

Open Set Gestenerkennung zur Interaktion zwischen Passanten und Fahrzeugen mit Automatisierungsfunktionen

Vollständiger Beitrag

Christian Reichhardt¹, Manuel Martin ² und Patrick Baier¹


Abstract: Die Zunahme an intelligenten Steuerungssystemen in Fahrzeugen wirft zunehmend die Frage auf, inwiefern ein Fahrzeug mit seiner Umgebung automatisch interagieren kann. Eine wichtige Aufgabe besteht dabei in der Kommunikation und Interaktion mit Passanten am Fahrbahnrand, welche eine anwendungsspezifische Gestenerkennung von Personen voraussetzt. Für die Erkennung solcher Gesten existieren mittlerweile Ansätze auf Basis maschineller Bilderkennungsverfahren, welche einfache Gesten erkennen können. Zwei wichtige Fragestellungen sind dabei, welche Aktionsklassen durch ein Fahrzeug zuverlässig erkennbar sind und wie mit nicht explizit trainierten Aktionsgesten umgegangen werden kann. Unter diesen Aspekten untersucht die vorliegende Arbeit die Fähigkeiten und Grenzen der Gestenerkennung im Kontext relevanter Aktionen von Passanten im Straßenverkehr. Hierzu wurde als Teil einer Verarbeitungskette das häufig verwendete tiefe neuronale Netz *ST-GCN++* betrachtet und mit Gelenkpunkten von Personen in unterschiedlichen Detailgraden trainiert. Um mit nicht explizit trainierten Gesten umgehen zu können, wurde eine Restklasse eingeführt, welche auf im Training ungesehenen Gesten hohe Genauigkeiten erzielte.

Keywords: Gestenerkennung, Passanten, Fahrzeug, Open Set Recognition, ST-GCN++

1 Einleitung

Nonverbale Kommunikation gehört zum Alltag im Straßenverkehr und ist maßgebend für den reibungslosen Verkehrsablauf auf deutschen und internationalen Straßen. Dies wird insbesondere in Situationen klar, in denen Fahrzeuge und Passanten aufeinandertreffen. So zeigte Šucha [Šu14] in einer Studie, dass Passanten mit Autofahrern sowohl über Augenkontakt, Handgesten als auch über Kopfbewegungen interagieren. Dass bestimmte Gesten einen Einfluss auf das Fahrverhalten haben können, wurde bereits mehrfach durch Studien [CVL11; ZW14] belegt, die den Einfluss von Handgesten auf die Rate anhaltender Fahrzeuge untersuchten. Obwohl in der Industrie bereits mehrere Ansätze [MB24; NM15; SA24] existieren, Intentionen automatisierter Fahrzeuge Passanten bekannt zu geben, ist es bislang nicht üblich, dass Gesten von Passanten erkannt, interpretiert und darauf automatisiert reagiert werden kann. Dies könnte potentiell zu Missverständnissen mit schwerwiegenden Folgen führen.

1 Hochschule Karlsruhe - Technik und Wirtschaft, Fakultät für Informatik und Wirtschaftsinformatik, Moltkestr. 30, 76133 Karlsruhe, Deutschland, rech1028@h-ka.de; patrick.baier@h-ka.de

2 Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB, Human-AI Interaction, Fraunhoferstraße 1, 76131 Karlsruhe, Deutschland, manuel.martin@iosb.fraunhofer.de,  <https://orcid.org/0000-0001-9473-8804>

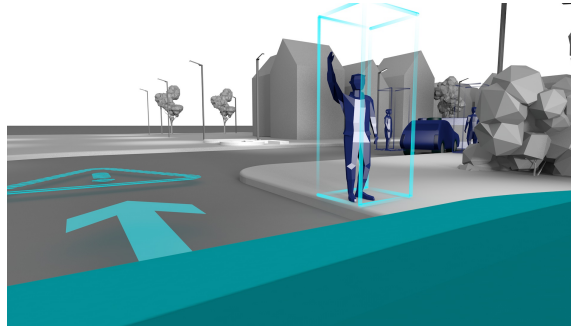


Abb. 1: Szenario des Forschungsprojekts *INITIATIVE* [PI21]. Interaktion zwischen Fahrzeug und Passant, wobei der Passant das Fahrzeug durch eine Geste vor einer Gefahrenstelle warnen möchte. [PIS21]

