, ohne andere Objekte oder die Umgebung zu beeinträchtigen.

Ein weiterer Vorteil der semantischen Segmentierung in der Industrierobotik besteht darin, dass sie die Roboter dabei unterstützt, die Absicht oder den Zustand von Objekten zu verstehen. Durch die Zuweisung semantischer Labels zu den erkannten Objekten kann der Roboter beispielsweise zwischen verschiedenen Arten von Produkten oder Materialien unterscheiden und entsprechend darauf reagieren. Dies ermöglicht eine adaptive und flexible Arbeitsweise, bei der der Roboter je nach Aufgabe oder Anforderung unterschiedliche Aktionen ausführen kann.

5.3 Augmented Reality

Die semantische Segmentierung spielt auch im Bereich der Augmented Reality (AR) eine wesentliche Rolle. AR-Anwendungen integrieren virtuelle Inhalte nahtlos in die reale Umgebung und erfordern eine präzise Wahrnehmung und Unterscheidung der physischen Welt. Die semantische Segmentierung ermöglicht es AR-Systemen, die Umgebung zu analysieren und virtuelle Inhalte entsprechend zu platzieren und zu interagieren.

Durch die semantische Segmentierung kann die AR-Anwendung die Szene in Echtzeit analysieren und verschiedene Objekte oder Regionen identifizieren. Dies ermöglicht eine präzise Verankerung von virtuellen Objekten an bestimmten Stellen in der realen Welt. Beispielsweise kann eine AR-Anwendung mit semantischer Segmentierung den Boden, Wände oder bestimmte Möbelstücke erkennen und virtuelle Objekte wie Möbel, Dekorationen oder Spielinhalte darauf platzieren. Dies schafft eine immersive Erfahrung und ermöglicht den Benutzern, virtuelle Inhalte nahtlos in ihre Umgebung einzubinden.

Ein weiterer Vorteil der semantischen Segmentierung in der AR liegt darin, dass sie die Inter-

aktion zwischen virtuellen und realen Objekten erleichtert. Durch die genaue Segmentierung von Objekten können AR-Anwendungen die virtuellen Inhalte auf bestimmte Bereiche oder Oberflächen beschränken. Dies ermöglicht eine präzise Kollisionserkennung und Interaktion zwischen virtuellen und realen Objekten. Beispielsweise kann eine AR-Anwendung mit semantischer Segmentierung verhindern, dass virtuelle Objekte durch physische Hindernisse hindurchgehen oder mit anderen Objekten in der Umgebung kollidieren.

5.4 Landwirtschaft

Die semantische Segmentierung spielt auch im Bereich der Landwirtschaft eine bedeutende Rolle. In der modernen Landwirtschaft werden fortschrittliche Technologien eingesetzt, um die Effizienz, Produktivität und Nachhaltigkeit zu verbessern. Die semantische Segmentierung ermöglicht es, landwirtschaftliche Flächen und Pflanzen präzise zu analysieren und spezifische Informationen zu extrahieren.

Durch die semantische Segmentierung können Landwirte und Agrarfachleute die Vegetation und den Zustand der Pflanzen genau erfassen. Mithilfe von Drohnen oder anderen Bildgebungssystemen kann die semantische Segmentierung verschiedene Klassen von Pflanzen, Unkräutern, Bodenarten und anderen landwirtschaftlich relevanten Merkmalen identifizieren. Dies ermöglicht eine detaillierte Kartierung und Überwachung von landwirtschaftlichen Flächen, um gezielte Maßnahmen wie Bewässerung, Düngung oder Unkrautbekämpfung durchzuführen.

Ein entscheidender Vorteil der semantischen Segmentierung in der Landwirtschaft besteht darin, dass sie es ermöglicht, gezielte Entscheidungen zu treffen und Ressourcen effizienter einzusetzen. Durch die genaue Identifizierung von Pflanzenarten und Unkräutern können Landwirte gezielt Pestizide und Herbizide einsetzen, um den Einsatz chemischer Substanzen zu minimieren und die Umweltbelastung zu reduzieren. Darüber hinaus ermöglicht die semantische Segmentierung eine gezielte Bewässerung und Düngung, um den Bedürfnissen der Pflanzen optimal gerecht zu werden und den Wasserverbrauch zu optimieren.

