

Generierung künstlicher Trainingsdaten für die Straßenschilderkennung in Fahrzeugen mittels Generative Adversarial Networks

Studienarbeit

Studiengang Informatik

an der Dualen Hochschule Baden-Württemberg Stuttgart

von

Frederik Esau

09.06.2023

Bearbeitungszeitraum
Matrikelnummer, Kurs
Betreuer

24.10.2022 - 09.06.2023
6526552, TINF20ITA
Prof. Dr. Monika Kochanowski

Erklärung

Ich versichere hiermit, dass ich meine Studienarbeit mit dem Thema: *Generierung künstlicher Trainingsdaten für die Straßenschilderkennung in Fahrzeugen mittels Generative Adversarial Networks* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Werther, 09.06.2023



Frederik Esau

Abstract

A traffic sign detection can nowadays be considered a common part of automotive vehicles. This is underpinned by the fact that, from the year 2024 on, an EU norm makes such systems mandatory for newly manufactured cars. Modern traffic sign detection software usually leverages machine learning techniques. Resultingly, images of real traffic signs are necessary for the development of such software. Various datasets exist for this purpose which contain traffic signs from different countries. One property of those datasets is that they can be significantly imbalanced. A reason for this the varying commonness of different traffic sign categories. Artificially generating such traffic sign images is a possible solution to mitigate this. Some papers already implement this. They base their machine learning models on recent types of generative modelling.

This work implements a similar approach. It uses a Cycle-Consistent Generative Adversarial Network to generate images of specific german traffic signs. In contrary to a previous publication, this work uses a Cycle-Consistent Generative Adversarial Network with a U-Net architecture and compares it to a ResNet architecture. The results show that the U-Net based model has a more stable training. Furthermore, a traffic sign classification model shows better performance on the test set when trained with images generated by the U-Net based model than with images generated by the ResNet based model. The question whether this only applies to this specific dataset and model implementation or whether this is a generalizable result is outside the scope of this work. The U-Net-based model is also faster to train, but it has about four times more parameters than the ResNet-based model.

In addition to this, this work implements functions for artificially augmenting generated traffic sign images. This should simulate external influences such as weather conditions or invalid traffic signs. The idea being that future research can enhance these augmentation techniques. This could be used to simulate edge cases for traffic sign detection software and thus reduce the need for real world data. The implemented augmentations include snow, motion blur and traffic signs that are marked as invalid.

Inhaltsverzeichnis

Abkürzungsverzeichnis	V
Abbildungsverzeichnis	VI
Tabellenverzeichnis	VIII
Listings	IX
1 Einleitung	1
1.1 Problemstellung	1
1.2 Vorgehensweise	2
1.3 Ziel der Arbeit	2
2 Stand der Technik	3
2.1 Straßenschilderkennung	3
2.2 Künstliche Neuronale Netze	5
2.2.1 Training	7
2.2.2 Convolutional Neural Networks	10
2.3 Bildgenerierung mit Künstlichen Neuronalen Netzen	13
2.3.1 Mathematischer Hintergrund	14
2.3.2 Pixel Recurrent Neural Networks	15
2.3.3 Autoencoder	17
2.3.4 Generative Adversarial Networks	19
2.4 Vorherige Arbeiten	27
2.4.1 Generierung Taiwanischer Straßenschilder mittels DCGAN	27
2.4.2 Generierung Deutscher Straßenschilder mittels CycleGAN	28
2.5 Machine Learning Frameworks	30
3 Konzeption des Modells	32
3.1 Datensatz	32
3.2 Framework	36
3.3 Architektur	37
3.4 Datenaugmentation	38
3.5 Training	39
4 Implementierung und Training	42
4.1 Modell	42
4.1.1 Konstruktor	43
4.1.2 fit-Methode	45
4.1.3 generate-Methode	47

4.2	Datenaugmentierung	47
4.2.1	Skalierung	47
4.2.2	Rotation	49
4.3	Training	50
4.3.1	Laden der Datensätze	50
4.3.2	Ausführen des Trainings	52
4.3.3	Logging	52
4.4	Trainingsergebnisse	52
4.4.1	U-Net	53
4.4.2	ResNet	54
4.5	Generierung	56
5	Augmentation der generierten Bilder	58
5.1	Bewegungsunschärfe	58
5.2	Ungültige Straßenschilder	60
5.3	Schnee	62
6	Evaluation	64
6.1	Evaluation der Generierung	64
6.1.1	Vorgehen	64
6.1.2	Ergebnisse	66
6.2	Evaluation der Augmentierung	68
6.3	Verbesserungsmöglichkeiten	69
7	Zusammenfassung	70
Anhang		76

Abkürzungsverzeichnis

CNN	Convolutional Neural Network
cGAN	Conditional Generative Adversarial Network
CycleGAN	Cycle-Consistent Generative Adversarial Network
DHBW	Duale Hochschule Baden-Württemberg
engl.	Englisch
GAN	Generative Adversarial Network
GPU	Grafikkarte (engl.: Graphics Processing Unit)
GTSRB	German Traffic Sign Recognition Benchmark
KNN	Künstliches Neuronales Netz
PixelRNN	Pixel Recurrent Neural Network
ResNet	Residual Neural Network
SSIM	Index struktureller Ähnlichkeit (engl.: Structural Similarity Index)
SVM	Support Vector Machine
TOML	Tom's Obvious Minimal Language
VAE	Variational Autoencoder

Abbildungsverzeichnis

2.1	Erschwerende Einflüsse auf die Straßenschilderkennung [4] [7]	4
2.2	Weitere mögliche Einflüsse auf die Straßenschilderkennung [4] [8]	4
2.3	Einzelnes Neuron eines Künstliches Neuronales Netzs (KNNs) [12]	6
2.4	Vollständiges KNN (<i>angelehnt an [12]</i>)	7
2.5	Beispiel für eine Faltung eines Convolutional Neural Network (CNN) (<i>angelehnt an [14]</i>)	12
2.6	Beispiel für eine CNN-Architektur [15]	13
2.7	Beispielhafter Vergleich von $\hat{p}(x)$ und $p(x)$ [17]	14
2.8	Taxonomie generativer Modelle <i>angelehnt an [20]</i>	15
2.9	Bestimmung von $\hat{p}(x)$ mit PixelRNNs [21]	16
2.10	Architektur eines Autoencoders (<i>angelehnt an [23]</i>)	18
2.11	Zusammenspiel zwischen Generator und Diskriminatior	20
2.12	Residual Block eines ResNet [30]	24
2.13	Architektur eines U-Net [32]	26
2.14	Beispielergebnisse der Generierung taiwanischer Schilder [33]	27
2.15	Beispielergebnisse der Generierung deutscher Schilder [34]	29
3.1	Beispielbilder aus dem GTSRB Datensatz [4]	32
3.2	Häufigkeitsverteilung der Klassen von Straßenschildern im präparierten German Traffic Sign Recognition Benchmark (GTSRB)	33
3.3	Beispielbilder aus der chinesischen Traffic Sign Recognition Database [38] . .	34
3.4	Beispielbild aus dem Mapillary Datensatz [40]	35
3.5	Häufigkeitsverteilung der Kategorien von Straßenschildern im präparierten Datensatz	35
3.6	Domänen für die Bild-zu-Bild Übersetzung	38
3.7	Rotation der Straßenschilder mittels eulerscher Winkel	39
3.8	Trainingsschritte des Cycle-Consistent Generative Adversarial Network (CycleGAN) .	40
4.1	Modal Collaps des ResNet nach 200 Trainingsepochen	54
4.2	Positiv herausstechende Bilder des ResNets verschiedener Epochen	55
4.3	Negativ herausstechende Bilder des ResNets verschiedener Epochen	55
4.4	Beispielbilder des ResNets mit 6 Residual Blocks	56
5.1	Horizontale, vertikale und diagonale Faltmatrix	59
5.2	Kreuz auf transparentem Hintergrund, das Schilder als ungültig kennzeichnet	60

5.3	Trainingsverlauf des CycleGAN um Bilder mit Schnee zu erzeugen	63
5.4	Generierte Bilder des CycleGAN mit Schnee	63
6.1	Vorgehen der Evaluation	65

Tabellenverzeichnis

2.1	Vergleich der Objekterkennung mit und ohne künstliche Trainingsdaten	28
2.2	Vergleich der Klassifikation mit echten und künstlichen Trainingsdaten [34] . .	30
4.1	Auswahl an Methoden aus der CycleGAN Klasse	43
4.2	Vergleich von U-Net und ResNet	53
4.3	Kommandozeilenargumente des Skripts generate.py	56
6.1	Ergebnisse des Trainings eines VGG16 Klassifikators	67
6.2	Genauigkeit des Klassifikators auf augmentierten Bildern des U-Net-basierten CycleGAN	68

Listings

4.1	model.py - Auswahl der Generator-Architektur	44
4.2	model.py - Initialisierung der Generatoren	45
4.3	model.py - fit-Methode	46
4.4	model.py - generate-Methode	47
4.5	utils.preprocess_image.py - Skalieren der Bild-Tensoren	48
4.6	utils.preprocess_image.py - Zufällige Generierung der Rotationswinkel α_z	49
4.7	train.py - Laden des Trainingsdatensatzes	51
4.8	train.py - Laden des Trainingsdatensatzes	52
4.9	Beispieldaufrufe des Skripts generate.py	57
5.1	utils.image_augmentation.py - Hinzufügen einer diagonalen Bewegungsunschärfe	59
5.2	utils.image_augmentation.py - Schilder als ungültig markieren	61
1	Augmentierung eines Batches von Bildern	87
2	Vollständige Trainingsfunktion	89
3	Anpassen des VGG16 Modells für die Klassifikation von der Straßenschilder [1]	89

1 | Einleitung

Während in großen Teilen des letzten Jahrhunderts Innovationen in der Fahrzeugentwicklung vor allem im Bereich der mechanischen Leistung und Effizienz stattgefunden haben, erwartet man zukünftige Verbesserungen im Automobil besonders softwareseitig. Unter anderem im Bereich der Fahrerassistenzsysteme und bezüglich autonomer Fahrzeuge. [2]

Ein solches Fahrerassistenzsystem, das auch für autonome Fahrzeuge eine Rolle spielt, ist die automatische Erkennung von Straßenschildern. In nicht-autonomen Fahrzeugen unterstützt das System Fahrzeugführende, indem es auf geltende Verkehrsregeln aufmerksam macht. Das ist beispielsweise relevant, wenn der Fahrzeugführende Verkehrsschilder übersieht oder absichtlich missachtet. In autonomen Fahrzeugen ist Software zur Straßenschilderkennung eine der Grundlagen für die Navigation, da die Entscheidungen der Fahrzeug-Software unter anderem auf den durch Straßenschilder kommunizierten Verkehrsvorschriften basieren. [3]

1.1 Problemstellung

Systeme zur Straßenschilderkennung nutzen mitunter maschinelles Lernen. Entwickelnde führen der Software reale Bilder von Straßenschildern zu. Anhand dessen lernt das System, Bilder von Straßenschildern zu klassifizieren, die es zuvor nicht gesehen hat. Die Qualität der Klassifizierung hängt dabei unter anderem von der Menge und Qualität der Bilddaten ab. [3]

Aus diesem Grund existieren verschiedene Datensätze, die Bilder von Straßenschildern aus unterschiedlichen Ländern enthalten [4]. Eine Eigenschaft solcher Datensätze ist, dass sie ungleichmäßig verteilte Daten enthalten können. Das liegt daran, dass bestimmte Arten von Straßenschildern, wie etwa Geschwindigkeitsbegrenzungen, häufiger im Straßenverkehr vorkommen als andere. Ungleichmäßig verteilte Daten können dazu führen, dass das System bestimmte Arten von Schildern fehlerhaft klassifiziert. Weiterhin existieren verschiedene Grenzfälle, die die Qualität der Klassifizierung beeinträchtigen können. Dazu zählen unter anderem bestimmte Wetterbedingungen wie etwa Schnee und Nebel. Idealerweise sollte ein System zur Straßenschilderkennung auch mit selten auftretenden Grenzfällen trainiert werden. Dafür müssen die Datensätze Bilder enthalten, die solche Fälle zeigen. Das ist allerdings, je nach Auftrittswahrscheinlichkeit des Grenzfalls, mit einem erhöhten Aufwand verbunden.

Diese Studienarbeit soll deshalb eine Methode implementieren, die es ermöglicht, Trainingsbilder für die Straßenschilderkennung künstlich zu erzeugen. Solche computergenerierten

Bilder sollen die Qualität der Klassifizierung verbessern, indem sie die Datensätze um weitere Bilder ergänzen. Auch kann hierdurch die Verteilung der Trainingsdaten gezielt gesteuert und ausbalanciert werden. Außerdem soll die Implementierung dieser Studienarbeit verschiedene Grenzfälle der Straßenschilderkennung simulieren.

1.2 Vorgehensweise

Zunächst prüft diese Arbeit den aktuellen Stand der Technik von serienmäßig eingesetzter Software zur Straßenschilderkennung. Insbesondere wird darauf Wert gelegt, unter welchen Bedingungen die Straßenschilderkennung fehlerhaft arbeitet. Das sind Fälle, die für diese Studienarbeit eine primäre Rolle spielen.

Anschließend erfolgt ein Überblick, welche Methoden zur computergestützten Generierung von Bildern existieren. Relevant sind hier mitunter bereits existierende Veröffentlichungen, die sich mit der Generierung von Bildern für die Straßenschilderkennung beschäftigen.

Aufbauend darauf wird die Umsetzung der Studienarbeit entworfen und implementiert. Die anschließende Evaluation prüft die Qualität der generierten Bilder. Darauf basiert eine Bewertung, ob sich die Bilder als Trainingsdaten für die Straßenschilderkennung eignen können.

1.3 Ziel der Arbeit

Damit die Bilder einen Mehrwert für die Straßenschilderkennung bieten, müssen sie vor allem fotorealistisch sein. Das ist deshalb eines der konkreten Ziele, die die Evaluation prüft. Dafür erfolgt ein indirekter Vergleich von realen mit künstlich erzeugten Bildern. Außerdem soll die Studienarbeit zeigen, dass Bilder, die Grenzfälle simulieren, Algorithmen zur Straßenschilderkennung zu Fehlinterpretationen verleiten können. Es ist insbesondere eine quantitative Beurteilung relevant, damit die Implementierung dieser Studienarbeit mit anderen Arbeiten vergleichbar ist.

Die generierten Bilder sollen dabei das Symbol eines einzelnen Straßenschildes zeigen sowie eine geringfügige Menge an Hintergrund. Eine künstliche Generierung von mehreren Schildern pro Bild oder von vollständigen Fahrsituationen ist nicht Ziel dieser Arbeit.

2 | Stand der Technik

2.1 Straßenschilderkennung

Eine Straßenschilderkennung zählt mittlerweile zu der Standardausstattung vieler Neuwagen. Im Jahr 2024 tritt zudem eine EU-Verordnung in Kraft, durch die sämtliche neu produzierten Fahrzeuge mit einer solchen Funktion ausgestattet werden müssen [5]. Daran zeigt sich, dass das Thema bereits weitreichend etabliert ist.

Straßenschilder werden zu folgendem Zweck eingesetzt: Sie sollen Informationen über die Verkehrssituation und über geltende Vorschriften des Gebiets, in dem sich das Fahrzeug zu einem gegebenen Zeitpunkt befindet, präsentieren. Straßenschilder kommunizieren unter anderem Geschwindigkeitsvorgaben, Gefahrenhinweise und allgemeine Verkehrsregeln. Dabei sind die Schilder so konzipiert, dass sie sich visuell möglichst von ihrem Hintergrund abheben und leicht voneinander zu unterscheiden sind. Automatische Straßenschilderkennungen können Fahrzeugführende in Situationen unterstützen, in denen sie Schilder übersehen oder gezielt missachten. Anstelle dass ein reales Schild beispielsweise nur für einige Sekunden sichtbar ist, bevor es außerhalb der Sichtweite des Fahrzeugführenden ist, ist eine durchgehende Anzeige auf den Displays eines Fahrzeugs möglich. Auch akustische Warnungen oder ein aktives Eingreifen von Sicherheitssystemen sind denkbar, beziehungsweise bereits in Serienfahrzeugen vorhanden. [3]

Eine Straßenschilderkennung erfolgt visuell und wird somit mittels Kameras umgesetzt. Dabei lassen sich viele Schilder durch eine bestimmte Form (Kreis, Dreieck, Achteck, etc.) und ein Symbol (Schneeflocke, Person, etc.) oder eine Zahl identifizieren. Somit existieren verschiedene Arten von Straßenschildern, die durch den Erkennungsalgorithmus identifiziert werden. Die praktische Umsetzung solcher Algorithmen verwendet häufig CNNs, welche in Kapitel 2.2 thematisch aufgeführt sind. Besonders relevant für das Thema dieser Studienarbeit ist, wie zuverlässig die momentan ausgelieferten Straßenschilderkennungen sind und welche Situationen die Algorithmen am ehesten zu falschen Aussagen verleiten. Auf dieser Grundlagen kann sich orientiert werden, welche Arten von Bildern vermehrt generiert werden sollen, um die Straßenschilderkennung verbessern zu können. [3]

Im Jahr 2019 hat eine Automobilzeitschrift die Straßenschilderkennung von unterschiedlichen Fahrzeugherstellern getestet [6]. Zudem existieren verschiedene Publikationen, die sich mit der Thematik befassen [3]. Die Ergebnisse des Zeitschriftenartikels weisen darauf hin, dass

die Straßenschilderkennung einiger Fahrzeuge bereits weitreichend funktioniert. Die Systeme sind dazu in der Lage, Geschwindigkeitsvorgaben überwiegend zu erkennen und dem Fahrzeugführenden auf einem Display anzuzeigen. Auch existieren bereits akustische Warnungen bei einer Überschreitung der Höchstgeschwindigkeit. Dennoch gibt es einige Situationen, die bei mehreren Fahrzeugen zu Problemen bei der Straßenschilderkennung geführt haben. Eine Auswahl davon zeigt Abbildung 2.1. [6]



Abbildung 2.1: Erschwerende Einflüsse auf die Straßenschilderkennung [4] [7]

Fahrzeuge verschiedener Hersteller haben in den Tests Aufhebungsschilder nicht korrekt interpretiert. Damit sind Schilder gemeint, die entweder Geschwindigkeitsbegrenzungen oder Überholverbote außer Kraft setzen. Des Weiteren sorgten mittels Klebestreifen als ungültig erklärte Schilder, in einigen Fällen Dunkelheit, beispielsweise in Tunneln, und sogenannte LED Wechselverkehrszeichen für falsche Aussagen. Auch erkennen die Systeme in einigen Situationen nicht, dass Schilder für eine kreuzende Straße gelten, statt für die Straße, auf der sich das Fahrzeug zu dem Zeitpunkt befinden. Weitere Aspekte, die in dem Artikel nicht explizit genannt sind, aber laut einer Publikation von 2014 in der Vergangenheit zu Schwierigkeiten geführt haben, sind mitunter die folgenden: [3]



Abbildung 2.2: Weitere mögliche Einflüsse auf die Straßenschilderkennung [4] [8]

Eine weitere Publikation aus dem Jahre 2019 zeigt, dass die Qualität der Straßenschilderkennung von der Stärke der äußeren Einflüsse abhängt. Vergleichsweise geringfügig verdeckte Schilder hat das Testfahrzeug hier in der Regel korrekt klassifizieren können. Sobald eine größere Fläche des Schilds verdeckt ist oder das Schild vermehrt beschmutzt ist, kann das System die Schilder zum Teil nicht mehr identifizieren. Auch hier schreiben die Autoren, dass das Wetter und somit die Sichtverhältnisse einen negativen Einfluss auf die Qualität der Erkennung zeigen. [9]

Die Erkenntnisse der Tests aus den genannten Artikeln geben Hinweise auf das folgende: In einigen genannten Situationen können sich Fahrzeugführende nicht vollständig auf die

Straßenschilderkennung ihrer Fahrzeuge verlassen. Ein Ziel der Hersteller ist das Anbieten von vollständig autonomen Fahrzeugen. Damit das möglich ist, muss die Software der Fahrzeuge auch solche Grenzfälle korrekt interpretieren. Das erfordert eine gewisse Menge an Daten, durch die diese Algorithmen trainiert werden.

Ziel dieser Arbeit ist ausgehend davon, gezielt Trainingsbilder erzeugen zu können, die einige dieser Aspekte simulieren. Es soll als alternative Möglichkeit dazu vorgeschlagen werden, sämtliche Trainingsdaten für Grenzfälle eigenständig in realen Fahrsituationen aufzunehmen.

2.2 Künstliche Neuronale Netze

Sowohl Software zur Straßenschilderkennung als auch die Implementierung dieser Studienarbeit basiert auf künstlichen neuronalen Netzen (KNNs). Aus diesem Grund soll dieses Kapitel die Grundlagen des Themas erläutern.

Durch künstliche neuronale Netze (KNNs) können Maschinen lernen, bestimmte Probleme zu lösen, ohne dass ein Mensch vorher explizite Regeln dafür definieren muss. Dies steht im Kontrast zur Methode, der Maschine vorher einen festen, vollständigen Regelsatz bereitzustellen. Letztgenannter Ansatz zeigt in einigen Gebieten nur begrenzten Erfolg, da es für Menschen herausfordernd sein kann, Regelsätze für Vorgänge zu definieren, die unbewusst im Gehirn stattfinden oder viel Kontext erfordern. Zu nennen sind hierbei die visuelle Objekterkennung oder menschliche Sprache. Außerdem können neue, nicht in den Regeln beachtete Situationen dazu führen, dass die Maschine das Problem nicht mehr lösen kann. [10]

Die Grundidee hinter dem sogenannten *maschinellen Lernen* ist deshalb, dass sich die Maschine selber einen Wissensschatz aufbaut, der ihr beim Lösen des Problems hilft. Dies geschieht, indem die Entwickler ihr reale Trainingsbeispiele zeigen. Möchte man einen Algorithmus trainieren, der Schach spielen soll, kann man ihm beispielsweise eine Vielzahl an realen Schachpartien zeigen. Anhand dessen lernt der Algorithmus verschiedene Strategien und baut ein Spielverständnis auf, das womöglich über die menschlichen Fähigkeiten hinausgeht. KNNs bauen auf diesem Prinzip auf. Sie implementieren lernbare Funktionen, die eine Abbildung zwischen einer Eingabe und der zugehörigen Ausgabe herstellen. [10]

In den letzten Jahrzehnten erlebte das maschinelle Lernen und damit auch das Gebiet der KNNs einen Aufschwung. Es existiert bereits seit Mitte des vergangenen Jahrhunderts, wird allerdings erst durch die zunehmende Rechenleistung und die Verfügbarkeit von großen Datenmengen flächendeckend eingesetzt. Einsatzgebiete für KNNs sind unter anderem die Objekterkennung, das Verstehen von natürlicher Sprache und die Generierung von Text und Bildern. [11]

Die Inspiration für KNNs bildet die Informationsverarbeitung des Gehirns in Lebewesen. Die kleinste hier betrachtete Einheit ist das Neuron. Neuronen in KNNs sind konzeptionell inspiriert von realen, biologischen Neuronen, besitzen aber eine abstrahierte Funktionsweise. In KNNs berechnen sie ein Skalarprodukt ihrer gewichteten Eingangswerte, addieren einen sogenannten *Schwellenwert* (engl.: *Bias*) hinzu und wenden auf das Ergebnis eine nichtlineare Funktion an. Letztere wird auch als Aktivierungsfunktion bezeichnet. Diese Aktivierungsfunktion kann analog dazu gesehen werden, dass biologische Neuronen einen Grenzwert (engl.: *threshold*) besitzen, der überschritten werden muss, damit das Neuron *feuert*, also einen Impuls an weitere Neuronen weitergibt. Aktivierungsfunktionen sind notwendig, damit neuronale Netze Probleme lösen können, die über die Fähigkeiten einer linearen Regression hinausgehen. Dadurch können Neuronen eine nichtlineare Abhängigkeit zwischen dem Eingang x und dem Ausgang y umsetzen. [12]

In Abbildung 2.3 ist ein einzelnes künstliches Neuron eines KNNs darstellt. Das *Plus-Zeichen* steht für die Berechnung des Skalarprodukts der Eingänge und der darauf addierte Schwellenwert. Die Aktivierungsfunktion wird durch das *Sigmoid-Zeichen* symbolisiert. Der Ausgang (*rechts*) ist das Ergebnis der Berechnung des Neurons. [12]

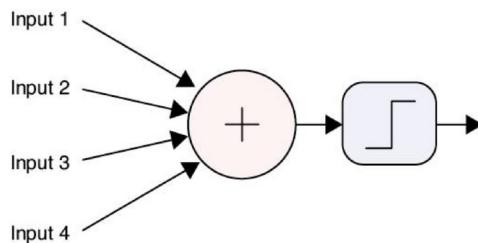


Abbildung 2.3: Einzelnes Neuron eines KNNs [12]

Um komplexe Probleme lösen zu können, besitzen KNN mehrere Neuronen, die miteinander verbunden und in Schichten angeordnet sind. Jede Schicht erhält die Ausgangswerte der vorherigen Schicht als Eingang und gibt die daraus berechneten, neuen Werte an die nächste Schicht weiter. Abbildung 2.4 zeigt ein vollständiges KNN. Die Neuronen sind durch Kreise dargestellt und ihre Verbindungen durch Pfeile. [11]

Eine Verbindung stellt dar, dass ein Neuron seinen berechneten Wert an das nachfolgende Neuron weitergibt. Dies geschieht hierbei ausschließlich von *links nach rechts*, womit das KNN als *Feedforward-Netzwerk* bezeichnet wird. Die Eingabeschicht erhält die Eingabewerte des Netzwerks, die Ausgabeschicht liefert die Vorhersage des Modells. Zwischen diesen beiden Schichten befinden sich beliebig viele verarbeitende Schichten, die als *Hidden Layer* bezeichnet werden. Die Neuronen der Hidden Layer sind in Abbildung 2.4 in weiß dargestellt. [11]

Die Vorhersage, gekennzeichnet durch die Werte der Ausgabeschicht, hängt von den jeweiligen Parametern der Neuronen des Netzwerks ab. Dies sind die Gewichte (engl.: *weights*) der

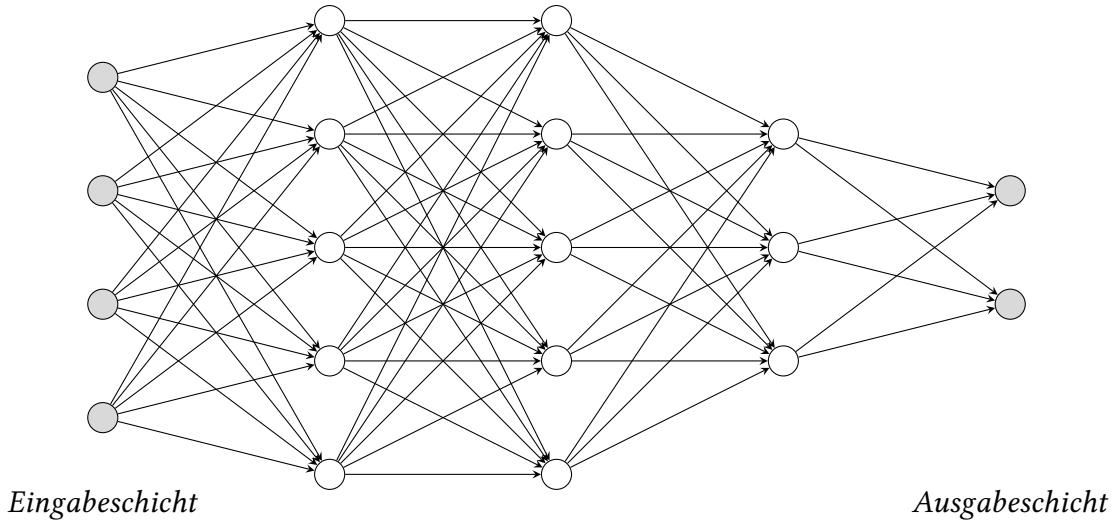


Abbildung 2.4: Vollständiges KNN (angelehnt an [12])

Verbindungen zwischen den Neuronen sowie der Schwellenwert der Neuronen. Es existieren auch trainierbare Aktivierungsfunktionen, diese sind jedoch vergleichsweise unüblich. Entwickler sind für den Entwurf der Netzwerkarchitektur zuständig, während das KNN die Parameter selbst lernt. Zu Beginn besitzt das Modell zufällige Parameter, wodurch es in der Regel nicht die gewünschte Abbildung zwischen Ein- und Ausgabe implementiert. Ein untrainierter Schachalgorithmus spielt demnach augenscheinlich willkürliche Züge. Ein untrainierter Katzenklassifikator besitzt keinen erkennbaren Wissensschatz darüber, welche Charakteristiken eine Katze optisch auszeichnen. Das Ziel des Trainings ist, dass die Parameter des Modells zunehmend gegen das Optimum konvergieren und so das Modell immer plausibler in seinen Vorhersagen wird. [11] [12]

2.2.1 Training

Das Training von KNNs benötigt Daten. Einerseits einen Satz an *Trainingsdaten* und andererseits einen Satz an *Testdaten*. Die Trainingsdaten dienen dazu, das Modell zu verbessern. Eine *Kostenfunktion* berechnet, wie genau die Vorhersagen des Modells auf den Trainingsdaten sind. Darauf basierend wird das Modell optimiert. Die Testdaten dienen zur Messung der Qualität des Modells. Es kann nämlich vorkommen, dass das KNN die Trainingsdaten *auswendig* lernt und deshalb hier gute Ergebnisse erzielt, aber eine unzureichende Performanz auf den Testdaten zeigt. Etwa wenn das KNN unerwartete Eigenschaften in den Trainingsdaten lernt. Aus diesem Grund werden trainierte Modelle anhand der Testdaten evaluiert. [11]

Die Kostenfunktion bestimmt die durchschnittliche Fehlerrate des KNN auf einem gegebenen Datensatz. Sie berechnet somit, gemittelt über alle m Beispiele aus dem Datensatz, die Abweichung des vorhergesagten Wertes \hat{y} von dem tatsächlichen Wert y . Für ein einzelnes

Beispiel aus dem Datensatz nutzt die Funktion dafür eine *Verlustfunktion* \mathcal{L} . Die Verlustfunktion bewertet somit eine einzelne Aussage des KNN, während die Kostenfunktion die durchschnittliche Qualität der Aussagen auf dem gesamten Datensatz misst. Folgende Gleichung zeigt die Kostenfunktion: [10]

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) \quad (2.1)$$

Wobei:

J : Wert der Kostenfunktion

θ : Trainierbare Parameter des Modells

m : Anzahl der Trainingsbeispiele

i : Index des momentan betrachteten Trainingsbeispiels

\mathcal{L} : Verlustfunktion

\hat{y} : Vorhersage des Modells

y : Erwarteter Wert

Die Kostenfunktion J ist abhängig von einem θ . Das θ ist ein Vektor, der alle Gewichte und Schwellenwerte und damit alle trainierbaren Parameter des KNN beinhaltet. Verändern sich die Parameter des KNN, dann liefert es andere Aussagen für das \hat{y} . Somit ändert sich der Wert der Kostenfunktion, wenn das KNN seine Parameter anpasst. [10]

Die Gleichung 2.1 kann auch mit dem Operator \mathbb{E} formuliert werden. Er beschreibt den Erwartungswert: [10]

$$J(\theta) = \mathbb{E}_{x,y}[\mathcal{L}(f(x; \theta), y)] \quad (2.2)$$

Was in Gleichung 2.1 das arithmetische Mittel der Verlustfunktionen ist, wird hier als Erwartungswert der Verlustfunktion geschrieben. Die Fragestellung lautet: *Wenn ein zufälliges Beispiel x und das zugehörige y aus dem Datensatz gezogen werden, was ist dann der erwartete Verlust des Modells?* Die Vorhersage \hat{y} des KNN wird hier als $f(x; \theta)$ geschrieben. Dabei ist f das KNN, das für ein gezogenes x eine bestimmte Vorhersage trifft. Diese Vorhersage ist zudem abhängig von θ . Diese alternative Darstellung der Kostenfunktion mit dem Operator \mathbb{E} ist relevant für Kapitel 2.3. [10]

Kostenfunktionen können verschiedene Arten von Verlustfunktionen verwenden. Für diese Arbeit sind sowohl die \mathcal{L}_1 und \mathcal{L}_2 Verlustfunktionen als auch der *Binary Crossentropy Loss* von besonderer Relevanz.

