



ulm university universität
ulm

Fakultät für Ingenieurwissenschaften, Informatik und Psychologie
Institut für Mess-, Regel- und Mikrotechnik

Segmentierung von Punktwolken mit neuronalen Netzen

Bachelorarbeit

von

Tarik Enderes

31.12.2001

Betreuer: Prof. Dr. rer. nat. Vasileios Belagiannis
1. Prüfer: Prof. Dr. rer. nat. Vasileios Belagiannis
2. Prüfer: Prof. Dr.-Ing. Klaus Dietmayer

Hiermit versichere ich, dass ich die vorliegende Arbeit mit dem Titel

Segmentierung von Punktwolken mit neuronalen Netzen

bis auf die offizielle Betreuung selbstständig und ohne fremde Hilfe angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind. Aus fremden Quellen direkt oder indirekt übernommene Gedanken sind jeweils unter Angabe der Quelle als solche kenntlich gemacht.

Ich erkläre außerdem, dass die vorliegende Arbeit entsprechend den Grundsätzen guten wissenschaftlichen Arbeitens gemäß der „Satzung der Universität Ulm zur Sicherung guter wissenschaftlicher Praxis“ erstellt wurde.

Ulm, den 31.12.2001

Tarik Enderes

Danksagung

Ich möchte ich bei allen bedanken, die mir geholfen haben, diese Arbeit anzufertigen. Insbesondere bedanke ich mich bei Prof. Dr. Vasileios Belagiannis für seine kompetente und geduldige Betreuung, sowie seine Vorschläge zur Gestaltung des Projekts. Ein besonderer Dank geht auch an Dipl.-Ing. Uwe Kerner und M. Sc. Nico Engel für ihre technische Unterstützung.

Inhaltsverzeichnis

1 Einleitung	1
1.1 Segmentierung	1
1.1.1 Semantische Segmentierung	1
1.1.2 Instanz-Segmentierung	2
1.1.3 Panoptische Segmentierung	2
1.1.4 Objekt-Segmentierung	2
1.2 Ziele und Anforderungen	2
2 Literatur	5
2.1 Digitale Bildverarbeitung	5
2.2 Digitale Bildsegmentierung	6
2.2.1 Klassische Methoden der Segmentierung	6
2.2.2 Segmentierung mit Neuronalen Netzen	7
2.3 Verwandte Arbeiten	8
2.3.1 PointNet	8
2.3.2 UPSNet	10
3 Grundlagentheorie	13
3.1 Convolutional Neural Networks	13
3.2 Atrous Convolution	14
3.3 Atrous Spatial Pyramid Pooling	16
3.4 Conditional Random Fields	17
3.5 Residual Networks	18
3.6 Kamerakalibrierung	18
4 Arbeitsmethodik und Entwicklung	23
4.1 Algorithmus	23
4.2 DeepLab	23
4.2.1 Anpassungen für Semantische Segmentierung	25
4.3 Integration von DeepLab	25
4.4 Backbones	27
4.4.1 Xception	27
4.4.2 MobileNetV2	28

4.5 Segmentierung von Punkt wolken der KITTI-Daten	28
5 Datensätze	33
5.1 Cityscapes	33
5.2 KITTI	35
5.3 COCO	35
5.4 Pascal VOC	37
5.5 WildDash	37
6 Experimente	39
6.1 Technische Daten des für die Experimente verwendeten Rechners	39
6.2 Backbones	39
6.2.1 MobileNetV2	40
6.2.2 Xception65	44
6.3 Verfeinerung mit KITTI	44
6.4 Aufgetretene Probleme und Lösungen	48
6.4.1 False Positives	48
6.4.2 Overfitting	48
7 Zusammenfassung	53
Literaturverzeichnis	55

1 Einleitung

Für zahlreiche Entwicklungsthemen der heutigen Zeit, wie beispielsweise autonomes Fahren, ist eine präzise Erkennung der Umweltbedingungen unerlässlich. Kameras und Laserscanner finden für diesen Zweck oft Verwendung, was die Verarbeitung von Bildern und Punktwolken zu einem verbreiteten Gegenstand moderner Forschung macht. Häufig wird zur Lösung dieser komplexen Probleme auf Elemente der Neuroinformatik zurückgegriffen.

Je nach Anwendungsfeld ist ein bestimmter Grad an Auswertung der gegebenen Daten erforderlich. Diese Arbeit befasst sich mit der Aufgabe, Bilder und Punktwolken zu segmentieren.

1.1 Segmentierung

Segmentierung bezeichnet einen Vorgang, bei dem ein Bild nach bestimmten Heterogenitätskriterien in inhaltlich zusammenhängende Regionen eingeteilt wird. Von den verschiedenen Ansätzen, die das erreichen sollen, befasst sich diese Arbeit mit pixelbasierten Verfahren, bei denen jedem Pixel in einem Bild eine Klasse zugeordnet wird. Man unterscheidet zwischen den in [ups] beschriebenen, semantische Segmentierung, Instanz-Segmentierung und panoptische Segmentierung und der in [YLX⁺19] ausgeführten Objekt-Segmentierung. Für weitere Informationen siehe [GW08].

1.1.1 Semantische Segmentierung

Bei der semantischen Segmentierung soll jeder Pixel eine valide Klasse erhalten. Es wird dabei nicht zwischen unterschiedlichen Instanzen einer Objektklasse unterschieden. Wenn beispielsweise auf einem Bild zwei Fahrzeuge zu sehen sind und bei der Segmentierung die Klasse „Fahrzeug“ zugeteilt werden soll, erhalten die Pixel beider Fahrzeuge das Label „Fahrzeug“. Die Anzahl valider Klassen bleibt somit bei jeden prozessierten Bild gleich.

Einige Anwendungsgebiete von semantischer Segmentierung sind autonomes Fahren im Gelände [STB⁺19], Zellanalyse in der Biomedizin [RFB15] und Auswertung von Satellitenbildern für Kartographie [NR19].

1.1.2 Instanz-Segmentierung

Im Gegensatz zur semantischen Segmentierung werden bei der Instanz-Segmentierung nurzählbare Objekte betrachtet und deren Instanzen berücksichtigt. Übertragen auf vorheriges Beispiel würden die Pixel eines Fahrzeuges ein Label wie „Fahrzeug1“ und die des anderen analog „Fahrzeug2“ erhalten.

Instanz-Segmentierung findet beispielsweise Anwendung zur Detektion von Personen in Videodaten für Verhaltensanalysen und Überwachung [VWLT11].

1.1.3 Panoptische Segmentierung

Die panoptische Segmentierung stellt eine Kombination der vorherigen Segmentationsarten dar. Zählbare Objekte werden demnach nach dem Prinzip der Instanz-Segmentierung und amorphe nach dem der semantischen Segmentierung segmentiert. Die Ergebnisse beider Verfahren werden anschließend kombiniert.

1.1.4 Objekt-Segmentierung

Bei der Objekt-Segmentierung soll für jeden Pixel eines Bildes entschieden werden, ob er Teil des Vorder- oder des Hintergrundes ist, weshalb sie häufig als Vordergrund-Hintergrund-Segmentierung bezeichnet wird. Von Interesse ist dabei nur, wo sich Objekte im Bild befinden, nicht, wie bei den anderen Disziplinen, worum es sich handelt. Oft ist das Ziel dabei die Erkennung von Bewegung in Videodaten.

Objekt-Segmentierung findet Verwendung im Bereich der Videoüberwachung [GM10].

1.2 Ziele und Anforderungen

Ziel der Arbeit ist es, ein System zu entwickeln, das mit Hilfe von neuronalen Netzen ein Bild semantisch segmentiert und aufgrund der so entstandene Labels auf Pixelebene

eine Punktwolke derselben Szene segmentiert. Der Anwendungsbereich des Systems soll autonomes Fahren sein, weshalb Entwicklung und Experimente mit Datensätzen für diesen durchgeführt werden. Konkret soll die Ausgabe des Systems Algorithmen für Einfädelvorgänge an Kreuzungen verbessern. Besondere Wichtigkeit kommt daher der Erkennung von Fahrzeugen, Personen und Straßen zu. Eine Kernanforderung ist dabei Echtzeitfähigkeit. Optimierung der Laufzeit ist also essentiell. Weiterhin soll das System transportabel, leicht zu verwenden, benutzerfreundlich und ressourcenschonend sein.

Die Entwicklung erfolgt in Python mit CUDA-Unterstützung unter Verwendung des von Google entwickelten Framework DeepLab, das zur Zeit der Entstehung dieser Arbeit als State-of-the-Art angesehen wird.

2 Literatur

2.1 Digitale Bildverarbeitung

Das Feld der digitalen Bildverarbeitung umfasst nach [GW08] die Verbesserung der Bildinformation für menschliche Betrachter, sowie die Verarbeitung von Daten aus Bildern zur Übertragung oder autonomen maschinellen Erkennung durch digitale Rechner.

Der erste bekannte Einsatz digitaler Bildverarbeitung fand 1964 im Zuge des Raumfahrprogramms der USA statt, wobei ein Computer eingesetzt wurde, um Störungen in Bilder von der Mondoberfläche zu Korrigieren. Eine weitere Nennenswerte Entwicklung auf dem Gebiet ist die Erstellung von Röntgenbildern im Jahr 1979. Die Bedeutung digitaler Bildverarbeitung in der Geschichte steht in einem proportionalen Verhältnis zu den Möglichkeiten der Datenübertragung und -speicherung. Analog dazu erreicht sie ihren Durchbruch mit dem Aufkommen des World Wide Web um 1989. Zu den Aufgabe digitaler Bildverarbeitung gehören:

Bildbearbeitung Ziel der Bildbearbeitung ist es, die Ästhetik oder Erkennbarkeit eines Bildes zu verbessern. Da zu diesem Zweck häufig Fouriertransformationen zum Einsatz kommen, unterscheidet [GW08] zwischen Bildbearbeitung im Frequenz- und räumlichen Bereich. Zu ihren typischen Methoden gehören beispielsweise Histogrammverarbeitung, also das Zuordnen von Farbwerten auf andere Werte, oder das Anwenden von Filtern, entweder direkt im Frequenzbereich oder als räumliche Faltung mit einem Kernel.

Bildaufbereitung Die Bildaufbereitung befasst sich mit dem Entfernen von Störungen wie Rauschen oder Verzerrungen. Oft werden dazu wiederum Filter eingesetzt.

Bildkompression Bei der Bildkompression wird versucht, die Bildinformation zu codieren, um den Speicheraufwand für Übertragung und Sicherung zu verringern.

Bildsegmentierung Bei der Segmentierung wird versucht, Punkte oder Bereiche von besonderem Interesse in einem Bild zu finden, um das Bild maschinell erkennbar

zu machen.

Representation und Beschreibung Dieser Bereich der Bildverarbeitung beschäftigt sich mit der Anfertigung mathematischer Beschreibungen von Bildern wie zum Beispiel eine statistische Verteilung gerichteter Linien.

Objekterkennung Ziel der Objekterkennung ist es, Objekte auf Bildern zu detektieren, sodass dargestellte Szenen maschinell erkannt werden können. Übliche Methoden sind Bildvergleiche und Deep Learning.

2.2 Digitale Bildsegmentierung

Die automatische Segmentierung digitaler Bilder zählt zu den kompliziertesten Problemen der digitalen Bildverarbeitung und ist, wie in [YLX⁺19] erläutert, unerlässlich für das Ziel der Objekterkennung. Aufgrund der Komplexität der Aufgabe verwenden viele moderne Methoden Neuronale Netze zu diesem Zweck. Als besonders geeignet gelten auf Convolutional Neural Networks basierende Architekturen. Nach wie vor kommen auch klassische Methoden der Bildsegmentierung zum Einsatz, wenn die Anforderungen an das System nicht hoch sind.

2.2.1 Klassische Methoden der Segmentierung

Als klassisch werden im Zuge dieser Arbeit alle Methoden der Segmentierung bezeichnet, die keine Prinzipien der Neuroinformatik ausnutzen und stattdessen auf den Grundlagen von Diskontinuität und Ähnlichkeit basieren. Dazu gehören nach [GW08]:

Detektion von Diskontinuitäten Auch das Detektieren von Punkten, Linien und Kanten, die unter dem Begriff Diskontinuitäten zusammengefasst werden können, gehört zum Bereich der Bildsegmentierung. Um das zu erreichen werden im Bild nach Stellen gesucht, an denen sich der Farbwert der Pixel räumlich rapide ändert. Im einfachsten Fall kann dies mittels Faltung mit einem Filter bewerkstelligt werden, der auf starke Änderungen des Farbwertes in eine bestimmte Richtung reagiert.

Segmentierung nach Schwellwerten Die trivialste Methode der Segmentierung ist die nach Schwellwerten. Dabei werden Pixel oder Superpixel anhand ihres Farbwertes eingeordnet.

Regionen-basierte Segmentierung Bei Regionen-basierter Segmentierung werden räumliche Homogenitätskriterien ausgenutzt, um ein Bild in Bereiche einzuteilen. Die zwei verbreitetsten klassischen Ansätze dafür sind "Region Growing" und "Region Splitting and Merging".

Beim *Region Growing* werden eine Anzahl von Pixel im Bild ausgewählt. Die an diese angrenzenden Pixel werden untersucht und der Region hinzugefügt, wenn sie die entsprechenden Bedingungen erfüllen. Dann werden die Nachbarpixel der hinzugefügten Pixel untersucht und auf diese Weise die Region rekursiv aufgebaut.

Beim *Region Splitting and Merging* wird das Bild in Bereiche aufgeteilt bis alle damit entstandenen Regionen intern die Homogenitätsbedingungen erfüllen (Splitting). Danach werden benachbarte Regionen auf Ähnlichkeit untersucht und gegebenenfalls zusammengeführt (Merging).

Wasserscheide-Verfahren Das Wasserscheide(Watershed)-Verfahren nutzt lokale Minima und Maxima aus, um Regionen zu bilden. Das (Graustufen-)Bild wird dazu als dreidimensional betrachtet, wobei der Farbwert als Höhe interpretiert wird. Man stelle sich vor, die so entstandene Struktur werde ausgehend von den lokalen Minima geflutet und dort abgegrenzt, wo sich zwei Wasserflächen treffen würden. Auf diese Weise entstehen Regionen im Bild anhand der Lage von eindimensionalen lokalen Maxima, die zweidimensionale miteinander verbinden.

Segmentierung anhand von Bewegung Bei der Verarbeitung von Videodaten kann auch die zwischen den einzelnen Bildern stattfindende Bewegung zur Segmentierung ausgenutzt werden. Dazu gibt es zahlreiche Ansätze. Die einfachsten Fälle sind Verfahren, bei denen Bilder pixelweise miteinander verglichen werden, um ein Differenzbild zu erstellen, wie es beispielsweise bei Background Subtraction der Fall ist. Dabei geht der Algorithmus davon aus, dass die aufnehmende Kamera statisch und der Hintergrund damit unbeweglich ist. Pixel deren Farbwerte sich rapide ändern werden dementsprechend als Vordergrundobjekte gewertet.

2.2.2 Segmentierung mit Neuronalen Netzen

Wie bereits erwähnt sind Neuronale Netze ein beliebtes Mittel zur digitalen Bildsegmentierung. Auf einige Ansätze in diesem Gebiet soll hier eingegangen werden. Oft ist es sinnvoll, Methoden die Machine Learning verwenden danach einzuteilen, wie intensiv der Lernvorgang von einer Person überwacht werden muss. An dieser Stelle wird unterschieden zwischen unüberwachtem, schwach überwachtem und überwachtem Lernen.

Unüberwachtes Lernen findet Anwendung im Bereich der Objekt-Segmentierung. Der

Lern-Prozess konzentriert sich dabei meistens auf Bewegungen in Videodaten. Ein Beispiel stellt die Architektur aus [GWP18] dar, die Arcade-Spiele durch Reinforcement Learning lernt. Das Netz erhält zwei Frames als Eingabe und berechnet für jeden eine bestimmte Anzahl Objektmasken und entsprechende Translationen, sowie die Kamerabewegung. Beim Lern-Vorgang wird der Optische Fluss anhand der Ergebnisse ermittelt und eine Fehlerfunktion danach berechnet, wie genau der erste Frame aus dem zweitem und dem Optischen Fluss hervorgeht.

Es existieren Ansätze zur Bildsegmentierung mit Neuronalen Netzen, die als schwach überwachtes Lernen bezeichnet werden können. Das Netz erhält dabei Bilder, von denen bekannt ist, ob sich ein bestimmtes Objekt darin befindet oder nicht. Die Idee ist, Objekte die in beiden Arten von Bildern vorkommen als Hintergrund zu erkennen und zu ignorieren. In [HGH⁺12] wird ein Verfahren vorgestellt, dass mit großen Mengen zufällig ausgewählter, eventuell verrauschter Videos der Plattform YouTube arbeitet. Dabei wird ein Bild nach klassischen Segmentierungsverfahren in Regionen unterteilt. Die einzelnen Regionen werden von einem Netz danach bewertet, mit welcher Wahrscheinlichkeit sie zu dem gelernten Objekt gehören und anhand der Segmente mit hohen Wahrscheinlichkeiten wird, ebenfalls mit klassischen Methoden, eine Maske erstellt, die das gesamte Objekt umfassen soll.

Wenn an ein Verfahren hohe Anforderungen gestellt werden, wie bei der panoptischen oder Instanz-Segmentierung ist eine überwachte Lern-Methode oft unerlässlich. Für jedes Bild eines Trainingssatzes muss dafür eine Ground Truth zur Verfügung stehen, die das gewünschte Ergebnis des Netzes darstellt. Für Beispiele für diese Art von Verfahren siehe Abschnitt 2.3.