5.5 Medizin

Die semantische Segmentierung spielt eine entscheidende Rolle im Bereich der Medizin und trägt dazu bei, die Diagnose, Behandlung und Forschung zu verbessern. Durch die präzise Analyse und Klassifizierung von medizinischen Bildern ermöglicht die semantische Segmentierung eine detaillierte Erfassung von Geweben, Organen oder Läsionen.

In der medizinischen Bildgebung, wie z. B. CT- oder MRT-Scans, kann die semantische Segmentierung verwendet werden, um verschiedene anatomische Strukturen oder pathologische Bereiche zu identifizieren und zu segmentieren. Dies ermöglicht eine genaue Visualisierung und quantitative Analyse von Organen oder Geweben, um Veränderungen oder Anomalien zu erkennen. Zum Beispiel kann die semantische Segmentierung in der Onkologie dabei helfen, Tumore oder Metastasen zu lokalisieren und ihre Ausdehnung zu bestimmen.

Ein weiterer Bereich, in dem die semantische Segmentierung in der Medizin von großer Bedeutung ist, betrifft die Bildanalyse in der Pathologie. Durch die Segmentierung von Zellen oder Geweben können Pathologen präzise diagnostische Informationen gewinnen und Krankheiten identifizieren. Dies erleichtert die Untersuchung von Proben und die Erkennung von Krankheiten wie Krebs oder anderen pathologischen Zuständen.

Darüber hinaus trägt die semantische Segmentierung zur Entwicklung und Verbesserung von medizinischen Bildgebungsverfahren und bildbasierten Interventionen bei. Sie ermöglicht die präzise Navigation und Ausrichtung von Instrumenten oder Implantaten während chirurgischer Eingriffe. Durch die genaue Segmentierung von anatomischen Strukturen können Chirurgen die präzise Positionierung von Implantaten sicherstellen und komplexe Eingriffe durchführen.

6 State-of-the-Art Verfahren zur semantischen Segmentierung von 3D-Daten

6.0.1 PointNet

PointNet++ ist eine fortschrittliche Methode zur Verarbeitung von Punktwolken in der 3D-Bildverarbeitung, die speziell für die Aufgaben der Klassifikation, Segmentierung und Erkennung von Objekten entwickelt wurde. Es basiert auf der PointNet-Architektur und erweitert sie durch die Integration einer hierarchischen Struktur, um lokale und globale Kontextinformationen in Punktwolken zu erfassen.

Die Funktionsweise von PointNet++ besteht aus mehreren aufeinanderfolgenden Schritten. Zunächst wird die ursprüngliche Punktwolke in eine hierarchische Struktur unterteilt. Dies wird erreicht, indem die Punktwolke in immer kleinere Unterteilungen aufgeteilt wird, wobei auf jeder Hierarchieebene ein separater PointNet-Block angewendet wird.

Auf jeder Hierarchieebene führt der PointNet-Block zwei grundlegende Operationen durch: Das Pooling und die Verarbeitung von Punktwolken. Das Pooling dient dazu, die relevanten Merkmale aus den Punkten auf jeder Hierarchieebene zu extrahieren. Dabei werden aggregierte Merkmale auf einer höheren Ebene erzeugt, um den globalen Kontext der Punktwolke zu erfassen.

Die Verarbeitung von Punktwolken auf jeder Hierarchieebene beinhaltet die Anwendung von PointNet auf die Punkte innerhalb der Teilbereiche der Punktwolke. Dies ermöglicht die Extraktion von lokalen Merkmalen, die spezifisch für bestimmte Regionen oder Strukturen in der Punktwolke sind.

Die hierarchische Struktur von PointNet++ ermöglicht es, sowohl lokale als auch globale Kontextinformationen zu erfassen und zu integrieren. Durch die Verwendung mehrerer PointNet-

Blöcke auf verschiedenen Hierarchieebenen können feinere Details auf lokaler Ebene berücksichtigt werden, während gleichzeitig der globale Kontext der Punktwolke erhalten bleibt.