\mathcal{L}_1 Verlustfunktion Die \mathcal{L}_1 Verlustfunktion berechnet für ein gegebenes Beispiel die absolute Abweichung der Vorhersage von dem erwarteten Wert. Sie wird auch als *mittlere absolute Abweichung* (engl.: *mean absolute error*) bezeichnet. [10]

$$\mathcal{L}_1 = |\hat{y} - y| \quad (2.3)$$

\mathcal{L}_2 Verlustfunktion Die \mathcal{L}_2 Verlustfunktion berechnet hingegen die quadratische Abweichung von dem erwarteten Wert. Hier haben somit große Abweichungen einen stärkeren Einfluss auf den *Verlust* (engl.: *loss*) als bei der \mathcal{L}_1 Funktion. Diese Verlustfunktion wird auch als *mittlere quadratische Abweichung* (engl.: *mean squared error*) bezeichnet. [10]

$$\mathcal{L}_2 = (\hat{y} - y)^2 \quad (2.4)$$

Binary Crossentropy Loss Die Binary Crossentropy Verlustfunktion ist auch als *logarithmische Verlustfunktion* bekannt. Sie eignet sich für binäre Klassifikationen, wo demnach das KNN eine Wahrscheinlichkeit zwischen 0 und 1 ausgibt. Ein Beispiel hierfür ist, wenn das KNN abschätzen soll, ob auf einem Bild eine Katze zu sehen ist oder nicht. Ein Schwellenwert für eine Wahrscheinlichkeit zwischen 0 und 1 gibt dann an, ob das Ergebnis als *Ja* oder *Nein* interpretiert wird. Die Binary Crossentropy Verlustfunktion ist in Gleichung 2.5 dargestellt. Die Verlustfunktion nutzt die Eigenschaften aus, dass $\log(1) = 0$ ist und der Logarithmus von Werten zwischen 0 und 1 negativ ist. Wenn der erwartete Wert y gleich 1 ist, dann reduziert sich die Verlustfunktion außerdem zu $\mathcal{L} = -\log(\hat{y})$. Entspricht \hat{y} dem erwarteten Wert von 1, dann ergibt die Verlustfunktion $-\log(1) = 0$. Je weiter \hat{y} gegen 0 strebt, desto größer wird der Wert für $-\log(\hat{y})$. Ist der erwartete Wert y hingegen gleich 0, dann reduziert sich die Verlustfunktion zu $\mathcal{L} = -\log(1 - \hat{y})$. Hier tritt der umgekehrte Effekt auf. [13]

$$\mathcal{L}_{BCE} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (2.5)$$

Das Training des Modells besteht nicht nur daraus, die Vorhersagen des Modells zu bewerten. Damit das Modell in dem nächsten Trainingsdurchlauf idealerweise einen geringeren Wert für die Kostenfunktion erreicht, müssen die trainierbaren Parameter θ des Modells angepasst werden. Von ihnen hängt der Wert der Kostenfunktion ab. Die Ausgangssituation ist hierbei die folgende: Mit einem gegebenen θ befindet sich das KNN in einem bestimmten Punkt der Kostenfunktion $J(\theta)$. Gesucht ist eine Parameteränderung $d\theta$, mit der sich das KNN am weitesten an das globale Minimum der Kostenfunktion annähert. Das globale Minimum ist die beste Lösung, die das Modell für die gegebenen Trainingsdaten finden kann und damit das Optimum. Da $J(\theta)$ eine viel-dimensionale Funktion ist, berechnet die Trainingsfunktion diesen

Idealwert für θ nicht numerisch. Stattdessen nähert sich das KNN mit jedem Trainingsschritt dem Optimum an. [11]

Nähern tut sich das KNN dem Optimum, indem es den Punkt θ in Richtung des negativen Gradienten der Kostenfunktion bewegt. Also die Richtung, in die, aus der momentanen Ausgangsposition, die Kostenfunktion den steilsten Abstieg besitzt. Das bezeichnet man als *Gradientenabstieg*. Pro Trainingsiteration bewegen sich die Parameter θ nur um einen kleinen Betrag in Richtung des negativen Gradienten. In einem Trainingsschritt kann das KNN auch auf mehreren annotierten Beispielen trainiert werden. Dann bezieht sich die Verlustfunktion \mathcal{L} nicht mehr nur auf ein einzelnes Trainingsbeispiel, sondern auf einen Teil des gesamten Datensatzes. Die Berechnungsformel ist in dem Fall identisch zu den Kostenfunktionen in Gleichungen 2.1 und 2.2. Das ist relevant für Kapitel 2.3.3. [11]

Der Gradientenabstieg wird so lange durchgeführt, bis der Gradient einen so geringen Betrag hat, dass sich die Parameter des KNN nicht mehr signifikant verändern. Das KNN befindet sich hier bestenfalls im globalen Optimum, womit das Training beendet ist. Der Betrag der Annäherung pro Gradientenabstieg ist durch eine sogenannte Lernrate α bestimmt. Die Lernrate ist ein *Hyperparameter*, und damit klassischerweise ein nicht-trainierbarer Parameter, da sie durch die Entwickler fest bestimmt wird und nicht durch das Modell selbst gelernt wird. Ist das KNN ein mal mit jedem Trainingsbeispiel trainiert worden, spricht man von einer Trainingsepoke (*kurz: Epoche*). Üblicherweise werden KNNs über mehrere Epochen trainiert bis die Parameter des Netzes gegen das Optimum konvergieren [11]

2.2.2 Convolutional Neural Networks

KNNs bewähren sich mitunter besonders im Bereich *Computer Vision*. Ein Bereich, der sich mit der Interpretation von Bild- und Videodaten beschäftigt. Hier spielt die Mustererkennung eine tragende Rolle. Es sollen Merkmale erkannt werden, die jedes Objekt eines bestimmten Typs auszeichnen, die jedoch nicht auf jedem Bild die exakt identischen Pixelwerte besitzen. Die typische Form von Katzenohren ist beispielsweise ein Muster, das bei der Katzenerkennung verwendet werden kann. [11]

Verwendet man hierfür jedoch die bisher beschriebene Netzwerkarchitektur, treten verschiedene Probleme auf. Jedes sogenannte *Merkmal* des Eingangs wird über die Eingangsschicht in das KNN gespeist. Bei einem Schachalgorithmus kann die Menge aller Merkmale beispielsweise durch die momentane Position aller Figuren auf dem Schachbrett beschrieben werden. Bei der Bildklassifizierung ist jeder Pixel des Bildes ein Merkmal. Ein Netz, das Bilder der Grö-

Se 1024x1024 Pixel mit drei Farbkanälen (rot, grün, blau) klassifizieren soll, muss demnach folgende Anzahl an Eingängen verarbeiten: [11]

$$1024 \cdot 1024 \cdot 3 = 3.145.728 \quad (2.6)$$

KNNs, wie sie bisher gezeigt sind, benötigen dafür eine Architektur mit vielen Neuronen. Das sorgt für mehr trainierbare Parameter und damit unter anderem für eine längere Trainingsdauer des Modells. Im Bereich Computer Vision wird deshalb auf eine spezielle Art von neuronalen Netzen namens CNN zurückgegriffen, da sie eine effizientere Verarbeitung von solchen Eingabedaten ermöglichen. [11]

Der Bereich Computer Vision basiert auf der Verarbeitung von Bildern. Ein digitales Bild kann als eine Matrix von Pixelwerten betrachtet werden. Aus diesem Grund ist der Eingang zu einem CNN eine Eingangsmatrix. Die Funktionsweise von CNNs basiert auf der Faltung (engl.: convolution) der Eingangsmatrix mit einer Faltmatrix. Solche Netze besitzen mindestens einen *Convolutional Layer*, der diese Faltung durchführt. Dabei schiebt das CNN die Faltmatrix nach und nach über die Eingangsmatrix. Bei jedem Schritt berechnet es dabei das Skalarprodukt der momentan betrachteten Werte der Eingangsmatrix mit den Parametern der Faltmatrix. Das Ergebnis hiervon ist eine neue Matrix. Folgt hierauf eine weitere Schicht, dann erhält sie die Ergebnismatrix der Faltung als Eingabewert und nutzt eine eigene Faltmatrix um erneut eine Faltung durchzuführen. Die trainierbaren Parameter sind dabei die Werte der Faltmatrizen aller Schichten. [10]

Folgender Algorithmus beschreibt die Funktionsweise einer Faltung: [10]

1. Positioniere die Faltmatrix mittig über einen Pixel der Eingangsmatrix
2. Berechne das Skalarprodukt aus den Werten der Faltmatrix und den Werten der Eingangsmatrix, die sich unter der Faltmatrix befinden
3. Schreibe das Ergebnis in eine neue Matrix
4. Wiederhole die Schritte 1-3, bis die Faltmatrix alle Teilmatrizen der Eingangsmatrix abgedeckt hat

Zur Verdutlichung soll ein beispielhafter Convolutional Layer betrachtet werden. Angenommen, die Eingangsmatrix der Form 7x7 sei ein schwarz-weiß-Bild. Jeder Pixel hat demnach den Wert 0 oder 1. Die Faltmatrix habe die Form 3x3 und zufällig gewählte Parameter. Dann ergibt sich zum Beispiel die Faltung, die in Abbildung 2.5 dargestellt ist.

0	1	1	1	0	0	0
0	0	1	1	1	0	0
0	0	0	1	1	1	0
0	0	0	1	1	0	0
0	0	1	1	0	0	0
0	1	1	0	0	0	0
1	1	0	0	0	0	0

*

1	0	1
0	1	0
1	0	1

=

1	4	3	4	1
1	2	4	3	3
1	2	3	4	1
1	3	3	1	1
3	3	1	1	0

Eingangsmatrix

Faltmatrix

Ergebnis

Abbildung 2.5: Beispiel für eine Faltung eines CNN (angelehnt an [14])

Zwei Rechenschritte der Faltung sind hier farblich hervorgehoben. Das Skalarprodukt aus dem **blau** markierten Teil der Eingangsmatrix und der Faltmatrix ergibt den Wert 4. Die Rechnung hierzu ist:

$$1 \cdot 1 + 0 \cdot 0 + 0 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 = 4 \quad (2.7)$$

Dieses Ergebnis fügt der Convolutional Layer an den Index (1, 4) der Ergebnismatrix ein. Startet die Faltmatrix in der oberen linken Ecke der Eingangsmatrix und wird nach jedem Rechenschritt um eine Spalte nach rechts verschoben, ist das die vierte Rechnung der Faltung. Das Skalarprodukt der Faltmatrix mit den Werten der **orange** markierten Teilmatrix berechnet der Convolutional Layer dann als letztes. Das Ergebnis ist hier 0, da der betrachtete Teil der Eingangsmatrix überall den Wert 0 besitzt.

In diesem Fall ist das Ergebnis der Faltung eine 5x5-Matrix. Die Größe der Ausgangsmatrix ist abhängig von der Größe der Eingangsmatrix und der Faltmatrix. Eine größere Faltmatrix führt zu einer kleineren Ergebnismatrix, da die Faltmatrix weniger oft über die Eingangsmatrix geschoben werden kann. Auch ist es möglich, dass die Ergebnismatrix die gleichen Dimensionen besitzt wie die Eingangsmatrix oder auch größer ist. Das hängt davon ab, ob die Faltmatrix auch über den Rand der Eingangsmatrix hinausgeschoben wird. [11]

Auch führt ein Convolutional Layer nicht nur eine einzelne Faltung durch, sondern meist mehrere. Jede Faltung hat ihre eigene Faltmatrix und liefert deshalb eine eigene Ergebnismatrix. Jede dieser Faltungen soll ein bestimmtes Merkmal aus dem Eingang erkennen. Beispielsweise kann eine Faltung runde Formen erkennen, während eine andere Faltung gerade Linien erkennt.

Zusätzlich zu Convolutional Layern besitzen CNN-Architekturen sogenannte *Pooling Layer* und *Fully Connected Layer*. Pooling Layer fassen Teile der Eingangsmatrix zu einzelnen Werten zusammen. Beispielsweise existiert das *Max-Pooling*, bei dem der Pooling Layer nur den größten Wert eines Teilbereichs der Eingangsmatrix weitergibt. Die Haupteigenschaft von Pooling Layern ist, dass kleine Schwankungen in den Werten der Eingangsmatrix einen vergleichsweise geringen Einfluss auf die Ausgabe der Schicht haben. Bei Convolutional Layern ist das unter

Umständen nicht der Fall, da jeder Wert einen Einfluss auf das jeweilige Skalarprodukt hat. Das ist relevant, da CNNs generalisieren können sollen. Zum Beispiel wenn das gleiche Objekt in zwei Bildern leicht unterschiedlich aussieht. [10]

Fully Connected Layer sind vergleichbar mit den Schichten eines klassischen KNNs. Sie bestehen aus Neuronen, die jeweils alle Werte der vorherigen Schicht als Eingang erhalten. [10]

Abbildung 2.6 zeigt eine beispielhafte CNN-Architektur. Das Netz besitzt abwechselnd Convolutional Layer und Pooling Layer. Das ist üblich für CNNs [10]. Auf jede Ergebnismatrix der Faltung führt das CNN zudem eine Aktivierungsfunktion aus, die sich *ReLU* (*rectified linear unit*) nennt. Die Schichten sind dazu da, einzelne Merkmale aus dem Eingabebild zu erkennen. Am Ende des CNN befindet sich ein Fully Connected Layer, der aus den erkannten Merkmalen eine Vorhersage berechnet. Vorher konvertiert das CNN die Ergebnismatrizen in einen einzelnen Vektor, der an den Fully Connected Layer übergeben werden kann. Das ist in der Grafik als *Flatten* bezeichnet. In diesem Beispiel ist das CNN ein Klassifikator, der entscheidet, welche Art von Objekt auf dem Bild zu sehen ist.

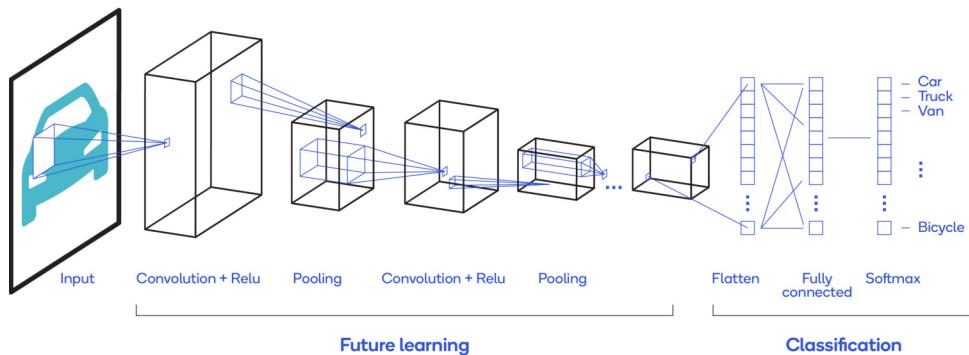


Abbildung 2.6: Beispiel für eine CNN-Architektur [15]

2.3 Bildgenerierung mit Künstlichen Neuronalen Netzen

Der Lernfortschritt von Klassifikatoren besteht darin, besser in der Aussage zu werden, ob eine Vorhersage y auf einen gegebenen Eingang x zutrifft. Das bezeichnet man auch als *diskriminative Modellierung*. Um das zu erlernen, benötigt der Klassifikator annotierte Trainingsdaten. Das bedeutet, dass jedes Trainingsbeispiel x eine erwartete Vorhersage y besitzt. Man bezeichnet das auch als *überwachtes Lernen*. [16]

Neben der diskriminativen Modellierung ist auch eine *generative Modellierung* möglich. Generative KNNs lernen, neue Daten zu erzeugen, die der Verteilung der Trainingsdaten ähneln. Dazu erlernen sie die statistische Verteilung der Trainingsdaten. Generative Netze zur Bildzeugung können demnach Bilder generieren, die vergleichbar mit den Trainingsbildern sind.

Beispielsweise wenn ein generatives Modell darauf trainiert ist, Katzenbilder zu erzeugen. Die Trainingsbilder sind nicht annotiert, wodurch generative Netze in der Regel in das *unüberwachte Lernen* einzuordnen sind. [16]

2.3.1 Mathematischer Hintergrund

Generative Netze zur Bilderzeugung sollen beurteilen können, wie wahrscheinlich es ist, dass ein gegebenes Bild aus der Verteilung der Trainingsdaten stammt. Wenn x für jedes mögliche existierende Bild steht, so bilden generative Netze folgende Wahrscheinlichkeitsverteilung ab: [16]

$$\hat{p}(x) \quad (2.8)$$

Für ein gegebenes Bild x gibt $\hat{p}(x)$ einen Schätzwert dafür an, wie wahrscheinlich es ist, dass das Bild aus den Trainingsdaten stammt. Diese Wahrscheinlichkeitsverteilung wird durch das Netz erlernt. Das Training des generativen Netzes optimiert, dass die geschätzte Verteilung der Daten $\hat{p}(x)$ möglichst ähnlich zu der tatsächlichen Verteilung der Trainingsdaten $p(x)$ ist. Ein beispielhafter Vergleich ist in Abbildung 2.7 dargestellt. Es ist erkennbar, dass sich die geschätzte und die tatsächliche Verteilung ähnlich sehen, jedoch nicht identisch sind. Die Abweichung zwischen diesen Verteilungen stellt dabei die Kosten dar. Die schwarzen Punkte kennzeichnen Trainingsdaten. Sie sollen die Verteilung $p(x)$ abbilden. Weniger diversifizierte Trainingsdaten würden sich beispielsweise nur in einem Teilbereich von $p(x)$ befinden. Dadurch könnte das Modell $p(x)$ weniger gut approximieren. [16] [17]

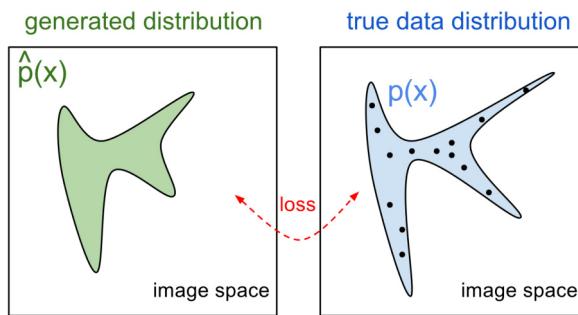


Abbildung 2.7: Beispielhafter Vergleich von $\hat{p}(x)$ und $p(x)$ [17]

Bei der Bildgenerierung versucht das Netz den Wahrscheinlichkeitswert für $\hat{p}(x)$ zu maximieren. Es lernt durch $\hat{p}(x)$, wie die Verteilung der Trainingsdaten aussieht und versucht anschließend ausschließlich Bilder zu generieren, die dieser Verteilung folgen. Bezogen auf Abbildung 2.7 befinden sich alle generierten Bilder des trainierten Netzes im grün markierten Bereich. [16] [17]

Es existieren verschiedene Arten generativer Netze. Die Taxonomie, also die Einteilung verschiedener Netze in bestimmte Kategorien, stellt Abbildung 2.8 dar. Einerseits existieren Architekturen, die die Wahrscheinlichkeitsverteilung $\hat{p}(x)$ explizit berechnen oder approximieren. Andere berechnen die Funktion nicht, verwenden sie jedoch implizit. Die Abbildung unterscheidet dahingehend zwischen den Kategorien *Explizite Wahrscheinlichkeitsdichte* und *Implizite Wahrscheinlichkeitsdichte*. Auf der untersten Ebene der Taxonomie befinden sich konkrete Architekturen generativer Netze. Dabei sind nur diejenigen aufgeführt, die diese Studienarbeit in betracht zieht. Die Auswahl basiert auf verschiedenen Veröffentlichungen, die Architekturen für generative Netze vergleichen [18] [19]. Abbildung 2.8 deutet jedoch an, dass darüber hinaus weitere Arten existieren. [20]

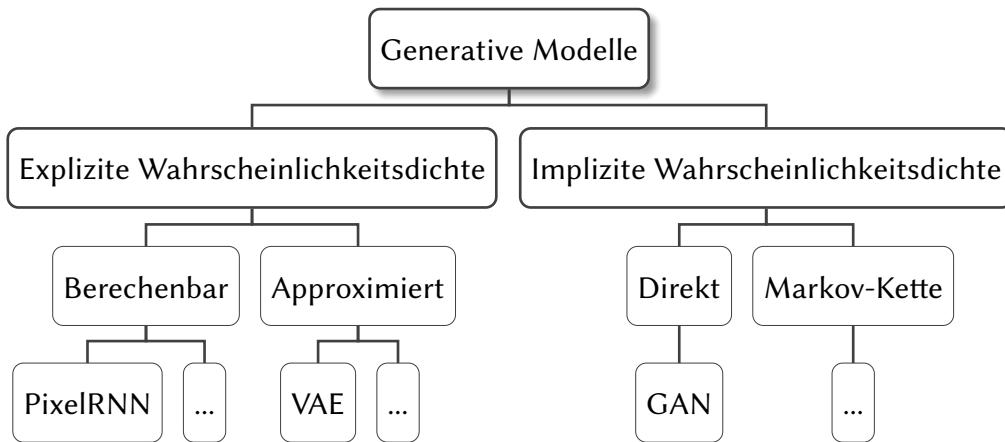


Abbildung 2.8: Taxonomie generativer Modelle *angelehnt an* [20]

Dieses Kapitel soll auf die gezeigten Architekturen eingehen, um ein Verständnis für die Funktionsweise generativer Netze zu schaffen. Darauf basierend trifft Kapitel 3.3 eine Auswahl für die Architektur, die für diese Studienarbeit verwendet wird.

2.3.2 Pixel Recurrent Neural Networks

Die Architektur der Pixel Recurrent Neural Networks (PixelRNNs) stammt aus dem Jahre 2016. Diese Netze stützen sich explizit auf die Maximierung der Maximum-Likelihood-Schätzung von $\hat{p}(x)$ für jeden Pixel. Sie sind in der genannten Taxonomie den Modellen zuzuordnen, die den tatsächlichen Schätzwert von $p(x)$ berechnen können. [21]

Im folgenden soll geklärt werden, wie ein PixelRNN den optimalen Wert für jeden Pixel eines generierten Bildes bestimmt. Ein betrachtetes Bild x der Auflösung $n \times n$ kann in seine einzelnen

Pixel $(x_1, x_2, \dots, x_{n^2})$ aufgeteilt werden. Gleichung 2.9 gibt an, wie die Wahrscheinlichkeit eines jeden Pixels in die gesamte Verteilung $\hat{p}(x)$ einfließt. [21]

$$\hat{p}(x) = \hat{p}(x_1, x_2, \dots, x_{n^2}) = \prod_{i=1}^{n^2} \hat{p}(x_i | x_1, \dots, x_{i-1}) \quad (2.9)$$

Jeder Pixel x_i besitzt eine eigene Wahrscheinlichkeitsverteilung $\hat{p}(x_i | x_1, \dots, x_{i-1})$. Sie ist abhängig von allen anderen Pixeln x_1, \dots, x_{i-1} des Bilds. Den optimalen Wert jedes Pixels kann das PixelRNN demnach nur dann berechnen, wenn die Werte aller anderen Pixel bekannt sind. Das Produkt aller Wahrscheinlichkeitswerte der einzelnen Pixel ergibt $\hat{p}(x)$. Soll $\hat{p}(x)$ maximiert werden, so müssen die Terme $\hat{p}(x_i | x_1, \dots, x_{i-1})$ möglichst hohe Werte liefern. Daraus ergibt sich dann unter gegebenem Kontext für jeden Pixel eine Maximum-Likelihood-Schätzung. Also der Pixelwert, für den $\hat{p}(x)$ möglichst weit gegen *eins* strebt. [21]

Die Idee von PixelRNNs ist, dass die Generierung in einer Ecke des Bilds startet. Das Bild wird zunächst auf einen Pixel reduziert, der im folgenden x_1 genannt wird. Für diesen Pixel generiert das PixelRNN einen Wert. Anschließend wird x_1 gemeinsam mit einem benachbarten Pixel x_2 betrachtet. Die Wahrscheinlichkeitsverteilung für x_2 ergibt sich dadurch zu $\hat{p}(x_2 | x_1)$. Der Wert für x_2 ist somit nur von x_1 abhängig. Da x_1 bekannt ist, kann das PixelRNN einen optimalen Wert für x_2 bestimmen. Die Wahrscheinlichkeitsverteilung von x_3 ergibt sich zu $\hat{p}(x_3 | x_1, x_2)$, die von x_4 zu $\hat{p}(x_4 | x_1, x_2, x_3)$. Das PixelRNN generiert das Bild sukzessive, wobei der momentane Pixelwert für x_i von allen bisher generierten Pixeln abhängt. Dieses vorgehen ist in Abbildung 2.9 dargestellt. Der Wert des rot markierten Pixels hängt von allen blau markierten Pixel ab. Ist für diesen ein Wert bestimmt, wird der rechtsseitig benachbarte Pixel als neues x_i gewählt.

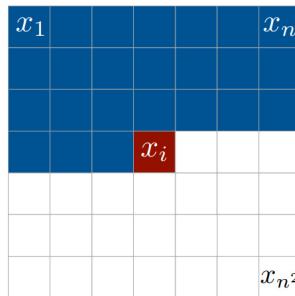


Abbildung 2.9: Bestimmung von $\hat{p}(x)$ mit PixelRNNs [21]

Um die beschriebene Abhängigkeit des momentan generierten Pixels zu allen bisher generierten Pixel umsetzen zu können, besitzen PixelRNNs eine Art *Erinnerung*. Bisher wurden in dieser Arbeit nur sogenannte *Feedforward* Netze behandelt, bei denen der Informationsfluss stets in eine Richtung erfolgt. Nämlich vom Eingang des Netzes zum Ausgang. Es existieren auch *Recurrent Neural Networks*. Sie werden besonders zur Verarbeitung von natürlicher Sprache eingesetzt (engl.: natural language processing).