Aus diesem Grund liegt das Ziel aktueller Forschungen [PIG21] darin, einen durch maschinelles Lernen gestützten Interaktionsprozess zwischen Fahrzeugen mit Automatisierungsfunktionen und anderen Verkehrsteilnehmern zu etablieren. Stanciu et al. [St18] führen in ihrer Arbeit aus, dass solche Fahrzeuge zukünftig in der Lage sein müssen, sowohl Strategien aller Verkehrsteilnehmer zu erkennen, als auch zu interpretieren. Um eine Interpretation der Intentionen von Passanten zu ermöglichen, ist jedoch eine Gestenerkennung Voraussetzung. Abbildung 1 zeigt ein solches Szenario, in dem ein Passant das entgegenkommende Fahrzeug durch Winken auf eine Gefahrenstelle aufmerksam machen möchte. Eine besondere Herausforderung stellt dabei die Diversität genutzter Gesten dar. Wie von Najamuddin [Na19] gezeigt, bestehen außerdem große Unterschiede was die Nutzung von Gesten in verschiedenen kulturellen Umfeldern angeht. Zur Lösung von Problemen dieser Art, schlagen Stanciu et al. [St18] in ihrer Arbeit eine Standardisierung von Kommunikationsstrategien im Straßenverkehr vor.

Um einen Überblick zu bieten, welche Aktionsgesten hierfür verwendbar sind, wurden in der vorliegenden Arbeit für diesen Anwendungsfall spezifische Aktionsklassen gewählt und auf deren Erkennbarkeit geprüft. Hierzu wurde die von Duan et al. [Du22] beschriebene Architektur *Spatial-Temporal Graph Convolutional Network ST-GCN++* zur Gestenerkennung in unterschiedlichen Konfigurationen trainiert und ausgewertet. Diese erhält zweidimensionale Gelenkpunkte von Personen über eine vordefinierte Zeit als Eingabe, um Aktionsgesten zu klassifizieren. Intern werden diese Gelenkpunkte als mehrdimensionale Graphen in Form kinematischer Skelette über einen vordefinierten Zeitraum dargestellt. Um für den Anwendungsfall einen sinnvollen Graphen zu finden, wurden zwei Graphtopologien untersucht und in jeweils zwei Konfigurationen trainiert. Diese Topologien unterscheiden sich jeweils in der Anzahl betrachteter Knotenpunkte. Zu diesem Zweck wurden aus einem öffentlichen Datensatz Gelenkpunkte von Personen in unterschiedlichen Detailgraden für das Training generiert. Üblicherweise wurde die Gestenerkennung im Straßenverkehr bisher unter einer geschlossenen Anzahl von Aktionsklassen betrachtet, wie beispielsweise bei der

Erkennung von Verkehrszeichen autorisierter Verkehrslotsen [Mi21]. Diese Annahme ist für die Interaktion mit Passanten aufgrund der Vielzahl möglicher Gesten nicht praktikabel. Aus diesem Grund wurden Gesten im Kontext unbekannter Bewegungen erfasst, was auch als *Open Set Recognition Problem* [Sc12] bezeichnet wird. Da sich dieser Ansatz grundlegend von üblichen Trainings- und Evaluationsmethoden von Gestenerkennungs-Algorithmen unterscheidet, wurde ein öffentlicher Datensatz umfassend an diese Anforderungen angepasst. Zur Auswertung wurden zwei unterschiedliche Ansätze zur Manipulation der Trainingsdaten verfolgt und die Ergebnisse eingeordnet. Dabei konnten hohe Genauigkeiten bezüglich im Training ungesehener Aktionsgesten der hierfür eingeführten Restklasse beobachtet werden. Explizit trainierte Aktionsklassen mit einem hohen Bewegungsanteil lieferten ebenfalls sehr hohe Genauigkeiten, während Handgesten vergleichsweise geringere Genauigkeiten erzielten.