Die Ausgabe von PointNet++ ist eine repräsentative Merkmalsdarstellung, die die Informationen über die semantische Struktur der Punktwolke enthält. Diese Merkmale können dann für verschiedene Aufgaben wie Klassifikation, Segmentierung oder Erkennung von Objekten verwendet werden.

Insgesamt ermöglicht die Funktionsweise von PointNet++ die effektive Verarbeitung von Punktwolken und die Erfassung von lokalen und globalen Kontextinformationen. Es hat sich als leistungsstarkes Verfahren erwiesen, um komplexe Strukturen und Muster in 3D-Daten zu erfassen und die Genauigkeit und Zuverlässigkeit bei der Klassifikation und Segmentierung von Punktwolken zu verbessern.

6.0.2 3D U-Net

3D U-Net ist ein leistungsstarkes Verfahren für die semantische Segmentierung von 3D-Daten. Es basiert auf der beliebten 2D U-Net-Architektur, die für die Bildsegmentierung entwickelt wurde, und wurde speziell für die Verarbeitung von Volumendaten angepasst.

Die Funktionsweise von 3D U-Net kann in zwei Hauptphasen unterteilt werden: den Encoder-Teil und den Decoder-Teil.

Im Encoder-Teil werden die Eingabedaten schrittweise abwärts durch das Netzwerk geleitet, um Merkmale auf verschiedenen Abstraktionsebenen zu extrahieren. Dies erfolgt durch eine Kombination aus 3D-Convolutional-Layern und Pooling-Operationen, die die räumliche Auflösung der Daten reduzieren. Dabei werden lokalisierte Merkmale erfasst, um wichtige Informationen über die Struktur und den Kontext der 3D-Daten zu gewinnen.

Der Decoder-Teil arbeitet in umgekehrter Richtung und verwendet Upsampling-Operationen, um die räumliche Auflösung schrittweise zu erhöhen. Dabei werden die Merkmale aus dem Encoder-Teil mit den entsprechenden Merkmalen auf höheren Auflösungsebenen kombiniert, um detailliertere und präzisere Segmentierungsergebnisse zu erzielen. Dieser Prozess wird durch sogenannte SSkip Connectionsërmöglicht, die es dem Decoder ermöglichen, sowohl lokale als auch globale Informationen zu nutzen.

Während des Trainings wird das 3D U-Net-Modell mit gelabelten Trainingsdaten trainiert,

um die Gewichtungen der Netzwerkparameter zu optimieren. Dies geschieht durch den Vergleich der vorhergesagten Segmentierungsergebnisse mit den tatsächlichen Labels. Durch den Einsatz von Verlustfunktionen wie der Kreuzentropie wird das Modell kontinuierlich verbessert und kann genaue semantische Segmentierungen von neuen, nicht-gelabelten Daten vorhersagen.

Die Funktionsweise von 3D U-Net hat sich in verschiedenen Anwendungen bewährt, darunter medizinische Bildgebung, Robotik, geografische Kartierung und mehr. Es ermöglicht eine präzise Segmentierung von 3D-Daten und eröffnet Möglichkeiten zur Analyse und Interpretation komplexer Strukturen und Merkmale in volumetrischen Daten.

6.0.3 OctNet

OctNet ist ein leistungsstarkes Verfahren zur Verarbeitung und Segmentierung von 3D-Daten, das auf der Verwendung von Octrees basiert. Octrees sind eine hierarchische Datenstruktur, die es ermöglicht, den Raum in kleine Voxelblöcke zu unterteilen und gleichzeitig eine adaptive Auflösung bereitzustellen.

Die Funktionsweise von OctNet beruht auf der effizienten Organisation und Verarbeitung von 3D-Daten. Zunächst wird die 3D-Datenrepräsentation in ein Octree umgewandelt, wobei jeder innere Knoten des Baums acht Unterknoten hat, die den Raum in kleinere Voxelblöcke unterteilen. Dadurch können Bereiche mit höherer Detailgenauigkeit mehr Unterknoten haben, während weniger detaillierte Bereiche weniger Unterknoten aufweisen.