In Recurrent Neural Networks spielen die *Zustände* eines Netzes eine besondere Rolle. Ein Zustand wird durch die Eingangs- und Ausgangswerte aller Neuronen zu einem gegebenen Zeitpunkt beschrieben. In PixelRNNs ist der Zustand des Netzes für den Pixel x_2 abhängig von dem Zustand des Netzes für x_1 . Um solche Beziehungen darstellen zu können, besitzt ein Recurrent Neural Network Neuronen, die den vorherigen Wert eines Neurons rekursiv auf seinen Eingang zurückführen. Somit wird der vorherige Zustand des neuronalen Netzes als zusätzlicher Eingang für die Berechnungen genutzt. PixelRNNs nutzen eine besondere Form der Recurrent Neural Networks. Sie arbeiten mit sogenannter *Long Short-term Memory*. Dadurch soll das Problem behoben werden, dass in klassischen Recurrent Neural Networks weit in der Vergangenheit liegende Zustände nur noch einen geringen Einfluss auf den momentanen Zustand haben. [19]

Da die durch ein PixelRNN umgesetzte Verteilung $\hat{p}(x)$ direkt erfassbar ist, wird ihnen nachgesagt, dass die Performanz solcher Netze gut evaluiert werden kann. Es gilt als vergleichsweise leicht, für solche Netze Metriken zur Messung der Performanz umzusetzen. Ein grundlegender Nachteil von PixelRNNs ist hingegen, dass sie die Bilder nur sequenziell generieren können. Es ist in dem beschriebenen Verfahren nicht möglich, mehrere Pixel parallel zu erzeugen, da der Wert eines Pixels von denen aller vorher generierten Pixel abhängig ist. Dies verlangsamt die Generierung, da keine Parallelisierung möglich ist. [19]

Es existieren auch sogenannte *PixelCNNs*, bei denen sich die Berechnung stets nur auf bestimmte Bildbereiche konzentriert. Diese Bildbereiche kann das PixelCNN parallel zueinander berechnen. Die Parallelisierung ist jedoch nur während des Trainings des Netzwerks oder während der Evaluation von $\hat{p}(x)$ für gegebene Bilder möglich. Die Bildgenerierung erfolgt auch hier, analog zu PixelRNNs, vollständig sequenziell. [21]

2.3.3 Autoencoder

Das Ziel von Autoencodern ist, den Eingang des Netzes am Ausgang zu rekonstruieren. Dazu setzen sich diese Netze aus drei Bestandteilen zusammen: dem Kodierer, dem latenten Raum und dem Dekodierer. In Abbildung 2.10 ist eine beispielhafte Autoencoderarchitektur dargestellt. [22]

Der **Kodierer** besitzt die Aufgabe, Merkmale aus dem Eingang des Netzes zu extrahieren. Diese Merkmale sollen daraufhin mit einer begrenzten Anzahl an Parametern durch den latenten Raum repräsentiert werden. Somit besteht die Aufgabe des Kodierers darin, den Eingang auf seine für das Netz wesentlichen Eigenschaften zu reduzieren. Und zwar erfolgt die Komprimierung dabei so, dass die Informationen gerade so durch den latenten Raum dargestellt werden können. [22]

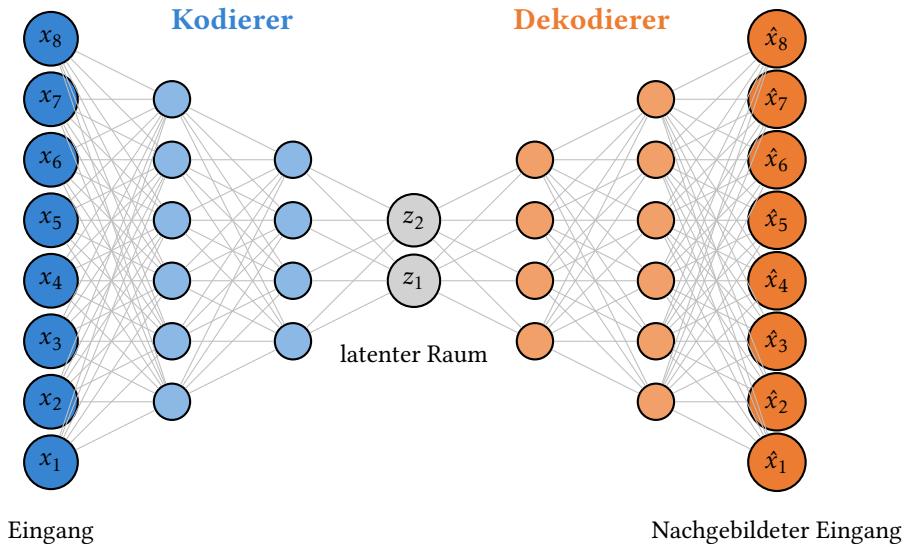


Abbildung 2.10: Architektur eines Autoencoders (angelehnt an [23])

Bei dem **latenten Raum** handelt es sich um eine einzelne Schicht von Parametern, oder in diesem Kontext: Neuronen. Je mehr Neuronen sich in dem latenten Raum befinden, desto mehr Informationen kann der Kodierer an den Dekorier übergeben. Ein Vektor \vec{z} beschreibt beispielsweise die Merkmale, die sich in dem latenden Raum befinden. Der latente Raum sollte klein genug sein, damit der Kodierer in \vec{z} nicht alle Merkmale des Eingangs speichern kann. So wird der Kodierer dazu gezwungen, vereinzelte Merkmale zu extrahieren. [22]

Der **Dekodierer** nutzt die Werte aus dem latenten Raum, um den Eingang nachzubilden. Diese Nachbildung stellt den Ausgang des Autoencoders dar. Die Kosten eines Autoencoders ergeben sich durch die Abweichung zwischen Ein- und Ausgang. [22]

Da der Zweck von Autoencodern darin besteht, einen Eingang auf seine relevanten Merkmale zu reduzieren, wird der Dekodierer nur während des Trainings verwendet. Beim praktischen Einsatz verwendet ein Autoencoder lediglich den Kodierer und den latente Raum. Im Gegensatz zu PixelRNNs basieren Autoencoder klassischerweise nicht auf Recurrent Neural Networks, sondern auf Feedforward Netzen. [19]

Diese beschriebene Architektur der Autoencoder eignet sich nicht für die generative Modellierung, da sie deterministisch ist. Erhält das Modell bestimmte Eingangswerte, so liefert es stets die gleichen Ausgangswerte. Es versucht den Eingang möglichst zu rekonstruieren, wobei der Inhalt \vec{z} des latenten Raums für ein gegebenes Eingangsbild stets gleich ist. Zufällige Bilder kann ein Autoencoder somit nicht erzeugen. Die sogenannten *Variational Autoencoder* sind hingegen eine Architektur, die für die Generierung neuer Daten verwendet werden können. [12] [19]

Variational Autoencoder (VAEs) verfolgen das Ziel, eine Zufallskomponente in die Bildzeugung einfließen zu lassen. Ein entscheidender Unterschied zu klassischen Autoencodern ist

deshalb der folgende: VAEs kodieren einen gegebenen Eingang x nicht auf ein festes \vec{z} , sondern auf eine Wahrscheinlichkeitsverteilung. Sie wird bezeichnet als: [24]

$$p(z|x) \quad (2.10)$$

Für ein gegebenes Bild x gibt $p(z|x)$ an, wie wahrscheinlich es ist, dass das Bild aus der Verteilung der Merkmale \vec{z} stammt. Im Gegensatz zu PixelRNNs versucht ein VAE somit nicht die Verteilung der Trainingsdaten $p(x)$ zu approximieren, sondern die Verteilung der Merkmale \vec{z} aus den Trainingsdaten x . Es wird angenommen, dass jedes Merkmal normalverteilt ist. Damit ist jede Komponente z_i des Vektors $\vec{z} = [z_1, z_2, \dots, z_n]^\top$ durch eine Gaußsche Normalverteilung $\mathcal{N}(\mu_i, \sigma_i^2)$ beschrieben. Aufgabe des Kodierers ist damit nicht mehr, aus einem gegebenen x eine Menge von Merkmalen \vec{z} zu bestimmen. Stattdessen soll der Kodierer die Vektoren $\vec{\mu}$ und $\vec{\sigma}$ bestimmen, durch die sich die einzelnen Normalverteilungen von \vec{z} beschreiben lassen. [19] [24]

Der latente Raum übergibt an den Dekodierer ein zufällig aus der Verteilung $p(z|x)$ entnommenes Set an Merkmalen \vec{z} . Der Dekodierer übersetzt dieses gegebene \vec{z} daraufhin in ein Bild. Im praktischen Einsatz werden bei einem VAE nur der latente Raum und der Dekodierer genutzt. Der Dekodierer erhält zufällige Werte für \vec{z} , also zufällige Merkmale, und generiert daraus ein Bild. [24] [12]

Variational Autoencoders (VAEs) können die Maximum-Likelihood-Schätzung, ob ein gegebenes Bild der Verteilung $\hat{p}(x)$ entstammt, nicht direkt berechnen. Sie können lediglich die Verteilung der Merkmale \vec{z} aus den Trainingsdaten x approximieren. Dadurch wird ihnen nachgesagt, dass sie weniger gut evaluiert werden können als PixelRNNs. Ein Vorteil von VAEs ist, dass der Merkmalsvektor \vec{z} gezielt manipulierbar ist. Kleine Änderungen der Merkmale führen in der Regel auch nur zu einem leicht veränderten Bild. [19]

2.3.4 Generative Adversarial Networks

Eine weitere Architektur, die zur Bildgenerierung verwendet werden kann, sind sogenannte Generative Adversarial Networks (GANs). Eine Arbeit aus dem Jahre 2014 stellt GANs erstmals vor. [25]. Zudem existiert eine Veröffentlichung aus dem Jahre 2020 [26]. Während die erste Veröffentlichung GANs als eine neuartige Architektur vorstellt, zeigt die neuere Arbeit vor allem Erfolge und Hindernisse auf, die sich im Laufe der Zeit bei der Verwendung von GANs herausgestellt haben.

Ein GAN besteht aus zwei Komponenten: Dem *Generator* und dem *Diskriminatator* (engl.: *Discriminator*). Der Generator erzeugt aus einem zufälligen Eingangsvektor ein Bild. Der Diskriminatator erhält ein Bild als Eingang und soll bewerten, ob das Bild echt oder künstlich generiert ist. Bei

beiden Komponenten handelt es sich um CNNs. Das Ziel des Trainings ist, dass der Generator Bilder erzeugen kann, die der Diskriminatior nicht von echten Trainingsbildern unterscheiden kann. Dabei wird der Generator besser in seiner Generierung, während der Diskriminatior besser in seiner Unterscheidung wird. Mit zunehmender Güte des Generators muss auch der Diskriminatior weitere Merkmale erlernen, anhand derer er künstliche Bilder erkennt. Dies gilt ebenso in die entgegengesetzte Richtung. Damit agieren Generator und Diskriminatior als direkte Gegenspieler, die versuchen einander zu überlisten. [26]

Die Güte des Generators ist dadurch gegeben, wie ähnlich die von ihm erzeugte Wahrscheinlichkeitsverteilung \hat{p} zu der Verteilung der Trainingsdaten p ist. Der Diskriminatior wird hingegen darin bewertet, wie erfolgreich er in seiner Aussage ist, ob ein gegebenes Bild entweder aus \hat{p} oder p stammt. [26]

Das Zusammenspiel zwischen Generator und Diskriminatior während des Trainings ist in Abbildung 2.11 dargestellt.

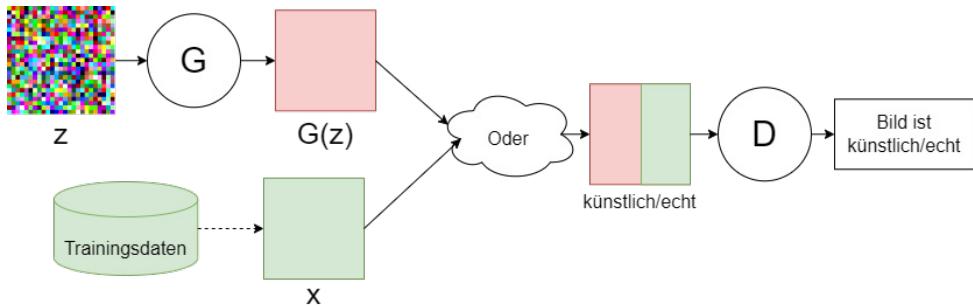


Abbildung 2.11: Zusammenspiel zwischen Generator und Diskriminatior

Der Eingangvektor heißt z . Daraus erzeugt Generator G ein Bild $G(z)$. Der Diskriminatior D erhält entweder ein Bild x aus den Trainingsdaten oder ein Bild $G(z)$ vom Generator. Darauf basierend trifft der Diskriminatior entweder eine Vorhersage $D(x)$ oder $D(G(z))$, wie wahrscheinlich es ist, dass das Bild real ist. Bei der Vorhersage handelt es sich um eine Wahrscheinlichkeit zwischen 0 und 1. Ein Wert von 0.8 bedeutet demnach zum Beispiel: Der Diskriminatior ist sich zu 80% sicher, dass das Bild real ist. [26]

Training Hieraus ergibt sich die Kostenfunktion für GANs. Sie nennt sich $V(D, G)$: [25]

$$\min_G \max_D V(D, G) = \mathbb{E}_x[\log(D(x))] + \mathbb{E}_z[\log(1 - D(G(z)))] \quad (2.11)$$

Bei dem Training handelt es sich um ein Min-Max-Problem. Der Generator versucht die Funktion $V(G, D)$ zu minimieren, wohingegen der Diskriminatior sie zu maximieren versucht. Das Ziel des Trainings ist damit nicht wie klassischerweise bei KNNs, dass die Funktion einen Wert von 0 anstrebt. Stattdessen soll sie im Verlauf des Trainings gegen einen Wert größer als 0

konvergieren. Hier können sich der Generator und der Diskriminator nicht weiter verbessern. Sie haben idealerweise optimale Parameter. Es wird davon gesprochen, dass G und D sich in einem *Nash-Gleichgewicht* befinden. Dieser Begriff stammt aus der Spieltheorie und kann deshalb verwendet werden, weil ein GAN als ein Spiel zwischen Generator und Diskriminator beschrieben werden kann. [26]

Die Funktion $V(D, G)$ basiert auf einer logarithmischen Verlustfunktion. Der Term 2.12 aus der Kostenfunktion stellt die Fehlerrate des Diskriminators auf den echten Trainingsdaten dar. Die Fragestellung lautet: *Wenn ich ein zufälliges Bild x aus den Trainingsdaten ziehe, wie wahrscheinlich ist es, dass der Diskriminator es als echt klassifiziert?* Der Wert für $\log(D(x))$ sollte möglichst nahe 0 sein, da für jedes echte Trainingsbild x zu erwarten ist, dass $D(x)$ einen Wert nahe 1 ausgibt.

$$\mathbb{E}_x[\log(D(x))] \quad (2.12)$$

Der Term 2.13 beschreibt hingegen, wie viele generierte Bilder der Diskriminator als echt klassifiziert. Hierfür werden zufällige z aus der Wahrscheinlichkeitsverteilung der Eingangsvektoren entnommen. Diese Verteilung kann beliebig definiert sein. Ist der Diskriminator besser trainiert als der Generator, dann ist zu erwarten, dass $D((G(z)))$ einen Wert nahe 0 hat. Somit ist der Logarithmus ebenfalls nahe 0. Im umgekehrten Fall nimmt der Logarithmus einen negativen Wert an. Der Betrag des Werts erhöht sich, je mehr $D(G(z))$ gegen 1 strebt.

$$\mathbb{E}_z[\log(1 - D(G(z)))] \quad (2.13)$$

Der Term 2.12 ist nicht vom Generator abhängig, sondern lediglich von D und x . Hierauf hat der Generator keinen Einfluss, wodurch er alleine durch diesen Teil der Kostenfunktion nicht trainiert wird. Der Diskriminator besitzt somit zwei Terme, die ihn trainieren, während der Generator nur einen besitzt. Aus diesem Grund werden dem Diskriminator in der Regel doppelt so viele unechte Daten wie echte Trainingsdaten gezeigt. Dies soll einem ungleichen Training der beiden Komponenten des GANs entgegenwirken. Der Optimalzustand eines GANs ist, dass der Diskriminator so gut wie möglich identifizieren kann, ob ein gegebenes Bild aus $p(x)$ stammt, während der Generator dennoch in der Lage ist, den Diskriminator zu überlisten. Das Training von GANs gilt als empfindlich gegenüber den gewählten Hyperparametern und der Netzwerkarchitektur. GANs wird nachgesagt, dass kleine Änderungen in den beiden genannten Aspekten die Qualität der Generierten Bildern signifikant beeinflussen können. [12]

Im praktischen Einsatz wird nur der Generator des GANs verwendet. Der Diskriminator wird ausschließlich dazu eingesetzt, mit $p(G(z))$ möglichst gut $p(x)$ zu approximieren, sodass die generierten Bilder im Optimalfall nicht von echten Trainingsdaten zu unterscheiden sind. [12]

Die bisher beschriebene Architektur von GANs wird auch als *Vanilla GAN* bezeichnet. Forscher und Anwender haben seit dieser Veröffentlichung verschiedene Limitationen und Probleme bei Vanilla GANs feststellen können. Insbesondere im Hinblick auf spezielle Einsatzgebiete. Ein hier häufig anzutreffender Begriff ist *Modal Collaps*. Damit ist die Situation gemeint, dass der Generator bei beliebigem Input stets dasselbe Bild generiert. Er lernt, dass ein bestimmtes Bild den Diskriminatoren überlistet kann und generiert es deshalb jedes Mal, egal welchen Input man ihm zuführt. In dem Anwendungsfall dieser Arbeit könnte das zum Beispiel dazu führen, dass der Generator nur eine einzelne Art von Straßenschild generiert. Lösen lässt sich das Problem des *Modal Collaps* beispielsweise mit sogenannten CycleGANs. [12]

CycleGANs CycleGANs eignen sich für eine bestimmte Art der generativen Modellierung: Der Bild-zu-Bild Übersetzung. Hierbei soll das Modell nicht ein völlig neues Bild generieren, sondern ein vorhandenes Bild in eine andere Domäne übersetzen. Ein Beispiel dafür ist das Umwandeln von Zeichnungen in fotorealistische Bilder oder das Einfärben von schwarz-weiß Bildern. Ein entscheidender Unterschied ist somit, dass der Generator keinen zufälligen Vektor z als Eingang enthält, sondern ein Bild, welches das Modell in eine andere Domäne übersetzen soll. Bei solchen Anwendungsfällen sollen CycleGANs einen Modal Collaps verhindern. Das Ziel ist, dass das erzeugte Bild des Modells immer abhängig ist von dem Eingangsbild. [27]

Ein CycleGAN besteht aus zwei miteinander gekoppelten GANs. Generator G übersetzt ein Bild x in ein Bild y . Aus diesem y soll dann Generator F den Eingang x rekonstruieren. Somit soll das gesamte Netzwerk nicht nur neue Bilder generieren können, sondern soll auch von einem generierten Bild zurück auf den zugeführten Eingang schließen können. Die Bilder x und y entstammen den Domänen X und Y . Dabei steht X zum Beispiel für Zeichnungen und Y für fotorealistische Bilder. Dies lässt sich mathematisch so darstellen, dass G und F folgende Abbildungen implementieren: [27]

$$G : X \mapsto Y \wedge F : Y \mapsto \hat{X} \quad (2.14)$$

Das Modell G erzeugt somit aus einem gegebenen x ein y , wohingegen F aus dem y auf das x schließen soll. Die Behebung des Modal Collaps findet dadurch statt, dass das Netzwerk den Output \hat{x} von F überprüft und diesen mit dem tatsächlichen Eingang x vergleicht. Es wird überprüft, wie ähnlich sich \hat{x} und x sind. Liegt eine zu hohe Diskrepanz vor, kann das Netzwerk darauf schließen, dass G Ausgaben erzeugt, die nicht in direkter Abhängigkeit zu x stehen. [27]

Das Training von CycleGANs basiert auf mehreren Kostenfunktionen. Die GANs G und F besitzen jeweils eigene Diskriminatoren D_y und D_x . Dadurch haben beide GANs einen eigenen *Adversarial Loss*. Damit ist die in Gleichung 2.11 beschriebene Kostenfunktion eines Vanilla

GANs gemeint. Es sei erwähnt, dass die CycleGAN-Veröffentlichung hier von Verlustfunktionen (*engl: losses*) mit dem Formelzeichen \mathcal{L} spricht, während hier der Begriff Kostenfunktion genutzt wird. Die Formeln sind jedoch identisch, da sich die Verlustfunktionen bereits auf mehrere Trainingsbeispiele beziehen. [27]

Im Kontext von CycleGANs ist der beschriebene Adversarial Loss für das GAN G wie folgt definiert: [27]

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_y[\log D_Y(y)] + \mathbb{E}_x[\log(1 - D_Y(G(x)))] \quad (2.15)$$

Die Funktion für GAN F ist identisch, mit dem Unterschied, dass sie von F und D_X statt von G und D_Y abhängt. Ein Unterschied zu den Vanilla GANs ist hierbei: Bei Vanilla GANs steht der Buchstabe x für die Trainingsdaten aus der Zieldomäne. Hier sind jedoch einzelne x die Eingabewerte für das CycleGAN. Was vorher z war, ist demnach hier x . Die Trainingsdaten aus der Zieldomäne werden mit dem Formelzeichen y abgekürzt. [27]

Als weitere Kostenfunktion besitzen CycleGANs einen *Cycle Consistency Loss* (Gleichung 2.16). Die beiden Summanden der Gleichung setzen sich daraus zusammen, wie weit die Pixelwerte von den generierten Bildern und den echten Bildern auseinander liegen. Mit $G(x)$ wird ein Bild \hat{y} generiert, F generiert anschließend aus diesem \hat{y} wieder ein \hat{x} . Wenn \hat{x} und x möglichst ähnlich sind, dann kann davon ausgegangen werden, dass die generierten Bilder des Netzwerks G in direkter Abhängigkeit von dem Input X stehen. Was hierbei berechnet wird, ist die durchschnittliche, absolute Abweichung der Pixelwerte. [27]

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_x[\mathcal{L}_1(F(G(x)) - x)] + \mathbb{E}_y[\mathcal{L}_1(G(F(y)) - y)] \quad (2.16)$$

Um die gesamten Kosten des CycleGANs zu erhalten, werden die bisher beschriebenen Kostenfunktionen addiert. Der Cycle Consistency Loss $\mathcal{L}_{cyc}(G, F)$ wird dabei mit einem absoluten Wert λ multipliziert, um die Kosten dieser Funktion im Vergleich zu den *Adversarial Losses* gewichten zu können. In der Veröffentlichung der CycleGANs wird ein λ von 10 verwendet. Somit wird mit einem vergleichsweise hohen Gewicht versehen, dass die generierten Bilder in direkter Abhängigkeit zu den Eingangswerten des CycleGANs stehen. Der Wert λ stellt einen Hyperparameter dar. [27]

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \cdot \mathcal{L}_{cyc}(G, F) \quad (2.17)$$

[27]

Einige Implementierungen teilen die Kostenfunktion auf einzelne Verluste für jeweils die Generatoren G und F sowie die Diskriminatoren D_y und D_x auf: [28] [29]

$$\mathcal{L}_G = \mathbb{E}_x[\log(D_Y(G(x)))] + \mathcal{L}_{cyc}(G, F) \quad (2.18a)$$

$$\mathcal{L}_F = \mathbb{E}_y[\log(D_X(F(y)))] + \mathcal{L}_{cyc}(F, G) \quad (2.18b)$$

$$\mathcal{L}_{D_y} = 0.5 \cdot \mathcal{L}_{GAN}(G, D_Y, X, Y) \quad (2.18c)$$

$$\mathcal{L}_{D_x} = 0.5 \cdot \mathcal{L}_{GAN}(F, D_X, Y, X) \quad (2.18d)$$

Das erlaubt ein separates Training der einzelnen KNNs des CycleGANs. Die Generatoren besitzen in ihrer Kostenfunktion lediglich den Teil des Adversarial Loss, den sie beeinflussen können. Dieser ist jedoch so abgeändert, dass die Generatoren einen möglichst hohen Wert für $D_y(G(x))$ beziehungsweise $D_x(F(y))$ anstreben. In diesem Fall führt dort ein Wert nahe 1 zu Kosten von Nähe 0. Das entspricht dem, dass die Generatoren versuchen, den entsprechenden Diskriminator zu Falschaussagen zu verleiten.

Die Qualität der Diskriminatoren wird anhand des gesamten Adversarial Loss gemessen. Diese Kosten werden bei den Diskriminatoren mit 0.5 multipliziert, damit die Diskriminatoren nicht schneller trainiert werden als die Generatoren [28]. Stattdessen wäre es auch möglich, die Trainingschritte der Generatoren doppelt so oft durchzuführen, wie die der Diskriminatoren [12].

Die Basis für G und F ist eine bestimmte CNN-Architektur, die sich Residual Neural Network (ResNet) nennt. Diese haben Forscher von Microsoft Research im Jahre 2015 eingeführt. Die Autoren der CycleGAN Veröffentlichung wählen diese Architektur, weil sie in einem ähnlichen Anwendungsfall bei anderen Autoren überzeugende Resultate gezeigt habe. Der Grundbaustein eines ResNet, der sogenannte *Residual Block*, ist in Abbildung 2.12 dargestellt. [30] [27]

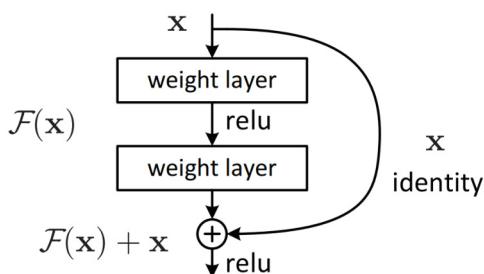


Abbildung 2.12: Residual Block eines ResNet [30]

Residual Blocks besitzen zwei Pfade. Ein Pfad enthält mindestens einen Convolutional Layer, der den Eingangswert x verändert. Das Ergebnis dieser Berechnung ist ein neuer Wert $\mathcal{F}(x)$. Der andere Pfad gibt den Eingangswert unverändert weiter und addiert ihn auf den Wert $\mathcal{F}(x)$. Das Ergebnis der beiden Pfade ist ein neuer Wert $\mathcal{F}(x) + x$. Die Aktivierungsfunktion des Residual Block erhält diese Summe als Eingabe. Residual Blocks beschleunigen laut den

Autoren das Training von tiefen neuronalen Netzen und erlauben damit Netzwerkarchitekturen mit mehr Schichten als klassische CNN. [30]

Conditional GANs Eine weitere Möglichkeit, die Generierung von GANs zu steuern, sind sogenannte Conditional Generative Adversarial Networks (cGANs). Dabei erhält der Generator zusätzlich zu dem zufällig erzeugten Eingangsvektor z noch eine Bedingung x . Das kann es zum Beispiel eine Zahl sein, die für ein bestimmtes Objekt steht, welches das cGANs erzeugen soll. Auch kann es sich beispielsweise um natürliche Sprache handeln, die das zu generierende Objekt beschreibt. Der Diskriminatator überprüft dann nicht nur, ob das Bild echt oder künstlich ist, sondern auch, ob es der Bedingung x entspricht.

Die Kostenfunktion lautet dann wie folgt: [31]

$$\min_G \max_D V(D, G) = \mathbb{E}_{x,y}[\log(D(x, y))] + \mathbb{E}_{x,z}[\log(1 - D(x, (G(x, z))))] \quad (2.19)$$

Wobei:

- x : Bedingung, die der Generator enthält beziehungsweise zu dem Bild y gehört
- y : Reales Bild aus den Trainingsdaten

Bei der Generierung von Straßenschildern könnte das x zum Beispiel eine Zahl sein, die für das Straßenschild steht, das der Generator erzeugen soll.