2 Stand der Forschung

Im Bereich der Gestenerkennung gehören tiefe neuronale Netze aktuell zum Standard und erzielen in unterschiedlichen Ausprägungen auch bei einer höheren Anzahl an Aktionsgesten gute Resultate. Dies zeigen unter anderem Ergebnisse von Duan et al. [Du22]. Aufgrund der dynamischen Natur vieler Gesten ist es bei der Erkennung menschlicher Aktivitäten oft unerlässlich, neben den räumlichen Informationen zudem zeitliche Informationen zu berücksichtigen. Gegenüber videobasierten Methoden [St20; Su17; Wa16; Zh21] existieren hierfür skelettbasierte Ansätze, wobei Gelenkpunkte als Knoten von Graphen genutzt werden. Für deren Klassifikation eignen sich neben Faltungsnetzen wie *PoseConv3D* [Du22] auch Graphfaltungsnetze [KW16] besonders gut. Diese ermöglichen durch die Repräsentation eines Graphen als Adjazenzmatrix Verbreitung und Faltungen von Graphen im Netz. Architekturen wie das von Duan et al. [Du22] vorgeschlagene *ST-GCN++* versprechen dabei echtzeitfähige Laufzeit bei einer hohen Genauigkeit. Da als Eingang solcher Netze meist eine Posenschätzung benötigt wird, wird in der Praxis häufig eine *Top-Down* Verarbeitungskette aus Objekterkennung, Posenschätzung und anschließender Gestenerkennung genutzt. Dabei bieten sich neuronale Netze, beispielsweise basierend auf den Architekturen *R-CNN* [Gi14] und *HRNet* [Su19] an.

Bei bisherigen Untersuchungen zur Gestenerkennung im Straßenverkehr [He20; MAA16; Mi21; Po19] wird a priori angenommen, dass alle zur Laufzeit auftretenden Gesten bekannt sind. Die Vielzahl an möglichen Aktionen im Straßenverkehr stellt jedoch die Anforderung an eine *Open Set Recognition (OSR)* [Sc12], welche darin besteht, zur Inferenzzeit im Training ungesehene Aktionsklassen von gesehenen unterscheiden zu können. Um zu erkennen, ob ein Datenelement einer trainierten Aktion entspricht, existieren in der Literatur verschiedene Ansätze [BIG18; BYK21; Ro20; Sh18]. Einen einfachen und effizienten Ansatz bietet die von Zhang und LeCun [ZL17] beschriebene *Dustbin*-Methode. Dabei wird eine zusätzliche Restklasse mit Samples einer Menge nicht beschrifteter, d.h. nicht explizit zu erkennender, Aktionen trainiert. Sollte eine Klasse zur Inferenzzeit mit höchster Wahrscheinlichkeit der

Restklasse zugeordnet werden, so wird diese als unbekannte Geste eingestuft. Nach Zhang; LeCun [ZL17] kann sich dieses Vorgehen durch einen Regularisierungseffekt außerdem positiv auf das Training auswirken. Im Gegensatz zu bisherigen Untersuchungen wird in dieser Arbeit daher die Gestenerkennung im Straßenverkehr unter dem Gesichtspunkt der *OSR* mithilfe der *Dustbin*-Methode betrachtet und ausgewertet.