Das OctNet-Modell besteht aus einer Kombination von Convolutional Neural Networks (CNNs) und speziellen Operationen, die auf die Struktur des Octrees abgestimmt sind. Durch die Verwendung von Convolutional-Operationen auf den Octree-Daten werden Merkmale auf verschiedenen Auflösungsebenen erfasst und repräsentiert. Dies ermöglicht eine effiziente und adaptive Verarbeitung von 3D-Daten, da die Operationen nur auf den relevanten Voxelblöcken des Octrees durchgeführt werden.

Während des Trainings wird das OctNet-Modell mit gelabelten Trainingsdaten trainiert, um die Gewichtungen der Netzwerkparameter zu optimieren. Dabei wird die Beziehung zwischen den Eingabedaten und den entsprechenden Segmentierungskarten erlernt. Durch die Anpassung der Gewichtungen können genaue Segmentierungsergebnisse erzielt werden, die die semantischen Strukturen in den 3D-Daten korrekt erfassen.

Die Funktionsweise von OctNet ermöglicht eine effiziente Verarbeitung und Segmentierung von 3D-Daten mit variabler Auflösung. Es hat sich in verschiedenen Anwendungen wie der Segmentierung von Punktwolken, der Analyse von 3D-Modellen und der medizinischen Bildgebung als wirksam erwiesen. Durch die Verwendung von Octrees bietet OctNet eine leistungsstarke Methode zur Erfassung und Analyse von komplexen Strukturen in 3D-Daten.

7 Herausforderungen und zukünftige Entwicklungen

7.1 Herausforderungen und Limitationen

Die semantische Segmentierung ist eine komplexe Aufgabe mit verschiedenen Herausforderungen und Limitationen, die ihre Anwendung beeinflussen können. Im Folgenden werden einige dieser Herausforderungen und Limitationen erläutert.

Eine der Hauptherausforderungen der semantischen Segmentierung liegt in der Verfügbarkeit hochwertiger annotierter Datensätze. Um semantische Segmentierungsalgorithmen zu trainieren, sind große Mengen an Daten erforderlich, die präzise mit den entsprechenden semantischen Labels annotiert wurden. Das Erstellen solcher Datensätze erfordert oft umfangreiche manuelle Arbeit und Expertenwissen, was teuer und zeitaufwendig sein kann.

Ein weiteres Problem ist die Bewältigung von Klassenungleichgewichten. In vielen Szenarien sind bestimmte Objektklassen in den Bilddaten seltener vertreten als andere. Dies kann dazu führen, dass semantische Segmentierungsmodelle dazu neigen, häufigere Klassen besser zu erkennen und seltene Klassen zu vernachlässigen. Der Umgang mit Klassenungleichgewichten erfordert spezielle Strategien wie Gewichtungsschemata oder Datenanreicherungstechniken, um die Leistung der semantischen Segmentierungsalgorithmen für seltene Klassen zu verbessern.

Die Komplexität und Vielfalt von Objekten und Szenen stellen eine weitere Herausforderung dar. Objekte können unterschiedliche Formen, Größen, Texturen und Beleuchtungsbedingungen aufweisen. Zudem können komplexe Szenen eine Überlappung oder Verschmelzung von Objekten beinhalten, was die korrekte Segmentierung erschwert. Die semantische Segmentierung muss robust gegenüber solchen Variationen sein und genaue Ergebnisse liefern, unabhängig von den spezifischen Bedingungen.

Ein weiterer Aspekt, der berücksichtigt werden muss, ist die Effizienz und Echtzeitfähigkeit von semantischer Segmentierung. In vielen Anwendungen wie autonomem Fahren oder Echtzeitanalyse medizinischer Bilder ist es erforderlich, dass die semantische Segmentierung in Echtzeit erfolgt. Dies erfordert leistungsfähige Algorithmen und optimierte Implementierungen, um die Rechenleistung und den Speicherbedarf zu minimieren.

Schließlich gibt es auch Herausforderungen im Zusammenhang mit der Generalisierung und der Anpassung an neue Umgebungen. Semantische Segmentierungsalgorithmen werden oft auf bestimmte Datensätze oder Szenarien trainiert und können Schwierigkeiten haben, sich auf neue oder unerwartete Situationen anzupassen. Eine erfolgreiche Anwendung der semantischen Segmentierung erfordert daher die Fähigkeit, Modelle zu entwickeln, die robust und generalisierbar sind und in verschiedenen Umgebungen effektiv arbeiten können.