Zusätzlich ist es möglich, solche cGANs zur Bild-zu-Bild Übersetzung zu verwenden. Darauf bezieht sich gezielt eine bestimmte Veröffentlichung [31]. Dabei sorgt der zufällige Vektor z dafür, dass die Ausgabe des Generators nicht-deterministisch ist. Die Autoren der Veröffentlichung schreiben jedoch, dass der Generator das z meistens weitgehend ignoriert und dadurch wenig Varianz in den generierten Bildern zeigt. Der Quellcode der Veröffentlichung trägt den Namen *pix2pix*. [31]

Innerhalb von *pix2pix* basiert der Generator auf einem *U-Net*. Dies ist eine Architektur aus dem Jahre 2015, die das Eingangsbild zunächst auf eine geringere Anzahl an Parametern Kodiert und anschließend zu einem neuen Bild dekodiert. Die Architektur ähnelt somit konzeptionell der eines Autoencoders. U-Nets sind ursprünglich für die Segmentierung, also das markieren von bestimmten Bereichen in Bildern, für medizinische Zwecke entwickelt worden. Die Architektur dieser Modelle stellt Abbildung 2.13 dar. [32] [31]

Ihren Namen erhalten U-Nets durch die u-förmige Architektur. Sie erhalten ein Bild, extrahieren mittels mehrer Convolutional Layer daraus Merkmale und geben diese dann an weitere Convolutional Layer weiter, die hieraus ein neues Bild erzeugen. U-Nets besitzen in jeder kodierenden Schicht Direktverbindungen zu der zugehörigen dekodierenden Schicht. Das ist in

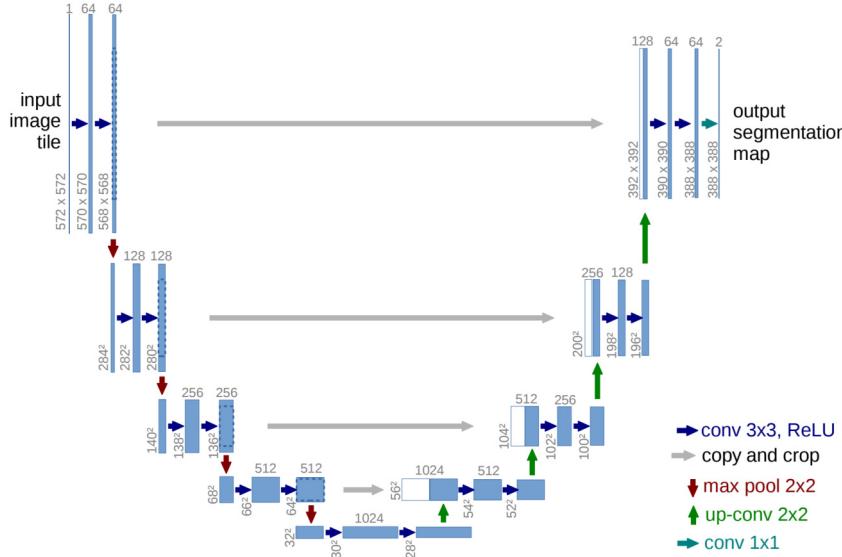


Abbildung 2.13: Architektur eines U-Net [32]

der Abbildung durch graue Pfeile dargestellt. Ein Vorteil dieser Verbindungen ist, dass dadurch weniger Informationen zwischen der Kodierung und der Dekodierung verloren gehen. [32]

Der Diskriminator des pix2pix cGAN verwendet eine Architektur, die sich *PatchGAN* nennt. Die Besonderheit ist hierbei, dass der Diskriminator nicht einzelne Pixel betrachtet, sondern Teile eines Bilds als echt oder unecht klassifizieren kann. Die gleiche Diskriminatorarchitektur verwendet auch die Veröffentlichung der CycleGANs. [31] [27]

Einer der Hauptunterschiede zwischen pix2pix und CycleGANs ist, dass pix2pix gepaarte Trainingsdaten benötigt. Zu jedem x müssen die Trainingsdaten das erwartete y enthalten. Nur so kann der Diskriminator lernen, welche Bilder zu einer bestimmten Bedingung gehören. CycleGANs brauchen das nicht. Sie benötigen Trainingsbilder aus der Domäne X und aus der Domäne Y, diese müssen jedoch in keiner Beziehung zueinander stehen. Deshalb fallen cGANs unter das überwachte Lernen, während CycleGANs dem unüberwachten Lernen zuzuordnen sind. [31] [27]

Analog zu den Vanilla GANs berechnen CycleGANs und cGANs keine direkte Wahrscheinlichkeitsverteilung der Trainingsdaten. Das Spiel zwischen Generator und Diskriminator sorgt implizit für eine Abbildung zwischen $\hat{p}(x)$ und $p(x)$. In der in Abbildung 2.8 gezeigten Taxonomie befinden sie sich aus diesem Grund in der Kategorie *Implizite Wahrscheinlichkeitsdichte*. [31] [27]

2.4 Vorherige Arbeiten

Durch eine Recherche haben sich zwei Arbeiten gezeigt, die sich, analog zu dieser Studienarbeit, mit der künstlichen Generierung von Straßenschildern mittels KNNs beschäftigen. Beide Arbeiten konzentrieren sich darauf, Bildausschnitte zu erzeugen, die ein Straßenschild zeigen und eine geringfügige Menge an Hintergrund um das Schild.

2.4.1 Generierung Taiwanischer Straßenschilder mittels DCGAN

Eine der beiden Arbeiten wurde im Jahr 2021 veröffentlicht. Sie konzentriert sich auf die Generierung taiwanischer Straßenschilder. Dafür verwenden die Autoren ein *DCGAN*. Das ist ein GAN, das im Generator und im Diskriminatator eine tiefe CNN Architektur besitzt. Die Autoren testen, inwiefern künstlich generierte Trainingsbilder die Erkennung von Straßenschildern verbessern können. Die Arbeit konzentriert sich auf die Generierung vier verschiedener Arten von Verkehrsschildern. [33]

Für jede der vier Klassen ist das DCGAN mit 350 Bildern trainiert. Die Bildgrößen variieren dabei, wobei die Maximalgröße bei 200x200 Pixel liegt. Die generierten Bildgrößen korrespondieren zu denen der Trainingsbilder. Es sollen in der Arbeit bewusst keine größeren Bilder als 200x200 Pixel erzeugt werden, da Straßenschilder laut den Autoren häufig nur einen kleinen Teil des Sichtfelds auf der Straße ausmachen. Das Training erstreckt sich auf bis zu 2000 Epochen. Die Qualität der generierten Bilder nimmt auch bei 1000 und 2000 Trainingsepochen zu. Das kann bei der Evaluierung dieser Studienarbeit einbezogen werden, da hier die Anzahl an Epochen um einen Faktor 10 geringer ist. [33]

Da die Anzahl an Trainingsbildern beschränkt ist, generiert das Modell keine völlig neuartigen Bilder, sondern für jede Klasse jeweils vergleichsweise ähnlich aussehende. Abbildung 2.14 zeigt beispielhaft einige der generierten Bilder. [33]



Abbildung 2.14: Beispieldarstellung von generierten taiwanischen Straßenschildern [33]

Zur Beurteilung der Generierung nutzen die Autoren mitunter den sogenannte Index struktureller Ähnlichkeit (engl.: Structural Similarity Index) (SSIM). Damit prüfen sie, wie ähnlich sich die

generierten Bilder und die Trainingsbeispiele sind. Statt dass die Differenz aller entsprechenden Pixelwerte berechnet wird, vergleicht der SSIM hier die Aspekte *Kontrast*, *Leuchtdichte* und *Struktur* der generierten und der echten Bilder. Dafür werden keine Berechnungen mit einzelnen Pixelwerten durchgeführt, sondern es wird mit den Mittelwerten und der Standardabweichung der Pixelwerte gerechnet. [33]

Den Nutzen der generierten Bildern testen die Autoren anhand eines Modells zur Objektdetection. In dem Fall ein sogenanntes *YOLO* Modell. Die Detektion erfolgt auf größeren Bildern, auf denen mehrere Straßenschilder zu sehen sind. Für das Training des Modells werden mit etwa gleicher Gewichtung die Trainingsdaten und generierte Daten des GANs verwendet. Zur Evaluation wurden hierbei Bilder verwendet, die insgesamt 40 Straßenschilder beinhalten. Die Ergebnisse können folgender Tabelle entnommen werden: [33]

Modell	Reale Trainingsbilder?	Künstliche Trainingsbilder?	Genauigkeit
DenseNet	Ja	Ja	92%
ResNet	Ja	Ja	91%
DenseNet	Ja	Nein	88%
ResNet	Ja	Nein	63%

Tabelle 2.1: Vergleich der Objekterkennung mit und ohne künstliche Trainingsdaten

Das Resultat ist, dass die Erkennung durch die generierten Trainingsdaten verbessert wird. Sowohl das DenseNet als auch das ResNet liefern durch sie genauere Ergebnisse. Diese Arbeit zeigt die Möglichkeiten, die künstlich erzeugte Bilder für die Straßenschilderkennung bieten können. Der konkrete Ansatz mit einem klassischen DCGAN spielt für diese Studienarbeit jedoch eine untergeordnete Rolle, da das Ziel ist, eine größere Bandbreite an Bildern zu erzeugen. [33]

2.4.2 Generierung Deutscher Straßenschilder mittels CycleGAN

Eine weitere Publikation, die sich mit der künstlichen Generierung von Bildern mit Straßenschildern konzentriert, verwendet einen Datensatz, der deutsche Straßenschilder enthält. Er ist unter dem Namen GTSRB bekannt. Auf den Datensatz wird näher in Kapitel 3 eingegangen, da er auch die Basis für diese Arbeit bildet. Die genannte Publikation ist aus einer Masterarbeit an der Ruhr Universität Bochum entstanden. Dort wurde auch der GTSRB veröffentlicht.

Die Veröffentlichung verwendet, im Gegensatz zu der bisher beschriebenen Generierung taiwanischer Schilder, ein CycleGAN statt eines *Vannila GANs*. Die beiden Generatoren basieren, wie in der CycleGAN-Veröffentlichung vorgeschlagen, auf einem ResNet [27]. [4] [34]

Für die Generierung erhält das Netzwerk das Piktogramm eines Straßenschildes als Eingang. Das CycleGAN soll einen möglichst realistisch wirkenden Hintergrund um das Straßenschild erzeugen. Vor der Generierung werden die Piktogramme der Straßenschilder zufällig rotiert und ein zufälliger einfarbiger Hintergrund erzeugt. Letzteres soll eine weitere stochastische Komponente für die Generierung bilden, damit eine größere Varianz an Hintergründen erzeugt wird. Für das Training des Netzes verwendet die Arbeit eine präparierte Version des GTSRB. Der Datensatz beinhaltet dadurch folgende Eigenschaften:

- Der Datensatz beinhaltet nur Bilder mit einer Mindestauflösung von 64x64 Pixeln
- Die Klassen sind ausbalanciert, um der asymmetrischen Verteilung an Trainingsbildern pro Klasse entgegenzuwirken
- Insgesamt besteht der präparierte Datensatz aus 12.212 Bildern

[34]

Abbildung 2.15 zeigt beispielhaft einige der generierten Bilder. Dabei zeigt die Abbildung paarweise links ein echtes Bild aus den Trainingsdaten und rechts daneben das generierte Bild, das die ähnlichsten Pixelwerte dazu besitzt. Es ist zu erkennen, dass die generierten Bilder den Trainingsdaten ähnlich sehen, während sie dennoch neuartig sind. [34]



Abbildung 2.15: Beispielergebnisse der Generierung deutscher Schilder [34]

Die Autoren führen eine Evaluation durch, inwiefern die künstlich generierten Trainingsbilder zwei verschiedene Klassifikatoren verbessern können. Dabei handelt es sich um eine sogenannte Support Vector Machine (SVM) und ein CNN. SVMs sind eine Art von trainierbaren Klassifikatoren, die nicht auf KNNs basieren [12]. Einerseits werden die Algorithmen mit realen Trainingsdaten trainiert und andererseits vollständig mit generierten Daten. Die Ergebnisse bezüglich der Genauigkeit der Klassifikation je nach der Art der Trainingsbilder ist in Tabelle 2.2 dargestellt. Es lässt sich erkennen, dass die Klassifikation um jeweils etwa 7-9% ungenauer ist, als mit realen Trainingsdaten. Es ist jedoch auch erwartbar, dass die Genauigkeit etwas

geringer ausfällt, da die generierten Bilder der Verteilung des echten Trainingsdatensatzes folgen sollen, dies jedoch nicht zu 100% möglich ist. [34]

Modell	Reale Trainingsbilder?	Künstliche Trainingsbilder?	Genauigkeit
CNN	Ja	Nein	95,42%
SVM	Ja	Nein	87,97%
CNN	Nein	Ja	87,57%
SVM	Nein	Ja	79,27%

Tabelle 2.2: Vergleich der Klassifikation mit echten und künstlichen Trainingsdaten [34]

Die Veröffentlichung spielt eine übergeordnete Rolle für diese Studienarbeit. Sie zeigt, dass die Generierung deutscher Straßenschilder mit einem CycleGAN möglich ist.

2.5 Machine Learning Frameworks

Es ist möglich, KNNs von Grund auf zu programmieren. Die Hauptaufgabe von Entwickelnden besteht jedoch darin, sowohl den Datensatz als auch die Architektur sowie die Hyperparameter des Modells zu entwerfen und anzupassen. Verschiedene *Frameworks* bieten für die Berechnung der Vorhersagen und das Optimieren eines KNN bereits Funktionen an. Sie sorgen zudem dafür, dass diese Funktionen möglichst performant sind. Hierfür spielen vor allem Grafikkarten (GPUs) eine bedeutende Rolle. Neben sogenannten Tensor Processing Units (TPUs) und Field Programmable Arrays (FPGAs) werden nämlich in erster Linie GPUs für das Berechnen von KNNs eingesetzt, da sie dafür performanter sind als Prozessoren. [35]

Einige Frameworks im Bereich des maschinellen Lernens unterstützen eine Berechnung auf GPUs. Dazu zählen unter anderem **TensorFlow**, **PyTorch**, **MXNet**, **Microsoft CNTK** und **Caffe**. Eine Veröffentlichung aus dem Jahre 2019 vergleicht dabei mitunter diese Frameworks. Das am meisten verbreitete Framework sei dabei TensorFlow. Es wurde im Jahre 2015 von der Firma Google entwickelt und ist, wie die meisten anderen genannten Frameworks, überwiegend in der Programmiersprache C++ geschrieben. PyTorch stammt von der Firma Facebook und basiert auf dem Framework *Torch*. Einige Frameworks wie Caffe sind für spezielle Anwendungsbereiche gedacht, wohingegen beispielsweise TensorFlow und PyTorch allgemein Anwendung finden. Anwendende können die Funktionen aller genannten Frameworks mit der Sprache Python nutzen, welche als die für das maschinelle Lernen am meisten eingesetzte Programmiersprache gilt. [35]

Eine bestimmte Art und Weise, wie Frameworks KNNs optimieren ist im Verlauf dieser Arbeit relevant: Viele der genannten Frameworks unterteilen Datensätze in sogenannte *Batches*. Jeder Batch beinhaltet einen Teil des Datensatzes. Eine *Batch Größe* (engl.: *batch size*) legt fest, wie viele Elemente sich in einem Batch befinden. Besitzt der Datensatz 1024 Bilder und das KNN eine Batch Größe von 16, dann wird der Datensatz in 64 Batches unterteilt, da $\frac{1024}{16} = 64$. Es ist möglich, alle Elemente eines Batches gleichzeitig in ein neuronales zu speisen. Bei einer Batch Größe von 16 erhält das KNN pro Trainingsschritt 16 Eingaben und trifft somit eben so viele Vorhersagen. In einem Trainingsschritt wird das KNN dann auf allen 16 Vorhersagen trainiert. Die Frameworks sorgen dafür, dass die Batches dynamisch in den Arbeitsspeicher geladen werden. Somit muss der Arbeitsspeicher keine Kapazität für den gesamten Datensatz besitzen. [36]

Weiterhin ist mitunter in TensorFlow und PyTorch der Begriff des *Tensors* relevant. Für diese Arbeit kann ein Tensor als ein mehrdimensionales Array betrachtet werden. Tensoren werden mit einer Stufe beschrieben, die ihre Dimensionalität angibt. Ein Skalar hat die Stufe 0, ein Vektor die Stufe 1 und eine Matrix die Stufe 2. Weiterhin besitzen Tensoren eine Form. Ein Tensor dritter Stufe der Form (256, 256, 3) besitzt eine Höhe und Breite von 256 sowie eine Tiefe von 3. Das kann zum Beispiel ein digitales Bild mit den Pixelmaßen 256x256 und drei Farbkanälen sein. Betrachtet man einen Batch solcher Bilder mit einer Batch Größe von 16, dann erhält man einen Tensor der Stufe vier mit der Form (16, 256, 256, 3). Vorstellen kann man sich das als 16 übereinander gestapelte Bilder. Würde man nun zwei solcher Batches in einem Tensor zusammenfassen, dann hätte dieser Tensor die Form (2, 16, 256, 256, 3) [37].

3 | Konzeption des Modells

Dieses Kapitel beschreibt, basierend auf den vorgestellten Grundlagen, das Konzept des Modells. Das Kapitel bezieht sich dabei ausschließlich auf die Generierung der Bilder, nicht auf die Augmentation der Bilder durch Grenzfälle für die Straßenschilderkennung. Letzteres ist in Kapitel 5 erläutert.

3.1 Datensatz

Analog zu der in Kapitel 2.4.2 beschriebenen Arbeit verwendet diese Studienarbeit den GTSRB als Datensatz. Dass dies mit 39.209 Bildern der größte veröffentlichte Datensatz für deutsche Straßenschilder ist, stellt hierbei den ausschlaggebenden Punkt dar.

Die Bilder des GTSRB verteilen sich auf 43 Klassen respektive 43 verschiedene Arten von Straßenschildern. Eine Auflistung aller Klassen ist im Anhang in Abbildung A.1 dargestellt. Beispielbilder aus dem GTSRB zeigt Abbildung 3.1. [4]



Abbildung 3.1: Beispielbilder aus dem GTSRB Datensatz [4]

Der GTSRB setzt sich aus Bildern zusammen, die unterschiedliche Seitenverhältnisse und verschiedene Auflösungen besitzen. Ein Großteil davon ist kleiner als 100x100 Pixel. Auf jedem Bild ist genau ein Straßenschild zu sehen. Die Bilder basieren auf Videos, die durch die Autoren tagsüber im Straßenverkehr aufgenommen wurden. Dabei sind die Trainingsbilder ungleich auf die Anzahl an Klassen verteilt. Dies hängt mitunter damit zusammen, dass die jeweiligen Schilder nicht gleich häufig im Straßenverkehr vorkommen. Zusätzlich zu den Trainingsbildern besitzt der GTSRB 12.630 Testbilder, welche in dieser Arbeit zur Evaluation des Modells verwendet werden können. Eine nennenswerte Eigenschaft des GTSRB ist, dass eine signifikante Anzahl an Bildern einer Klasse sich ähnlich sehen. Das hängt damit zusammen, dass sie mit zeitlicher Verzögerung aus der selben Fahrsituation stammen. [4]

Die Bilder, die durch diese Studienarbeit generiert werden sollen, haben eine Auflösung von 256x256 Pixel. Unter anderem deshalb ist der GTSRB für diese Arbeit zunächst so präpariert, dass nur Bilder verwendet werden, die mindestens 50 Pixel breit oder hoch sind. Dies wird im Verlauf geändert, sodass die Mindestgröße 75 Pixel beträgt. Das verringert die Anzahl an verfügbaren Trainingsbildern signifikant. Der präparierte Datensatz besteht aus 4.510 Bildern, wodurch nur etwa 11% des GTSRB genutzt werden. Eine qualitative Analyse hat ergeben, dass sich Bilder ab einer Größe von etwa 75 Pixel für den Trainingsdatensatz eignen.

Die Verteilung der Daten ist auch im präparierten Datensatz nicht homogen. Das nachfolgende Diagramm zeigt hierfür die Anzahl an Trainingsbildern pro Klasse.

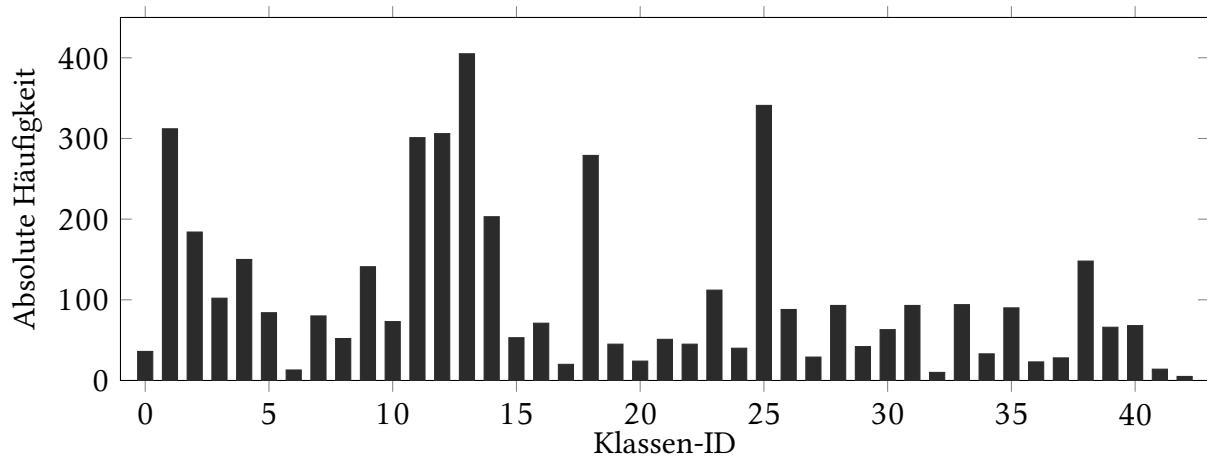


Abbildung 3.2: Häufigkeitsverteilung der Klassen von Straßenschildern im präparierten GTSRB

Eine ungleiche Verteilung der Trainingsdaten kann die Qualität der generierten Bilder negativ beeinflussen. Die in Kapitel 2.4.2 beschriebene Veröffentlichung gibt hierauf bereits Hinweise [34]. Aus diesem Grund, und da die Größe des präparierten Datensatzes als zu gering bewertet werden kann, ist der Datensatz derart erweitert, dass er einerseits mehr Trainingsbilder enthält und andererseits eine gleichmäßige Verteilung der Klassen vorliegt. Dabei wird nicht darauf geachtet, jede einzelne Klasse möglichst gleich oft zu repräsentieren, sondern jede Kategorie von Klassen. Die 43 Klassen werden dazu in die Kategorien *Geschwindigkeitsbegrenzungen*, *Richtungsweiser*, *Aufhebungen*, *Verbotszeichen*, *Gefahrzeichen* und *Einzigartig* unterteilt. Es zeigt sich nämlich, dass das Modell zwischen Straßenschildern, die eine ähnliche Bedeutung und damit auch äußerliche Ähnlichkeiten besitzen, recht gut transferieren kann. Diese Einteilung der Schilder in unterschiedliche Kategorien erfolgt auch in der ursprünglichen Veröffentlichung des GTSRB. Die Kategorisierung für diese Studienarbeit ist identisch, verwendet jedoch deutsche Bezeichnungen für die Kategorien. Zu der Kategorie *Einzigartig* zählen die Kategorien mit den Nummern 12, 13, 14 und 17 (siehe Abbildung A.1).

Der Großteil an hinzugefügten Trainingsdaten stammt aus der chinesischen *Traffic Sign Recognition Database*, die ein Teil der *Chinese Traffic Sign Database* ist. Dieser Datensatz ist bedeutend

kleiner als der GTSRB, bietet jedoch auch einige Bilder mit einer höheren Auflösung als der GTSRB. Somit ist hier ein größerer Anteil des Datensatzes nutzbar. Allgemein ähneln diese Bildern denen des GTSRB. Mit dem Unterschied, dass sie chinesische Straßenschilder zeigen. Für den präparierten Datensatz werden nur die Bilder verwendet, die in eine der genannten Kategorien fallen. Nachfolgend sind Beispield Bilder aus dem Datensatz dargestellt: [38]



Abbildung 3.3: Beispielbilder aus der chinesischen Traffic Sign Recognition Database [38]

Zu sehen ist in Abbildung 3.3 eine Geschwindigkeitsbegrenzung, ein Richtungsweiser, ein Verbotszeichen und ein Schild der Kategorie *Einzigartig*. Der präparierte Datensatz beinhaltet auch Schilder, die nicht durch das Modell dieser Studienarbeit generiert werden sollen. Wie etwa die in Abbildung 3.3 vorhandene Geschwindigkeitsbegrenzung von $15 \frac{km}{h}$. Die Idee ist, dass das Modell diese Bilder dennoch nutzen kann, um die Generierung von anderen Geschwindigkeitsbegrenzungen zu optimieren. Es sind allgemein Unterschiede zu deutschen Straßenschildern vorhanden, die jedoch in dieser Arbeit als vernachlässigbar angenommen werden. Zumindest dann, wenn deutsche Straßenschilder weiterhin den größten Teil des präparierten Datensatzes ausmachen. Eine gewisse Ähnlichkeit ist vorhanden, auch da das Aussehen von Straßenschildern durch das Wiener Übereinkommen über Straßenverkehrszeichen in vielen Ländern weltweit vereinheitlicht ist [39]. [38]

Zusätzlich setzt sich der präparierte Datensatz aus Bildern weiterer Datensätze zusammen. Hier ist jedoch die Anzahl an Bildern signifikant geringer als die der chinesischen Traffic Sign Recognition Database. Zwei der Datensätze bestehen aus Bildern, die eine vollständige Sicht außerhalb des Fahrzeugs zeigen. Hier sind demnach gegebenenfalls mehrere Straßenschilder pro Bild zu sehen, wobei zusätzlich andere Fahrzeuge, Gebäude, Personen und weitere Objekte sichtbar sind. Die Datensätze nennen sich *Mapillary Traffic Sign Dataset* und *BelgianTS Dataset* [40] [41].

Da diese Bilder manuell so zugeschnitten werden müssen, dass sie ähnlich zu dem GTSRB und der chinesischen Traffic Sign Recognition Database einzelne Schilder mit wenig Hintergrund zeigen, macht diese Menge an Bildern einen geringeren Anteil aus. Aus einem dritten Datensatz werden weniger als 50 Bilder verwendet [42]. Hier sind größtenteils Stoppschilder zu sehen mit einer Bildauflösung von etwa 190 bis zu 300 Pixel in der Höhe oder Breite.



Abbildung 3.4: Beispielbild aus dem *Mapillary* Datensatz [40]

Für diese Studienarbeit existiert ein Ordner `data`, der alle für die Arbeit relevanten Bilddateien enthält. Der Ordner ist öffentlich unter [diesem Link](#) verfügbar (Stand: 09.06.2023). In dem Unterordner `Train` befindet sich dort der gesamte Trainingsdatensatz.

Nachfolgende Abbildung zeigt die Häufigkeitsverteilung dieser Trainingsdaten (x-Achse) je Straßenschild-Kategorie (y-Achse). Die drei letztgenannten Datensätze sind in der Farbkodierung der Kategorie *Sonstige* zugeordnet.

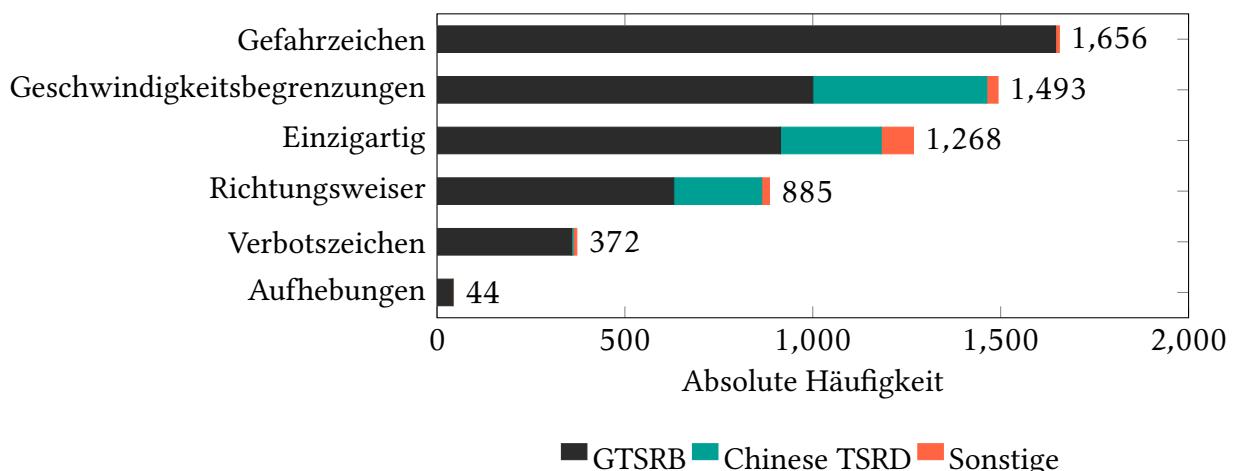


Abbildung 3.5: Häufigkeitsverteilung der Kategorien von Straßenschildern im präparierten Datensatz

Der Datensatz besteht damit aus 5.809 Bildern.

Die Kategorie *Aufhebungen* ist nach wie vor signifikant unterrepräsentiert. Eine mögliche Lösung, für die der zeitliche Rahmen nicht ausreicht, wäre, manuell auf [Mapillary](#) nach solchen Bildern zu suchen. Mapillary ist eine Webseite, auf der Nutzer Bilder zu bestimmten GPS-Koordinaten hochladen können. Das Ziel von Mapillary ist, Straßenansichten für alle Straßen auf der Welt bereitzustellen. Hier ist es möglich, explizit nach bestimmten Arten von Straßenschildern zu suchen. Einer der genannten Datensätze stammt aus einer Mischung von weltweiten Mapillary-Bildern. [40]

3.2 Framework

Für das Modell ist nicht nur relevant, auf welchem Datensatz es trainiert wird. Die Wahl eines Frameworks hat einen signifikanten Einfluss die Art der Implementierung. Die Wahl liegt darauf, **TensorFlow** für die Umsetzung zu nutzen. Das hängt einerseits damit zusammen, dass es als das am meisten verbreitete Framework für maschinelles Lernen gilt [35]. Somit existieren hier einige vorgefertigte Implementierungen, die als Basis für die Studienarbeit genutzt werden können. Erwähnenswert ist, dass das bei anderen Frameworks wie etwa PyTorch jedoch ebenso der Fall ist. Ein weiterer Vorteil von TensorFlow ist, dass es standardmäßig von **Google Colab** unterstützt wird. Bei Google Colab ist es möglich, Rechner von Google zu benutzen, um Modelle zu trainieren. Außerdem ist TensorFlow nicht nur in der Forschung, sondern auch in der Industrie verbreitet. Diese Studienarbeit soll ein Verfahren darlegen, das in der Art auch in der Industrie eingesetzt werden könnte. Somit soll hier eine Basis genutzt werden, die dort bereits Verwendung findet. Es wäre jedoch ebenso möglich, ein anderes Framework zu nutzen um vergleichbare Ergebnisse zu erzielen.