2.3 Verwandte Arbeiten

In diesem Abschnitt wird auf vorhandene Arbeiten eingegangen, die sich mit der Problematik der Segmentierung von Bildern oder Punktwolken mit neuronalen Netzen befassen.

2.3.1 PointNet

Das 2017 in [pnet] vorgestellte PointNet ist ein neuronales Netzwerk zum Auswerten von Punktwolken. Das Netz bietet dabei sowohl eine Architektur zur Klassifizierung als auch eine zur Segmentierung. Der Strukturelle Aufbau ist in Abbildung 2.1 dargestellt. Problematisch an der Auswertung von Punktwolken mit Technologien der Neuroinformatik ist vor allem, dass die Daten im Allgemeinen ungeordnet sind. Das Netzwerk

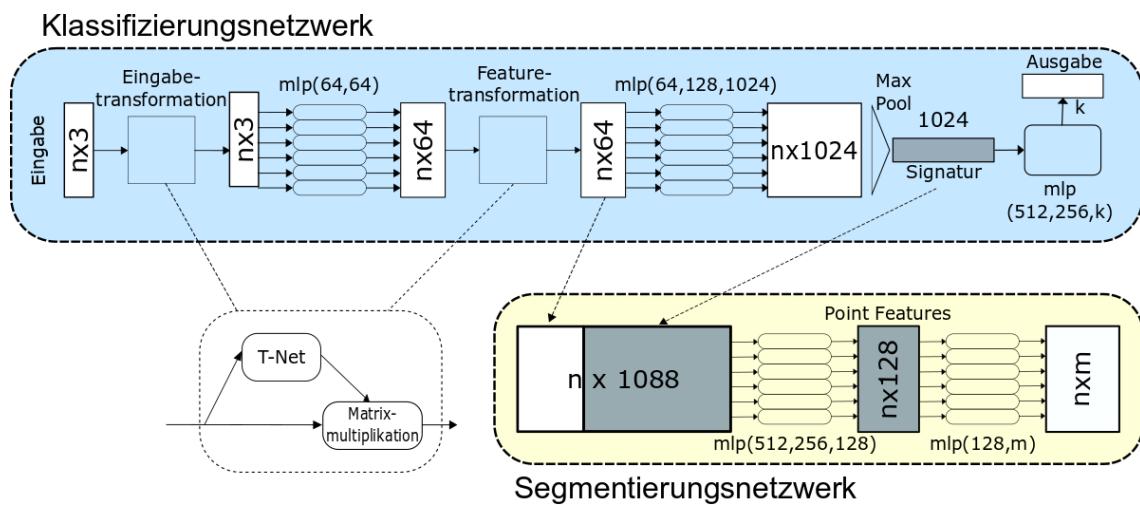


Abbildung 2.1: Architektur von PointNet nach [pnet]. Aus der Eingabe, die aus einer im Allgemeinen ungeordneten Menge an Punkten besteht, wird durch eine Reihe von Transformationen mittels Neuronaler Netze eine globale Signatur erstellt. Zu Klassifizierungsaufgaben kann dieser globale Feature-Vektor von einem weiteren Netz direkt weiterverarbeitet werden. Um eine Segmentierung durchzuführen, wird aus den lokalen und globalen Informationen ein neuer Vektor gebildet, der dann von einem Netz ausgewertet werden kann.

erzeugt darum aus einem Eingabevektor, der aus einer Menge von Koordinaten gebildet wird zunächst durch Feature Transformation eine globale Signatur, also einen Feature-Vektor, der unabhängig von der Reihenfolge der Eingabegrößen ist. PointNet erreicht dies durch den Einsatz von Max-Pooling. Um Invarianz bezüglich bestimmter räumlicher Transformationen wie z.B. Rotation zu erreichen, wird bei der Erstellung dieses globalen Feature-Vektor mit einem kleinen neuronalen Netz, dem „T-Net“, eine Transformationsmatrix angenähert und auf die Eingabedaten angewandt. Mit der so berechneten Signatur kann ein weiteres Netz, in diesem Fall ein MLP, trainiert werden, das diese klassifiziert. Da der globale Feature-Vektor keine Ortsinformationen enthält, ist es nicht möglich, damit eine Segmentierung durchzuführen. Soll das Netzwerk also für diesen Zweck verwendet werden, wird der globale Feature-Vektor mit dem Eingabevektor kombiniert, um einen Vektor zu erzeugen, der sowohl globale als auch lokale Eigenschaften repräsentiert. Anschließend kann ein Label für jeden Punkt geschätzt werden.

2.3.2 UPSNet

Das 2019 in [ups] vorgestellte UPSNet (Unified Panoptic Segmentation Network) ist ein neuronales Netz für panoptische Segmentierung. Dazu führt das Netzwerk parallel eine semantische Segmentierung und eine Instanzsegmentierung des Eingabebildes durch und erstellt mit den kombinierten Ausgaben beider Methoden einen Tensor von Wahrscheinlichkeiten für jede Klasse und Instanz. Aus diesem Tensor wird in einem letzten Schritt ein Ausgabebild erzeugt. Der Aufbau des Netzwerks ist in Abbildung 2.2 dargestellt.

UPSNNet verwendet als Backbone das in [rcnn] beschriebene Mask R-CNN, das sich aus dem ResNet und dem Feature Pyramid Network [fpn] ableitet. Mask R-CNN führt eine Instanz-Segmentierung des Bildes durch und erzeugt parallel dazu eine Maske für jedes erkannte Objekt. Die Ausgabe des Backbones wird von zwei leichtgewichtigen Netzen, dem „Semantic Segmentation Head“, der semantisch segmentiert und dem „Instance Segmentation Head“, der eine Instanzsegmentierung durchführt unabhängig voneinander weiterverarbeitet. Die Implementierung eines einzelnen Backbones spart Rechenzeit und Speicherplatz gegenüber Architekturen mit zwei getrennten Netzen. Die daraus entstandenen Ergebnisse werden von dem „Panoptic Segmentation Head“ anhand einer Heuristik ausgewertet, um die Netzwerkausgabe zu erstellen.

Der „Instance Segmentation Head“ folgt dem Konzept von Mask R-CNN und erzeugt eine Anzahl von Bounding Boxes und Masken. Der „Semantic Segmentation Head“ ist ein CNN, das einen vom Backbone erzeugten Feature-Vektor als Eingabe erhält und mittels Soft-Max die Klasse jedes Pixels schätzt. Der „Panoptic Segmentation Head“ ermittelt zuerst die Anzahl von im Bild vorhandener Instanzen und erstellt einen Tensor aus den von den beiden vorherigen Köpfen errechneten Wahrscheinlichkeiten

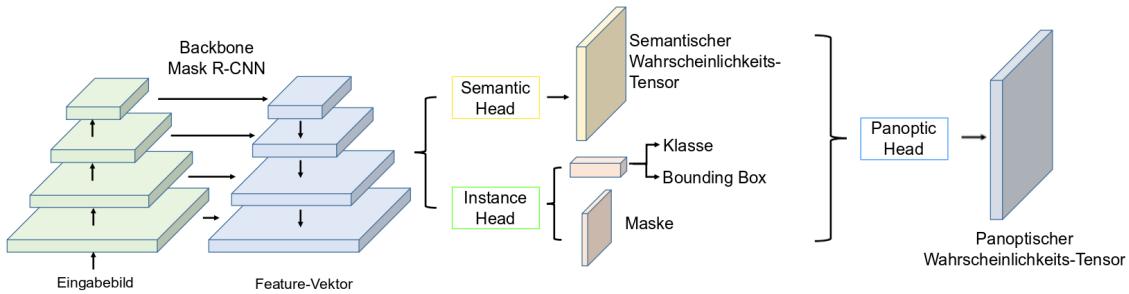


Abbildung 2.2: Architektur von UPSNet nach [ups]. Ein Mask R-CNN fungiert als Backbone des Netzes. Es erstellt einen Feature Vektor aufgrund der Eingabe, der dann von zwei leichtgewichtigen Netzen, dem "Semantic Head" und dem "Instance Head" verarbeitet wird. Die so entstandenen Ausgaben werden kombiniert und in dem "Panoptic Head" anhand einer Heuristik ausgewertet.

und führt eine Soft-Max Berechnung durch. Aus dem Resultat wird anhand einer Heuristik für jeden Pixel entschieden, ob er einer Instanz eineszählbaren Objekts und welcher Klasse er angehört. Es ist dabei möglich, dass Pixel als unbekannt klassifiziert werden, was den IoU der Ergebnisse durch die Verminderung von False Positives verbessert.

3 Grundlagentheorie

3.1 Convolutional Neural Networks

Wie in [GBC16] beschrieben, handelt es sich bei Convolutional Neural Networks (CNNs) um Neuronale Netze, die in mindestens einer Verarbeitungsschicht Faltung an Stelle von Matrixmultiplikation als mathematische Operation durchführen. Der Begriff Faltung bezieht sich dabei nicht auf die streng mathematischen Definition. In der Regel wird eine Variation eingesetzt. Verwendet das Netz ausschließlich Faltung spricht man von einem Fully Convolution Neural Network. CNNs eignen sich zur Anwendung auf rasterförmige Datenstrukturen und werden aufgrund ihrer im Folgenden beschriebenen Eigenschaften häufig zur Bildverarbeitung eingesetzt.

Ein Vorteil von Faltung gegenüber Matrixmultiplikation ist, dass Größe der Eingabematrix variabel ist. Im Fall von Bildbearbeitung bedeutet das, dass ein Fully Convolution Neural Network Bilder unabhängig von deren Größe und Auflösung verarbeiten kann. Es ist zu beachten, dass damit nicht Größeninvarianz erreicht wird. Bei der Faltung einer Matrix mit einem Kernel ist jeder Wert des Ergebnisses nur abhängig von bestimmten Werten der Eingabematrix, nicht unbedingt von allen, wie bei einer Matrixmultiplikation. Für semantische Segmentierung bedeutet das, dass der Ausgabewert für einen Pixel nur von Pixeln in einem begrenzten Bereich des Eingabebildes, dem Sichtfeld, bestimmt wird. Durch Verknüpfung mehrerer Faltungsschichten wird dieses Sichtfeld vergrößert. Außerdem wird jeder Wert der Eingabematrix auf dieselbe Weise verarbeitet. Damit werden die Ergebnisse der Faltungsschichten in einem Netzwerk Equivariant gegenüber Translation. Das bedeutet, wenn die Eingabe verschoben ist, tritt die gleichen Verschiebung in der Ausgabe auf. Die Größe des Faltungskernels kann theoretisch frei gewählt werden und ist im Fall von Bildverarbeitung vernachlässigbar klein verglichen mit den Eingabedaten, was CNNs deutlich effizienter im Bezug auf Laufzeit und besonders Speicherbedarf macht.

Üblicherweise wird in CNNs eine Pooling genannte Operation eingesetzt. Beim Pooling beziehungsweise Downsampling wird aus einer Matrix eine andere, meistens kleinere erstellt, die eine Zusammenfassung der Originalmatrix darstellt. Es gibt verschiedene Arten von Pooling. Häufig verwendet wird so genanntes Max-Pooling, bei dem jeder Eintrag der Ausgabematrix das Maximum eines rechteckigen Bereichs der Eingabema-

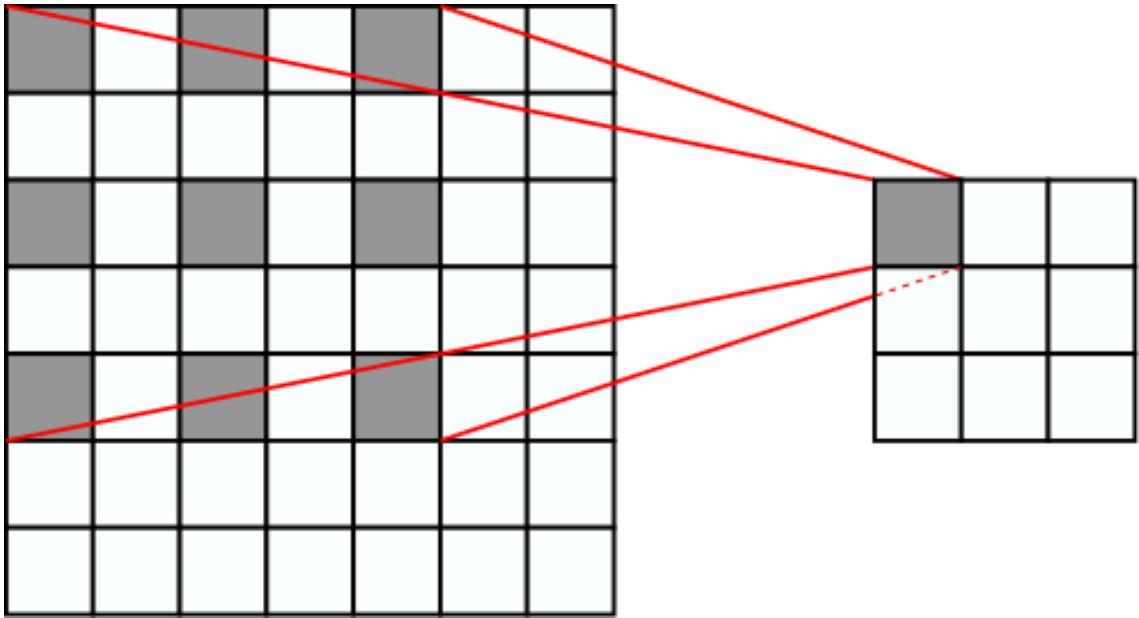


Abbildung 3.1: Prinzip von Atrous Convolution. Das Vorgehen bei der Faltung ist beispielhaft bei einer Erweiterungsrate von zwei dargestellt. Der Wert an der eingefärbten Stelle der Ausgabematrix (rechts) hängt von denen an den eingefärbten Stellen der Eingabematrix (links) ab, statt, wie bei einer herkömmlichen zweidimensionalen Faltung, von benachbarten. Die roten Strahlen markieren die Ecken des Sichtfelds.

trix ist. Durch Pooling soll das Netz Resistenter gegenüber kleinen Änderungen der Eingabedaten werden und die Größe für weitere Verarbeitungsschichten verringert werden, um die Laufzeit zu verbessern. Typischerweise folgt eine Pooling-Schicht auf eine oder mehrere Faltungsschichten.

In der Regel enthalten CNNs auch "Fully Connected Layers", bei denen, wie in einem klassischen Netzwerk, jeder Wert der Ausgabe von jedem Wert der Eingabe abhängt. Diese Schichten werden oft aus MLPs aufgebaut und befinden sich meistens am Ende des Netzes.

3.2 Atrous Convolution

Atrous Convolution, auch Dilated Convolution genannt, beschreibt eine Technik bei der eine Matrix mit einem spärlich bestückten Kernel gefaltet wird, wie in Abbildung 3.1 illustriert.

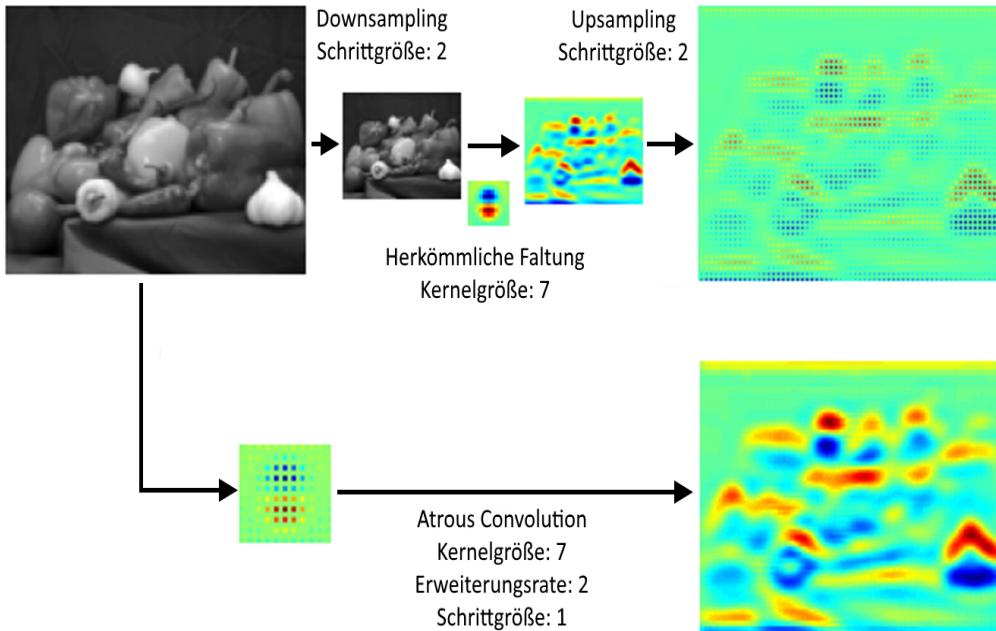


Abbildung 3.2: Beispielhaft dargestellte Vorteile von Atrous Convolution nach [dl2]. In der oberen Reihe ist das übliche Vorgehen bei CNNs dargestellt, in dem durch Down- und Upsampling die Effizienz verbessert wird. Der Vorgang entspricht einer Faltung mit erhöhter Schrittgröße. Unten dargestellt ist die Anwendung von Atrous Convolution mit einer Schrittgröße von eins und Erweiterungsrate zwei. Ein Vergleich der Ergebnisse der beiden Algorithmen zeigt, dass die Ausgabe von Atrous Convolution eine höhere Auflösung aufweist.