7.2 Potenziale und Trends für zukünftige Entwicklungen

Ein bedeutendes Potenzial liegt in der Verbesserung der Genauigkeit der semantischen Segmentierungsalgorithmen. Obwohl bereits beeindruckende Fortschritte erzielt wurden, besteht weiterhin Raum für Verbesserungen. Zukünftige Entwicklungen werden sich auf die Verfeinerung der Modellarchitekturen, die Optimierung der Datenpräparation und die Anwendung fortschrittlicher Optimierungstechniken konzentrieren. Durch die Steigerung der Genauigkeit können feinere Details in den Segmentierungsergebnissen erfasst werden, was zu einer noch präziseren Analyse von visuellen Szenen führt.

Ein weiteres Potenzial liegt in der Echtzeitsegmentierung. Die Fähigkeit, Bilder oder Videos in Echtzeit zu segmentieren, eröffnet neue Anwendungsmöglichkeiten in Bereichen wie Robotik, Überwachung und erweiterter Realität. Zukünftige Entwicklungen werden darauf abzielen, leistungsfähige Algorithmen zu entwickeln, die komplexe Szenen in Echtzeit analysieren und segmentieren können. Dies erfordert die Optimierung von Rechenleistung und Energieeffizienz, um Echtzeitsegmentierung auch auf eingebetteten Systemen oder mobilen Plattformen zu ermöglichen.

Ein weiteres Potenzial besteht in der Erweiterung der Anwendungsbereiche der semantischen Segmentierung über die bloße Objekterkennung hinaus. Bisher lag der Fokus hauptsächlich auf der Identifizierung bestimmter Objekte in Bildern. Zukünftige Entwicklungen könnten sich auf

die Entwicklung von Algorithmen konzentrieren, die eine robuste Erkennung und Segmentierung einer breiten Palette von Objekten ermöglichen, einschließlich solcher, die bisher nicht im Trainingssatz enthalten waren. Dies eröffnet Möglichkeiten für eine breitere Anwendung in Bereichen wie Umweltüberwachung, Landwirtschaft und Industrie.

Im Hinblick auf Trends wird der Einsatz von Deep Learning in der semantischen Segmentierung weiterhin eine bedeutende Rolle spielen. Deep-Learning-Modelle haben bereits große Fortschritte ermöglicht, da sie eine hohe Kapazität zur Mustererkennung und Modellierung komplexer Zusammenhänge bieten. Zukünftige Entwicklungen werden darauf abzielen, die Leistungsfähigkeit und Effizienz von Deep-Learning-Modellen weiter zu verbessern, um eine noch bessere Segmentierungsgenauigkeit zu erreichen.

Ein weiterer vielversprechender Trend liegt in der Integration der semantischen Segmentierung mit anderen Techniken wie Objekterkennung, Instanzsegmentierung und Tiefenwahrnehmung. Durch die Kombination von Informationen aus verschiedenen Quellen können integrierte Systeme geschaffen werden, die eine umfassendere und genauere Analyse von visuellen Szenen ermöglichen. Diese Kombinationstechniken haben das Potenzial, die semantische Segmentierung auf ein neues Niveau zu heben und den Einsatz in komplexen Anwendungen wie autonomen Fahrzeugen und erweiterter Realität zu verbessern.

Ein weiterer vielversprechender Trend besteht in der Nutzung von unüberwachtem Lernen für die semantische Segmentierung. Anstatt auf eine große Menge gelabelter Daten angewiesen zu sein, könnten Algorithmen entwickelt werden, die aus unbekannten Daten lernen und semantische Segmentierungsaufgaben ohne explizite Annotation durchführen können. Dies würde die Skalierbarkeit und Anwendbarkeit der semantischen Segmentierung erheblich verbessern, da der Aufwand für die Datenannotation entfällt.