Ein Nachteil, den die Literatur bei TensorFlow nennt, ist auf folgendes zurückzuführen: TensorFlow 1.0 übersetzt den Quellcode in Graphen. Das verbessert die Performanz der Berechnungen, sorgt aber dafür, dass die Berechnungen während der Laufzeit statisch sind. Außerdem benötigen einige Operationen dadurch eine spezielle Syntax. Auch erschwert das eine Anbindung anderer Python-Bibliotheken. TensorFlow 2.0, eine neuere Version des Frameworks, erlaubt jedoch zusätzlich eine *Eager Execution*. Hierbei erstellt das Framework standardmäßig keine Graphen. Das Ziel von TensorFlow 2.0 ist unter anderem, einfachere Programmierschnittstellen zu bieten und somit den genannten Nachteil zu beheben. [37]

TensorFlow ist analog zu anderen Frameworks für das maschinelle Lernen auf eine performante Ausführung ausgelegt. Die Geschwindigkeit der Berechnungen hat Auswirkungen auf die Trainingsdauer des Modells. Aus diesem Grund sollen möglichst viele Funktionen in dieser Studienarbeit ausschließlich mit TensorFlow implementiert werden. Andere Python-Pakete werden dann eingesetzt, wenn TensorFlow zu einem bestimmten Problem keine Funktion bietet. Zur Bildverarbeitung stellt TensorFlow die Bibliothek **TensorFlow Graphics** zur Verfügung [43]. Hiermit können nicht nur einzelne Bilder bearbeitet werden, sondern gesamte Batches von Bildern. Die Bibliothek **TensorFlow Addons** bietet zudem zusätzliche Funktionen an, die standardmäßig nicht in TensorFlow vorhanden sind. In dem Code dieser Arbeit werden TensorFlow, TensorFlow Graphics und TensorFlow Addons als `tf`, `tfg` und `tfa` referenziert.

In Fällen, in denen TensorFlow Graphics keine geeigneten Funktionen besitzt, nutzt diese Studienarbeit **OpenCV** oder **Pillow**. Beides sind Python-Pakete zur Bildverarbeitung, wobei erstere vor allem für Computer-Vision-Aufgaben gedacht ist, während Pillow das Manipulieren von Bilddateien erlaubt. [44] [45]

Außerdem existiert die Bibliothek **NumPy**, die mathematische Operationen auf Vektoren und Matrizen (*hier NumPy-Arrays*) ermöglicht. Einige Funktionen aus TensorFlow, Pillow und OpenCV erlauben NumPy-Arrays als Parameter. Auch bestitzt TensorFlow Funktionen, um Tensoren und NumPy-Arrays ineinander umzuwandeln. [46]

3.3 Architektur

In Kapitel 2.3 sind verschiedene generative Netzwerkarchitekturen vorgestellt. Jede dieser Architekturen besitzt ihre Vor- und Nachteile. Die Entscheidung liegt darauf, kein PixelRNN zu benutzen, da die ausschließlich sequentielle Generierung die Performanz des Modells beeinträchtigt. GANs gelten von den vorgestellten Modellen als die am schwierigsten zu trainierende Kategorie. Das hängt vor allem damit zusammen, dass die Kostenfunktion nicht gegen *null* streben soll, sondern gegen einen Wert größer als *null* konvergieren soll. Das Training gilt als instabil, da die Kostenfunktion oszillieren kann und in dem Fall nicht konvergiert. Weiterhin ist es möglich, dass der Generator oder der Diskriminatior jeweils den Gegenspieler soweit überlistet, dass das Modell nicht mehr lernt. Da die Literatur GANs nachgesagt, qualitativ hochwertigere Bilder zu erzeugen als Variational Autoencoder, sollen trotzdem GANs die Basis für diese Studienarbeit bilden. [19] [18]

Die Problemstellung dieser Arbeit soll jedoch nicht als eine reine Bildgenerierung interpretiert werden, sondern als eine Bild-zu-Bild Übersetzung. Damit muss das Netzwerk nicht von alleine lernen, die Symbole der Straßenschilder zu erzeugen. Das ist aus einem bestimmten Grund von Vorteil: Es heißt in der Literatur, dass GANs nicht sonderlich gut darin seien, bestimmte Formen exakt zu erzeugen [19]. Eine reine Bildgenerierung mit einem zufälligen Eingangsvektor könnte also dazu führen, dass das GAN verschwommene oder verformte Schilder erzeugt. Außerdem soll das Netzwerk lernen, verschiedene Arten von Straßenschildern zu erzeugen. Am besten so, dass Anwendende die Arten der generierten Schilder selbst bestimmen können. Das wäre mit einem klassischen cGAN möglich. Da jedoch Straßenschilder genormt sind und somit die Schilder eines Landes stets identisch aussehen, ist es auch möglich, dem GANs das zu erzeugende Schild bereits in minimaler Form als Eingangsbild zu geben. Das Modell muss dann lernen, dieses Bild in die Zieldomäne Y zu übersetzen, die das Schild in einer möglichst fotorealistischen Umgebung zeigt. Was in dem Fall variabel ist, ist die Perspektive des Schildes, die Helligkeit des Bilds sowie das Aussehen der Umgebung. Hier soll das Modell eine möglichst Große Varianz erzeugen.

Die Basis für die Bild-zu-Bild Übersetzung bilden Piktogramme von Straßenschildern. Entnommen sind diese Piktogramme von der Internetseite der *Bundesanstalt für Straßenwesen*. Sie zeigen das jeweilige Straßenschild-Symbol auf einem hellgrauen Hintergrund. Daraus wird ein

zusätzlicher Datensatz an Bildern erstellt, der die Domäne X darstellt. Er beinhaltet die Piktogramme für alle 43 Klassen von Straßenschildern, die in dem GTSRB vorkommen. Im Anhang befindet sich eine Abbildung, die alle Piktogramme zeigt. Abbildung 3.6 soll verdeutlichen, was die Domänen X und Y sind, die das Modell ineinander übersetzen soll. [47]



Abbildung 3.6: Domänen für die Bild-zu-Bild Übersetzung

Eine letzte Fragestellung ist, ob cGANs analog zu pix2pix verwendet werden sollen oder aber CycleGANs. Wie bereits beschrieben, benötigt pix2pix gepaarte Trainingsdaten. Zu jedem echten Bild aus Y muss hinterlegt werden, welches Piktogramm aus X dazu gehört. Das kann insofern als unproblematisch betrachtet werden, als dass die Bilder des GTSRB nach ihren Klassen sortiert sind. Soll diese Studienarbeit jedoch mit einem größeren Datensatz fortgeführt werden, ist es von Vorteil, wenn die Trainingsdaten nicht annotiert werden müssen. Außerdem schreiben die Autoren von pix2pix, dass die erzeugten Bilder keine sonderlich hohe Varianz aufzeigen würden, da, wie bereits erwähnt, der Vektor z einen geringen Einfluss auf die Generierung habe. Die Veröffentlichung der CycleGANs baut auf pix2pix auf. Es wird davon ausgegangen, dass die Bildgenerierung deshalb nicht signifikant schlechter, oder noch besser ist als mit pix2pix, während das Modell die genannten Vorteile besitzt. Aus diesem Grund basiert das Modell für diese Studienarbeit auf einem CycleGANs.

3.4 Datenaugmentation

Bevor die Piktogramme an den Generator übergeben werden, werden sie zufällig rotiert. Dadurch muss der Generator die Rotation nicht eigenständig lernen und dieser Aspekt der Generierung lässt sich deterministisch bestimmen. Dabei soll die Rotation nicht nur in x-y-Richtung erfolgen, sondern auch eine dreidimensionale Rotation simuliert werden. Und zwar so, als sei das Schild aus einer beliebigen Frontalperspektive aufgenommen worden.

Um bestimmte Transformationen eines Bilds mittels einer Matrixmultiplikation darstellen zu können, wird häufig ein sogenanntes *homogenes Koordinatensystem* verwendet. Dabei wird das Koordinatensystem um eine weitere Dimension erweitert. Ein Punkt $p = [x, y]^T$ kann somit um einen beliebigen Wert in z-Richtung verschoben werden. Dadurch wird ein Punkt \tilde{p} im homogenen Koordinatensystem durch drei Koordinaten \tilde{x} , \tilde{y} und \tilde{z} beschrieben. Transformationen werden in der homogenen Darstellung durchgeführt und anschließend

werden daraus die kartesischen Koordinaten x und y bestimmt. Somit erhält man aus der Transformation erneut ein zweidimensionales Bild. [48] [49]

Dies wird für eine dreidimensionale Rotation der Piktogramme benötigt. Die Rotation soll durch drei *eulersche Winkel* beschrieben werden. Das bedeutet, dass sie sich aus einer Rotation um die z-Achse, einer um die y-Achse und einer um die x-Achse zusammensetzt. Dies ist in Abbildung 3.7 gezeigt. Die bläulichen Balken zeigen dabei die Achse an, um die gedreht wird. Die erste Rotation ist um die z-Achse, wodurch der Balken in die dritte Bildebene geht. [49]



Abbildung 3.7: Rotation der Straßenschilder mittels eulerscher Winkel

Jede Rotation ist durch einen einzelnen Winkel um die jeweilige Achse bestimmt. Kombiniert man die Rotationen, kann die resultierende Transformation somit durch drei Winkel ($\alpha_z, \alpha_y, \alpha_x$) eindeutig beschrieben werden. Für die Erzeugung einer zufälligen Rotation müssen randomisierte Werte für diese Winkel bestimmt werden. [49]

Zusätzlich zu der Rotation, soll das Modell die Piktogramme zufällig in ihrer Größe skalieren. Die genannten Augmentationen dienen dazu, die Verteilung der real aufgenommenen Schilder abbilden zu können. Im Datensatz besitzen die Schilder eine unterschiedliche Größe und sind aus verschiedenen Perspektiven aufgenommen. Dadurch dass die Augmentation deterministisch ist, kann sie dazu genutzt werden, um gezielt nur Bilder durch das Modell zu generieren, die aus bestimmten Perspektiven und mit festgelegten Größen generiert wurden. Alternativ kann auch die randomisierte Augmentation beibehalten werden, um eine möglichst große Bandbreite an unterschiedlichen Bildern zu erzeugen.

3.5 Training

Das Training basiert auf den in Kapitel 2.3.3 vorgestellten Verlustfunktionen für CycleGANs. In Abbildung 3.8 sind hierfür die drei Trainingsschritte des CycleGAN dargestellt. Die ersten beiden Schritte berechnen den *Adversarial Loss* von Generator G und Diskriminatoren D_Y , beziehungsweise von Generator F und Diskriminatoren D_X . Was hier trainiert wird, ist die Übersetzung von Domäne X in Y, beziehungsweise von Domäne Y in X. Im Anschluss daran erfolgt die Berechnung des *Cycle Consistency Loss*. Das ist der Trainingsschritt der überprüfen soll, dass die von Generator G erzeugten Bilder das erwartete Straßenschild zeigen. Dazu erzeugt

G aus einem Piktogramm das Bild eines Straßenschildes woraus F wiederum das Piktogramm erzeugen soll. Der Generator G ist hierbei grün hervorgehoben, da dies das einzige KNN ist, dass für die praktische Generierung von Bildern verwendet wird. Die weiteren KNNs sollen lediglich den Generator G trainieren. [27]

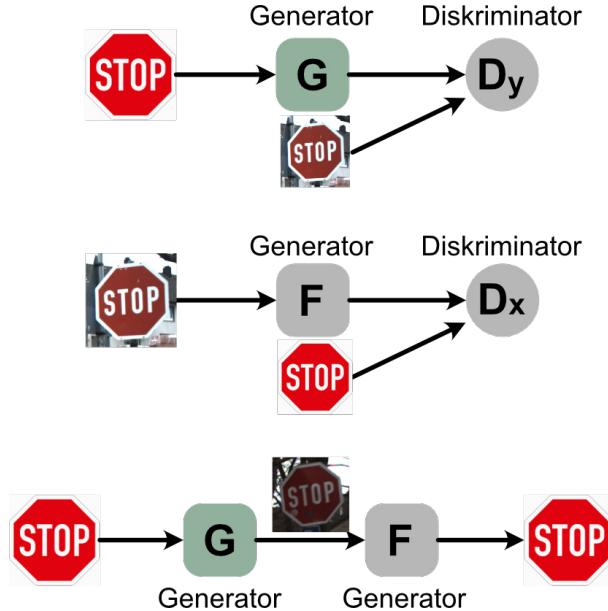


Abbildung 3.8: Trainingsschritte des CycleGAN

Für bestimmte Anwendungsfälle schlägt das CycleGAN Paper vor, einen *Identity Loss* hinzuzufügen. Dabei wird Generator G ein echtes Bild eines Straßenschildes und Generator F ein echtes Bild eines Piktogramms zugeführt. Da das Eingabebild für G und F bereits aus der Zieldomäne entstammt, wird hier von den Generatoren erwartet, dass sie das Eingabebild möglichst wenig verändern. Die Veröffentlichung schlägt das vor, wenn die Generatoren beispielsweise die Farben der Eingangsbilder beibehalten sollen. [27]

Die Vermutung ist, dass der Identity Loss auch für diese Studienarbeit sinnvoll sein kann. Hierdurch könnte das Netzwerk dazu gebracht werden, das Straßenschild möglichst wenig zu verformen und es könnte aus den Eingabebildern erlernen, verschiedene Hintergründe um die Schilder zu erzeugen. Deshalb wird der Identity Loss in diesem Projekt erprobt. Eine Veröffentlichung deutet außerdem darauf hin, dass der Identity Loss die allgemeine Qualität der generierten Bilder verbessern kann [50]. [27]

Zusätzlich schreiben die Autoren der CycleGAN Veröffentlichung, dass es sinnvoll sein kann, den *Adversarial Loss* mit einer \mathcal{L}_2 Verlustfunktion zu berechnen statt mit einer Binary Cross-entropy Verlustfunktion. Das soll das Training stabilisieren. Die pix2pix-Veröffentlichung erwähnt das ebenfalls [31]. Streng genommen handelt es sich dabei dann bei den Paaren G und D_y , beziehungsweise F und D_x nicht mehr um klassische GANs, sondern um sogenannte *least squared GANs (LS-GANs)*. In dieser Studienarbeit soll das Training zunächst mit einer

klassischen CycleGAN Architektur erfolgen. Zeigt sich ein instabiles Training, soll getestet werden, ob sich das Training mittels \mathcal{L}_2 Verlustfunktionen stabilisieren lässt. [27]

Das Training gilt als beendet, wenn die Verlustfunktionen des Modells gegen einen Wert konvergieren. Um das zu messen, ist eine Form des *Loggings* notwendig. Es müssen demnach über den Verlauf des Trainings die Werte der Kostenfunktionen gespeichert werden.

4 | Implementierung und Training

Die Implementierung des CycleGAN basiert auf den bisher beschriebenen Entscheidungen. In einem Ordner `src` befinden sich alle Dateien der Implementierung. Hier existieren neben Dateien für das Training und die Anwendung des Modells auch eine Konfigurationsdatei und verschiedene Hilfsfunktionen. Die Ordnerstruktur der Implementierung ist im Anhang abgebildet. Im wesentlichen geht dieses Kapitel auf folgende Bestandteile ein:

- Die Konfigurationsdatei für das Projekt befindet sich im Pfad `config/config.toml`
- Selbst-definierte Hilfsfunktionen befinden sich im Python-Modul `utils`
- Die Datei `model.py` implementiert das CycleGAN Modell
- Das Modell wird mittels des Python-Skripts `train.py` trainiert
- Das Modell kann mittels des Python-Skripts `generate.py` zum Generieren von Bildern genutzt werden

Das CycleGAN ist als Klasse implementiert und lässt sich somit innerhalb der Skripte instanzieren. Die Konfigurationsdatei ist eine TOML-Datei. Tom's Obvious Minimal Language (TOML) ist eine Konfigurationssprache, die Daten als Schlüssel-Werte-Paare speichert. Die Konfigurationsdatei ist in die Kategorien `paths`, `model` und `training` unterteilt. Anwendende können hier die Pfade zu den Trainingsdaten oder aber Parameter wie `batch_size` oder `number_of_epochs` angeben.

4.1 Modell

Das Modell ist in der Datei `model.py` implementiert. Innerhalb dieser Datei existiert eine Klasse `CycleGan`, die im wesentlichen die in Tabelle 4.1 aufgeführten Methoden besitzt. Der Inhalt dieser Klasse basiert zum Teil auf einer Beispiel-Implementierung von TensorFlow [28]. Teile des Quellcodes stellen die Basis für die Umsetzung der Studienarbeit dar. In erster Linie basieren die Verlustfunktionen des CycleGAN auf dem Beispiel sowie der Fakt, dass das Modell die Bilder auf Pixelwerte zwischen -1 und 1 skaliert. Die Skalierung kann das Training des Modells beschleunigen und stabilisieren [51]. Außerdem implementiert der Quellcode bereits eine Möglichkeit, die trainierten Parameter des Modells in sogenannten *Checkpoints* zu

¹ Vollständiger Methodename: `restore_latest_checkpoint_if_exists`

Methode	Aufgabe
<code>__init__</code>	Initialisieren des CycleGAN mittels der Konfigurationsdatei
<code>generate</code>	Generieren eines einzelnen Batches von Bildern
<code>fit</code>	Trainieren des CycleGANs über mehrere Epochen
<code>train_step</code>	Durchführen eines einzelnen Trainingsschritts
<code>restore_..._checkpoint_...¹</code>	Laden der aktuell gespeicherten Parameter des Modells

Tabelle 4.1: Auswahl an Methoden aus der CycleGAN Klasse

speichern. Das Modell aus dem TensorFlow Beispiel verwendet einen U-Net Generator, wie es normalerweise bei pix2pix statt bei CycleGANs der Fall ist. Die Implementierung dieser Studienarbeit erprobt ein CycleGAN mit einem U-Net Generator, verwendet jedoch zusätzlich einen ResNet Generator und den dazugehörigen Diskriminatoren. Diese befinden sich in dem Pfad `src/external/resnet.py` und entstammen ebenfalls einer bereits existierenden CycleGAN Implementierung [29]. Die Architekturen der KNNs entsprechen dabei den Vorgaben der CycleGAN Veröffentlichung [27]. Die Klasse `CycleGAN` erlaubt eine Auswahl zwischen der U-Net- und der ResNet-Architektur. Das Ziel dieser Studienarbeit ist damit unter anderem zu prüfen, ob ein U-Net- oder ein ResNet-basierter Generator für diesen Anwendungsfall besser geeignet ist.

Nachfolgend soll näher auf einige Methoden der Klasse CycleGAN eingegangen werden.

4.1.1 Konstruktor

Die `__init__`-Funktion stellt in Python den Konstruktor einer Klasse dar. Die Klasse `CycleGAN` erwartet hier den Parameter `config`. Dies ist ein Python-Dictionary, das die Werte der Konfigurationsdatei enthält. Es existieren hier somit Schlüssel-Werte-Paare für die verschiedenen Trainingseinstellungen des Modells. Ein Vorteil dieses Ansatzes ist, dass Anwendende des Modells die Parameter in der TOML-Datei ändern können, ohne den Python-Quellcode aufrufen zu müssen. Allgemein verfolgt die Implementierung des CycleGAN das Ziel, dass Anwendende sich mit keinem Python Quellcode auseinander setzen müssen, um das Modell zu trainieren und zu verwenden.

Eines der Attribute der Klasse `CycleGAN` heißt `generator_type`. Es handelt sich dabei um ein Enum, das bestimmt, ob der Generator ein U-Net oder ein ResNet ist. Der Wert für das Attribut ist durch die Konfigurationsdatei und somit durch den Parameter `config` bestimmt. In Listing 4.1 ist der Codeabschnitt zu sehen, der die `generator_type`-Eigenschaft der Klasse `CycleGAN` initialisiert.

```

class GeneratorType(Enum):
    RESNET = 0
    U_NET = 1

class CycleGan:
    ...
    def __init__(self, config):
        ...
        if config['model']['generator_type'] == 'unet':
            self.generator_type = GeneratorType.U_NET
        elif config['model']['generator_type'] == 'resnet':
            self.generator_type = GeneratorType.RESNET
        ...

```

Listing 4.1: model.py - Auswahl der Generator-Architektur

Das Modell verwendet den U-Net Generator mit drei Farbkanälen und einer *Instance Normalization* (hier `'instancenorm'`) statt einer *Batch Normalization*. Die pix2pix-Veröffentlichung schlägt dies für die Bildgenerierung vor [31]. Beides sind bestimmte Schichten in CNNs zur Normalisierung von Werten. Die Klasse `CycleGan` import die Generator- und Diskriminatiorarchitekturen von pix2pix aus dem GitHub Repository von *TensorFlow Examples* als Python-Modul `pix2pix` [52]. Das ist analog zu dem Quellcode des genannten TensorFlow Beispiels [28].

Der ResNet Generator erhält als einen Parameter die Dimensionen des Bilds, das er generieren soll. Hier ist demnach nicht nur die Anzahl an Farbkanälen variabel, sondern auch die Höhe und Breite des Bilds. Der U-Net Generator erzeugt hingegen Bilder mit einer Höhe und Breite von 256 Pixeln. Ein weiterer Parameter nennt sich `n_blocks`. Er gibt die Anzahl an Residual Blocks an. Die CycleGAN-Veröffentlichung schlägt hier für eine Bildgröße von 256x256 einen Wert von 9 vor [27]. Falls dieser Wert zu keinen zufriedenstellenden Ergebnissen führt, soll er verändert werden. Listing 4.2 zeigt die Initialisierung der Generatoren abhängig von der gewählten Generator-Architektur.

```

...
# Generators
if self.generator_type == GeneratorType.U_NET:
    self.generator_g = pix2pix.unet_generator(3,
                                              norm_type='instancenorm')
    self.generator_f = pix2pix.unet_generator(3,
                                              norm_type='instancenorm')
else:
    image_dimensions = (self.image_size, self.image_size, 3)
    self.generator_g = resnet.ResnetGenerator(image_dimensions,
                                               n_blocks=9)
    self.generator_f = resnet.ResnetGenerator(image_dimensions,

```

```

    ...
    n_blocks=9)

```

Listing 4.2: model.py - Initialisierung der Generatoren

Analog dazu initialisiert die Klasse `CycleGan` die Diskriminatoren D_y und D_x . Die Klasse besitzt weitere Attribute, die für die weiteren Methoden der Klasse von Relevanz sind.

4.1.2 fit-Methode

Bevor das Modell zur Generierung von Bildern genutzt werden kann, muss es trainiert werden. Hierfür existiert die `fit`-Methode. Die vollständige Methode ist aufgrund ihrer Länge im Anhang in Listing 2 abgebildet. Als Parameter benötigt die Methode zum einen den Datensatz an Piktogrammen und zum anderen den Trainingsdatensatz mit realen Straßenschildaufnahmen. Diese Parameter tragen die Bezeichnungen `pictograms` und `real_images`. Die beiden Datensätze müssen dabei als `tf.data.Dataset` Objekte übergeben werden [36]. Die Klasse `tf.data.Dataset` ist Teil des `tf.data` Frameworks. Sie ist explizit auf die Performanz beim Laden großer Datensätze ausgelegt. Über `tf.data.Dataset` Objekte kann beispielsweise mittels einer `for`-Schleife iteriert werden. Bei jeder Iteration gibt der Datensatz dabei einen Batch zurück. Somit muss sich nicht manuell um das Laden einzelner Elemente des Datensatzes gekümmert werden. Auch erfolgt das Laden der Batches asynchron. Batches werden dann in den Arbeitsspeicher geladen, wenn sie benötigt werden. [36]

Ein optionaler Parameter ist zusätzlich die Anzahl an Epochen, die das Modell trainieren soll. Die Anzahl an Epochen ist standardmäßig auf 1 gesetzt. Listing 4.3 zeigt die Funktionsdeklaration sowie einen Ausschnitt aus der Implementierung.

```

def fit(self, pictograms, real_images, epochs=1):
    ...
    for epoch in range(epochs):
        ...
        # Single training step
        for image_batch in tqdm(real_images):
            ...
            # Transform the pictograms
            pictograms.shuffle(buffer_size=100,
                                reshuffle_each_iteration=True)
            single_pictogram_batch =
                pictograms.take(1).get_single_element()
            single_pictogram_batch, _, _ =
                utils.preprocess_image.randomly_transform_image_batch(
                    single_pictogram_batch)

```

```
# Train the model
losses = self.train_step(single_pictogram_batch,
                         image_batch)
...
```

Listing 4.3: model.py - fit-Methode

Die `fit`-Methode iteriert zunächst über die Anzahl an Trainingsepochen. In jedem Durchlauf trainiert das CycleGAN über den gesamten Trainingsdatensatz. Das realisiert die zweite `for`-Schleife. Hier iteriert die Funktion über das `tf.data.Dataset` Objekt `real_images`. Bei jeder Iteration gibt das Objekt einen neuen Batch an Bildern zurück. Die Funktion `tqdm` stammt aus dem gleichnamigen Python-Paket. Sie dient dazu, eine Fortschrittsanzeige (*engl.: progress bar*) in der Konsole anzuzeigen, die sich nach jeder Epoche aktualisiert.

Bevor das Modell auf einem einzelnen Batch trainiert, muss die `fit`-Methode die Piktogramme zunächst augmentieren. Das besteht aus verschiedenen Schritten. Als erstes wird der Datensatz der Piktogramme gemischt. Dies geschieht mit Hilfe der Methode `shuffle`. Anschließend erzeugen die Methoden `take(1)` und `get_single_element` daraus einen neuen Datensatz, der nur einen Batch aus dem Piktogramm-Datensatz beinhaltet und geben diesen einzigen Batch zurück. Zusammengefasst dient dieser Codeabschnitt dazu, die Piktogramme zufällig zu mischen und einen einzelnen Batch an Piktogrammen daraus zu entnehmen.

Zur Augmentierung der Piktogramme ruft die `fit`-Methode dann eine Methode namens `randomly_transform_image_batch` auf. Diese Methode ist in dem für diese Studienarbeit erstelltem Modul `utils` implementiert. Diese Methode gibt nicht nur die augmentierten Bilder zurück, sondern auch sowohl eine Liste der Rotationsmatrizen als auch der zufälligen Skalierung der Piktogramme. Für das Training werden nur die augmentierten Bilder benötigt. Die zusätzlich zurückgegebenen Daten sind hierfür nicht notwendig. Dass die Rückgabewerte der Methode `randomly_transform_image_batch` nicht benötigt werden, soll der Variablenname `_` signalisieren.

Anschließend dazu führt die Methode `train_step` den eigentlichen Trainingsdurchlauf durch. Sie erhält dabei den zufälligen Batch an Piktogrammen sowie den Batch an realen Straßenschildaufnahmen aus der `for`-Schleife. Damit berechnet sie die Verlustfunktionen des CycleGAN und führt basierend darauf Gradientenabstiege für G , F , D_y und D_x aus. Die Methode implementiert die Kostenfunktionen 2.18 sowie den Identity Loss. Hat die `train_step`-Methode die Verlustfunktionen berechnet, führt sie den Gradientenabstieg für das CycleGAN durch.

4.1.3 generate-Methode

Ist ein CycleGAN trainiert, dient die `generate`-Methode zur Generierung von Bildern. Dafür wird lediglich der Generator G benötigt, der Bilder von Piktogrammen in Bilder von Straßenschildern übersetzt. Die `generate`-Methode besteht deshalb nur aus einer Zeile Code. Als Parameter erhält die Methode einen Batch an Piktogrammen, welche sie an Generator G übergibt. Aus diesen Piktogrammen erzeugt der Generator Bilder von Straßenschildern. Die Methode gibt diese Bilder anschließend zurück. Listing 4.4 zeigt die `generate`-Methode.

```
def generate(self, pictograms):
    return self.generator_g(pictograms)
```

Listing 4.4: `model.py` - generate-Methode

4.2 Datenaugmentierung

In Kapitel 3.4 ist das Vorgehen für die Datenaugmentierung beschrieben. Wie bereits gezeigt, nutzt beispielsweise die `fit`-Methode die Datenaugmentierung, um dem Generator G die Größe und Perspektive des Straßenschildes vorzugeben. Hier soll auf die konkrete Implementierung dessen eingegangen werden. Zur Augmentierung müssen die Piktogramme der Straßenschilder zufällig rotiert und skaliert werden. Hierzu dient die Bibliothek Tensorflow Graphics.

Die Augmentierung ist in der Datei `utils/preprocess_image.py` implementiert. Die Funktion, die hierbei aus der CycleGAN Klasse aufgerufen wird, ist `randomly_transform_image_batch`. Das Listing 1 im Anhang zeigt die vollständige Implementierung. Die Funktion erhält einen vierdimensionalen Tensor `img_tensor_batch` als Eingang. Dieser Tensor beinhaltet den Batch an Bildern, der transformiert werden soll. Zunächst skaliert die Funktion zufällig den Inhalt dieser Bilder. Anschließend führt sie darauf eine zufällige dreidimensionale Rotation aus und gibt die transformierten Bilder zurück. Was sie ebenfalls zurückgibt, sind die Listen `content_sizes` und `rotation_matrices`. Die Anzahl an Elementen der Listen entspricht der Größe des übergebenen Batches, also der Anzahl an transformierten Bildern. Hierdurch kann die aufrufende Funktion für jedes Bild identifizieren, welche Zufallswerte für die Transformation generiert wurden. Dies kann genutzt werden, um die Transformation zu replizieren. Genutzt wird das in Kapitel 5, um Bilder von als ungültig markierten Schildern zu erzeugen.

4.2.1 Skalierung

Die Skalierung des Bildinhalts besteht aus mehreren Schritten. Zunächst generiert die Funktion `randomly_transform_image_batch` mittels Numpy eine Liste an zufälligen `content_sizes`.