Die Abstände der zu berücksichtigenden Werte in der Matrix wird dabei durch die so genannte Dilation Rate bzw. Erweiterungsrate festgelegt. Das Tatsächliche Sichtfeld des Filters wird also durch die Größe des Kernels und die Rate bestimmt. ein Filter mit einem Kernel der Größe 3x3 und einer Rate von 2, was dem Einfügen einer leeren Zeilen und Spalte zwischen den Werten entspricht, hat demnach ein Sichtfeld der Größe 5x5. Dadurch wird das effektive Sichtfeld des Filters erhöht und es kann eine höhere Auflösung bei gleichen Rechenaufwand erreicht werden. Vor allem kann Downsampling damit vermieden werden. Die Vorteile der Verwendung von Atrous Convolution für Bildsegmentierung sind in Abbildung 3.2 dargestellt.

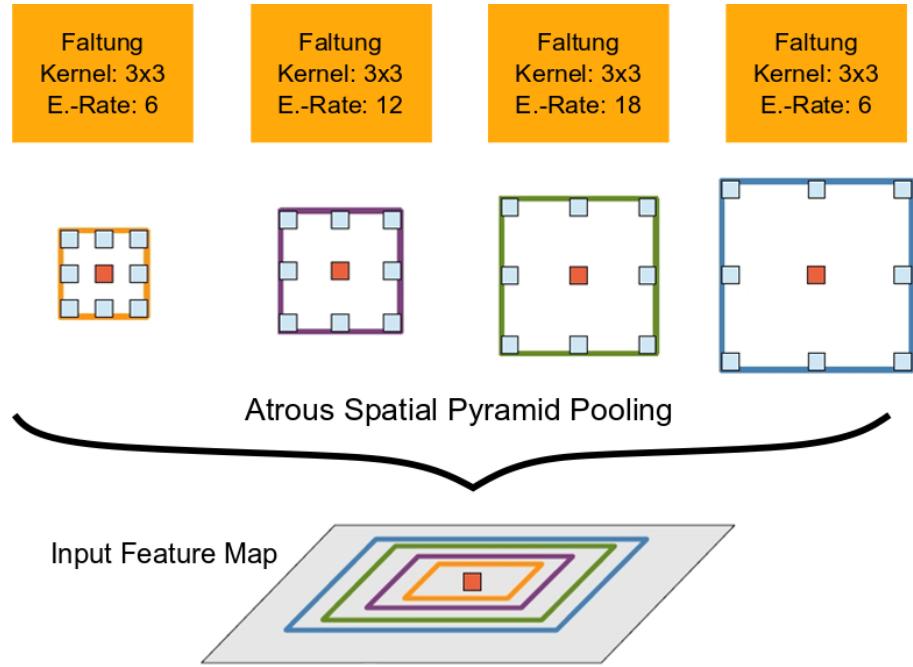


Abbildung 3.3: Prinzip von Atrous Spatial Pyramid Pooling wie in [dl2]. Die Eingabe wird mit verschiedenen Filtern mit unterschiedlicher Erweiterungsrate gefaltet, was dazu führt, dass die Stellen in den Ausgabematrizen verschieden große Sichtfelder aufweisen. Die so entstandene Ergebnisse werden zu einer Feature Map kombiniert.

3.3 Atrous Spatial Pyramid Pooling

Beim Atrous Spatial Pyramid Pooling werden mehrere parallele Convolutional Layers, die Atrous Convolutional Layers mit unterschiedlicher Erweiterungsrate verwenden, in das DCNN eingebaut. Aus den Ergebnissen der Verarbeitungszweige wird ein Tensor gebildet, der anschließend einer Dimension-übergreifenden 1x1 Faltung unterzogen wird, um die endgültigen Wahrscheinlichkeiten zu berechnen. Das Prinzip ist in Abbildung 3.3 dargestellt. Durch dieses Vorgehen soll Größeninvarianz erreicht werden.

3.4 Conditional Random Fields

Ein Conditional Random Field (CRF) ist ein Modell, das eine Datensequenz erhält und eine Sequenz gleicher Länge und Art ausgibt. CRFs werden zur Segmentierung und zum Labeln genutzt. Das Modell setzt, wie in [crf] und [McC03] beschrieben, überwachtes Lernen ein, um Parameter für eine Distribution zu bestimmen, die die Verteilung $d(X|Y)$ beschreibt, wobei X und Y Zufallsvariablen sind. X beschreibt die beobachtete Eingabesequenz, Y die zu bestimmende Ausgabesequenz.

Man betrachte eine Familie von Distributionen $p(z)$ und eine Menge von Feature-Funktionen ϕ_1, \dots, ϕ_N , die alle Daten ausdrücken, die in dem Modell berücksichtigt werden sollen. Es soll nun eine Sammlung von Parametern ω gefunden werden, sodass durch $p(y|x, \omega)$ die reale Verteilung $d(y|x)$ möglichst gut angenähert wird. p soll dabei so gewählt werden, dass dessen Entropie $H(p)$ maximal ist. Modelliert man, was in der Regel der Fall ist, p als Markov-Kette erster Ordnung, heißt, man nimmt an, dass vergangene Zustände den Ausgangszustand nicht beeinflussen, hat die Distribution mit maximaler Entropie die Form:

$$p(z) = \frac{1}{Z} \exp\left(\sum_i \omega_i \phi_i(z)\right) \quad (3.1)$$

mit einem festen Normalisierungsfaktor Z .

Idealerweise wird für eine Eingabesequenz x eine Ausgabesequenz y so gewählt, dass $p(y|x)$ nach dem berechneten Random Field $p(X|Y)$ maximal wird. Eine naheliegende Möglichkeit ist das „Ausprobieren“ aller möglichen y . Allerdings wäre das für lange Sequenzen zu aufwändig, weshalb in praktischen Anwendungen andere Optimierungsalgorithmen eingesetzt werden. Modelliert man, wie zuvor, p als Markov-Kette, ist eine effiziente Berechnung auf Grundlage des Viterbi-Algorithmus [RN93] möglich. Der Lernprozess für CRFs beruht auf Gradient Descend. Da die Wahrscheinlichkeitsfunktion bei eindeutigen Trainingsdaten wegen der Eigenschaften der Exponentialfunktion konvex ist, ist ein lokales Minimum dabei garantiert ein globales. Das Berechnen des Gradienten ist allerdings nicht-trivial und ineffizient, weshalb auf iterative und stochastische Lernalgorithmen, wie das Newtonverfahren zurückgegriffen wird. Wird für die Berechnung eines Wertes der Ausgabe alle Einträge der Eingabe betrachtet, spricht man von einem Fully Connected Conditional Random Field. Für weitere Informationen und eine präzise Definition siehe [LMP01].

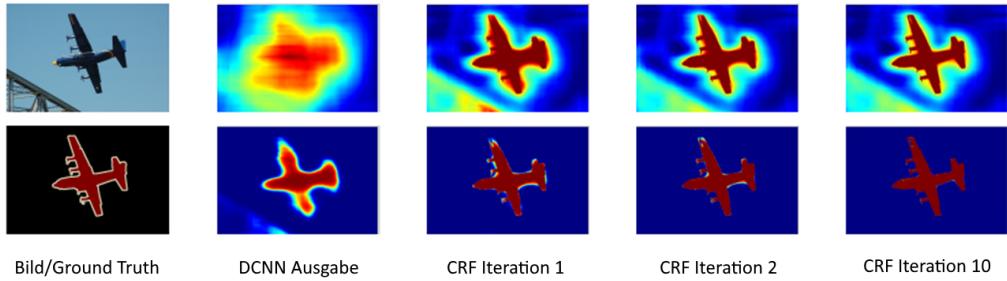


Abbildung 3.4: CRFs können eingesetzt werden, um die Ergebnisse von CNNs zu raffinieren, wie in [dl1] gezeigt. In der oberen Reihe werden die Score Maps, die Ergebnisse vor Anwendung einer Soft-Max-Funktion, gezeigt, in der unteren die Ausgabe der Soft-Max-Funktion.

3.5 Residual Networks

Ein Residual Neural Networks (ResNet) ist ein neuronales Netz, das das in [HZRS15] vorgestellte Residual Learning implementiert. Dabei werden, wie in 3.5 dargestellt, „Abkürzungen“ in das Netz eingebaut, über die die Ausgabewerte einer Schicht eine oder mehrere nachfolgende Schichten überspringen und eine tiefere Schicht unverändert erreichen und mit den Ergebnissen der übersprungenen Schichten addiert werden.

Mit ResNets wird ein Problem von Deep Neural Networks gelöst, bei dem der Trainingsfehler durch Hinzufügen zusätzlicher Verarbeitungsschichten vergrößert wird. Die Überlegung dabei ist, dass weitere Schichten die Resultate nicht verschlechtern können, wenn die Eingabe vorheriger Schichten noch unverändert vorhanden ist. Experimente bestätigen diese These. Residual Learning ist heute ein gebräuchliches Mittel beim Einsatz von vielschichtigen Netzwerken.

3.6 Kamerakalibrierung

Eine Kamera ermöglicht es, ein dreidimensionales Objekt auf eine zweidimensionale Ebene zu projizieren. Diese Projektion kann, wie in [HZ03] beschrieben, mit dem in Abbildung 3.6 dargestellten Modell angenähert werden, in dem von Punkten im dreidimensionalen Raum ein Strahl durch einen bestimmten Fixpunkt, das Projektionszentrum, verläuft und dabei eine vorgegebene Ebene, die Bildebene, schneidet. Der Schnittpunkt dieses Strahls mit der Bildebene entspricht dem projizierten Punkt auf dem entstehenden Bild.

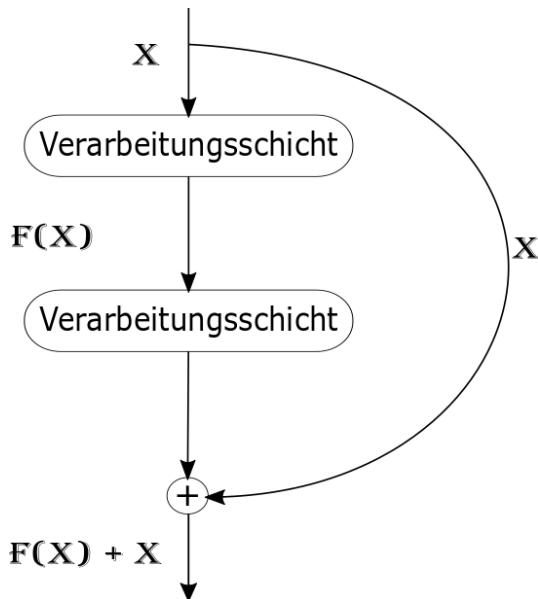


Abbildung 3.5: Prinzip von Residual Learning nach [HZRS15]. Über spezielle Abkürzungs-Schichten werden Daten unverändert in weiter darunter liegende Schichten geleitet und auf deren Ausgabe addiert.

Offensichtlich werden alle Punkte im dreidimensionalen Raum, die auf einem Strahl durch das Projektionszentrum liegen, auf denselben Punkt in der Bildebene projiziert. Das Bild lässt sich also praktisch als eine Menge von Strahlen auffassen.

Ist eine Kamera kalibriert, ist es möglich, aus zwei Punkten auf einem damit aufgenommenen Bild den Winkel zwischen den beiden Strahlen zu bestimmen, durch die sie entstanden sind. Analog dazu kann beispielsweise aus dem Bild auf die Größe einer fotografierten Fläche geschlossen werden oder bestimmt werden, ob eine Ellipse auf dem Bild die Projektion eines Kreises ist. Um solche Berechnungen anzustellen sind zusätzliche Informationen notwendig, da bei der Kameraprojektion Informationen über Entfernung, Längen, Winkel, Verhältnisse und dementsprechend Formen nicht erhalten bleiben.

Dass eine Kamera kalibriert ist, bedeutet im Praktischen Sinn, dass eine Matrix P berechenbar ist, für die gilt:

$$zm = PM. \quad (3.2)$$

M bezeichnet dabei die homogenen Koordinaten eines Punktes im dreidimensionalen Raum und m diejenigen von dessen Projektion auf der Bildebene, z ist ein reeller Faktor. Homogene Koordinaten unterscheiden sich von Euklidischen durch einen Zusätzlichen Parameter, der oft mit T bezeichnet wird. Eine Koordinate im zweidimensionalen Raum hat also die Form: $(X, Y, Z, T)^T$. Es gilt dabei, dass der Punkt $(x, y, 1)$ in homogenen Koordinaten äquivalent ist zum Punkt (x, y) in euklidischen Koordinaten.

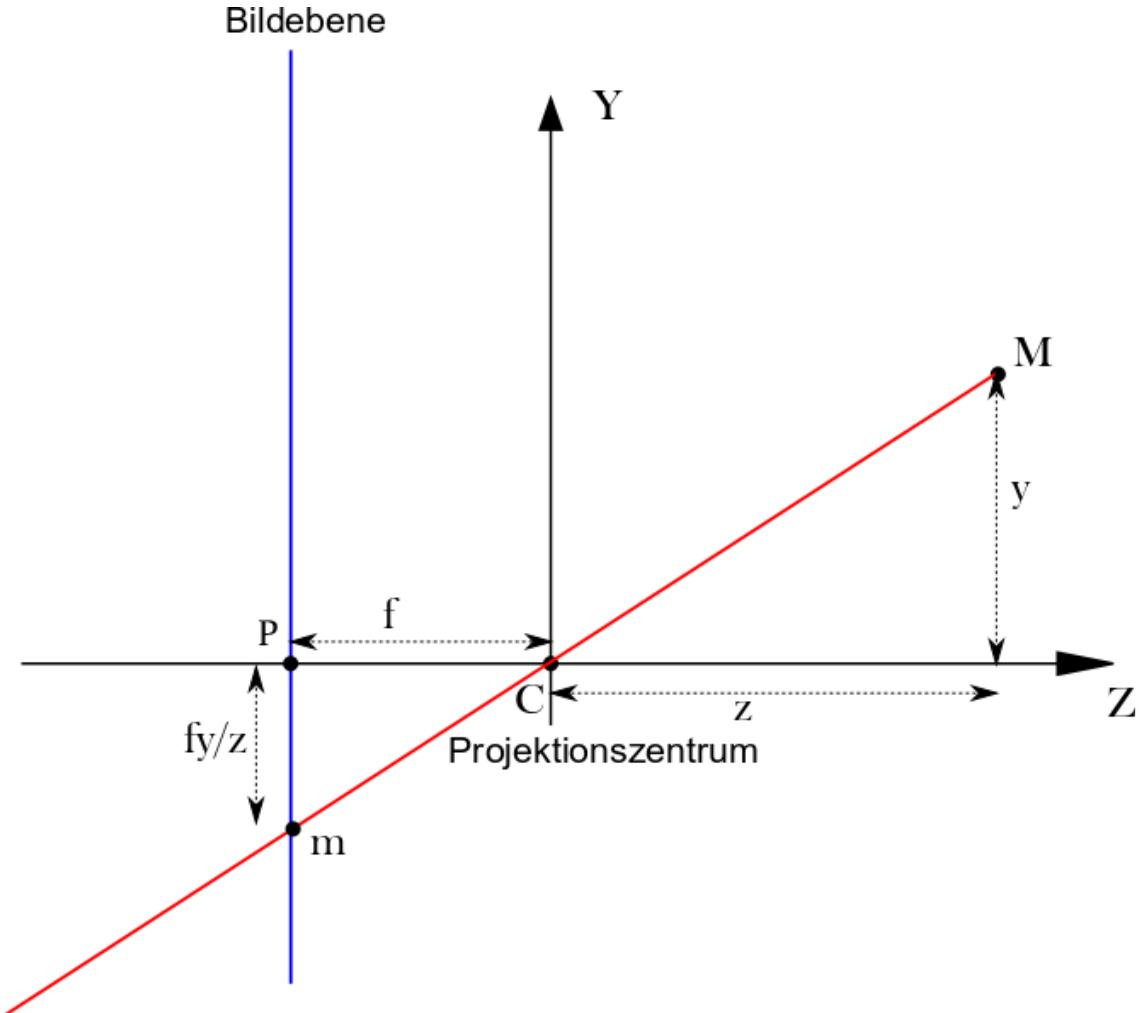


Abbildung 3.6: Prinzip einer Lochkamera zur Projektion vom dreidimensionalen in den zweidimensionalen Raum nach [Fus06]. Von einem realen Punkt M verläuft ein Strahl durch das Projektionszentrum C und schneidet die Bildebene im Bildpunkt m . Das Koordinatensystem ist so gewählt, dass sich C im Ursprung befindet und die **Y**-Achse parallel zur Bildebene verläuft. Die **Z**-Achse schneidet die Bildebene im Bildzentrum P . Hat M die Entfernung z auf der **Z**-Achse und y auf der **Y**-Achse und der Abstand zwischen C und der Bildebene ist f , so beträgt die Entfernung von m zur **Z**-Achse $\frac{fy}{z}$.

Es gilt also:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \hat{=} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (3.3)$$

Für homogene Koordinaten mit $T \neq 0$ verhält es sich für alle $k \neq 0$ so, dass:

$$\begin{pmatrix} x \\ y \\ \omega \end{pmatrix} = \begin{pmatrix} kx \\ ky \\ k\omega \end{pmatrix} \quad (3.4)$$

Eine Umrechnung in euklidische Koordinaten erfolgt folglich durch:

$$\begin{pmatrix} x \\ y \\ \omega \end{pmatrix} \hat{=} \begin{pmatrix} \frac{x}{\omega} \\ \frac{y}{\omega} \\ 1 \end{pmatrix} \quad (3.5)$$

Ein Punkt, in dem $T = 0$ gilt, bezeichnet man als „Punkt im Unendlichen“. Was hier beispielhaft für zwei Dimensionen dargestellt ist, kann offensichtlich für beliebig viele Dimensionen erweitert werden. Für weitere Informationen über homogene Koordinaten siehe [HZ03].