3 Modell

In dieser Arbeit werden Modelle des tiefen neuronalen Netzes *ST-GCN++* auf unterschiedlichen Daten trainiert, wobei die Architektur, die Vorverarbeitung sowie die Netzkonfiguration der Implementierung von Duan et al. [Du22] folgt. Die Modelle unterscheiden sich dabei lediglich in den Graphtopologien, beschrieben in Abschnitt 4.1, was zu einer Änderung der Adjazenzmatrix führt. Die Gelenkpunkte des Datensatzes, welche durch zweidimensionale Koordinaten gegeben sind, wurden zunächst relativ zur Bildmitte auf den Bereich $[-1, 1]$ normiert und anschließend mit deren Erkennungswahrscheinlichkeiten konkateniert. Fünfdimensionale Tensoren der Form $N \times M \times T \times V \times C$ wurden als Eingabe verwendet. Diese bestehen aus N Samples pro Batch mit T Frames pro Sample. Dabei existieren pro Sample M detektierte Personen aus jeweils V Knotenpunkten mit C Dimensionen. In diesem Fall ist $C = 3$, wobei die dritte Dimension der Knotenpunkte die Erkennungswahrscheinlichkeit der Posenschätzung ist. Zu beachten ist, dass jedes geladene Video mit $T = 100$ gesamplet wird. Zu diesem Zweck wurden Videos in $n = 100$ Segmente der selben Länge geteilt und pro Segment zufällig ein Frame gewählt. Für den Fall, dass der Eingang kleiner als T ist, startet die Indizierung an einem zufälligen Frame des Videos und dieses wird solange wiederholt, bis die festgelegte Anzahl der Frames erreicht wurde. Die Zahl erkannter Personen wird insgesamt auf $M = 2$ beschränkt. Die zeitliche Faltung des trainierten Modells *ST-GCN++* wird mit einer Kernelgröße $\Gamma \times 1$ und $\Gamma = 3$ für jeden der vier *Convolutional Streams* durchgeführt. Dies gilt ebenfalls für die Kernelgröße des *MaxPooling Streams*. Als Partitionierungsstrategie für *ST-GCN++* wurde die von Duan et al. [Du22] beschriebene *Spatial Configuration* gewählt, mit Knotenpunkt null als Körperschwerpunkt für beide untersuchten Graphtopologien.

4 Experiment

Die beschriebene Architektur wurde in mehreren Konfigurationen trainiert und evaluiert, wobei zwei Graphtopologien für die Abbildung von Gesten miteinander verglichen wurden. Hierfür wurde ein öffentlich verfügbarer Datensatz umfassend entsprechend den Anforderungen einer *OSR* modifiziert. Um Modelle ebenfalls auf im Training ungesesehenen Gesten evaluieren zu können, wurden zwei verschiedene Versionen eines Trainingsdatensatzes erstellt und jeweils unterschiedliche Modelle trainiert. Als Basis für die Implementierung diente der Programmcode der Veröffentlichung von Duan et al. [Du22].

Klassendefinitionen					
ID	Klasse	NTU RGB+D	ID	Klasse	NTU RGB+D
0	Restklasse	-	13	Fallen	A43
1	Brille abnehmen	A19	14	Auf Person zeigen	A54
2	Winken mit einer Hand	A23	15	Ball prellen	A64
3	Hüpfen auf einem Bein	A26	16	„Leise“-Zeichen	A67
4	Aufspringen	A27	17	Daumen hoch	A69
5	Telefonieren	A28	18	Daumen runter	A70
6	Handy/ Tablet bedienen	A29	19	„OK“-Zeichen	A71
7	Mit Finger zeigen	A31	20	„Victory“-Zeichen	A72
8	Kopf nicken	A35	21	Faust schütteln	A93
9	Kopf schütteln	A36	22	Beide Hände oben	A95
10	Salut	A38	23	Armkreisen	A97
11	Arme kreuzen („Stopp“)	A40	24	Armschwingen	A98
12	Taumeln	A42			

Tab. 1: Trainierte Klassen mit deren neuen Labels (ID) sowie deren Bezeichnung im *NTU RGB+D 120* Datensatz.

4.1 Datensatz

Aufgrund des Mangels an frei verfügbaren Videodatensätzen im Straßenverkehr wurde als Basis für den finalen Datensatz der Trainings- und Testdatensatz des *Cross-Subject* Evaluationsprotokolls von *NTU RGB+D 120* [Li19] verwendet. Dieser wurde an die Anforderung einer *OSR* sowie für die Erkennung spezifischer Gesten angepasst.