Danach skaliert die Funktion die Piktogramme der Straßenschilder auf die in `content_sizes` gespeicherten Pixelgrößen mittels der Hilfsfunktion `resize_content_of_img`. Das ist in Listing 4.5 gezeigt.

```
content_sizes_tmp = content_sizes[ : ]
transformed_imgs = tf.map_fn(
    lambda img: resize_content_of_img(
        img, target_size, content_sizes_tmp.pop(0)),
    img_tensor_batch)
```

Listing 4.5: `utils.preprocess_image.py` - Skalieren der Bild-Tensoren

Als ersten Schritt dupliziert die Funktion die Liste `content_sizes` in einer Variable namens `content_sizes_tmp`. Anschließend skaliert die Funktion die Bilder. Dazu nutzt sie die von TensorFlow bereitgestellte Funktion `tf.map_fn`. Diese Funktion erhält als Parameter einen Tensor der Stufe n und eine Funktion. Sie führt die Funktion auf jedem Element der Stufe $n - 1$ des Tensors aus. Beispielsweise auf jedem Bild eines Batches von Bildern. In diesem Fall übergeben wir einen Tensor der Stufe vier und der Form (*Batch Größe, Breite, Höhe, Anzahl Farbkanäle*) an die Funktion `tf.map_fn`. Die Funktion erstellt daraus eine Menge an Tensoren der Stufe drei, wobei jeder dieser Tensoren ein Bild mit der Form (*Breite, Höhe, Anzahl Farbkanäle*) ist. Auf jedem Bild führt die Funktion `tf.map_fn` anschließend die Funktion `resize_content_of_img` aus. Letztere Funktion erhält ein Bild und die Zielgröße, auf die der Inhalt des Bilds skaliert werden soll. Anschließend fügt die Funktion `tf.map_fn` die Bilder wieder zu einem einzelnen Tensor der Stufe vier zusammen und gibt ihn zurück. [53]

Es wäre ebenso möglich, über die einzelnen Bild-Tensoren des Batches mittels einer `for`-Schleife zu iterieren. Die Dokumentation von `tf.map_fn` gibt jedoch explizit an, dass sie eine parallele Ausführung ermöglicht. Das ist hier der ausschlaggebende Vorteil gegenüber einer `for`-Schleife. Es ist jedoch dennoch weniger performant als eine Funktion zu verwenden, die eine einzelne Operation vektorisiert auf dem gesamten Tensor ausführt. Warum die Funktionen dennoch mit `tf.map_fn` arbeitet, statt alle Bilder des Batches gleichzeitig zu transformieren, soll der folgende Abschnitt klären. [53]

Die Funktion `resize_content_of_img` verwendet zwei Funktionen aus dem TensorFlow Framework: Die Funktion `tf.image.resize` um das Bild zu skalieren und die Funktion `tf.image.resize_with_crop_or_pad` um das Bild zurück auf die ursprüngliche Größe zu bringen. Wird das Piktogramm verkleinert, dann fügt letztere Funktion einen weißen Rand um das Piktogramm hinzu. Wird es hingegen vergrößert, schneidet die Funktion die Pixel ab, die über die Zielgröße hinausgehen. Obwohl es möglich wäre, den vierdimensionalen Tensor an beide TensorFlow Funktionen zu übergeben, um alle Bilder gleichzeitig zu skalieren, erhält `resize_content_of_img` lediglich dreidimensionale Tensoren, sprich einzelne Bilder, als Parameter. Das hängt damit zusammen, dass die Piktogramme in einem Batch unterschiedliche

Skalierungen besitzen sollen. Die TensorFlow Funktionen sind jedoch nur dazu in der Lage, eine bestimmte Skalierung auf allen Bildern des Batches auszuführen.

Die Liste `content_sizes_tmp` fungiert für die Funktion `tf.map_fn` als Warteschlange engl.: *Queue*. Bei jedem Durchlauf der Lambda-Funktion wird das erste Element der Liste gelesen und anschließend entfernt. Dazu dient die Listenfunktion `pop`.

4.2.2 Rotation

Der zweite Schritt der Augmentation ist, dass das Modell die Piktogramme rotiert. Es existieren Rotationsmatrizen, die Rotationen mittels eulerscher Winkel beschreiben. Die Drehungen um die x-, y- und z-Achse besitzen jeweils eigene Rotationsmatrizen R_x , R_y und R_z . Der Aufbau dieser Matrizen ist der Literatur entnommen [49]. Um daraus eine einzelne Rotationsmatrix R zu erhalten, werden die Rotationsmatrizen miteinander multipliziert. Dazu dient die Funktion `create_rotation_matrix`. Sie erhält die drei Winkel α_x , α_y und α_z als Parameter und gibt eine einzelne Rotationsmatrix R zurück.

Bevor die Funktion `randomly_transform_image_batch` die Funktion `create_rotation_matrix` aufruft, muss sie ausgehend davon zunächst zufällige Winkel $(\alpha_x, \alpha_y, \alpha_z)$ erzeugen. An dieser Stelle entstammen die zufälligen Winkel jedoch nicht einer Gleichverteilung, sondern einer gaußschen Normalverteilung. Das hängt damit zusammen, dass die Piktogramme in den meisten Fällen nur leicht rotiert sein sollen. Nur ein vergleichsweise geringer Prozentsatz der Piktogramme soll stark Augmentiert sein. Dies soll in etwa nachbilden, aus welchen Perspektiven die Straßenschilder im Trainingsdatensatz aufgenommen sind. Listing 4.6 zeigt die zufällige Generierung der Rotationswinkel α_z :

```
alpha_z_values = np.random.normal(loc=0.0, scale=3.5,
                                   size=batch_size)
```

Listing 4.6: `utils.preprocess_image.py` - Zufällige Generierung der Rotationswinkel α_z

Der Parameter `loc` gibt den Erwartungswert der Winkel an, während `scale` die Standardabweichung setzt. Durch das Angeben der `batch_size` wird deutlich, dass die Numpy Funktion hier nicht nur einen Winkel erzeugt wird, sondern ein *Numpy Array* das für jedes Piktogramm einen zufälligen Winkel enthält. Die Implementierung ist somit hier vektorisiert. Im Mittel liegen die Winkel der Rotation bei $0,0^\circ$. Der Wert für die Standardabweichung ist empirisch bestimmt, da die Werte für die Winkel nicht den tatsächlichen Winkeln in Grad entsprechen. Die Funktion `randomly_transform_image_batch` erzeugt die Winkel α_y und α_x analog hierzu, mit dem Unterschied, dass sie dort andere Standardabweichungen (`scale`) als Parameter setzt. Dass eine gaußsche Normalverteilung verwendet wird, bedeutet, dass die meisten

Piktogramme zu einer Aufnahme aus der Frontalperspektive führen. Der Großteil der Rotationswinkel ist demnach vergleichsweise gering. Einige wenige Schilder hingegen sind stärker rotiert. Würde die Funktion eine Gleichverteilung zur Erzeugung der Winkel verwenden, wäre der Anteil an starken Rotationen in etwa gleich zu dem Anteil an geringen Rotationen.

Die eigentliche Rotation setzt die TensorFlow Graphics Funktion `perspective_transform` um. Sie erhält einen Tensor der Stufe vier an Bildern und einen Tensor der Stufe vier an Rotationsmatrizen. Das bedeutet, dass der gesamte Batch an Bildern samt seiner Rotationsmatrizen übergeben wird. Somit erfolgt die Augmentation der Bilder hier vektorisiert und damit parallel. Die Codezeile `transformed_imgs = 1 - transformed_imgs` kehrt vor der Rotation zunächst die Farbwerte des Bilds um. Das ist nötig, da der Hintergrund, den die genannte TensorFlow Funktion erzeugt, schwarz ist. Kehrt man zunächst die Farbwerte um, führt die Rotation aus und setzt die Farbwerte auf ihren ursprünglichen Wert zurück, so wird der schwarze Hintergrund durch einen weißen ersetzt.

4.3 Training

Anwendende können das CycleGAN mit dem Skript `train.py` trainieren. Prinzipiell besitzt dieses Skript zwei Aufgaben: Es lädt sowohl den Trainingsdatensatz als auch die Piktogramme und ruft die Trainingsfunktion des Modells CycleGAN auf.

4.3.1 Laden der Datensätze

Die Konfigurationsdatei enthält in der Kategorie `paths` den Eintrag `train_data`. Das ist der relative oder absolute Pfad zu dem Trainingsdatensatz. Das Skript `train.py` lädt den Datensatz in ein `tf.data.Dataset` Objekt. Dafür stellt TensorFlow eine Funktion bereit, mit der ein Datensatz an Bildern aus einem Dateipfad geladen werden kann. Anhand der Ordnerstruktur sortiert die Funktion die Bilder automatisch in ihre Klassen ein. Die Funktion nennt sich `load_image_dataset_from_directory` [54]. In folgendem Listing ist der Teil des Skripts dargestellt, der mittels dieser Funktion den Datensatz lädt.

```
training_path = config['paths']['train_data']

train_set =
    tf.keras.utils.image_dataset_from_directory(training_path,
    batch_size=BATCH_SIZE, image_size=(IMAGE_SIZE, IMAGE_SIZE),
    labels=None, shuffle=True, crop_to_aspect_ratio=True)
```

```
train_set_processed = utils.load_data.normalize_dataset(train_set)
```

Listing 4.7: train.py - Laden des Trainingsdatensatzes

An die genannte TensorFlow Funktion übergibt das Skript mitunter den Pfad zu den Trainingsdaten, die Batch Größe und die Bildauflösung. Die Auflösung muss deshalb übergeben werden, da die Funktion `load_image_dataset_from_directory` alle Bilder auf diese Größe skaliert. Hierzu nutzt die Funktion standardmäßig *bilineare Interpolation*. Dadurch erscheint das Bild nicht als *verpixelt*, sondern fehlende Pixel, die bei der Vergrößerung unweigerlich auftreten, werden durch eine Kombination der benachbarten Pixel aufgefüllt. Dadurch wirkt das Bild statt *verpixelt* eher *verwaschen*. Das CycleGAN benötigt die Daten nicht nach ihren Klassen sortiert, da es mit unüberwachtem Lernen arbeitet. Deshalb wird zusätzlich der Parameter `labels` auf `None` gesetzt.

Durch den nächsten Parameter `shuffle` erfolgt die Einstellung, dass der Datensatz zufällig durchmischt werden soll. Abschließend folgt ein entscheidender Parameter, der einer näherer Erläuterung bedarf. Wie bereits in Kapitel 3.1 beschrieben, besitzen die Trainingsbilder des Datensatzes verschiedene Auflösungen. Das bedeutet, dass Bilder die nicht quadratisch sind, durch die Funktion `load_image_dataset_from_directory` verzerrt würden, damit sie in ein quadratisches Seitenverhältnis von 256x256 Pixel passen. Tests haben ergeben, dass die meisten Bilder des Datensatzes nur gerüfügig verzerrt werden. Einige Bilder besitzen jedoch signifikant mehr Pixel in der Höhe als in der Breite oder umgekehrt. Um dafür zu sorgen, dass alle Bilder ohne Verzerrung in das Modell gespeist werden, existiert der Parameter `crop_to_aspect_ratio`. Dieser Parameter sorgt dafür, dass das Bild derart zugeschnitten wird, dass es in das angegebene Bildformat passt. Hierbei wird das Bild so zugeschnitten, dass es gerade in das Seitenverhältnis passt. Was stets erhalten bleibt, ist der zentrale Teil des Bilds. Da sich die Straßenschilder in den meisten Bildern mittig befinden, ist dies genau das gewünschte Verhalten.

Was die Funktion `load_image_dataset_from_directory` zurückgibt, ist ein Objekt vom Typ `tf.data.Dataset`. Es kann somit direkt für die `CycleGan.fit`-Methode verwendet werden. Ein letzter Schritt ist, die Bilder zu normalisieren. Dies geschieht mittels der Funktion `normalize_dataset` aus dem eigens definierten Modul `utils.load_data`. Diese Funktion normalisiert die Pixelwerte der Bilder auf den Bereich von -1 bis 1. Dies ist notwendig, um die Bilder in die KNNs einzuspeisen.

Damit ist das Laden der Trainingsdaten abgeschlossen. Die Piktogramme werden unter der Verwendung des gleichen Schemas geladen. Es ergibt sich ein `tf.data.Dataset` Objekt mit dem Namen `pictograms_processed`.

4.3.2 Ausführen des Trainings

Das Training wird vollständig durch die Klasse `CycleGan` durchgeführt. Sind sowohl die Piktogramme als auch die Trainingsbilder geladen, sind folgende Codezeilen notwendig, um das Training zu starten:

```
cycle_gan = model.CycleGan(config)
cycle_gan.restore_latest_checkpoint_if_exists()
cycle_gan.fit(pictograms_processed, train_set_processed,
    epochs=config['training']['number_of_epochs'])
```

Listing 4.8: `train.py` - Laden des Trainingsdatensatzes

Das Skript `train.py` instanziert zunächst ein `CycleGan` Objekt. Falls vorherige Parameter in einem Checkpoint gespeichert sind, werden diese anschließend geladen. Ansonsten werden die Parameter durch TensorFlow zufällig initialisiert. Abschließend erfolgt der Aufruf der `fit`-Methode mit der in der Konfigurationsdatei angegebenen Anzahl an Epochen.

4.3.3 Logging

Ebenfalls bietet die Implementierung die Möglichkeit, den Verlauf der Verlustfunktionen über das Training zu betrachten. Dazu ist in der Implementierung zusätzlicher Code, der dieses sogenannte *Logging* ermöglicht. Das Logging erfolgt mittels der Bibliothek `TensorBoard`. Diese Bibliothek ist Teil des TensorFlow Frameworks. Mit TensorBoard ist es ebenso möglich, diesen Verlauf zu visualisieren. Die hierzu notwendigen Konsolenbefehle für jeweils das U-Net- und ResNet-basierte CycleGAN zeigt folgendes Listing:

```
$ tensorboard --logdir ./logs/unet
$ tensorboard --logdir ./logs/resnet
```

4.4 Trainingsergebnisse

Für das Training der U-Net- und ResNet-basierten CycleGANs dienen zwei verschiedene Systeme. Dabei zum einen Google Colab. Hier bietet Google eine kostenfreie Version sowie ein Premium-Abonnement an. In der kostenfreien Version ist jedoch kein Training über Nacht möglich, da das System nach einer etwa zwanzig-minütigen Inaktivität die Verbindung zum Rechner trennt. Unter anderem aus diesem Grund wird das Training zusätzlich auf einem Server

der DHBW durchgeführt. Dieser besitzt eine Grafikkarte mit 25 Gigabyte Speicher und ist damit leistungsstärker als die bei Google Colab verwendeten Systeme. Dort standen maximal 20 GB zur Verfügung. Die Trainingsdauer pro Epoche ist für die trainierten Modelle in der nachfolgenden Tabelle aufgeführt. Es zeigt sich hier außerdem, dass das U-Net signifikant schneller trainiert als das ResNet, jedoch Checkpoints besitzt, die mehr Speicherplatz verbrauchen. Das bedeutet, dass das Training des U-Net-basierten CycleGAN weniger Rechenaufwand benötigt, obwohl es mehr Parameter besitzt.

Trainingsdauer pro Epoche				
Modell	Google Colab	DHBW Server	Parameter	Checkpoint Größe
U-Net	30 min. ¹	5 min.	114 Millionen	1.340.240 Byte
ResNet	90 min.	30 min.	28 Millionen	331.709 Byte

Tabelle 4.2: Vergleich von U-Net und ResNet

4.4.1 U-Net

Das U-Net-basierte CycleGAN verbessert sich während des Trainings nahezu kontinuierlich. Das Modell ist mit einer Anzahl von 200 Epochen trainiert. Der Verlauf der Verlustfunktionen über die Anzahl der Trainingsschritte ist im Anhang in Abbildung A.2 gezeigt. Ein Trainingsschritt entspricht einem Durchlauf der `train_step` Funktion des CycleGAN über einem Batch. Es lassen sich verschiedene Dinge aus den Graphen ablesen: Die Verluste der Generatoren scheinen gegen einen Wert zu konvergieren. Die Verluste der Diskriminatoren zeigen hingegen deutliche Schwankungen ohne ein erkennbares Muster. Die Verluste der Generatoren sind beinahe um einen Faktor 10 größer als die der Diskriminatoren. Aus diesem Grund konvergiert der Gesamtverlust des CycleGAN gegen einen Wert. Dieser liegt bei 6.

Abbildung A.3 im Anhang zeigt zudem für die Epochen 1 bis 100, wie sich die Qualität der Generierung mit den Trainingsepochen verbessert. Hier lässt sich außerdem erkennen, dass das Modell durch die Wahl einer Bild-zu-Bild-Übersetzung die Schilder nicht selber erzeugen muss. Verschiedene Trainingsdurchläufe haben ergeben, dass der Identity Loss für das U-Net keinen signifikanten Einfluss hat. Damit kann argumentiert werden, dass er für dieses CycleGAN nicht nötig sei.

Die generierten Bilder zeigen verschiedene Hintergründe. Diese Varianz an Hintergründen ist allgemein pro Kategorie von Straßenschild gleich. Somit besitzen beispielsweise alle Schilder

¹ Minuten

der Kategorie Geschwindigkeitsbegrenzung die gleichen Arten von Hintergründen. Die Schilder der Kategorie Aufhebung können allgemein als wenig fotorealistisch bewertet werden. Das ist vermutlich auf die geringere Anzahl an Trainingsdaten für diese Klassen zurückzuführen.

Abbildung A.5 im Anhang zeigt Bilder, die das U-Net-basierte Modell nach 200 Epochen generiert. Ausgehend von den Erkenntnissen der in Kapitel 2.4.1 vorgestellte Veröffentlichung könnte sich die Qualität der Generierung über noch mehr Trainingsepochen weiter verbessern. Aus folgenden Gründen ist das U-Net-basierte Modell genau 200 Epochen trainiert:

- Die Verlustfunktionen scheinen zu konvergieren.
- Für ein weiteres Training muss vermutlich die Lernrate verringert werden
- Das Training von 200 Epochen dauert bereits 16 Stunden auf dem System der DHBW
- U-Net und ResNet sollten für die Evaluation in Kapitel 6 in etwa die gleiche Anzahl an Epochen trainiert sein

4.4.2 ResNet

Für das ResNet-basierte CycleGAN zeigen die Verlustfunktionen einen ähnlichen Verlauf wie bei dem U-Net-basierten CycleGAN. Aus diesem Grund sind hierfür die Graphen nicht im Anhang abgebildet. Der Unterschied ist, dass der Verlauf der Verlustfunktionen hier eine geringere Aussagekraft für die Qualität der generierten Bilder zu haben scheint. Das Training des ResNet-basierten Modells ist oszillierend, da das CycleGAN einige Klassen in nachfolgenden Epochen besser erzeugt, während andere Klassen eine schlechtere Generierungsqualität als davor aufweisen. Um diesem Verhalten entgegenzuwirken, verwendet dieses CycleGAN-Modell eine \mathcal{L}_2 Verlustfunktion für den Adversarial Loss, während das U-Net-basierte Modell weiterhin eine logarithmische Verlustfunktion verwendet. Auch diese Veränderung der Verlustfunktion, die laut der Literatur das Training stabilisieren kann, behebt das oszillierende Verhalten nicht.

Eine weitere Eigenschaft dieses Modells ist, dass bei 200 Epochen ein Modal Collaps auftritt. Das ist in Abbildung 4.1 dargestellt. Hier und in den folgenden Abbildungen steht die Abkürzung *Ep.* für das Wort *Epochen* und beschreibt damit die Anzahl an Trainingsepochen, aus der das Bild stammt.



Abbildung 4.1: Modal Collaps des ResNet nach 200 Trainingsepochen

Für jedes Piktogramm und jede Perspektive erzeugt das Modell durch den Modal Collaps einen beinahe identischen Hintergrund. Das gibt Hinweise darauf, dass die Generatoren die Diskriminatoren derart überlisten, dass letztere nicht mehr lernen.

Eine Lösung ist, das Training vorzeitig abzubrechen (*engl.: early stopping*). Diese Lösung wird gewählt. Dabei muss für jede infrage kommende Epoche manuell geprüft werden, welche davon das zufriedenstellendste Ergebnis zeigt. Auch da die Werte der Verlustfunktionen hierauf, wie bereits erwähnt, nur eine begrenzte Aussagekraft haben. Jede der Epochen 120 bis 190 erzeugt eine Bandbreite an Generierungsqualitäten. Abbildung 4.2 zeigt positiv herausstechende Bilder für verschiedene Epochen, während 4.3 negativ herausstechende Bilder zeigt. Die finale Entscheidung ist, das auf 180 Epochen trainierte Modell zu verwenden.



Abbildung 4.2: Positiv herausstechende Bilder des ResNets verschiedener Epochen

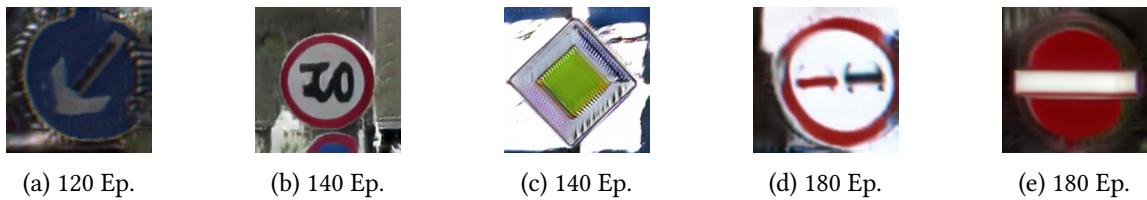


Abbildung 4.3: Negativ herausstechende Bilder des ResNets verschiedener Epochen

Abbildung A.4 im Anhang zeigt Bilder, die das ResNet-basierte Modell nach 180 Epochen generiert.

Eine zweite Lösung im Gegensatz zum vorzeitigen Abbruch des Trainings ist, dass die Generatoren ein anderes ResNet-Modell verwenden. Ein Modell, das verhindern kann, dass die Generatoren die Diskriminatoren vollständig überlisten können. Aus diesem Grund besitzen die Generatoren in einem nächsten Trainingsdurchlauf 6 Residual Blocks, statt der durch die CycleGAN Veröffentlichung vorgeschlagene Anzahl an 9 für Bilder der Auflösung 256x256. Dadurch haben die Generatoren eine weniger komplexe Architektur mit einer geringeren Anzahl an lernbaren Parametern. Während hier kein Modal Collaps auftritt, ist das Training weiterhin oszillierend. Auch hier zeigt die Epoche 200 nicht die höchste Generierungsqualität. Bei diesem ResNet-basierten Modell wird deshalb Epoche 150 für die Evaluation in Kapitel 6 gewählt. Damit soll geprüft werden, welche Anzahl an Residual Blocks sich für das Modell eignet. Die Abbildung 4.4 zeigt Beispielbilder des ResNet-basierten Modells mit 6 Residual Blocks für verschiedene Epochen.



Abbildung 4.4: Beispielbilder des ResNets mit 6 Residual Blocks

Generierte Bilder des ResNet-basierten Modells mit 6 Residual Blocks befinden sich unter **dem Link des Datensatzes** in dem Ordner '**Generated Images**'. Hier sind zudem weitere generierte Bilder des ResNet-basierten Modells mit 9 Residual Blocks sowie des U-Net-basierten Modells.

4.5 Generierung

Das Skript `generate.py` dient dazu, Bilder von Straßenschildern mittels eines trainierten CycleGAN Modells zu generieren. Eine Design-Entscheidung ist hierbei, dass das Skript vollständig mittels der Kommandozeile konfiguriert werden kann. Dies folgt dem Leitprinzip dieser Studienarbeit, dass Anwendende keinen Python-Code anpassen müssen. Das Skript besitzt die in Tabelle 4.3 gezeigten Kommandozeilenargumente.

Argument	Parameter	Aufgabe
--num-imgs	Ganzzahl	Anzahl der zu generierenden Bilder
--model	' unet ' oder ' resnet '	Art des Modells
--motion-blur	-	Fügt Bewegungsunschärfe hinzu
--make-invalid	-	Markiert Schilder als ungültig
--snow	-	Fügt Schnee hinzu

 Tabelle 4.3: Kommandozeilenargumente des Skripts `generate.py`

Standardmäßig nutzt das Skript `generate.py` das Modell aus der Konfigurationsdatei, mit dem Kommandozeilenargument `--model` können Anwendende diesen Wert jedoch überschreiben. Die Argumente `--motion-blur`, `--make-invalid` und `--snow` können außerdem miteinander kombiniert werden. Dann erzeugt das Skript mehrere Augmentationen gleichzeitig. Auf die Implementierung dieser Augmentationen geht Kapitel 5 ein.

Beispielhafte Aufrufe des Skripts `generate.py` mit verschiedenen Kommandozeilenargumenten zeigt Listing 4.9.

```
$ python generate.py --num-imgs 10 --motion-blur
$ python generate.py --num-imgs 10 --model 'resnet' --make-invalid
$ python generate.py --num-imgs 10 --snow --motion-blur
$ python generate.py --num-imgs 50 --model 'unet'
```

Listing 4.9: Beispieldaufrufe des Skripts `generate.py`

Ein spezieller Anwendungsfall dieses Modells könnte der folgende sein: Anwendende möchten nur für bestimmte Arten von Straßenschildern Bilder generieren, statt für alle 43. Beispielsweise nur für Stopp-Schilder. Das ist insbesondere dann relevant, wenn das Modell dazu genutzt werden soll, einen bestehenden Datensatz auszugleichen. Wenn etwa bestimmte Klassen unterrepräsentiert sind. Das Skript `generate.py` erlaubt explizit das folgende Vorgehen: Anwendende können aus dem Ordner, in dem sich die Bilder der Piktogramme befinden, alle Arten von Straßenschildern löschen, die nicht generiert werden sollen. Befindet sich in dem Ordner beispielsweise nur ein Piktogramm für Stopp-Schilder, dann werden auch nur Bilder von Stopp-Schildern generiert. Dafür kann zum Beispiel ein zweiter Ordner für die Piktogramme angelegt werden, der dann in der TOML-Konfigurationsdatei unter dem Wert `'pictograms'` innerhalb der Kategorie `'paths'` angegeben wird.

In dem Pfad `experimental/generate_single_classes.py` existiert ein Skript, dass dieses Vorgehen für alle Klassen automatisiert. Es ruft das Skript `generate.py` nacheinander für jede der 43 Arten von Straßenschildern auf und sortiert die generierten Bilder in separate Ordner ein.

5 | Augmentation der generierten Bilder

Wie bereits erwähnt ist das Ziel dieser Studienarbeit nicht alleine die Generierung von Bildern, die Straßenschilder zeigen. Zusätzlich soll die Arbeit einige der in Kapitel 2.1 genannten Probleme für die Straßenschilderkennung simulieren. Das Skript `generate.py` erlaubt deshalb die Methoden der Augmentierung, die bereits in Tabelle 4.3 aufgeführt sind. Diese sind eine **Bewegungsunschärfe**, das Hinzufügen von **Schnee** und das **markieren von Schildern als ungültig**. Dafür dient ein Modul `utils.image_augmentation`. Für jede der Augmentierungen befinden sich Beispiel-Abbildungen im Anhang.

Die Augmentierung ist weitgehend prozedural implementiert, da keine ausreichend großen Datensätze bekannt sind, die solche Grenzfälle der Straßenschilderkennung zeigen. Diese wären als Trainingsgrundlage für das CycleGAN notwendig.

5.1 Bewegungsunschärfe

Verwackelte Bilder können insbesondere dann entstehen, wenn sich das Fahrzeug mit einer hohen Geschwindigkeit bewegt. Hierbei entsteht eine Bewegungsunschärfe. Dies tritt bei einer Kamera auf, wenn sich das Bild während der Belichtungszeit deutlich verändert. Wenn also die Geschwindigkeit des Fahrzeugs groß ist im Vergleich zur Belichtungszeit. Da sich das fotografierte Objekt zu unterschiedlichen Zeitpunkten an verschiedenen Positionen im Bild befindet, erscheint es als verschwommen. Während bei der Straßenschilderkennung eine Bewegungsunschärfe die Erkennung erschwert, existieren Bereiche, in denen Fotografen absichtlich versuchen sie zu erzeugen. Dazu zählt beispielsweise die Sportfotografie, in der Objekte schneller erscheinen, wenn sie verschwommen zu sehen sind. Die Literatur gibt jedoch Hinweise darauf, dass es nicht trivial sei, die Bewegungsunschärfe mit einer Kamera exakt zu steuern. Beispielsweise wenn die Unschärfe eine bestimmte Intensität haben soll. Aus diesem Grund besteht Interesse daran, Bewegungsunschärfe computergestützt zu erzeugen. [55]

Ähnlich zu der Funktionsweise von CNNs basiert die künstliche Erzeugung von Bewegungsunschärfe auf einer Faltung des Bilds mit einer Faltmatrix. Die Idee ist, jeden Pixelwert durch einen Durchschnitt der umliegenden Pixelwerte zu ersetzen. Dabei jedoch linear entlang einer bestimmten Richtung, um eine lineare Bewegung zu simulieren. Je mehr Pixel in die Berechnung des Durchschnitts einbezogen werden, desto stärker erscheint die Bewegungsunschärfe.