Kenntnis über die so genannte Projektions- oder Kameramatrix P ermöglicht also die Berechnung der zum Projektionszentrum relativen Koordinaten des projizierten Punktes aus denen des aufgenommenen Punktes und umgekehrt. Offensichtlich kann damit berechnet werden, wo ein mit einer kalibrierten Kamera aufgenommener Punkt im Bild einer anderen erscheint. Zum Ermitteln der Projektionsmatrix sind, wie in [Fus06] beschrieben, folgende Informationen notwendig:

- Die Entfernung f zwischen dem Projektionszentrum und der Bildebene.
- Die Koordinaten des Bildzentrums $P = (p_x, p_y)$.
- Höhe s_x und Breite s_y der Pixel.
- Der Scherungswinkel Θ zwischen den Achsen, der für Gewöhnlich $\frac{\pi}{2}$ beträgt.

Diese Informationen lassen sich in der so genannten Kalibrierungsmatrix K folgendermaßen Zusammenfassen:

$$K = \begin{bmatrix} \frac{f}{s_x} & \frac{f}{s_x} \cot \Theta & p_x \\ 0 & \frac{f}{s_y} & p_y \\ 0 & 0 & 1 \end{bmatrix} \quad (3.6)$$

Ist die Kalibrierungsmatrix bekannt, muss noch die Position der Kamera durch die Rotationsmatrix R und den Translationsvektor t ausgedrückt werden. Dann kann die

Projektionsmatrix berechnet werden mittels:

$$P = K [R|T] \quad (3.7)$$

4 Arbeitsmethodik und Entwicklung

4.1 Algorithmus

Es gibt mehrere Ansätze zum Segmentieren von Punktwolken mit Neuronalen Netzen, die unterschiedliche Leistungen bezüglich Laufzeit und Qualität erzielen. In dieser Arbeit wird ein möglichst einfacher und zeiteffizienter Algorithmus angewandt. Dabei wird, wie in Abbildung 4.1 dargestellt, zuerst ein Eingabebild mit Hilfe neuronaler Netze semantisch segmentiert. Dazu wird an dieser Stelle DeepLabV3+ benutzt, da es mit seinen Ergebnissen auf populären Datensätzen (79.7% auf Pascal VOC 2012 und 70.4% auf Cityscapes in mIOU-Metrik) laut [dl2], als State-of-the-Art angesehen wird und gleichzeitig eine große Fülle an Informationen und öffentlichen Implementierungen zur Verfügung stehen.

Die durch die Segmentierung ermittelten Labels werden anschließend auf eine Punktwolke projiziert, die dieselbe Szene wie das segmentierte Bild darstellt. Dazu werden ein Bild und eine Punktwolke der zu segmentierenden Szene, sowie eine Projektionsmatrix benötigt. die verwendete Kamera muss also kalibriert sein. Offensichtlich ist es damit nur möglich, den Bereich der Punktwolke zu segmentieren, der im Bild zu sehen ist. Außerdem werden auf diese Weise keine in der Punktwolke vorhandenen Tiefen-Information ausgenutzt. Stattdessen können Farb-Informationen berücksichtigt werden.

4.2 DeepLab

DeepLab ist ein von Google entwickeltes, 2015 in [dl1] vorgestelltes Modell für semantische Segmentierung. Bei der in [dl2] vorgestellten Methode wird ein Deep Convolutional Neural Network (DCNN) zum Erzeugen einer Score Map benutzt, die anschließend mit einem Conditional Random Field (CRF) zur endgültigen Ausgabe weiterverarbeitet wird. Das Verfahren wird in Abbildung 4.2 grob dargestellt.

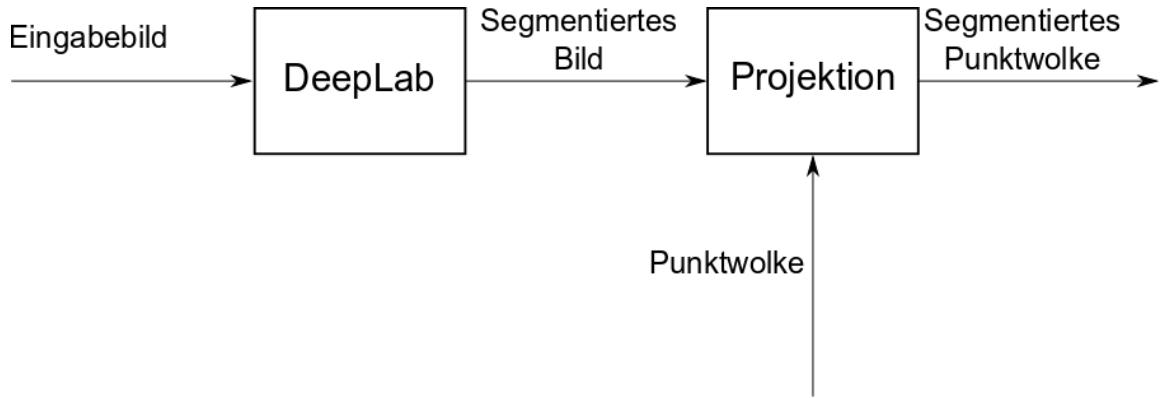


Abbildung 4.1: Schematische Arbeitsweise des Systems. Ein Eingabebild wird zunächst mit DeepLab segmentiert. Die so entstandenen Labels werden unter Verwendung einer Projektionsmatrix auf eine zum Eingabebild gehörende Punktfolke projiziert.

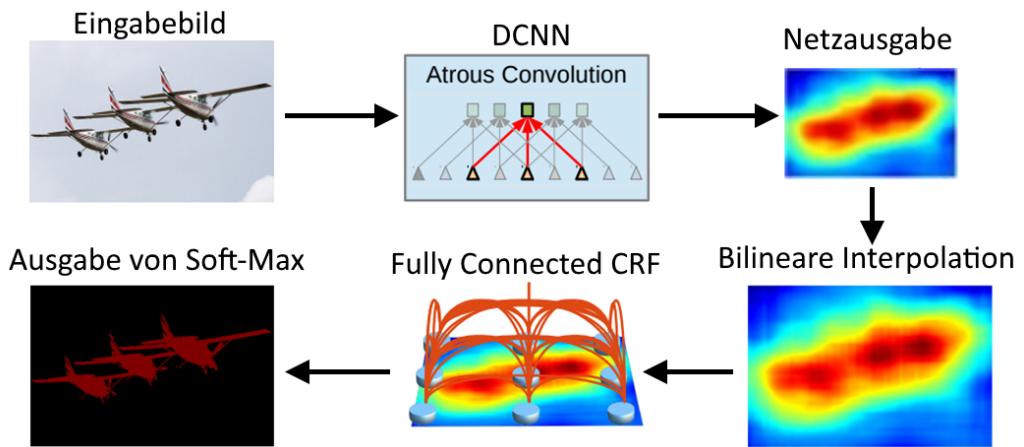


Abbildung 4.2: Arbeitsfluss von DeepLab nach [dl2]. Das Eingabebild wird von einem DCNN segmentiert. Das Ergebnis wird durch Bilineare Interpolation auf die Größe der Eingabe vergrößert und in einem Fully Connected CRF raffiniert.

4.2.1 Anpassungen für Semantische Segmentierung

Klassische DCNNs haben Eigenschaften, die sie für die Verwendung zur Bildsegmentierung nicht ideal machen.

- Der Einsatz von Downsampling führt zu verringerter Auflösung, die bei Klassifizierungsaufgaben nicht ins Gewicht fällt, für die Segmentierung aber essentiell ist.
- Neuronale Netze sind in der Regel gut geeignet, um Objekte unterschiedlicher Größe zu erkennen, wenn solche in der Lernphase präsentiert werden. Die Eigenschaften der Faltung, insbesondere dem begrenzten Sichtbereich beim Berechnen eines einzelnen Pixels ist allerdings für diese Problematik ungünstig.
- Der wiederholte Einsatz von Convolutional Layers führen zu einem Verlust an Ortsinformation. Infolgedessen produzieren DCNNs bei Segmentierungsaufgaben verschwommene, oft verrauschte Ergebnisse ohne klare Kanten.

Um diese Probleme zu lösen, erhält das von DeepLab verwendete DCNN einige Anpassungen. Zunächst werden alle Fully Connected Layers durch Convolutional Layers ersetzt, um ein Fully Convolutional Network zu bilden. Noch dazu wird anstatt von Pooling Layers in den unteren Schichten Atrous Convolution eingesetzt, womit die Auflösung der Ausgabe erhöht wird. In den höheren Schichten werden auch hier Pooling Layers eingesetzt, um Speicherbedarf und Rechenzeit zu verbessern. Um Größeninvarianz zu erreichen wird bei den unteren Schichten Atrous Spatial Pyramid Pooling verwendet und um die Ergebnisse, vor allem an den Kanten von Objekten zu verbessern und Rauschen zu reduzieren, wird die Ausgabe des DCNN mit einem Fully Connected CRF weiterverarbeitet.

4.3 Integration von DeepLab

Als Basis für die Integration von DeepLab dient eine öffentlich zugänglichen Implementierung in Python mit dem Pytorch-Framework.

Das Projekt implementiert die zwei Netzwerke Xception65 und MobileNetV2. Es wird die Möglichkeit angeboten, vor-trainierte Encoder zu verwenden, die in dieser Arbeit aber nicht genutzt wird, da derzeit Kompatibilitätsprobleme mit den bereitgestellten vor-trainierten MobileNetV2-Models auftreten. Neben den bereits beschriebenen Technologien werden die klassischeren Konzepte der Neuroinformatik Dropout, L2-

Regularisierung und Momentum verwendet.

Bezüglich der Entwicklung ist das erste Ziel die Verwendung dieses Projekts mit einem vor-trainierten Model zur Evaluierung frei gewählter Bilder. Die ursprüngliche Implementierung erwartet Test- und Trainingsdaten im Format von entweder Cityscapes oder Pascal. Der erste Schritt ist folglich die Erstellung eines Python-Skripts zur Evaluierung beliebiger Bilder. Dieses Skript soll Ausgaben im PNG-Format erzeugen, die das Originalbild, die Auswertung des Netzes, eine Legende, ein RGB-Histogramm und die erforderliche Rechendauer zeigen. Zur Erzeugung der Plots wird das Modul Matplotlib verwendet. Für den Fall, dass die Eingabebilder einen schwarzen Rand aufweisen, wird dieser Rand auf die gelabelten Bilder übertragen, wozu das Bildverarbeitungs-Framework OpenCV verwendet wird. Die Legende wird automatisch erstellt und richtet sich nach der Klasse mit dem höchsten Label, die in dem Eingabebild erkannt wurde. Sie zeigt außerdem den IoU-Wert jeder Klasse. Das RGB-Histogramm wird mit einer Funktion von OpenCV automatisch erstellt. Ein Beispiel ist in Abbildung 4.3 dargestellt.

In einem nächsten Schritt soll die Möglichkeit geschaffen werden, ein eigenes Model mit eigenen Trainingsdaten zu trainieren. Das bereits vorhandene Trainings-Skript kann dazu verwendet werden. Die Verwendung von Nebenläufigkeit beim Laden der Trainingsdaten führt allerdings unter Windows dazu, dass das Training nach einer Epoche abgebrochen wird. Auch hier wird erwartet, dass die Daten im Format von Cityscapes oder Pascal sind. Da hier vor allem der Cityscapes-Datensatz zum Trainieren verwendet wird und dieser einfach zu erweitern ist, wird das als sinnvoll eingeschätzt und mit einer Änderung in den Nomenklatur-Regeln beibehalten. Zum Beginnen eines Trainings muss eine Konfigurationsdatei im YAML-Format übergeben werden. Darin können folgende Trainingsparameter festgelegt werden:

- Encoder-Type
- Decoder-Typ
- Format des Datensatzes
- Zielgröße der Trainingsbilder
- Anzahl Trainingsepochen
- Batch-Größe
- Verwendung von FP16
- Fehler-Typ

- zu ignorierender Index (255 in Cityscapes)
- Optimierer
- Grundlernrate

Außerdem kann der Pfad zu einem vor-trainierten Model angegeben werden und bestimmt werden, ob ein bereits existierendes Model weiter trainiert werden soll, was zum Nach-Training benutzt werden kann.

Als letztes muss noch ein Skript erstellt werden, das ein Model auf dem Cityscapes-Datensatz testet und in IoU-Metrik bewertet. Eine Funktion zur Berechnung des IoU ist bereits vorhanden, muss aber noch in einer für diese Arbeit sinnvolle Weise aufgerufen werden. Das dazu geschriebene Skript erwartet Evaluierungsdaten im Cityscapes-Format, da die Experimente auf diesem Datensatz durchgeführt werden. Es wertet alle Bilder dieses Datensatzes mit dem zu testenden Model aus und berechnet den durchschnittlichen IoU für jede von dem Netz erkennbare Klassen, sowie die durchschnittliche Rechendauer pro Eingabebild. Zu beachten ist dabei, dass NaN-Werte bei der Berechnung speziell behandelt werden müssen.

Zusätzlich soll noch die Möglichkeit geschaffen werden, eigene Daten zu annotieren. Dazu wird der von der University of Oxford bereitgestellte VGG Image Annotator (VIA) genutzt. Dieser bietet die Möglichkeit, Polygone in ein Bild einzutragen und diese als CSS-Dateien zu exportieren. Ein Python-Skript, das OpenCV verwendet verarbeitet diese Daten dann zu Cityscapes-konformen PNG-Dateien, die zum Training oder zur Validierung verwendet werden können. Im Rahmen dieser Arbeit wird das allerdings nicht eingesetzt.

4.4 Backbones

Die hier verwendete Implementierung von DeepLab umfasst zwei Backbones: Xception65 und MobileNetV2.

4.4.1 Xception

Das in [xce] präsentierte Xception-Netzwerk ist ein einfaches aber leistungsfähiges DCNN, das auf der Idee von Depthwise Separable Convolution basiert. Es wird also angenommen, dass Korrelationen, die mehrere Kanäle umfassen von räumlichen Korrelationen entkoppelt werden können. Die einzelnen Module des Netzes verarbeiten

zunächst alle Kanäle separat und führen dann eine einfache Faltung über alle Dimensionen durch. Dadurch werden Laufzeit- und Speichereffizienz verbessert. Das Prinzip ist in Abbildung 4.4 dargestellt.

Das vorgestellte Netz hat insgesamt 36 Faltungsschichten, die in 14 Module gegliedert sind, die als ResNet miteinander verbunden sind. In jeder Schicht werden nach dem oben erklärten Prinzip mehrere Faltungen mit 3x3 Filtern parallel auf allen Kanälen durchgeführt. Der so entstandene Tensor wird anschließend mit einem 1x1 Filter gefaltet, der alle Dimensionen umfasst.

4.4.2 MobileNetV2

Bei MobileNetV2 [mn2] handelt es sich um ein leichtgewichtiges DCNN für die Implementierung auf mobilen Geräten. Da Laufzeit und Ressourcenlastigkeit für das Ziel dieser Arbeit von kritischer Bedeutung sind, wird hier vorrangig mit diesem Backbone gearbeitet. Wie Xception verwendet es das Prinzip von Depthwise Separable Convolution und Residual Connections. Für den Entwurf des Netzes wird zusätzlich die Annahme gemacht, dass die entscheidenden Eigenschaften eines Eingabetensors in einem Subraum mit weniger Dimensionen zusammengefasst werden können.

Begründet auf diesen Konzepten ist das grundsätzliche Architekturelement von MobileNet der so genannte „Bottleneck Residual Block“. Ein- und Ausgabetensoren dieser Verarbeitungsschichten haben eine niedrigere Dimension als Tensoren dazwischen. Innerhalb des Blocks wird zuerst der Eingabetensor zu einem mit mehr Dimensionen erweitert. Der so entstandene Tensor wird dann nach dem Prinzip von Depthwise Separable Convolution zuerst Kanalweise, dann Kanalübergreifend mittels Faltungsoperationen und ReLU6 weiterverarbeitet, wobei die Dimension wieder reduziert wird. Das Ergebnis davon wird nach dem ResNet-Prinzip mit dem Eingabetensor verrechnet. Der Vorteil dieses Verfahrens liegt in einem geringeren Speicherbedarf. Insgesamt besteht das Netz aus einem Convolutional Layer mit 32 Filtern am Anfang, gefolgt von 19 Residual Bottleneck Layer und einigen weiteren Schichten am Ende, die von der zu erfüllenden Aufgabe abhängig sind.

4.5 Segmentierung von Punktwolken der KITTI-Daten

Der KITTI-Datensatz bietet sowohl Aufnahmen von kalibrierten Farbkameras mit der Projektionsmatrix als auch Laserscans in Form von Punktwolken im Velodyne-Format

und eine, wenn auch verhältnismäßig geringe, Menge an fein-annotierten Bildern für semantische Segmentierung. Das macht ihn zu einer logischen Wahl für die Generierung von Beispielergebnissen in diesem Projekt.