8 Zusammenfassung und Anwendung

In dieser Arbeit wurden die Grundlagen der semantischen Segmentierung erläutert. Hierfür gibt es verschiedene Verfahren, die auf erweiterten neuronalen Netzen basieren. Zu diesen Verfahren gehören Convolutional Neural Networks (CNNs), Fully Convolutional Networks (FCNs), Region-based Convolutional Neural Networks (R-CNNs) und Encoder-Decoder-Architekturen. CNNs sind speziell für die Verarbeitung von Bildern konzipiert und werden für die Klassifikation von Bildern in vorbestimmte Kategorien eingesetzt. FCNs wurden speziell für die semantische Segmentierung von Bildern entwickelt und können die Pixel jedes Eingabebildes direkt klassifizieren, wodurch die räumliche Information beibehalten wird. R-CNNs wurden für die Objekterkennung in Bildern entwickelt und verwenden eine Region Proposal Technik, um Regions of Interest (ROI) innerhalb des Bildes zu detektieren. Encoder-Decoder-Architekturen bestehen aus einem Encoder, der das Eingabebild in eine kompakte, abstrakte Repräsentation umwandelt, und einem Decoder, der aus dieser Repräsentation eine Semantikkarte erzeugt.

Literatur

[1] Kaihong Yang, Sheng Bi und Min Dong. „Lightningnet: Fast and Accurate Semantic Segmentation for Autonomous Driving Based on 3D LIDAR Point Cloud“. In: *2020 IEEE International Conference on Multimedia and Expo (ICME)*. 2020, S. 1–6. DOI: 10.1109/ICME46284.2020.9102769.

[2] Mohammad Hosein Hamian u. a. „Semantic Segmentation of Autonomous Driving Images by the Combination of Deep Learning and Classical Segmentation“. In: *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*. 2021, S. 1–6. DOI: 10.1109/CSICC52343.2021.9420573.

[3] Tuan Pham. „Semantic Road Segmentation using Deep Learning“. In: *2020 Applying New Technology in Green Buildings (ATiGB)*. 2021, S. 45–48. DOI: 10.1109/ATiGB50996.2021.9423307.

[4] Liuhao Ge u. a. „3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images“. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, S. 5679–5688. DOI: 10.1109/CVPR.2017.602.

[5] Jingyun Liu u. a. „TOF Lidar Development in Autonomous Vehicle“. In: *2018 IEEE 3rd Optoelectronics Global Conference (OGC)*. 2018, S. 185–190. DOI: 10.1109/OGC.2018.8529992.

[6] Emre DANDIL und Kerim Kürşat ÇEVİK. „Computer Vision Based Distance Measurement System using Stereo Camera View“. In: *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. 2019, S. 1–4. DOI: 10.1109/ISMSIT.2019.8932817.

[7] Yosef Dalbah, Stephan Rohr und Friedrich M. Wahl. „Detection of dynamic objects for environment mapping by time-of-flight cameras“. In: *2014 IEEE International Conference on Image Processing (ICIP)*. 2014, S. 971–975. DOI: 10.1109/ICIP.2014.7025195.

[8] Inzamam Anwar und Sukhan Lee. „High performance stand-alone structured light 3D camera for smart manipulators“. In: *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. 2017, S. 192–195. DOI: 10.1109/URAI.2017.7992709.

[9] José Gomes da Silva Neto u. a. „Comparison of RGB-D sensors for 3D reconstruction“. In: *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*. 2020, S. 252–261. DOI: 10.1109/SVR51698.2020.00046.

[10] Yin Zhou und Oncel Tuzel. „VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection“. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, S. 4490–4499. DOI: 10.1109/CVPR.2018.00472.

[11] Karl Thurnhofer-Hemsi u. a. „Super-Resolution of 3D MRI Corrupted by Heavy Noise With the Median Filter Transform“. In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020, S. 3015–3019. DOI: 10.1109/ICIP40778.2020.9191237.

[12] Li Chen, Hui Lin und Shutao Li. „Depth image enhancement for Kinect using region growing and bilateral filter“. In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. 2012, S. 3070–3073.

[13] Keiron O'Shea und Ryan Nash. *An Introduction to Convolutional Neural Networks*. URL: https://arxiv.org/pdf/1511.08458.pdf.

[14] Vijay Badrinarayanan, Alex Kendall und Roberto Cipolla. „SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), S. 2481–2495. DOI: 10.1109/TPAMI.2016.2644615.

[15] Kaiming He u. a. „Mask R-CNN“. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, S. 2980–2988. DOI: 10.1109/ICCV.2017.322.