Die Bewegungsunschärfe in dieser Studienarbeit ist entweder horizontal, vertikal oder diagonal. Es ergeben sich dadurch drei Arten von Faltmatrizen, die in Abbildung 5.1 gezeigt sind. [55]

$\begin{array}{ c c c } \hline 0 & 0 & 0 \\ \hline 0.3 & 0.3 & 0.3 \\ \hline 0 & 0 & 0 \\ \hline \end{array}$	$\begin{array}{ c c c } \hline 0 & 0.3 & 0 \\ \hline 0 & 0.3 & 0 \\ \hline 0 & 0.3 & 0 \\ \hline \end{array}$	$\begin{array}{ c c c } \hline 0.3 & 0 & 0 \\ \hline 0 & 0.3 & 0 \\ \hline 0 & 0 & 0.3 \\ \hline \end{array}$
---	---	---

Abbildung 5.1: Horizontale, vertikale und diagonale Faltmatrix

Die Größe der Faltmatrix gibt die Stärke der Unschärfe an. Bei einer größeren Faltmatrix fließen nämlich mehr Pixel in die Berechnung des Durchschnitts ein. Da die Faltung hier jeden Pixel gleich stark gewichtet, sind alle Parameter der Faltmatrix entlang der simulierten Bewegungsrichtung identisch. Damit jeder Pixel dieselbe Helligkeit besitzt, müssen die Parameter zudem zusammen den Wert *eins* ergeben [55]. Das ist der Grund, wieso die 3x3-Matrizen in Abbildung 5.1 die Werte 0.3 haben. Eine 4x4-Matrix hingegen hätte die Werte 0.25. In dem Modul `utils.image_augmentation` existiert die Funktion `apply_motion_blur` um auf einen einzelnen Bild-Tensor der Stufe drei eine Bewegungsunschärfe auszuführen. Zusätzlich besitzt die Funktion verschiedene Parameter zur Steuerung der Intensität und der Richtung des Effekts. Listing 5.1 zeigt, wie die Funktion die Bewegungsunschärfe in diagonaler Richtung durchführt.

```
kernel = np.identity(kernel_size)
kernel = kernel / kernel_size
transformed_img = cv2.filter2D(img_tensor.numpy(), -1, kernel)
```

Listing 5.1: `utils.image_augmentation.py` - Hinzufügen einer diagonalen Bewegungsunschärfe

Die Variable `kernel` steht für die Faltmatrix. Numpy besitzt eine Funktion `identity`, die eine Einheitsmatrix erzeugt. Diese Matrix besitzt auf der Hauptdiagonalen Einsen und sonst Nullen. Die `kernel_size` gibt die Größe der Matrix an. Damit alle Werte der Matrix summiert den Wert *eins* ergeben, teilt die Funktion die Matrix durch die `kernel_size`. In der letzten Codezeile in Listing 5.1 faltet die Funktion `apply_motion_blur` das Eingangsbild mit der Faltmatrix. Dazu nutzt sie die Funktion `filter2D` aus der OpenCV Bibliothek. Die Funktion wird genutzt, da sich keine TensorFlow Funktion hat finden lassen, mit der eine derart definierte Faltmatrix auf einem Bild angewendet werden kann, ohne Convolutional Layer zu verwenden. Die Funktion erhält drei Parameter: Den Bild-Tensor, jedoch konvertiert in ein Numpy Array, den Wert -1 und die Faltmatrix. Der Wert -1 gibt an, dass die Funktion die Anzahl an Farbkanälen des Bilds beibehalten soll.

5.2 Ungültige Straßenschilder

In Kapitel 2.1 ist bereits beschrieben, dass als ungültig markierte Schilder eine Herausforderung für heutige Straßenschilderkennungen darstellen können. Aus diesem Grund implementiert diese Studienarbeit den Anwendungsfall. Ungültige Schilder sind im Straßenverkehr meist durch ein orangefarbenes Kreuz gekennzeichnet.

Bei der Umsetzung bieten sich verschiedene Möglichkeiten. Zum einen kann das CycleGAN darauf trainiert werden, solche Bilder eigenständig zu generieren. Dafür benötigt das Modell Trainingsdaten mit ungültigen Schildern. Der Datensatz müsste somit um reale Bilder ergänzt werden, was nicht ohne weiteres möglich ist. Es wäre jedoch auch denkbar, bereits vorhandene Trainingsbilder mit einer Bildbearbeitungssoftware so anzupassen, dass sie ungültige Schilder zeigen.

Eine weitere Möglichkeit ist, das Kreuz, das die Ungültigkeit eines Schilds markiert, nachträglich in die generierten Bilder des CycleGAN einzufügen. Die Besonderheit ist hierbei, dass die Straßenschilder zufällig rotiert und skaliert sind. Die Augmentierungsfunktion muss das Kreuz vorher so transformieren, dass es sich stets zentral und mit einer angepassten Rotation auf dem Schild befindet. Da hierfür keine zusätzlichen Trainingsdaten nötig sind, implementiert das Modul `utils.image_augmentation` dieses Vorgehen. Die Funktion dafür heißt `make_street_sign_invalid`.

Für zukünftige Arbeiten ist ein weiterer Ansatz, die Kreuze vor der Bild-zu-Bild Generierung auf die Piktogramme einzufügen. Dafür können Entwickelnde die Funktion `make_street_sign_invalid` verwenden. Sie müssen das entstehende Bild anschließend skalieren und rotieren und an das CycleGAN als Eingangsdomäne X übergeben. Auch hierfür benötigt das Modell zusätzliche Trainingsdaten. Es muss jedoch nicht von alleine lernen, das Kreuz zu erzeugen, da dies bereits im Eingangsbild enthalten ist. Außerdem kann der Quellcode genutzt werden, um andere Objekte auf die Schilder einzufügen. Beispielsweise wenn Entwickelnde das CycleGAN erweitern wollen, um Vandalismus auf Straßenschildern zu simulieren.

In dem Ordner `Augmentation` unter **dem Link des Datensatzes** (Stand: 09.06.2023) befindet sich das rohe Bild eines orangefarbenen Kreuzes auf einem transparenten Hintergrund. Abbildung 5.2 zeigt dieses Bild. Prinzipiell implementiert die Funktion `make_street_sign_invalid`

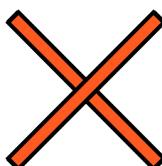


Abbildung 5.2: Kreuz auf transparentem Hintergrund, das Schilder als ungültig kennzeichnet

folgendes: Sie augmentiert das Bild genau wie das übergebene Straßenschild und fügt es dann auf das Schild ein. Aufrufende Funktionen müssen deshalb nicht nur ein generiertes Schild übergeben, sondern auch den Wert für die Skalierung sowie die Transformationsmatrix, die zu diesem generierten Bild geführt hat. Die Funktion `make_street_sign_invalid` kann das Kreuz dann mit den gleichen Parametern skalieren und rotieren wie das Straßenschild. Dadurch fügt die Funktion das Kreuz zentriert auf das Straßenschild ein. Die Funktion `make_street_sign_invalid` besteht in verkürzter Form aus dem Code in Listing 5.2.

```
cross = preprocess_image.transform_image(cross, content_size,
                                         transformation_matrix, bg_is_white=False)
img_tensor = tf.concat([img_tensor,
                       tf.ones_like(img_tensor[:, :, 0:1])], axis=-1)
# img_tensor and cross are numpy arrays here
img_tensor.paste(cross, (0, 0), cross)
```

Listing 5.2: `utils.image_augmentation.py` - Schilder als ungültig markieren

Die Variable `cross` ist ein Bild-Tensor der Stufe drei, der das Kreuz enthält. Dabei hat der Tensor jedoch nicht wie in dieser Studienarbeit üblich die Form $(256, 256, 3)$ sondern $(256, 256, 4)$. Das liegt daran, dass jeder der $256 \cdot 256$ Pixel zusätzlich zu den Werten für rot, grün und blau auch einen *Alphakanal* besitzt. Dieser gibt an, wie transparent der Pixel ist. Der Wert 0 bedeutet, dass der Pixel vollständig transparent ist, während der Wert 1 bedeutet, dass der Pixel vollständig deckend ist. Der transparente Hintergrund um das Kreuz besitzt damit überall einen Alpha-Wert von 0. In der ersten Codezeile in Listing 5.2 wird das Bild des Kreuzes augmentiert. Dazu wird die gleiche Funktion genutzt wie für die Datenaugmentierung des CycleGAN: `transform_image` aus `utils.preprocess_image`. Neben dem Tensor `cross` übergibt die Funktion hier die beiden Transformationsgrößen sowie den Parameter `bg_is_white=False`. Damit füllt die Funktion `tranform_image` den Hintergrund des Bilds mit vollständig transparenten statt mit weißen Pixeln auf.

In einem nächsten Schritt fügt die Funktion dem generierten Eingangsbild `img_tensor` einen Alphakanal mit den Werten 1 hinzu. Dafür existiert die Funktion `tf.concat`, die mehrere Tensoren zusammenfügen kann. Einer ihrer Parameter ist eine Liste von Tensoren, die miteinander konkateniert werden sollen. Der erste Tensor ist der Eingangstensor der Form $(256, 256, 3)$. Der zweite Tensor hat die Form $(256, 256, 1)$ und enthält nur Einsen als Werte. Der Code `img_tensor[:, :, 0:1]` entnimmt aus dem Eingangsbild einen einzelnen Farbkanal, während die Funktion `tf.ones_like` diesen Farbkanal komplett mit Einsen auffüllt. Der Parameter `axis=-1` legt fest, dass der neu erzeugte Alphakanal an die letzte Achse des Tensors angefügt wird, an der sich die Farbkanäle befinden. Der so erzeugte Tensor hat die Form $(256, 256, 4)$ und enthält neben den Farbwerten auch einen Alphakanal.

Das Bild wird dadurch nicht verändert, da beide Bilder nun die gleiche Anzahl an Farbkanälen haben, kann die Funktion `make_street_sign_invalid` das Kreuz jedoch nun auf das Eingangsbild einfügen. Dazu dient die Funktion `paste` aus der Bibliothek Pillow. Sie fügt das Kreuz an die Koordinaten $(0, 0)$ ein. Der dritte Parameter mit dem Wert `cross` sorgt dafür, dass die transparenten Pixel des Kreuzes erhalten bleiben und damit das Eingangsbild nicht überdecken. Vor der Anwendung der Pillow-Funktion werden die Bilder in Numpy Arrays konvertiert. [45]

5.3 Schnee

Schnee ist eine Wetterbedingung, die, ausgehend von Kapitel 2.1, eine Auswirkung auf die Straßenschilderkennung haben kann. Es ließen sich keine Veröffentlichungen finden, die das künstliche Einfügen von Schnee auf Bilder in eine in diesem Rahmen umsetzbare Art und Weise aufzeigen. Aus diesem Grund basiert die Implementierung auf folgendem Vorgehen: Ein Bildbearbeitungsprogramm wurde genutzt, um ein Bild aus künstlichem Schnee zu erstellen. Anschließend daran wurde das Vorgehen in dem Bildbearbeitungsprogramm als ein Algorithmus formuliert, der das Vorgehen in Python automatisieren kann.

Der Algorithmus besteht aus folgenden Schritten:

1. Erstelle ein Bild, das aus zufälligen schwarzen und weißen Pixeln besteht. Die Anzahl an weißen Pixel soll kleiner sein als die der schwarzen Pixel.
2. Führe auf dem Bild ein gaußsches Weichzeichen aus. Hierdurch verschmieren die einzelnen weißen Pixel zu größeren Punkten.
3. Führe auf dem Bild eine Bewegungsunschärfe aus. Dadurch wird die Bewegung der Schneeflocken entlang einer Windrichtung simuliert.
4. Mache den schwarzen Hintergrund transparent und füge das erstellte Bild auf ein generiertes Straßenschild-Bild ein.

Die Funktion `add_snow` implementiert diesen Algorithmus. Sie erzeugt die zufälligen schwarzen und weißen Pixel mittels einer Binomialverteilung. Für jeden Pixel existieren zwei Mögliche *Ziehungen* aus der Verteilung: Schwarz (0) oder weiß (1). Ein p gibt an, wie hoch die Wahrscheinlichkeit dafür ist, dass ein weißer Pixel gezogen wird. Die Problemstellung lässt sich für jeden Pixel als ein Zufallsexperiment mit zwei möglichen Ausgängen beschreiben. Das ist der Grund, wieso die Funktion hier eine Binomialverteilung implementiert. Dafür bietet TensorFlow die Funktion `tf.random.stateless_binomial`. Das auf dem Bild ausgeführte gaußsche Weichzeichen sorgt dafür, dass jeder Pixel *verschwommen* wird. Das kann mit dem vorgehen für die Bewegungsunschärfe verglichen werden. Mit dem Unterschied, dass

die Unschärfe hier nicht linear in eine Richtung zeigt, sondern kreisförmig um jeden Pixel ist. Damit wird aus jedem weißen Pixel ein kreisförmiger Punkt, der eine Schneeflocke darstellen soll. Für das gaußsche Weichzeichen bietet die Bibliothek TensorFlow Addons die Funktion `tfa.image.gaussian_filter2d`. Für die anschließende Bewegungsunschärfe ruft die Funktion `add_snow` die in Abschnitt 5.1 beschriebene Funktion `add_motion_blur` auf. [48]

Diese Augmentierung der Bilder ist zusätzlich für eine weitere Fragestellung genutzt: **Kann das CycleGAN lernen, verschneite Bilder eigenständig zu generieren?** Dazu existiert eine *Checkpoint*-Datei, die Parameter des U-Net-Modells enthält, wenn es auf verschneiten Bildern trainiert ist. In dem **Link des Datensatzes** befinden sich die Trainingsbilder unter dem Pfad '`Train with Snow`'.

Die Basis für das Training stellt das für 200 Epochen trainierte U-Net-basierte CycleGAN dar. Auf den verschneiten Bildern ist das Modell zusätzliche 20 Epochen trainiert. Abbildung 5.3 zeigt den Verlauf des Trainings. Die Bilder stammen aus verschiedenen Trainingsepochen, wobei das erste Bild zu der ersten Epoche korrespondiert und das letzte Bild zu der zwangigsten Epoche.



Abbildung 5.3: Trainingsverlauf des CycleGAN um Bilder mit Schnee zu erzeugen

Abbildung 5.4 zeigt Bilder, die das Modell am Ende des Trainings erzeugt. Es zeigt sich, dass das CycleGAN eigenständig verschneite Bilder generieren kann. 20 Epochen werden hier als eine akzeptable Dauer für ein zusätzliches Training betrachtet. Ausgehend davon können Anwendende die vortrainierten Modelle dieser Studienarbeit nutzen, um es auf eigene Augmentation zu trainieren. Ebenso kann erprobt werden, inwiefern das CycleGAN lernen kann, Bewegungsunschärfe und ungültige Schilder zu erzeugen.



Abbildung 5.4: Generierte Bilder des CycleGAN mit Schnee

Abbildung A.10 im Anhang zeigt weitere verschneite Bilder, die durch das CycleGAN generiert sind.

6 | Evaluation

Die Evaluation soll prüfen, inwieweit die Implementierung dieser Studienarbeit die in Kapitel 1.3 definierten Ziele erreicht. Somit beurteilt die Evaluation die folgenden Fragestellungen:

- Wie fotorealistisch sind die generierten Bilder des Modells?
- Wie neuartig sind die generierten Bilder des Modells?
- Wie sehr eignen sich die augmentierten Grenzfälle als Trainingsdaten für Straßenschilderkennungs-Software?

Dahingehend kann die Beurteilung in eine **Evaluation der Generierung** und eine **Evaluation der Augmentierung** unterteilt werden.

6.1 Evaluation der Generierung

Die Evaluation der Generierung bewertet die Qualität der generierten Bilder des CycleGAN. Die Augmentierung der Bilder ist hier nicht einbezogen.

6.1.1 Vorgehen

Das Modell dieser Studienarbeit verfolgt das Ziel, die statistische Verteilung der Trainingsdaten möglichst genau abzubilden. In dem Trainingsdatensatz befinden sich ausschließlich reale Aufnahmen von Straßenschildern. Daher kann argumentiert werden, dass das Modell genau dann fotorealistische Bilder erzeugt, wenn die generierte Verteilung $\hat{p}(x)$ ähnlich ist zu der tatsächlichen Verteilung der Trainingsdaten $p(x)$. Das Ziel der Evaluation ist somit ein Vergleich, wie ähnlich sich die beiden Verteilungen $\hat{p}(x)$ und $p(x)$ sind.

In der Beurteilung generativer Modelle sind dafür der *Inception Score* sowie die *Fréchet Inception Distance* üblich. Beide Verfahren nutzen ein Klassifizierungsmodell namens *Inception* und einen Datensatz namens *ImageNet*. Der Datensatz ImageNet enthält 14 Millionen Bilder, die in 20.000 Kategorien eingeteilt sind [56]. Das Klassifizierungsmodell Inception ist ein CNN, das mitunter von Google entwickelt wurde [57]. Sowohl der Inception Score als auch die Fréchet Inception Distance basieren auf folgendem Verfahren: [58]

- Ein Inception Modell wird auf dem ImageNet Datensatz trainiert

- Das generative Modell wird darauf trainiert, Bilder zu generieren, die dem ImageNet Datensatz ähnlich sehen
- Das Inception Modell soll die generierten Bilder klassifizieren
- Je höher die Genauigkeit des Inception Modells ist, desto ähnlicher sind die generierten Bilder dem ImageNet Datensatz

Die Klassifizierungsgenauigkeit schafft dabei einen quantitativen Wert für die Ähnlichkeit der beiden Verteilungen $\hat{p}(x)$ und $p(x)$. Hiermit können generative Modelle verglichen werden.

Das Verfahren kann für die Evaluation dieser Studienarbeit nicht gewählt werden, da das CycleGAN darauf trainiert ist, Bilder von Straßenschildern zu erzeugen. Es wäre notwendig, das Modell zusätzlich auf den ImageNet Datensatz zu trainieren. Stattdessen verwendet die Evaluation ein adaptiertes Vorgehen, das in Abbildung 6.1 dargestellt ist.

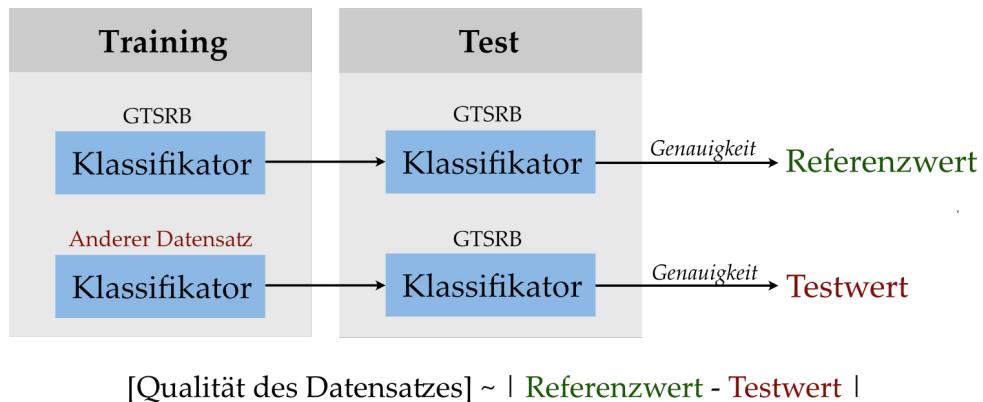


Abbildung 6.1: Vorgehen der Evaluation

Das Verfahren prüft folgendes: Wie genau kann ein Klassifikator die Testbilder des GTSRB klassifizieren, wenn er mit einem bestimmten Datensatz trainiert ist? Der GTSRB wird stellvertretend als die Verteilung $p(x)$ der Bilder von deutschen Straßenschildern betrachtet. Je ähnlicher dazu die Verteilung $\hat{p}(x)$ des Datensatzes ist, desto besser kann der Klassifikator daraus lernen, die Bilder des GTSRB zu klassifizieren.

Damit ist die Genauigkeit des Klassifikators, wenn er mit den Trainingsdaten des GTSRB trainiert wird der **Referenzwert**. Die Genauigkeit des Klassifikators, wenn er mit dem zu evaluierenden Datensatz trainiert wird ist hingegen der **Testwert**. Die absolute Differenz zwischen Referenz- und Testwert stellt die Evaluationsmetrik dar. Je geringer die Differenz ist, desto ähnlicher sind die generierten Bilder dem GTSRB Datensatz.

Im Unterschied zu dem Inception Score und der Fréchet Inception Distance wird der Klassifikator auf dem Datensatz trainiert und nicht getestet. Die Evaluation basiert auf diesem Vorgehen,

weil die generierten Daten als **Trainingsdaten** für die Straßenschilderkennung dienen sollen. Die Metrik prüft damit zudem, wie geeignet der jeweilige Datensatz als Trainingsdaten sind.

Der Klassifikator ist ein VGG16 Modell. Das ist eine CNN-Architektur aus dem Jahre 2014, die an der Oxford Universität entwickelt wurde. Das VGG16 besitzt weniger Schichten als das Inception Modell und damit auch weniger Parameter. Die Evaluation verwendet ein VGG16, da dadurch die Wahrscheinlichkeit verringert wird, dass der Klassifikator den GTSRB *auswendig lernt*. Damit ist gemeint, dass das Modell beispielsweise lernt, dass in den Trainingsdaten ein bestimmter Hintergrund nur bei einer einzelnen Art von Straßenschildern vorkommt. Das Modell könnte hierdurch womöglich die gelernte Klassifizierung weder auf die Testdaten des GTSRB noch auf andere fotorealistische Datensätze transferieren. Eine Veröffentlichung hat beispielsweise für ihren Anwendungsfall gezeigt, dass das VGG16 die Daten besser generalisieren kann als das Inception Modell. [59]

Der Quellcode der Evaluation befindet sich in dem Pfad `classifier/run.py`. Der Klassifikator basiert auf einem VGG16, das auf den ImageNet Datensatz vortrainiert ist. Aus der Literatur wird Code verwendet, der diesem Modell eine neue trainierbare Schicht hinzufügt [1]. Die vorherigen Schichten zur Merkmalsextraktion für den ImageNet Datensatz behalten ihre Parameter bei, während die hinzugefügte Schicht auf dem entsprechenden Datensatz (zum Beispiel dem GTSRB) trainiert wird. Auf Englisch heißt dieses Verfahren *transfer learning*. [1]

6.1.2 Ergebnisse

Tabelle 6.1 zeigt die Ergebnisse der Evaluation. Das VGG16 ist auf jeden der aufgelisteten Datensätze bis zu einer Trainingsgenauigkeit von nahezu oder genau 100% trainiert. Das entspricht in den meisten Fällen 20 Trainingsepochen. Die Testgenauigkeit gibt an, wie viel Prozent der **GTSRB Testbilder** das VGG16 korrekt klassifiziert, wenn es **ausschließlich** mit dem jeweiligen Datensatz trainiert ist.

Die Referenzgenauigkeit des GTSRB beträgt laut den Ergebnissen 82%. Der gesamte in Kapitel 3.1 beschriebene Datensatz hat eine vergleichbare Genauigkeit von 83% auf den Testdaten des GTSRB. Das deutet darauf hin, dass die hinzugefügten Datensätze die Verteilung $p(x)$ von deutschen Straßenschildern nicht verzerren.

Die drei Varianten des CycleGAN – U-Net-basiert, ResNet-basiert mit 9 Residual Blocks und ResNet-basiert mit 6 Residual Blocks – erzeugen Bilder, die zu einer signifikant niedrigeren Genauigkeit führen. Das bedeutet, dass sich die Bilder zu einem gewissen Grad von denen des GTSRB unterscheiden. Analog zu den Beobachtungen in Kapitel 4.4 zeigt dabei das U-Net-basierte CycleGAN die geringste Abweichung zum Referenzwert. Somit trifft auf dieses Modell mindestens eine der beiden Aussagen zu:

1. Das U-Net-CycleGAN erzeugt **fotorealistischere** Bilder als die ResNet-basierten Modelle
2. Das U-Net-CycleGAN erzeugt eine **größere Varianz** an unterschiedlichen Bildern pro Klasse

Trainingsdatensatz	Trainingsbilder (#)	Testgenauigkeit ¹ (%)
Präparierter GTSRB	4.554	82
Gesamte Trainingsdaten	5.685	83
U-Net	4.300	62
ResNet (9 residual blocks)	4.300	53
ResNet (6 residual blocks)	4.300	46
Augmentierte Piktogramme	4.300	19
Piktogramme	43	12
Gemischt	8.868	87

Tabelle 6.1: Ergebnisse des Trainings eines VGG16 Klassifikators

Von den beiden ResNet-basierten CycleGANs zeigt die Variante mit 9 Residual Blocks die höhere Testgenauigkeit.

Für eine Beurteilung der **Varianz** an generierten Bildern verwendet diese Evaluation kein quantitatives Verfahren. Es zeigt sich, dass die Hintergründe der generierten Bilder meist mindestens einem Hintergrund aus den Trainingsdaten ähnlich sehen. Dennoch erzeugt das CycleGAN neuartige Bilder. Das U-Net-basierte CycleGAN erzeugt pro Klasse etwa 5-10 verschiedene Hintergründe. Die beiden ResNet-basierten Modelle zeigen eine geringere Varianz verschiedener Hintergründe von maximal fünf verschiedenen Hintergründen.

Eine subjektive Beurteilung kommt zu dem Schluss, dass die Bilder des U-Net-basierten CycleGAN sowohl fotorealistischer sind als auch eine größere Varianz besitzen als die ResNet-basierten Varianten. Die Ergebnisse aus Tabelle 6.1 stützen diese Beobachtung.

Eine allgemeine Fragestellung ist, ob die generierten Bilder eine höhere Genauigkeit erzielen als die augmentierten Piktogramme. Ist das nicht der Fall, dann würde das zeigen, dass die von dem CycleGAN implementierte Funktion keinen Mehrwert bietet. Es wäre ebenso möglich, den Datensatz um Bilder von Piktogrammen zu erweitern. Hier zeigt sich jedoch ein signifikanter Unterschied. Die augmentierten Piktogramme erzielen eine Testgenauigkeit von 19% und damit weniger als die Varianten des CycleGAN.

¹ Des Klassifikators auf dem GTSRB, wenn er mit dem jeweiligen Trainingsdatensatz trainiert ist

Ein zentraler Aspekt ist zudem, ob ein erweiterter Datensatz, der aus realen und künstlich erzeugten Datensätzen besteht, eine höhere Genauigkeit erzielt als ein Satz an ausschließlich realen Bildern. Der Datensatz mit der Bezeichnung *Gemischt* besteht aus Bildern aller drei Varianten des CycleGAN sowie den gesamten Trainingsdaten des CycleGAN. Die Klassifizierungsgenauigkeit liegt vier Prozentpunkte über der Genauigkeit der gesamten realen Trainingsdaten. Das zeigt, dass die künstlich erzeugten Bilder die Trainingsdaten erweitern können, um die Genauigkeit der Straßenschilderkennung zu verbessern.

6.2 Evaluation der Augmentierung

Das Ziel der augmentierten Bilder ist, dass sie sich als Trainingsbilder für die Straßenschilderkennung eignen. Die Evaluation soll deshalb zeigen, dass ein Klassifikator eine geringere Genauigkeit auf augmentierten Bildern erzielt als auf nicht-augmentierten Bildern. Wenn die augmentierten Bilder fotorealistisch sind, dann zeigt das, dass Grenzfälle wie etwa Bewegungsunschärfe und Schnee die Genauigkeit der Straßenschilderkennung verringern. In dem Fall können solche Bilder verwendet werden, um Software zur Straßenschilderkennung für solche Grenzfälle zu trainieren.

Tabelle 6.2 zeigt die Ergebnisse dieser Evaluation. Der Klassifikator ist in jedem Fall auf dem GTSRB *trainiert* und auf Bildern mit der jeweiligen Augmentierung *getestet*. Der Referenzwert ist die Genauigkeit des Klassifikators auf nicht-augmentierten Bildern des U-Net-basierten CycleGAN.

Augmentierung	Testbilder (#)	Testgenauigkeit ¹ (%)
Keine Augmentierung	645	60
Bewegungsunschärfe	300	39
Ungültige Schilder	300	11
Schnee	300	10
Alle Augmentierungen	300	4

Tabelle 6.2: Genauigkeit des Klassifikators auf augmentierten Bildern des U-Net-basierten CycleGAN

Eine Bewegungsunschärfe zeigt den geringsten Einfluss auf die Genauigkeit. Ungültige Schilder und Schnee besitzen einen signifikant größeren Einfluss. Die Kombination aller drei Augmentierungen reduziert die Genauigkeit auf 4%.

¹ Des Klassifikators auf dieser Augmentierung, wenn er auf dem GTSRB trainiert ist

Da die Evaluation hier nicht den Fotorealismus der Bilder quantitativ beurteilt, kann daraus nicht eindeutig geschlussfolgert werden, dass sich die augmentierten Daten als Trainingsbilder eignen. Die Ergebnisse geben jedoch Hinweise darauf, dass Entwickelnde die Bilder als zusätzliche Trainingsdaten für die Straßenschilderkennung in Betracht ziehen können.

6.3 Verbesserungsmöglichkeiten

Es existieren verschiedene Verbesserungsmöglichkeiten, damit das CycleGAN fotorealistischere Bilder erzeugt. Eine Möglichkeit ist, dass das Training einen sogenannten *Learning Rate Decay* verwendet. Damit ist gemeint, dass die Lernrate, die die Größe der Parameteränderungen pro Trainingsschritt angibt, mit der Zeit abnimmt. Das schlägt auch die CycleGAN-Veröffentlichung vor.

Eine weitere Möglichkeit ist, ein ResNet-Modell speziell für diesen Anwendungsfall zu entwickeln. Eigentlich ist ausgehend von der CycleGAN-Veröffentlichung nämlich zu erwarten, dass ein ResNet-basiertes CycleGAN zufriedenstellende Ergebnisse erzielt.

Ein zusätzlicher Punkt ist, das Erzeugen von ungültigen Schildern fotorealistischer zu gestalten. Beispielsweise indem das Kreuz an die Helligkeit des Hintergrunds angepasst wird.

7 | Zusammenfassung

Diese Arbeit befasst sich mit der künstlichen Erzeugung von Bildern, die Straßenschilder zeigen. Zusätzlich ist eine Augmentierung der generierten Bilder auf Grenzfälle der Straßenschilderkennung implementiert.