Für die Implementierung bietet sich das für KITTI entwickelte Python-Modul Pykitti an, das Funktionen zur Verwaltung der Bild- und Velodyne-Daten und zum Umrechnen der Koordinaten anhand der Projektionsmatrizen anbietet. Das macht es einfach, die mit DeepLab gewonnen Labels auf die Punktwolken zu projizieren. Zu beachten ist, dass dabei nur Punkte beachtet werden können, die sich im Sichtfeld des Bildes befinden. Für die Visualisierung der Ergebnisse wird eine Python-Integration von Mayavi verwendet. Abbildung 4.5 zeigt ein Beispiel.

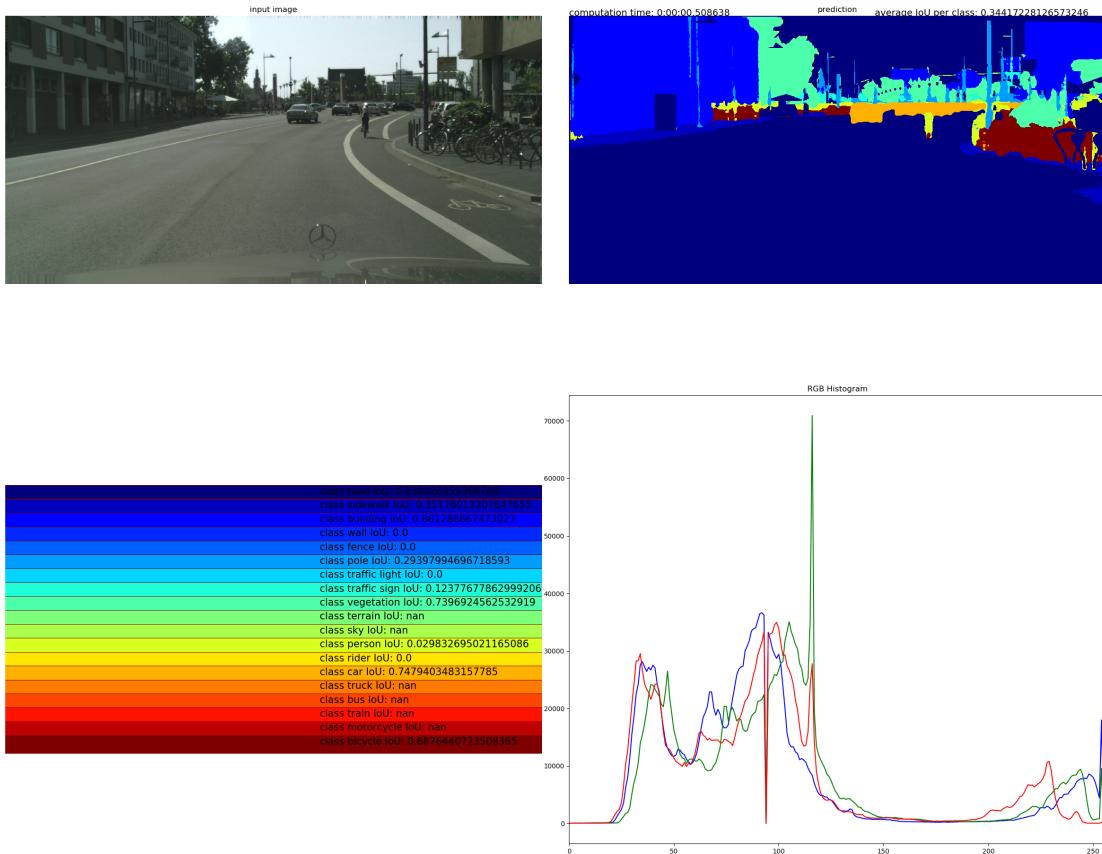


Abbildung 4.3: Beispiel für die Ausgabe des Evaluierungs-Skripts. Zu sehen ist das Eingabebild (oben links), das segmentierte Bild (oben rechts), eine Legende, die den IoU für jede Klasse enthält (unten links) und ein RGB-Farbhistogramm (unten rechts). Die Rechenzeit des Netzes und der durchschnittliche IoU werden über dem segmentierten Bild angezeigt.

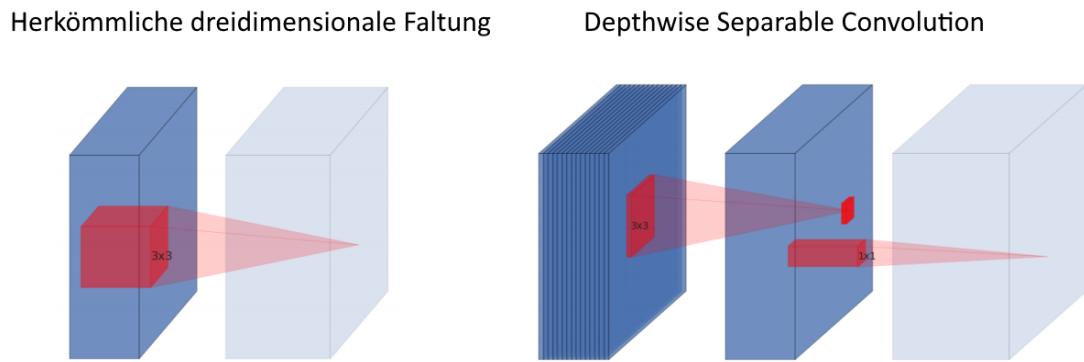


Abbildung 4.4: Schematische Darstellung des Prinzip von Depthwise Separable Convolution zur Tensorverarbeitung nach [mn2]. Statt einer Faltung mit einem 3×3 -Kernel über alle Dimensionen, wird für jede Dimension eine Faltung mit einem 3×3 -Kernel durchgeführt, gefolgt von einer Faltung mit einem 1×1 -Kernel über alle Dimensionen.

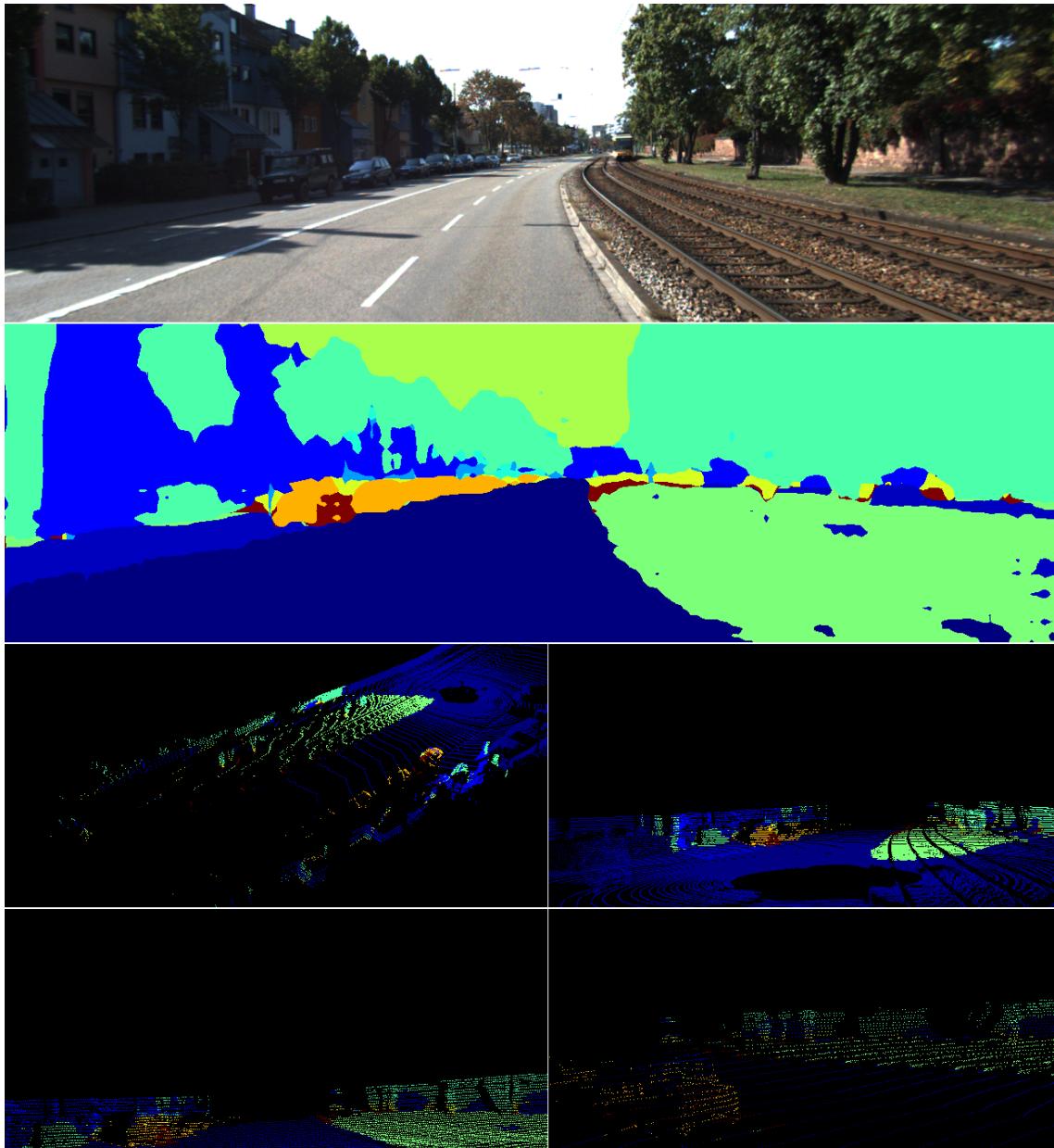


Abbildung 4.5: Beispiel für segmentierte Punktfolke aus dem KITTI-Datensatz. Zu sehen ist von oben nach unten: Originalbild, Ergebnis von DeepLab, segmentierte Punktfolke aus verschiedenen Perspektiven.

5 Datensätze

Methoden, die auf überwachtes Lernen zurückgreifen, erfordern in der Regel eine große Menge an Trainingsdaten, um verwendbare Resultate zu erzielen. Das Erstellen von Labeln für semantische Segmentierung ist besonders aufwändig, da jedem Pixel des Bildes eine Klasse zugeordnet werden muss. Deshalb werden in dieser Arbeit die öffentlich zugänglichen Datensätze Cityscapes und KITTI genutzt. In dem Kapitel soll näher auf verfügbare Datensätze eingegangen und erläutert werden, weshalb diese beiden ausgewählt wurden.

5.1 Cityscapes

Cityscapes ist ein öffentlich zugänglicher Datensatz ausgelegt für Bildsegmentierung zum autonomen Fahren. Er bietet 5000 fein und 20000 grob auf Pixel-Ebene annotierte Bilder für semantische oder Instanzsegmentierung. Der Satz an fein annotierten Aufnahmen, der in den Experimenten verwendet wird, ist unterteilt in einen Trainingssatz aus 2975 Bildern, einem Evaluierungssatz von 500 Bildern und einem Testsatz aus 1525 Bildern. Aufgenommen ist der Datensatz von einem Auto aus in 50 größtenteils deutschen Städten, jeweils am Tag bei sonnigem oder bewölktem Wetter um Frühling, Sommer und Herbst. Die Bilder zeigen ausschließlich Szenen, die sich auf vielbefahrenen Straßen abspielen.

Die Daten sind aufgenommen mit einer Stereo-Kamera, die hinter der Windschutzscheibe des Fahrzeugs angebracht ist, bei einer Frame-Rate von 17Hz. Die Bilder des Datensatzes sind kalibriert, Bayer-gefiltert und rektifiziert. Abbildung 5.1 zeigt Beispiele. Für weitere Informationen siehe [COR⁺16].

Die große Anzahl an qualitativ hochwertigen Bildern und Annotationen, sowie deren Verfügbarkeit machen Cityscapes zu einer logischen Wahl für diese Arbeit. 2975 Bilder scheinen eine vergleichbar kleine Datenmenge zu sein, doch die Information, die in einem Bild enthalten ist, ist ungleich größer als beispielsweise in für Klassifizierung annotierten Bildern.

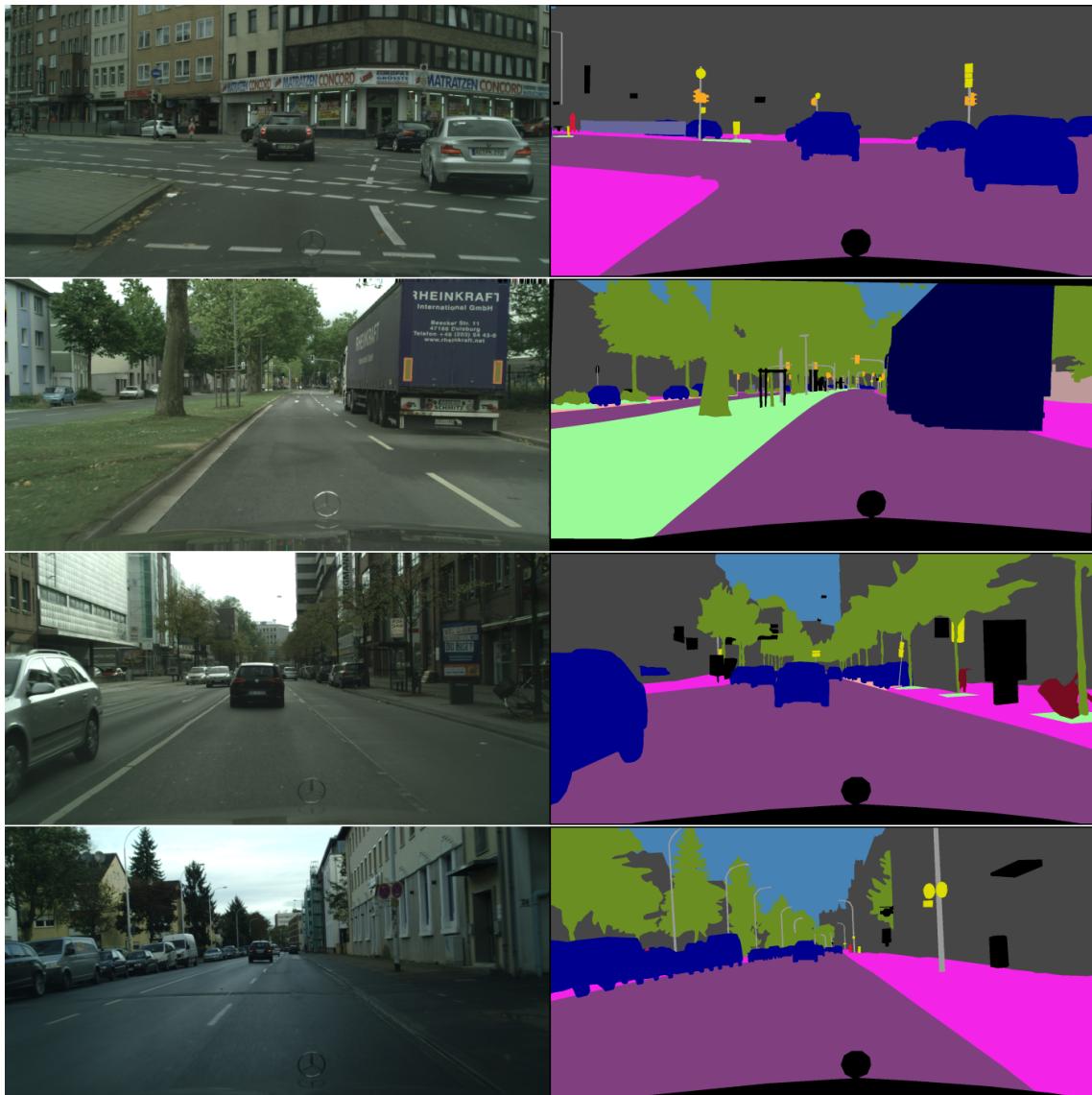


Abbildung 5.1: Hier zu sehen sind beispielhaft fein annotierte Trainingsdaten für semantische Segmentierung aus dem Cityscapes-Datensatz. Wie man sieht, werden die Bilder (links) von einer hinter der Windschutzscheibe platzierten Kamera aufgenommen. Die rechte Spalte zeigt die eingefärbte Ground Truth für semantische Segmentierung. Die schwarz markierten Bildflächen gehören keiner im Datensatz gelabelten Klasse an und werden während des Trainings ignoriert.

5.2 KITTI

Das in [kit] beschriebene KITTI ist ein Datensatz für Forschung in den Bereichen mobile Robotik und autonomes Fahren. Es werden darin Kamerabilder, Laserscans, GPS- und IMU-Daten zur Verfügung gestellt. Die Kamerabilder werden von zwei Stereo-Kamera-Rigs aufgenommen, eines für Farbaufnahmen, eines für Graustufenbilder und liegen sowohl als Rohdaten als auch rektifiziert vor. Die Laserscan-Daten sind im Velodyne LiDAR-Format gespeichert. Die Kalibrierungs-Matrizen sind im Rohdatensatz ebenfalls angegeben.

Der Datensatz umfasst insgesamt 6 Stunden an Aufnahmen mit zwischen 10Hz und 100Hz aus Karlsruhe. KITTI bietet außerdem 200 für Segmentierung annotierte Bilder. Die Messgeräte sind auf einer mobilen Plattform auf einem Auto angebracht. Die Perspektive unterscheidet sich also geringfügig von der der Cityscapes-Daten. Für weitere Informationen über die verwendeten Messgeräte siehe [kit].