Wie bereits eine vorherige Arbeit gezeigt hat, ist die Generierung solcher Bilder mit CycleGANs möglich. In dieser Arbeit erzielt ein U-Net-basiertes CycleGAN dabei bessere Ergebnisse und ist schneller zu trainieren. Allgemein zeigt sich anhand der Evaluation, dass sich die erzeugten Datensätze von realen Trainingsdaten unterscheiden. Trotzdem können sie einen Klassifikator verbessern, indem sie als **zusätzliche** Trainingsdaten genutzt werden. Das deutet darauf hin, dass sie sich als zusätzliche Trainingsdaten eignen. Verschiedene Möglichkeiten der Verbesserung existieren dennoch.

Diese Arbeit zeigt, dass CycleGANs ebenso lernen können, bestimmte Augmentierungen für die Straßenschilderkennung eigenständig zu lernen. So etwa das Generieren von Bildern mit Schnee. Dabei sind in dieser Arbeit in etwa 20 zusätzliche Trainingsepochen nötig. Auf dem verwendeten System der DHBW entspricht das einer Trainingsdauer des U-Net-basierten Modells von etwa 1,7 Stunden.

Die prozedural erzeugten Augmentierungen von Bewegungsunschärfe, ungültigen Schildern und Schnee verringern zudem die Genaugkeit eines Klassifikators für die Straßenschilderkennung. Das gibt Hinweise darauf, dass solche Bilder Datensätze für die Straßenschilderkennung erweitern können.

Zusammenfassend lässt sich demnach sagen, dass die Arbeit einige der festgelegten Ziele erfüllt. Jedoch sind die generierten Bilder nicht vollständig fotorealistisch. Das gilt auch für die Augmentierungen.

Zukünftige Arbeiten können hier anknüpfen, um insbesondere weitere Grenzfälle der Straßenschilderkennung zu simulieren. Auch um die momentanen Augmentierungen realistischer zu gestalten. Weitere Arbeiten auf dem Gebiet können sich zudem damit beschäftigen, Bilder für die Straßenschilderkennung zu erzeugen, die eine größere Menge an Hintergrund um die Schilder zeigen. Insbesondere um Grenzfälle der Straßenschilderkennung weiter zu simulieren. Auch ist denkbar, Videosequenzen künstlich zu generieren.

Literatur

- [1] T. Amaratunga, „Transfer Learning,“ in *Deep Learning on Windows: Building Deep Learning Computer Vision Systems on Microsoft Windows*. Berkeley, CA: Apress, 2021, S. 146–154. doi: [10.1007/978-1-4842-6431-7_7](https://doi.org/10.1007/978-1-4842-6431-7_7).
- [2] M. Staron, „AUTOSAR (AUTomotive Open System ARchitecture),“ in *Automotive Software Architectures: An Introduction*. Cham: Springer International Publishing, 2021, 97ff. ISBN: 978-3-030-65939-4. doi: [10.1007/978-3-030-65939-4_5](https://doi.org/10.1007/978-3-030-65939-4_5).
- [3] A. Gudigar, S. Chokkadi und R. U, „A review on automatic detection and recognition of traffic sign,“ *Multimedia Tools and Applications*, Jg. 75, S. 333–364, 2016. doi: [10.1007/s11042-014-2293-7](https://doi.org/10.1007/s11042-014-2293-7).
- [4] J. Stallkamp et al., „The German Traffic Sign Recognition Benchmark: A multi-class classification competition,“ in *IEEE International Joint Conference on Neural Networks*, 2011, S. 1453–1460.
- [5] EU-Kommission, *Regulation (EU) 2019/2144 of the European Parliament and of the Council*, Art. 6 Abs. 2c, 5. Sep. 2021.
- [6] H. Ippen und M. Bach, „Verkehrsschild-Erkennung Test,“ *Autozeitung*, 2. Apr. 2019.
- [7] bussgeldkatalog.org, *Wechselverkehrszeichen: Anzeigenänderung im Bedarfsfall*, o.D. Adresse: <https://www.bussgeldkatalog.org/wechselverkehrszeichen/> (besucht am 18.05.2023).
- [8] A. Borbe, *Sind verschneite Verkehrsschilder eigentlich noch gültig?* o.D. Adresse: <https://www.tz.de/auto/verkehrsschilder-gueltig-schnee-winter-verkehrsrecht-zr-9989863.html> (besucht am 18.05.2023).
- [9] H. Lengyel und Z. Szalay, „Test Scenario for Road Sign Recognition Systems with Special Attention on Traffic Sign Anomalies,“ in *2019 IEEE 19th International Symposium on Computational Intelligence and Informatics*, 2019, S. 000 193–000 198. doi: [10.1109/CINTI-MACRo49179.2019.9105238](https://doi.org/10.1109/CINTI-MACRo49179.2019.9105238).
- [10] I. Goodfellow, Y. Bengio und A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>. (besucht am 11.05.2023).
- [11] D. Sonnet, *Neuronale Netze Kompakt, Vom Perceptron zum Deep Learning*. Wiesbaden: Springer Vieweg Wiesbaden, 2020. doi: [10.1007/978-3-658-29081-8](https://doi.org/10.1007/978-3-658-29081-8).
- [12] A. S. Glassner, „Deep Learning: A Visual Approach,“ in San Francisco: No Starch Press, 2021, ISBN: 978-1-7185-0072-3.

- [13] Y. Ho und S. Wookey, „The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling,“ *IEEE Access*, Jg. 8, S. 4806–4813, 2020. doi: [10.1109/ACCESS.2019.2962617](https://doi.org/10.1109/ACCESS.2019.2962617).
- [14] J. Riebesell, *Convolutional Operator*, o.D. Adresse: <https://tikz.net/conv2d/> (besucht am 20.05.2023).
- [15] Qualcomm Developer Network, *Deep Learning and Convolutional Neural Networks for Computer Vision*, o.D. Adresse: <https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk/learning-resources/cnn-architectures/deep-learning-convolutional-neural-networks-computer-vision>.
- [16] M. T. Cicero, „Generatives Deep Learning,“ in übers. von M. Fraaß und K. Mach. Heidelberg: O'Reilly Verlag, 2020, Kap. Einführung ins Generative Deep Learning, ISBN: 9781492041948.
- [17] A. Karpathy, P. Abbeel, G. Brockman u. a., *Generative Models*, <https://openai.com/blog/generative-models/>, 16. Juni 2016. (besucht am 13.01.2023).
- [18] S. Bond-Taylor, A. Leach, Y. Long und C. G. Willcocks, „Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models,“ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Jg. 44, Nr. 11, S. 7327–7347, 2022. doi: [10.1109/tpami.2021.3116668](https://doi.org/10.1109/tpami.2021.3116668).
- [19] A. Oussidi und A. Elhassouny, „Deep generative models: Survey,“ in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2018, S. 1–8. doi: [10.1109/ISACV.2018.8354080](https://doi.org/10.1109/ISACV.2018.8354080).
- [20] S. I. Nikolenko, „Generative Models in Deep Learning,“ in *Synthetic Data for Deep Learning*. Cham: Springer International Publishing, 2021, S. 97–137. doi: [10.1007/978-3-030-75178-4_4](https://doi.org/10.1007/978-3-030-75178-4_4).
- [21] A. van den Oord, N. Kalchbrenner und K. Kavukcuoglu, „Pixel Recurrent Neural Networks,“ *CoRR*, 2016. doi: [10.48550/arXiv.1601.06759](https://doi.org/10.48550/arXiv.1601.06759).
- [22] J. Zhai, S. Zhang, J. Chen und Q. He, „Autoencoder and Its Various Variants,“ in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, S. 415–419. doi: [10.1109/SMC.2018.00080](https://doi.org/10.1109/SMC.2018.00080).
- [23] R. O'Connor, *Introduction to Variational Autoencoders Using Keras*, 3. Jan. 2022. Adresse: <https://www.assemblyai.com/blog/introduction-to-variational-autoencoders-using-keras/> (besucht am 20.05.2023).
- [24] D. Bank, N. Koenigstein und R. Giryes, „Autoencoders,“ *CoRR*, 2020. doi: [10.48550/arXiv.2003.05991](https://doi.org/10.48550/arXiv.2003.05991).

- [25] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza u. a., *Generative Adversarial Networks*, 2014. doi: [10.48550/ARXIV.1406.2661](https://doi.org/10.48550/ARXIV.1406.2661).
- [26] I. Goodfellow, J. Pouget-Abadie, M. Mirza u. a., „Generative Adversarial Networks,“ *Commun. ACM*, Jg. 63, Nr. 11, 139–144, 2020. doi: [10.1145/3422622](https://doi.org/10.1145/3422622).
- [27] J.-Y. Zhu, T. Park, P. Isola und A. A. Efros, *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*, 2017. doi: [10.48550/ARXIV.1703.10593](https://doi.org/10.48550/ARXIV.1703.10593).
- [28] TensorFlow, *CycleGAN*, <https://www.tensorflow.org/tutorials/generative/cyclegan>, o.D. (besucht am 01.04.2023).
- [29] Z. He, *CycleGAN-Tensorflow-2*, 13. Dez. 2021. Adresse: <https://github.com/LynnHo/CycleGAN-Tensorflow-2> (besucht am 06.05.2023).
- [30] K. He, X. Zhang, S. Ren und J. Sun, *Deep Residual Learning for Image Recognition*, 2015. arXiv: [1512.03385 \[cs.CV\]](https://arxiv.org/abs/1512.03385).
- [31] P. Isola, J.-Y. Zhu, T. Zhou und A. A. Efros, *Image-to-Image Translation with Conditional Adversarial Networks*, 2018. arXiv: [1611.07004 \[cs.CV\]](https://arxiv.org/abs/1611.07004).
- [32] O. Ronneberger, P. Fischer und T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, 2015. arXiv: [1505.04597 \[cs.CV\]](https://arxiv.org/abs/1505.04597).
- [33] D. Christine et al., „Synthetic Data generation using DCGAN for improved traffic sign recognition,“ *Neural Computing and Applications*, S. 1–16, Apr. 2021. doi: [10.1007/s00521-021-05982-z](https://doi.org/10.1007/s00521-021-05982-z).
- [34] D. Spata, D. Horn und S. Houben, „Generation of Natural Traffic Sign Images Using Domain Translation with Cycle-Consistent Generative Adversarial Networks,“ in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, S. 702–708. doi: [10.1109/IVS.2019.8814090](https://doi.org/10.1109/IVS.2019.8814090).
- [35] G. Nguyen, S. Dlugolinsky, M. Bobák u. a., „Machine Learning and Deep Learning Frameworks and Libraries for Large-Scale Data Mining: A Survey,“ *Artif. Intell. Rev.*, Jg. 52, Nr. 1, 77–124, 2019. doi: [10.1007/s10462-018-09679-z](https://doi.org/10.1007/s10462-018-09679-z).
- [36] TensorFlow, *tf.data.Dataset*, https://www.tensorflow.org/api_docs/python/tf/data/Dataset, o.D. (besucht am 30.03.2023).
- [37] P. Singh und A. Manure, *Learn TensorFlow 2.0, Implement Machine Learning and Deep Learning Models with Python*. Berkeley, CA: Apress, 2020. doi: [10.1007/978-1-4842-5558-2_1](https://doi.org/10.1007/978-1-4842-5558-2_1).
- [38] L. Huang, *Chinese Traffic Sign Database: Traffic Sign Recognition Database*, <http://www.nlpr.ia.ac.cn/pal/trafficdata/recognition.html>, o.D. (besucht am 25.03.2023).
- [39] United Nations Economic Commission for Europe, *Convention on Road Traffic*, 28. März 2006.

- [40] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold und Y. Kuang, *The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale*, 2020. doi: [10.48550/arXiv.1909.04422](https://doi.org/10.48550/arXiv.1909.04422).
- [41] R. Timofte, K. Zimmermann, und L. van Gool, „Multi-view traffic sign detection, recognition, and 3D localisation,“ *Journal of Machine Vision and Applications (MVA 2011)*, 2011. doi: [10.1007/s00138-011-0391-3](https://doi.org/10.1007/s00138-011-0391-3).
- [42] *Road Signs Dataset*, o.D. Adresse: <https://makeml.app/datasets/road-signs>.
- [43] J. Valentin und S. Bouaziz, *Introducing TensorFlow Graphics: Computer Graphics Meets Deep Learning*, https://www.tensorflow.org/api_docs/python/tf/keras/utils, o.D. (besucht am 31.03.2023).
- [44] OpenCV, *Introduction*, o.D. Adresse: <https://docs.opencv.org/4.x/d1/dfb/intro.html> (besucht am 08.05.2023).
- [45] J. A. Clark u.a., *Pillow, Overview*, o.D. Adresse: <https://pillow.readthedocs.io/en/stable/handbook/overview.html> (besucht am 08.05.2023).
- [46] E. Betzalel, C. Penso, A. Navon und E. Fetaya, *NumPy user guide*, o.D. Adresse: <https://numpy.org/doc/stable/user/index.html> (besucht am 01.06.2023).
- [47] Bundesanstalt für Straßenwesen (BASt), *Verkehrszeichen und Symbole, Verkehrszeichenkatalog 2017*, <https://www.bast.de/DE/Verkehrstechnik/Fachthemen/v1-verkehrszeichen/vz-start.html>, o.D. (besucht am 27.03.2023).
- [48] W. Burger und M. J. Burge, *Digital Image Processing: An Algorithmic introduction*. Springer Cham, 2022, S. 601–637. doi: [10.1007/978-3-031-05744-1](https://doi.org/10.1007/978-3-031-05744-1).
- [49] F. Dunn und I. Parberry, „3D Math Primer for Graphics and Game Development,“ in A K Peters/CRC Press, 2011, Kap. Rotation in Three Dimensions. Adresse: <https://gamedev.math.com/book/orient.html>.
- [50] S. Liu, „Study for Identity Losses in Image-to-Image Domain Translation with Cycle-Consistent Generative Adversarial Network,“ *Journal of Physics: Conference Series*, Jg. 2400, Nr. 1, S. 012030, 2022. doi: [10.1088/1742-6596/2400/1/012030](https://doi.org/10.1088/1742-6596/2400/1/012030).
- [51] R. Varma, *Downsides of the sigmoid activation and why you should center your inputs*, o.D. Adresse: <https://rohanvarma.me/inputnormalization> (besucht am 11.05.2023).
- [52] R. Varma, *TensorFlow Examples*, o.D. Adresse: <https://rohanvarma.me/inputnormalization> (besucht am 11.05.2023).
- [53] TensorFlow, *tf.map_fn*, o.D. Adresse: https://www.tensorflow.org/api_docs/python/tf/map_fn (besucht am 11.05.2023).

- [54] TensorFlow, *tf.keras.utils*, https://www.tensorflow.org/api_docs/python/tf/keras/utils, o.D. (besucht am 31.03.2023).
- [55] X. Luo, N. Z. Salamon und E. Eisemann, „Adding Motion Blur to Still Images,“ in *Proceedings of the 44th Graphics Interface Conference*, Toronto, Canada: Canadian Human-Computer Communications Society, 2018, 108–114. doi: [10.20380/GI2018.15](https://doi.org/10.20380/GI2018.15).
- [56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li und L. Fei-Fei, „ImageNet: A large-scale hierarchical image database,“ in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, S. 248–255. doi: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848).
- [57] C. Szegedy, W. Liu, Y. Jia u. a., *Going Deeper with Convolutions*, 2014. arXiv: [1409.4842](https://arxiv.org/abs/1409.4842).
- [58] E. Betzalel, C. Penso, A. Navon und E. Fetaya, *A Study on the Evaluation of Generative Models*, 2022. arXiv: [2206.10935 \[cs.LG\]](https://arxiv.org/abs/2206.10935).
- [59] J. Akther, M. Harun-Or-Roshid, A.-A. Nayan und M. Kibria, „Transfer learning on VGG16 for the Classification of Potato Leaves Infected by Blight Diseases,“ Dez. 2021. doi: [10.1109/ETCCE54784.2021.9689792](https://doi.org/10.1109/ETCCE54784.2021.9689792).

Anhang

A. Abbildungen

B. Listings

Abbildungen



Abbildung A.1: Alle 43 Piktogramme zu den generierten Klassen von Straßenschildern [47]

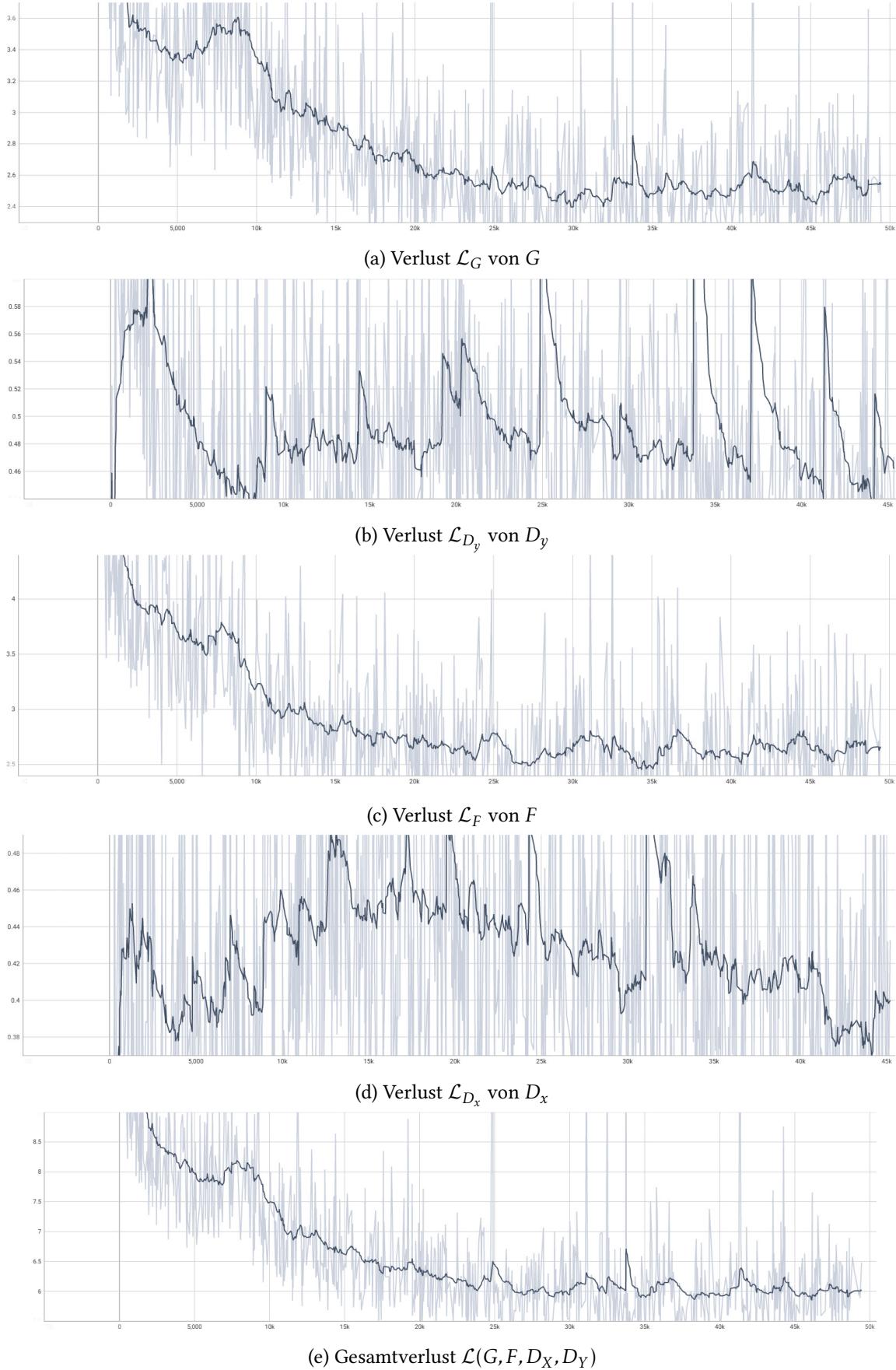


Abbildung A.2: Trainingsverlauf des U-Net bis Epoche 200 (x-Achse: Trainingsschritt, y-Achse: Verlust)



Abbildung A.3: Trainingsverlauf des U-Net-basierten CycleGAN bis Epoche 100



Abbildung A.4: Trainingsverlauf des ResNet-basierten CycleGAN bis Epoche 100



Abbildung A.5: Beispielbilder des U-Net-basierten CycleGAN nach 200 Epochen

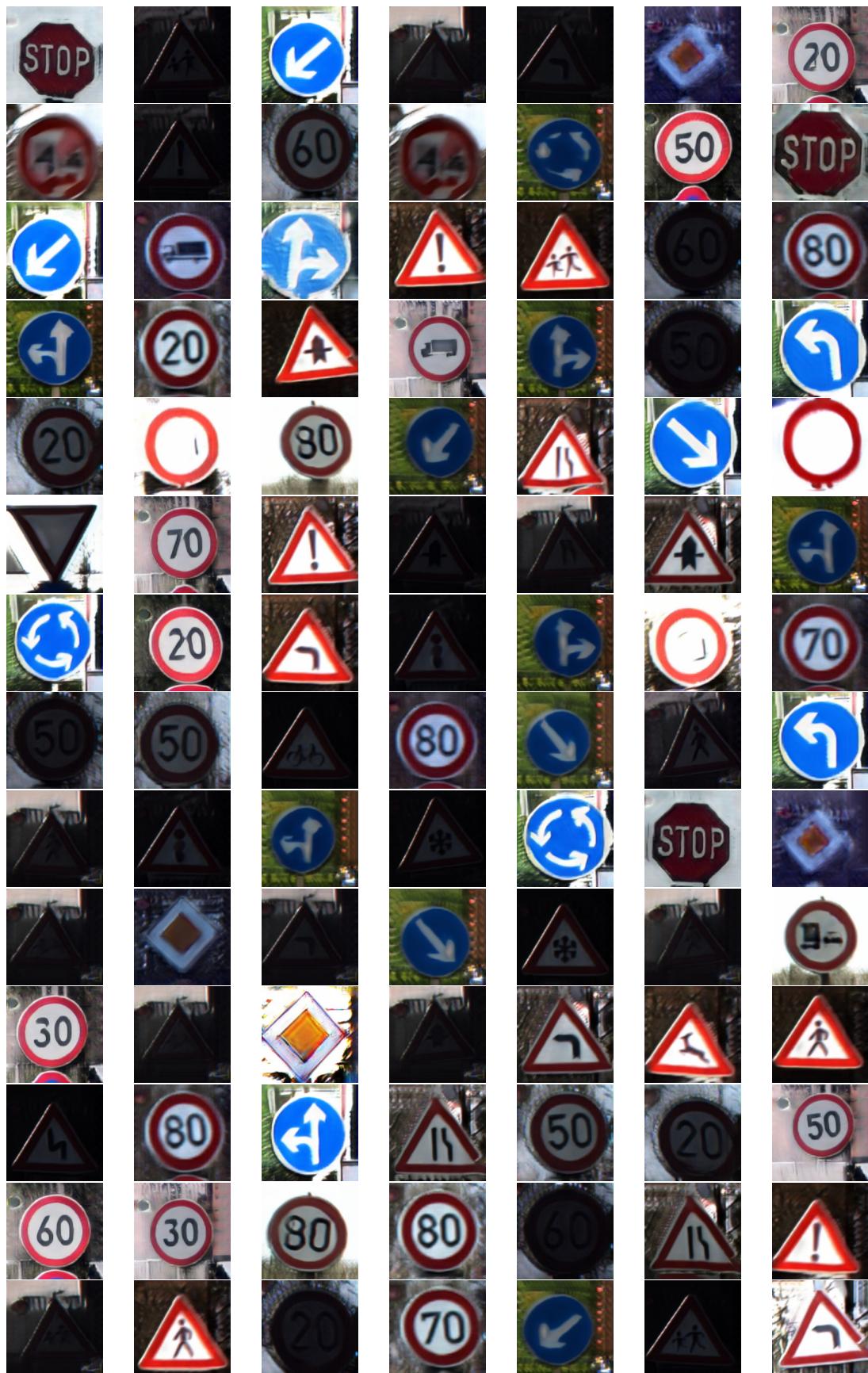


Abbildung A.6: Beispielbilder des ResNet-basierten CycleGAN mit 9 Residual Blocks nach 180 Epochen



Abbildung A.7: Beispielbilder mit Bewegungsunschärfe



Abbildung A.8: Beispielbilder mit als ungültig markierten Schildern



Abbildung A.9: Beispielbilder mit Schnee



Abbildung A.10: Bilder mit Schnee, dessen Augmentierung das CycleGAN **eigenständig generiert** hat



Abbildung A.11: Beispielbilder mit allen drei Augmentierungen

Listings

```
src
├── checkpoints
│   ├── cyclegan_resnet
│   └── cyclegan_u_net
├── classifier
│   ├── checkpoints
│   └── run.py
├── config
│   ├── config.toml
│   └── validate_config.py
├── experimental
│   └── generate_single_classes.py
├── external
│   └── resnet.py
└── logs
    ├── resnet
    └── unet
└── utils
    ├── image_augmentation.py
    ├── load_data.py
    ├── misc.py
    └── preprocess_image.py
    ├── generate.py
    ├── model.py
    └── train.py
```

utils/preprocess_image.py

```

def randomly_transform_image_batch(img_tensor_batch,
                                    target_size=256):

    batch_size = img_tensor_batch.shape[0]
    # resize content
    min_content_size = target_size / 1.5
    # we have to work with default python lists because we need the
    → pop function
    content_sizes = [
        np.random.randint(low=min_content_size, high=target_size) for el
        → in range(batch_size)
    ]
    # copy of content_sizes; will be used to pop the elements
    content_sizes_tmp = content_sizes[:]
    transformed_imgs = tf.map_fn(
        lambda img: resize_content_of_img(
            img, target_size, content_sizes_tmp.pop(0)),
        img_tensor_batch)

    # randomly rotate the image in x,y and z direction; scale values
    → are empirically chosen
    alpha_z_values = np.random.normal(loc=0.0, scale=3.5,
                                       size=batch_size)
    alpha_y_values = np.random.normal(loc=0.0, scale=0.01,
                                       size=batch_size)
    alpha_x_values = np.random.normal(loc=0.0, scale=0.01,
                                       size=batch_size)
    transform_matrices = np.zeros((batch_size, 3, 3))
    for i in range(batch_size):
        transform_matrices[i] =
            → create_rotation_matrix(alpha_z_values[i],
                                       alpha_y_values[i], alpha_x_values[i])
    transform_matrices = tf.convert_to_tensor(transform_matrices,
                                              dtype=tf.float32)

    transformed_imgs = 1 - transformed_imgs
    transformed_imgs =
        → tfg_image_transformer.perspective_transform(transformed_imgs,
                                                       transform_matrices)
    transformed_imgs = 1 - transformed_imgs

    return transformed_imgs, content_sizes, transform_matrices

```

Listing 1: Augmentierung eines Batches von Bildern

model.py

```

def fit(self, pictograms, real_images, epochs=1):
    """Train the model for a specific number of epochs. Checkpoints
    are saved every epoch.

    Args:
        pictograms: 4d tensor containing the raw pictograms
        ↵ (batch_size, height, width, channels).
        real_images: 4d tensor containing the training images of
        ↵ street signs (batch_size, height, width, channels).
        epochs: Number of epochs to train the model.

    """
    print('Training...')

    with self.summary_writer.as_default():
        for epoch in range(epochs):
            self.total_epochs.assign_add(1) # increment
            print(f'Epoch: {int(self.total_epochs)} / '
                  ↵ {int(self.total_epochs) + epochs-(epoch+1)}')

            # Single training step
            for image_batch in tqdm(real_images):
                self.total_steps.assign_add(1) # increment

                # Transform the pictograms
                pictograms.shuffle(buffer_size=100,
                    ↵ reshuffle_each_iteration=True)
                single_pictogram_batch =
                    pictograms.take(1).get_single_element()
                single_pictogram_batch, _, _ =
                    utils.preprocess_image.randomly_transform_image_batch(
                        single_pictogram_batch)

                # Train the model
                losses = self.train_step(single_pictogram_batch,
                    image_batch)

                # For Tensorboard; Log the losses
                for loss_name in losses:
                    tf.summary.scalar(loss_name, losses[loss_name],
                        int(self.total_steps))

                # After each epoch: Generate an image
                ... # not relevant for the training itself

                self.summary_writer.flush() # write tensorboard logs to
                ↵ log file

```

```
if ((epoch + 1) % 10 == 0) and ((epoch + 1) < epochs):
    self.checkpoint_manager.save()
    print('Checkpoint saved for epoch
          ↵ {}'.format(epoch + 1))
self.checkpoint_manager.save()
print('Checkpoint saved for this training')
```

Listing 2: Vollständige Trainingsfunktion

classifier/run.py

```
def get_model():
    num_classes = 43
    base_model = VGG16(weights='imagenet', include_top=False,
                        input_shape=(256, 256, 3))

    # Freeze all layers
    for layer in base_model.layers:
        layer.trainable = False

    x = base_model.output
    x = GlobalAveragePooling2D()(x)
    x = Dense(512, activation='relu')(x)
    predictions = Dense(num_classes, activation='softmax')(x)

    # Create a new model with the modified architecture
    model = Model(inputs=base_model.input, outputs=predictions)

    # Compile the model with a low learning rate
    optimizer = tf.keras.optimizers.Adam(lr=1e-5)
    model.compile(loss='categorical_crossentropy',
                  optimizer=optimizer, metrics=['accuracy'])

    return model
```

Listing 3: Anpassen des VGG16 Modells für die Klassifikation von der Straßenschilder [1]