Im Gegensatz zum Cityscapes-Datensatz, der sich auf Straßenverkehr spezialisiert, wird im KITTI-Datensatz versucht, eine möglichst große Szenenvielfalt anzubieten. Beispiele für Bilder des Datensatzes sind in Abbildung 5.2 zu sehen. KITTI wird in dieser Arbeit verwendet, da darin, im Gegensatz zu Cityscapes, neben der Bilder Punktwolken und Projektionsmatrizen zur Verfügung gestellt werden. Die mit Cityscapes konformen Trainingsdaten für semantische Segmentierung erlauben noch dazu Experimente mit der Verfeinerung des Netzes durch Nachtraining und Auswertung der vom Netz produzierten Ergebnisse mit anderen Daten als dem Cityscapes-Datensatz.

5.3 COCO

Bei dem in [LMB⁺14] von Microsoft vorgestellten COCO (Common Objects in Context) handelt es sich um einen umfangreichen Datensatz für Instanz-Segmentierung. Es werden darin 328.000 Bilder mit insgesamt 2,5 Millionen annotierten Instanzen geboten. Das Bildmaterial für COCO stammt aus dem Internet, wobei nicht-ikonische Bilder, die Objekte aus untypischen Perspektiven zeigen, bevorzugt wurden. Da sich die Annotationen aufzählbare Objekte beschränken, ist der COCO-Datensatz für den Zweck dieser Arbeit nicht geeignet.



Abbildung 5.2: Beispiele aus dem KITTI-Datensatz für semantische Segmentation. Auf die Abbildung der Ground Truth wird verzichtet, da keine eingefärbte Version bereitgestellt wird und es sich nicht in erster Linie um einen Datensatz für Segmentierung handelt. Die Bilder sind von einer mobilen Plattform aufgenommen, die sich auf dem Fahrzeug befindet, weshalb die Perspektive sich von der Cityscapes-Daten unterscheidet.

5.4 Pascal VOC

Wie in [EVW⁺10] beschrieben, ist Pascal VOC (Visual Object Classes) eine Sammlung öffentlich zugänglicher Datensätze für Bildklassifizierung, Objekt-Detektion, Segmentation und Personen-Layout (Detektierung von Körperteilen), sowie ein jährlicher Wettbewerb in diesen Bereichen im Zeitraum von 2007 bis 2012. Die Hauptdisziplinen sind dabei Bildklassifizierung und Objekt-Detektion. Der letzte Stand von Pascal aus dem Jahr 2012 umfasst insgesamt 11.530 Bilder mit 27.450 Region-of-Interest-annotierten Objekten und 6.929 Segmenten aus 20 verschiedenen Klassen. Die Bilder für die Datensätze stammen aus dem Internet und sind ohne Präferenz hinsichtlich Bildqualität, Perspektive, Beleuchtung und Ähnlichem ausgewählt.

In [MCL⁺14] wird ein Datensatz vorgestellt, der den Anforderungen dieser Arbeit entspricht und eine valide Alternative zu Cityscapes darstellen könnte. Im gegebenen Kontext wird trotzdem Cityscapes verwendet wegen dessen stärkeren Praxisbezugs und konsequenteren Label-Politik. Dazu kommt, dass Pascal Daten aus allgemeinen Szenen anbietet, während sich Cityscapes auf den Straßenverkehr konzentriert und für dieses Thema leichter LiDAR-Daten gefunden werden können.

5.5 WildDash

WildDash [ZHM⁺18] ist ein Datensatz, der spezifisch für das Testen von Methoden zur Bildsegmentierung mit Szenen aus dem Straßenverkehr zusammengestellt ist. Ähnlich wie bei KITTI ist der Satz an annotierten Bildern nicht ausreichend, um einen adäquaten Trainingssatz zu bilden, ist aber konform mit Cityscapes und kann eine sinnvolle Erweiterung von dessen Trainingssatz darstellen.

Die Bilder in WildDash stammen von „YouTube-Autoren“ und sind nach speziellen Kriterien ausgewählt. Zu diesen gehört unter anderem, dass in den Daten Situationen gezeigt werden, die erfahrungsgemäß schwierig auszuwerten sind und potentiell Risiken darstellen, wie zum Beispiel Tiere auf der Fahrbahn und Tunnelausfahrten. Außerdem enthält der Datensatz negative Testfälle, also Bilder bei denen ein schlechtes Ergebnis erwartet wird, wie beispielsweise ein gedrehtes Bild.

Im Rahmen dieser Arbeit wird der KITTI-Datensatz als geeigneter angesehen, da darin auch Punktwolken gestellt und hauptsächlich städtische Szenen gezeigt werden. Außerdem erscheinen Testergebnisse auf dem WildDash-Datensatz weniger repräsentativ.

6 Experimente

6.1 Technische Daten des für die Experimente verwendeten Rechners

- Prozessor: Intel(R) Core(TM) i7-7700HQ CPU @ 2.860GHz
- Grafikkarte: NVIDIA GeForce GTX 1070
- Arbeitsspeicher: 16GB RAM

6.2 Backbones

Für die Durchführung folgender Experimente wird der fein annotierte Cityscapes-Datensatz verwendet. Die Netze werden auf dem aus 2975 Bildern bestehenden Trainingssatz trainiert und auf dem 500 Bilder fassenden Validierungssatz ausgewertet.

Für jede Trainingsepoke werden aus dem Trainingssatz 1487 Bilder zufällig ausgewählt. Als Grundlernrate wird 0.007 gewählt. Der Trainingsfehler wird über die in [lov] beschriebene Lovasz-Funktion berechnet. Weitere Hyperparameter sind eine Dropout-Rate von 0.1, eine L2 Regularisierungsrate von $4 * 10^{-5}$ und ein Momentum-Faktor von 0.9.

Die Netzwerkausgaben werden mittels der im Bereich Bildsegmentierung üblichen Intersection-over-union-Metrik (IoU) ausgewertet, die sich folgendermaßen berechnet:

$$IoU = \frac{TP}{TP + FP + FN} \quad (6.1)$$

Dabei steht TP für True Positives, also richtig erkannte Pixel, FP für False Positives und FN für False Negatives. Da in diesem System jedem Pixel eine gültige Klasse zugeordnet wird, sind FP und FN hier stets identisch.

6.2.1 MobileNetV2

Da Echtzeitfähigkeit eine wichtige Rolle spielt, konzentrieren sich die Experimente auf das leichtgewichtige MobileNetV2, statt dem leistungsfähigeren Xception65. Das erste Experiment beschäftigt sich mit dem Finden der idealen Trainingsdauer.

Epochen trainiert	1	5	10	15	25	30
IoU Straße	0.6260	0.6498	0.6727	0.6729	0.6774	0.6675
IoU Gehsteig	0.1835	0.3788	0.4330	0.4616	0.5193	0.5078
IoU Gebäude	0.5563	0.6402	0.6573	0.6771	0.6931	0.7068
IoU Mauer	0.0	0.0	0.0	0.0	0.0	0.0
IoU Zaun	0.0	0.0	0.0	0.0	0.0	0.0
IoU Pfahl	0.0791	0.1931	0.2159	0.2561	0.2835	0.2728
IoU Ampel	0.0	0.0	0.0	0.0	0.0	0.0
IoU Verkehrszeichen	0.0	0.1056	0.1785	0.1859	0.2233	0.2291
IoU Vegetation	0.5581	0.7153	0.7273	0.7518	0.7728	0.7663
IoU Gelände	0.0	0.0915	0.1274	0.1355	0.1593	0.1445
IoU Himmel	0.5353	0.6606	0.6867	0.6977	0.7153	0.7081
IoU Person	0.0	0.1160	0.1582	0.1458	0.1680	0.1863
IoU Radfahrer	0.0	0.0	0.0	0.0	0.0	0.0
IoU Auto	0.3537	0.5438	0.6014	0.5863	0.6419	0.6103
IoU Lastwagen	0.0	0.0	0.0	0.0	0.0	0.0
IoU Bus	0.0	0.0	0.0	0.0	0.0	0.0
IoU Zug	0.0	0.0	0.0	0.0	0.0	0.0
IoU Motorrad	0.0	0.0	0.0	0.0	0.0	0.0
IoU Fahrrad	0.0	0.0	0.0	0.0750	0.1340	0.1542
Durchschnitt IoU	0.1522	0.2155	0.2346	0.2444	0.2625	0.2607
Durchschnitt IoU ≠ 0	0.4131	0.4094	0.4458	0.4222	0.4534	0.4503

Tabelle 6.1: Die Tabelle zeigt die Bewertung von Models in Intersection-over-union (IoU) Metrik in Abhängigkeit der Trainingsdauer.

Wie in Tabelle 6.1 und Abbildung 6.1 zu erkennen ist, verbessert sich die Bewertung des Models bis zu einem Punkt, der in etwa bei Epoche 25 liegt und verschlechtert sich bei Verlängerung der Trainingsdauer wieder. Es tritt also trotz der L2 Regularisierung der Netzparameter und der Nutzung des Dropout-Verfahrens wahrscheinlich Overfitting auf. Im Folgenden wird das für 25 Epochen trainierte Netz verwendet. Die Bewertung dieses Models ist in Abbildung 6.2 dargestellt.

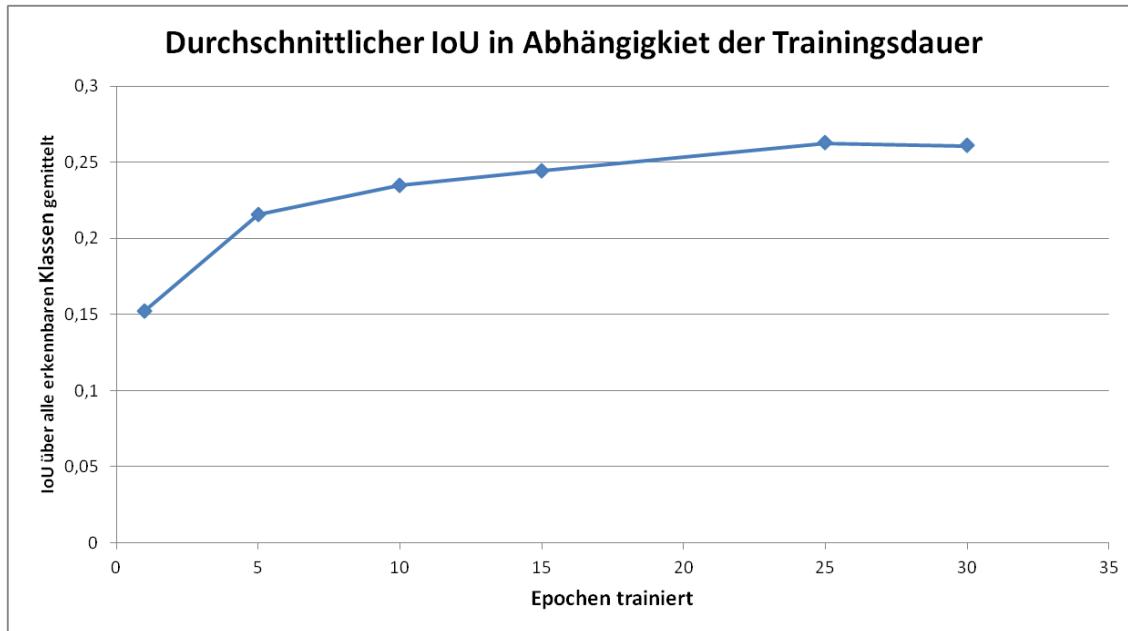


Abbildung 6.1: Durchschnittlicher IoU in Abhängigkeit der Anzahl trainierter Epochen. Der Graph erreicht sein Maximum bei 25 Epochen und fällt danach leicht ab, was auf Overfitting schließen lässt.

Die Models weisen durchwegs ihre höchsten Ergebnisse beim Erkennen der amorphen Klassen Straße, Gebäude, Vegetatio und Himmel auf, wie bei einem semantischen Segmentierungsverfahren zu erwarten. Das niedrige Ergebnisse bei der Klasse Gelände lässt sich dadurch erklären, dass der Cityscapes-Datensatz in städtischen Umgebungen aufgenommen wird, wo diese Klasse selten auftritt. Vergleichsweise hoch ist auch der IoU-Wert für die Klasse Auto, die in dem Datensatz besonders häufig ist.

Die niedrigsten positiven IoU-Werte weisen die Ergebnisse bei kleineren,zählbaren Objekten wie Personen, Pfähle, Fahrräder und Verkehrszeichen auf.

Die in Abbildung 6.2 dargestellten Ergebnisse lassen außerdem erkennen, dass das Netz bestimmte Klassen praktisch nicht erkennt. Dies lässt vermuten, dass es nur eingeschränkt fähig ist, Bildsegmente anhand ihres Kontextes zu bewerten und beispielsweise zwischen einem Fußgänger und einem Fahrradfahrer oder zwischen einer Mauer und einem Gebäude zu unterscheiden und die häufiger auftretende Variante auswählt. Die Ergebnisse der Klasse Fahrrad lassen vermuten, dass ein längeres Training dieses Verhalten verbessern könnte. Da ein zu langes Training sich, wie vorher erwähnt, negativ auf den durchschnittlichen IoU auswirkt, wird in diesem Experiment aber davon abgesehen. Eine weitere Möglichkeit wäre, dem Trainingssatz mehr Daten hinzuzufügen, die vermehrt die entsprechenden Objekte enthalten.

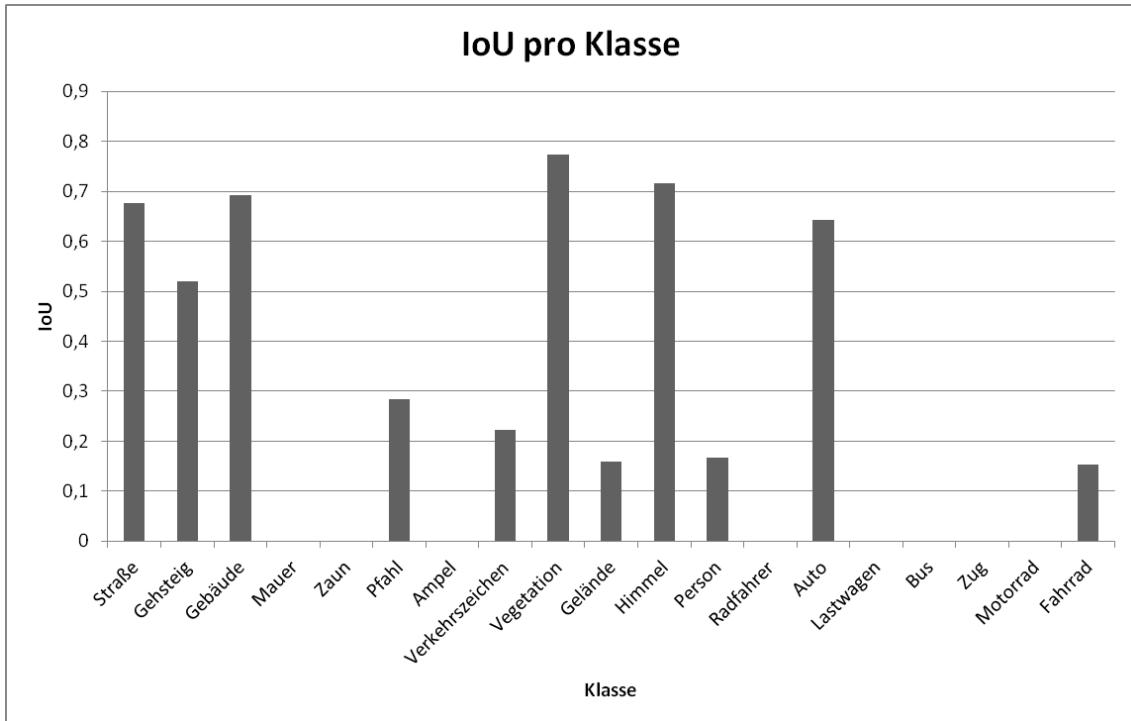


Abbildung 6.2: IoU für jede Klasse vom Test des für 25 Epochen trainierten Netzes. Das Diagramm zeigt gute Ergebnisse für amorphe Klassen und schlechtere für kleine Objekte. Zu sehen ist auch, dass einige Klassen, die in den Trainingsdaten selten vorkommen oder Ähnlichkeit mit anderen Klassen aufweisen, nicht erkannt werden.

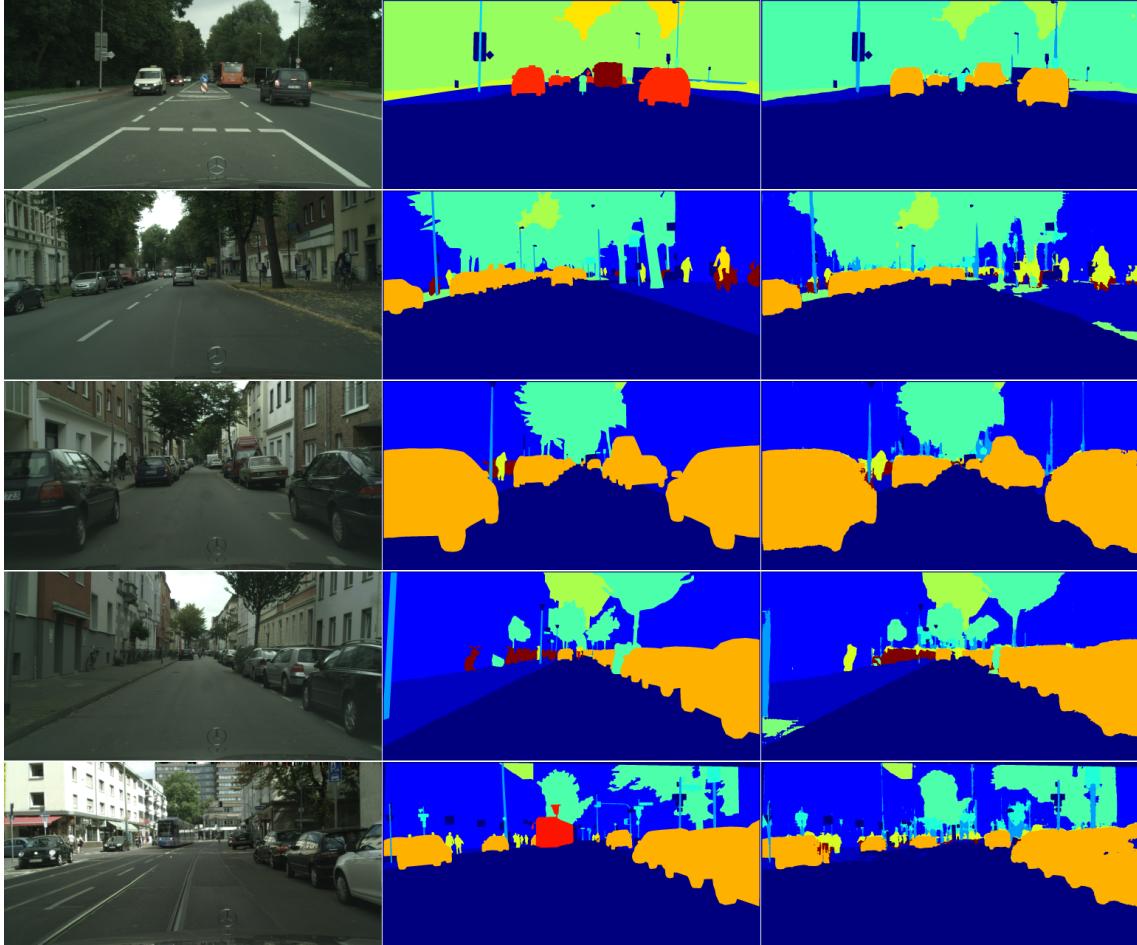


Abbildung 6.3: Zu sehen sind Beispiele für Ausgaben des Models mit der besten Bewertung im vorherigen Experiment (25 Epochen trainiert).
Von links nach rechts gezeigt sind Eingabebild, Ground Truth, Ausgabe von DeepLab.

Abbildung 6.3 zeigt ausgewählte Ausgaben des für 25 Epochen lang trainierten Netzes, das die besten Ergebnisse in der IoU-Metrik liefert. Sie spiegeln die Bewertung aus Tabelle 6.1 wieder, mit hoher Genauigkeit bei großen und amorphen Objekten und niedriger bei kleineren, zählbaren. Besonders auffällig sind False Positives der Klasse Person, die das Model oft an Fahrräder oder Bäume in der Nähe von Personen vergibt. Die Beispiele lassen erkennen, dass das Netz teilweise die Trainingsdaten memorisiert. So wird dem oberen Teil eines Fahrrades häufig die Klasse Person zugewiesen und umgekehrt der untere Teil einer Person als Fahrrad erkannt. Genauso werden Bereiche, die sich über einem als Straße erkannten Segments befinden tendenziell eher als Auto klassifiziert. Das lässt sich auf die Funktion des CRF zurückführen.

6.2.2 Xception65

Wir betrachten den Xception65-Backbone im Vergleich zu MobileNetV2. Auf ein Experiment mit unterschiedlichen Trainingsepochen wird an dieser Stelle aufgrund der langen Trainingsdauer der Netze verzichtet. Das verwendete Model ist 60 Epochen lang trainiert mit denselben Hyperparametern wie die Models mit MobileNetV2.

Wie in Tabelle 6.2 und Abbildung 6.4 zu sehen ist, erzeugt das Model mit Xception65 durchwegs bessere Ergebnisse als das mit MobileNetV2, benötigt aber im Durchschnitt 97% mehr Zeit für die Verarbeitung eines Eingabebildes.

Durch die Verwendung von Xception65 ist das Model außerdem in der Lage, alle Klassen des Datensatzes zu erkennen, auch wenn die IoU-Werte für die von MobileNetV2 nicht erkannten Klassen vergleichsweise niedrig sind. Eine besonders große Steigerung der Bewertung im Vergleich mit MobileNetV2 weist das Netz bei kleineren, zählbaren Objekten auf wie Personen (311%) und Verkehrszeichen (295%) auf.

Abbildung 6.5 zeigt beispielhafte Ergebnisse des Models aus dem Validierungs-Datensatz von Cityscapes.

6.3 Verfeinerung mit KITTI

In diesem Experiment wird das am besten bewertete, mit Cityscapes trainierte Model mit MobileNetV2 als Backbone mehrere Epochen mit den 200 Bildern umfassenden Trainingsdaten für semantische Segmentierung von KITTI nachtrainiert. Die Resultierenden Models werden anschließend mit anderen Bildern aus dem KITTI-Datensatz getestet. Da KITTI nur für diese 200 Bilder eine Ground Truth zur Verfügung stellt und ein Test auf den Trainingsdaten wenig aussagekräftig ist, begnügt sich das Expe-

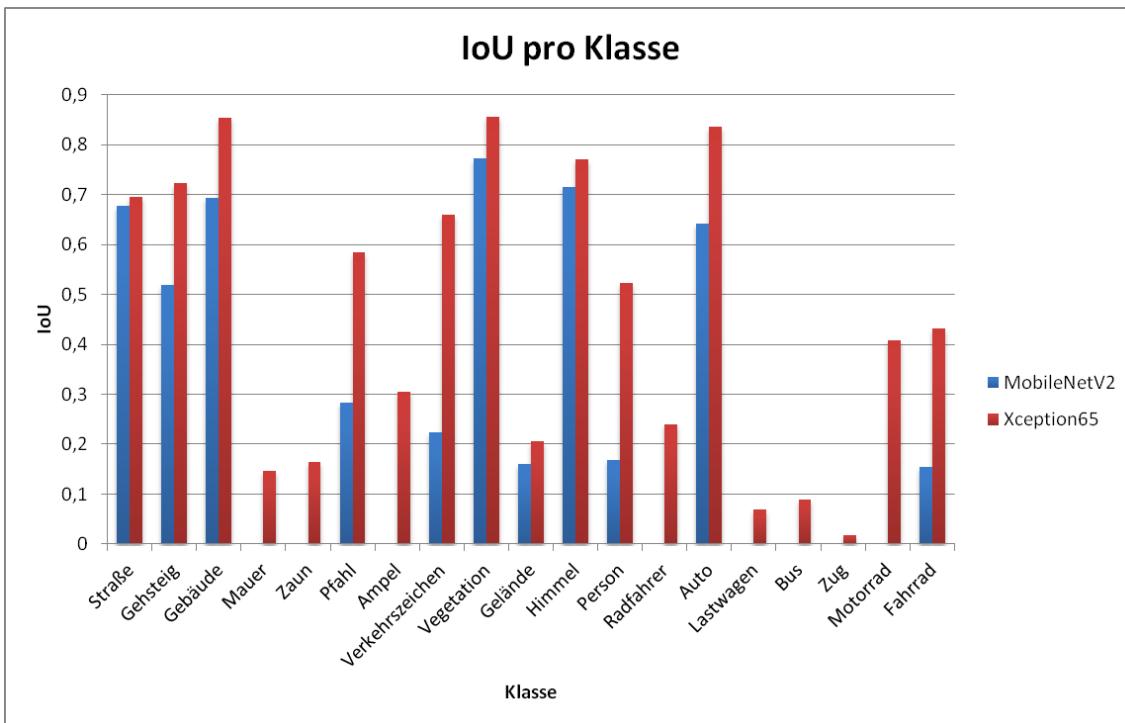


Abbildung 6.4: Der Vergleich von Xception65 und MobileNetV2 in IoU-Metrik zeigt, dass Xception die besten Ergebnisse bei denselben Klassen wie MobileNetV2 produziert, aber deutlich besser beim Erkennen von Objekten und Details ist.

	Xception65	MobileNetV2	Differenz
IoU Straße	0.6951	0.6774	0.0177
IoU Gehsteig	0.7238	0.5193	0.2045
IoU Gebäude	0.8533	0.6931	0.1602
IoU Mauer	0.1461	0.0	0.1461
IoU Zaun	0.1641	0.0	0.1641
IoU Pfahl	0.5843	0.2835	0.3008
IoU Ampel	0.3049	0.0	0.3049
IoU Verkehrszeichen	0.6592	0.2233	0.4359
IoU Vegetation	0.8558	0.7728	0.0830
IoU Gelände	0.2068	0.1593	0.0475
IoU Himmel	0.7707	0.7153	0.0554
IoU Person	0.5234	0.1680	0.3554
IoU Radfahrer	0.2386	0.0	0.2386
IoU Auto	0.8369	0.6419	0.1950
IoU Lastwagen	0.0691	0.0	0.0691
IoU Bus	0.0893	0.0	0.0893
IoU Zug	0.0185	0.0	0.0185
IoU Motorrad	0.0685	0.0	0.0685
IoU Fahrrad	0.4080	0.1340	0.4080
Durchschnitt IoU	0.4324	0.2625	0.2346
Durchschnittliche Verarbeitungszeit [ms]	764	387	377

Tabelle 6.2: Gezeigt ist ein Vergleich zwischen den Ergebnissen von DeepLab mit Xception65 und MobileNetV2 in IoU-Metrik und der Verarbeitungszeit.

periment in diesem Fall mit einer empirischen Auswertung.

Abbildung 6.6 zeigt beispielhaft je ein Bild aus beiden Datensätzen und die Ausgaben von DeepLab für verschieden lang nachtrainierte Models. Wie man sieht, tritt bei den Ergebnissen auf dem KITTI-Datensatz bereits nach einer Epoche Verfeinerung mit KITTI-Daten eine deutliche Verbesserung auf. Weitere Trainingsepochen verbessern die Ergebnisse weiter, aber weniger erheblich. Gleichzeitig verschlechtert sich die Ausgabe für ein Bild aus dem Cityscapes-Datensatz in ähnlichem Maße.

Das unterste Bild von Abbildung 6.6 zeigt die Ausgabe eines Models, das für 25 Epochen auf Cityscapes-Daten trainiert, für 5 Epochen mit KITTI-Daten verfeinert und anschließend für eine Epoche mit Cityscapes-Daten nachtrainiert ist. Wie man sieht, führt die Verfeinerung mit dem ursprünglichen Datensatz wieder zu einer Verschlechterung der Ergebnisse auf KITTI-Bildern und einer Verbesserung auf Cityscapes-Bildern.

Trainingsverhalten	Durchschnittlicher IoU auf Cityscapes-Bildern	Durchschnittlicher IoU auf KITTI-Bildern
25 Epochen mit Cityscapes-Daten	0.2625	0.1681
24 Epochen mit gemischten Daten	0.2605	0.2002
24 Epochen mit gemischten Daten + 1 Epoche mit KITTI-Daten	0.2045	0.2350
24 Epochen mit gemischten Daten + 1 Epoche mit KITTI-Daten + 1 Epoche mit gemischten Daten	0.2718	0.2105

Tabelle 6.3: Die Tabelle zeigt die Bewertung von Netzwerken mit unterschiedlich verfeinerten Trainingsdaten und -verhalten auf Evaluierungssätzen von Cityscapes und KITTI.

Diese Beobachtungen bestärken die Vermutung, dass beim Trainieren des Netzes Overfitting auftritt.

In einem nächsten Schritt soll untersucht werden, wie die Berücksichtigung der KITTI-Daten beim gesamten Lernprozess die Ergebnisse auf beiden Datensätzen beeinflusst. Dazu wird der annotierte KITTI-Datensatz geteilt in einen Trainings- und Evaluierungssatz zu je 100 Bildern. Dadurch wird es möglich, eine Aussagekräftige Bewertung in IoU-Metrik zu berechnen. Aufbauend auf dem vorherigen Experiment wird ein Model mit gemischten Daten für 24 Epochen trainiert und ausgewertet. Danach wird es für eine Epoche ausschließlich mit KITTI-Daten verfeinert und dann für eine weitere Epoche mit dem gesamten, gemischten Trainingssatz nachtrainiert. Damit ist das letzte Model 25 Epochen auf den gemischten Daten trainiert und hat damit genauso viele Lernschritte auf Cityscapes-Bilder absolviert wie das am besten Bewertete Model von Abschnitt ???. Die Ergebnisse sind in Tabelle ?? eingetragen.

6.4 Aufgetretene Probleme und Lösungen

6.4.1 False Positives

Models, die MobileNetV2 benutzen neigen dazu, einige Klassen zum Großteil mit einer anderen zu klassifizieren. Beispiele dafür sind die bereits angesprochene Erkennung von Fahrrädern als Personen und die Klassifizierung großer Fahrzeuge als Gebäude. Bei Verwendung des Xception65-Backbones kommt es reproduzierbar bei bestimmten Bildern zu einem Phänomen, bei dem eine große Anzahl Pixel nahe einer der Ecken als Straße beziehungsweise die Klasse mit dem niederwertigsten Label klassifiziert wird. Es besteht kein erkennbarer Zusammenhang zwischen den Bildern, bei denen dieses Verhalten auftritt. Abbildung 6.7 zeigt Beispiele dafür.

6.4.2 Overfitting

Overfitting hat sich als eines der hauptsächlichen Probleme bei der Verwendung von MobileNetV2 herausgestellt. Schon kleine Änderungen in der Perspektive, wie sie beispielsweise im KITTI-Datensatz auftritt, verursachen eine spürbare Verschlechterung der Ergebnisse.

Um bessere Ergebnisse auf dem KITTI-Datensatz zu erzeugen, hat es sich als vorteilhaft herausgestellt, den Trainingssatz für semantische Segmentierung der KITTI-Daten beim Trainieren des Netzes mitzuberücksichtigen. Das Netz mit jenen Daten nachzutrainieren verbessert ebenfalls die Ergebnisse, kann aber wiederum zu Overfitting auf diesen Datensatz führen. Es kann hilfreich sein, nach dem Nachtraining noch eine Epoche mit allen verfügbaren Daten nachzutrainieren.

Weitere Möglichkeiten, gegen Overfitting vorzugehen, mit denen in dieser Arbeit aber nicht experimentiert wird, sind eine Erhöhung der Dropout-Rate und dem L2 Regularisierungsfaktor.

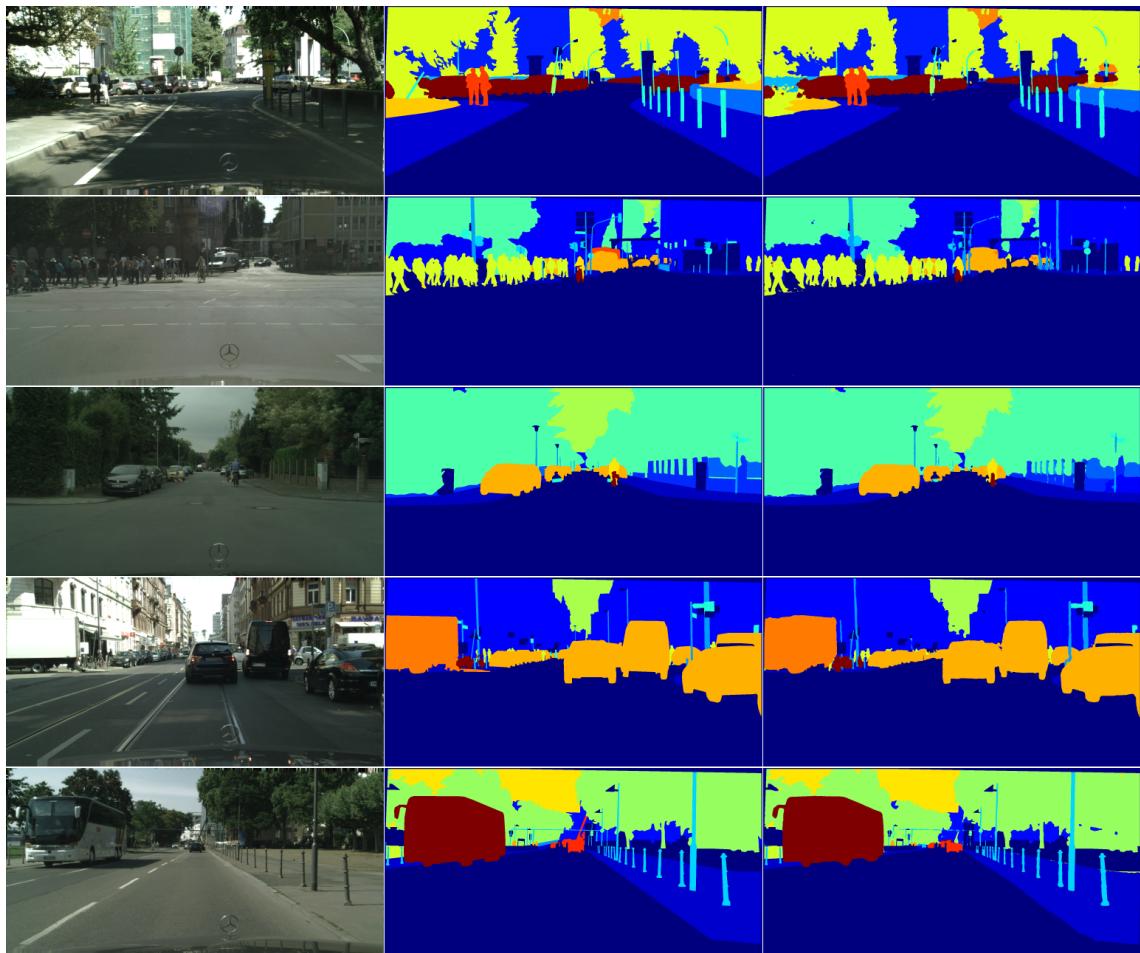


Abbildung 6.5: Gezeigt sind Beispiele für Ausgaben des Models, das Xception65 nutzt.

Von links nach rechts zu sehen ist Eingabebild, Ground Truth, Ausgabe von DeepLab.

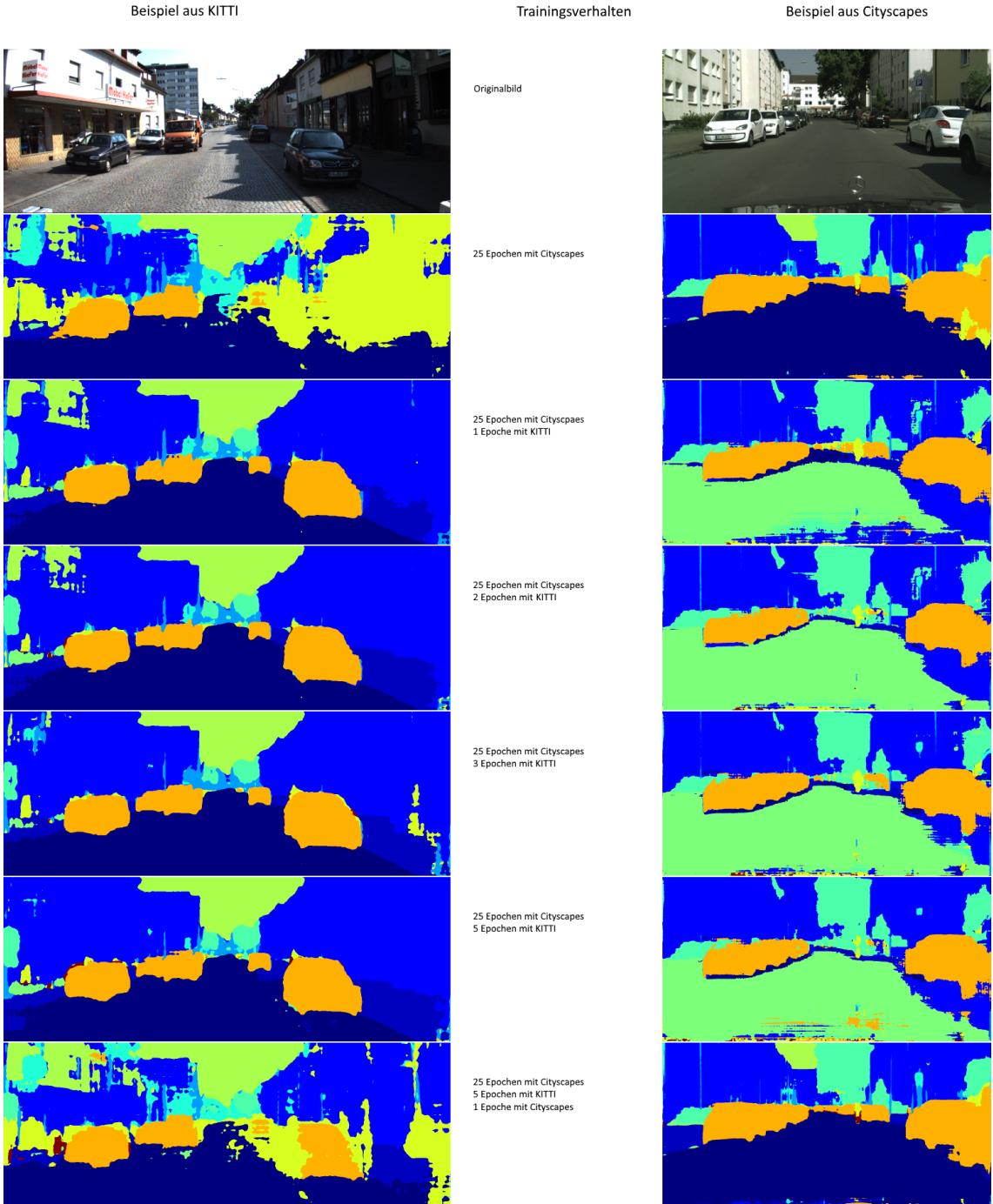


Abbildung 6.6: Beispiel für Ausgabe von mit KITTI-Daten verfeinerter Models. Die erste Epoche der Verfeinerung zeigt die größte Verbesserung auf den KITTI-Daten und gleichzeitig eine deutliche Verschlechterung bei Cityscapes-Bildern. Dieses Verhalten ist durch ein Nachtraining mit Cityscapes reversibel. Das Experiment zeigt, dass Overfitting ein Problem des Netzes ist.

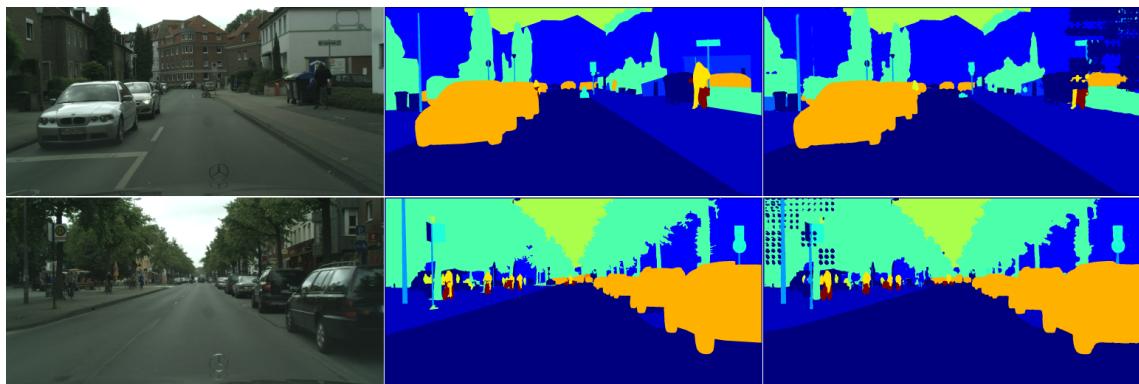


Abbildung 6.7: Hier zu sehen sind Beispiele für Ausgaben des Models mit Xception65, die das oben beschriebene Phänomen aufweisen. Von links nach rechts gezeigt sind Eingabebild, Ground Truth, Ausgabe von DeepLab. Das Verhalten ist auf den jeweiligen Bildern reproduzierbar.

7 Zusammenfassung

Ziel der Arbeit war die Entwicklung eines Systems zur semantischen Segmentierung von Bildern und Punktwolken mit neuronalen Netzen. Es soll also jedem Pixel eines Bildes, beziehungsweise jedem Punkt einer Punktwolke eine Klasse zugeteilt werden, ohne zwischen Instanzen vonzählbaren Objekten zu unterscheiden wie das bei Instanz- oder panoptischer Segmentierung der Fall wäre.

Als Netzarchitektur wurde das von Google entwickelte DeepLab verwendet. Dabei handelt es sich um ein Deep Fully-Convolutional Neural Network, das durch das Adaptieren von Atrous Convolution, Atrous Spatial Pyramid Pooling un Conditional Random Fields auf den Bereich der semantischen Bildsegmentierung angepasst ist.

Wegen der Anforderungen an die Laufzeit wurde als Backbone das leichtgewichtige MobileNetV2 gewählt, das in Experimenten mit dem Leistungsfähigeren Xception65 verglichen wurde. Beide Netze sind als Residual Networks strukturiert, was bedeutet, dass über so genannte Rasidual Connections Daten über mehrere Verarbeitungsschichten hinweg unverändert weitergeleitet und auf die Ergebnisse der übersprungenen Schichten addiert werden. Dadurch wird verhindert, dass ein Hinzufügen weiterer Schichten die Ergebnisse verschlechtert. Eine weitere Technik, die beide Architekturen verwenden ist Depthwise Separable Convolution, bei der zuerst eine räumliche und dann eine dimensionsübergreifende Faltung durchgeführt wird. MobileNetV2 zeichnet sich durch die Verwendung von Bottleneck-Blöcken aus. Darin werden die Daten zuerst Expandiert, dann Komprimiert, um Speicherplatz zu sparen.

Da diese Arbeit das Ziel hat, Algorithmen für Autonomes Fahren zu unterstützen und verbessern, wurde die Implementierung mit Hilfe der Datensätze Cityscapes und KITTI ausgewertet, die explizit für diesen Bereich konzipiert sind. Dabei dienen die 5000 fein annotierten Bilder für Segmentierung von Cityscapes zu Training und Evaluierung des Netzes. Der KITTI-Datensatz liefert Laserscans beziehungsweise Punktwolken und Bilder, die mit kalibrierten Kameras aufgenommen wurden, was es ermöglicht, die auf den Bildern erkannten Labels durch Berücksichtigung der Projektionsmatrix auf die zugehörigen Punktwolken zu projizieren.

Die ideale Trainingsdauer von 25 Epochen wurde für ein Netz mit MobileNetV2 experimentell ermittelt. Die Ergebnisse der verschiedenen Models wurden dafür in IoU-Metrik bewertet. Das Netzwerk erzielte dabei die besten Ergebnisse beim Erkennen amorpher Objekte, schlechtere beim Erkennen von Details im Bild. Der Vergleich mit einem Xception65-Model ergab, wie zu erwarten war, dass Xception bessere Ergebnisse

bei längerer Rechenzeit liefert. Als größtes Problem stellte sich Overfitting heraus, wie ein Versuch mit Verfeinerung mit Trainingsdaten aus dem KITTI-Datensatz zeigt. Bereits nach einer Trainingsepoch mit einem vergleichsweise kleinen Datensatz verschlechterten sich die Ergebnisse auf dem Cityscapes-Datensatz merklich. Zukünftige Experimente könnten zu Ziel haben, die Hyper-Parameter wie Dropout- und Regularisierungs-Rate zu anzupassen. Auch das Hinzufügen und Entfernen von Verarbeitungsschichten im Netzwerk könnte eine Möglichkeit sein, die Ergebnisse zu verbessern. Zur Verbesserung der Laufzeit könnten Bilder mit unterschiedlicher Auflösung getestet werden.

Literaturverzeichnis

- [COR⁺16] Cordts, Marius; Omran, Mohamed; Ramos, Sebastian u. a.: *The Cityscapes Dataset for Semantic Urban Scene Understanding*. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [crf] Sutton, Charles und McCallum, Andrew: *An Introduction to Conditional Random Fields*. URL: <https://arxiv.org/pdf/1011.4088.pdf> (zuletzt besucht am 02.09.2019).
- [dl1] Chen, Liang-Chieh; Papandreou, George; Kokkinos, Iasonas; Murphy, Kevin und Yuille, Alan L.: *Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs*. URL: <https://arxiv.org/pdf/1412.7062.pdf> (zuletzt besucht am 24.07.2019).
- [dl2] Chen, Liang-Chieh; Papandreou, George; Kokkinos, Iasonas; Murphy, Kevin und Yuille, Alan L.: *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*. URL: <https://arxiv.org/pdf/1606.00915.pdf> (zuletzt besucht am 24.07.2019).
- [EVW⁺10] Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J. und Zisserman, A.: *The Pascal Visual Object Classes (VOC) Challenge*. In: *International Journal of Computer Vision*, Band 88, Nr. 2, Seiten 303–338, 2010.
- [fpn] Lin, Tsung-Yi; Dollár, Piotr; Girshick, Ross B. u. a.: *Feature Pyramid Networks for Object Detection*. URL: <https://arxiv.org/pdf/1612.03144.pdf> (zuletzt besucht am 09.09.2019).
- [Fus06] Fusiello, Andrea: *Elements of Geometric Computer Vision*. 2006.
- [GBC16] Goodfellow, Ian; Bengio, Yoshua und Courville, Aaron: *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [GM10] Girisha, R. und Murali, S.: *Object Segmentation from Surveillance Video Sequences*. In: *Proceedings - 1st International Conference on Integrated Intelligent Computing, ICIIC 2010*, 2010.
- [GW08] Gonzalez, Rafael C. und Woods, Richard E.: *Digital image processing*. Prentice Hall, Upper Saddle River, N.J., 2008.

- [GWP18] Goel, Vik; Weng, Jameson und Poupart, Pascal: *Unsupervised Video Object Segmentation for Deep Reinforcement Learning*. In: *CoRR*, Band abs/1805.07780, 2018.
- [HGH⁺12] Hartmann, Glenn; Grundmann, Matthias; Hoffman, Judy u. a.: *Weakly Supervised Learning of Object Segmentations from Web-Scale Video*. In: Fusillo, Andrea; Murino, Vittorio und Cucchiara, Rita (Herausgeber): *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Seiten 198–208, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [HZ03] Hartley, Richard und Zisserman, Andrew: *Multiple View Geometry in Computer Vision*. 2. Auflage, Cambridge University Press, New York, NY, USA, 2003.
- [HZRS15] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing und Sun, Jian: *Deep Residual Learning for Image Recognition*. In: *CoRR*, Band abs/1512.03385, 2015.
- [kit] Geiger, Andreas; Lenz, Philip; Stiller, Christoph und Urtasun, Raquel: *Vision meets Robotics: The KITTI Dataset*. URL: <http://www.cvlibs.net/publications/Geiger2013IJRR.pdf> (zuletzt besucht am 14.08.2019).
- [LMB⁺14] Lin, Tsung-Yi; Maire, Michael; Belongie, Serge J. u. a.: *Microsoft COCO: Common Objects in Context*. In: *CoRR*, Band abs/1405.0312, 2014.
- [LMP01] Lafferty, John D.; McCallum, Andrew und Pereira, Fernando C. N.: *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In: *Proceedings of the Eighteenth International Conference on Machine Learning*, Seiten 282–289. ICML ’01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.
- [lov] Berman, Maxim; Triki, Amal Rannen und Blaschko, Matthew B.: *The Lovasz-Softmax loss: A tractable surrogate for the optimization of the ‘intersection-over-union’ measure in neural networks*. URL: https://zpzascal.net/cvpr2018/Berman_The_LovaSz-Softmax_Loss_CVPR_2018_paper.pdf (zuletzt besucht am 05.08.2019).
- [McC03] McCallum, Andrew: *Efficiently Inducing Features of Conditional Random Fields*. In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, Seiten 403–410. UAI’03, Morgan Kaufmann Publishers Inc., Acapulco, Mexico, 2003.
- [MCL⁺14] Mottaghi, Roozbeh; Chen, Xianjie; Liu, Xiaobai u. a.: *The Role of Context for Object Detection and Semantic Segmentation in the Wild*. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

- [mn2] Sandler, Mark; Howard, Andrew; Zhu, Menglong; Zhmoginov, Andrey und Chen, Liang-Chieh: *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. URL: <https://arxiv.org/pdf/1801.04381.pdf> (zuletzt besucht am 05.09.2019).
- [NR19] Nivaggioli, Adrien und Randrianarivo, Hicham: *Weakly Supervised Semantic Segmentation of Satellite Images*. In: *CoRR*, Band abs/1904.03983, 2019.
- [pnet] Qi, Charles R.; Su, Hao; Mo, Kaichun und Guibas, Leonidas J.: *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*. URL: <https://arxiv.org/pdf/1612.00593.pdf> (zuletzt besucht am 12.08.2019).
- [rcnn] He, Kaiming; Gkioxari, Georgia; Dollár, Piotr und Girshick, Ross: *Mask R-CNN*. URL: <https://arxiv.org/pdf/1703.06870.pdf> (zuletzt besucht am 12.08.2019).
- [RFB15] Ronneberger, Olaf; Fischer, Philipp und Brox, Thomas: *U-Net: Convolutional Networks for Biomedical Image Segmentation*. In: *CoRR*, Band abs/1505.04597, 2015.
- [RN93] Ryan, Matthew S. und Nudd, Graham R.: *The Viterbi Algorithm*. Techn. Ber. Coventry, UK, UK, 1993.
- [STB⁺19] Sharma, Suvash; Tang, Bo; Ball, John u. a.: *Semantic Segmentation with Transfer Learning for Off-Road Autonomous Driving*. In: *Sensors*, Band 19, 2019.
- [ups] Xiong, Yuwen; Liao, Renjie; Zhao, Hengshuang u. a.: *UPSNNet: A Unified Panoptic Segmentation Network*. URL: <https://arxiv.org/pdf/1901.03784.pdf> (zuletzt besucht am 24.07.2019).
- [VWLT11] Vineet, Vibhav; Warrell, Jonathan; Ladický, L und Torr, Philip: *Human Instance Segmentation from Video using Detector-based Conditional Random Fields*. In: *BMVC 2011 - Proceedings of the British Machine Vision Conference 2011*, 2011.
- [xce] Chollet, François: *Xception: Deep Learning with Depthwise Separable Convolutions*. URL: <https://arxiv.org/pdf/1610.02357.pdf> (zuletzt besucht am 05.09.2019).
- [YLX⁺19] Yao, Rui; Lin, Guosheng; Xia, Shixiong; Zhao, Jiaqi und Zhou, Yong: *Video Object Segmentation and Tracking: A Survey*. In: *CoRR*, Band abs/1904.09172, 2019.

- [ZHM⁺18] Zendel, Oliver; Honauer, Katrin; Murschitz, Markus; Steininger, Daniel und Fernandez Dominguez, Gustavo: *WildDash - Creating Hazard-Aware Benchmarks*. In: *The European Conference on Computer Vision (ECCV)*, 2018.