Der finale Datensatz muss zunächst für den Anwendungsfall relevante Gesten annotieren, die das Modell explizit erkennen soll. Aus diesem Grund dienten als Grundlage für die Auswahl betrachteter Aktionsgesten Untersuchungen zur Nutzung von Gesten im Straßenverkehr von Zhuang; Wu [ZW14] und Brand; Schmitz [BM23]. Die dort untersuchten Gesten wurden mit den im Ausgangsdatsatz verfügbaren Klassen abgeglichen. Vergleichbare Gesten, sowie einige weitere Klassen des Datensatzes, wurden nach Ermessen der Autoren aus dem *NTU RGB+D 120* Datensatz ausgewählt und mit neuen IDs versehen. Tab. 1 liefert eine Übersicht über alle Klassen des finalen Datensatzes sowie deren ursprüngliche Bezeichnung im Ausgangsdatsatz. Während sich davon Manche primär über die Position der Gelenkpunkte der Arme definieren (Klassen ID 1, 2, 5, 6, 7, 10, 11, 14, 15, 16, 21, 22, 23 und 24), beziehen sich Andere auf Bewegungen des gesamten Körpers (3, 4, 12, 13). Weitere Gesten zeichnen sich hauptsächlich durch Kopfbewegungen (8, 9) oder Fingerstellungen (17, 18, 19, 20) aus. Auch der Grad der Bewegung in den Datenproben variiert teilweise stark zwischen einzelnen Klassen. Besonders hohe Bewegungsgrade sind bei den Gesten *Aufspringen*, *Fallen*, *Armkreisen* sowie *Armschwingen* zu beobachten.

Zur Umsetzung der *OSR*, in Anlehnung an die Methodik der *Dustbin*-Methode von Zhang; LeCun [ZL17], wurde eine zusätzliche Restklasse mit ID null erstellt. Dieser

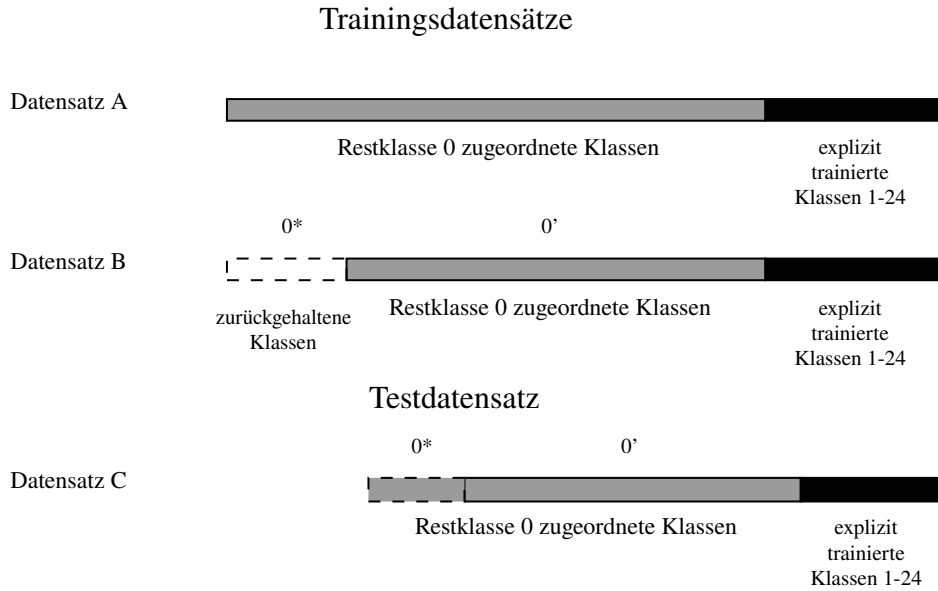


Abb. 2: Modifikationen des *NTU RGB+D 120* Datensatzes. Als Ausgangspunkt dienten die Trainings- und Testdatensätze des *Cross-Subject* Evaluationsprotokolls.

Klasse werden alle Datenproben des *NTU RGB+D 120* Datensatzes zugeordnet, deren Erkennung nicht den anderen Klassen in Tab. 1 entsprechen. Insgesamt wurden durch Modifikationen des *Cross-Subject* Evaluationsprotokolls zwei Trainingsdatensätze und ein Testdatensatz generiert, deren Aufbau schematisch in Abb. 2 gezeigt wird. Datensatz A weist allen Samples im Trainingsdatensatz, welche nicht den Klassen 1–24 zugeordnet wurden, der Restklasse zu. Das Training mit Datensatz A wird im Folgenden als Training *ohne Rückhaltung* bezeichnet. Diese Methodik ermöglicht jedoch keine Evaluation bezüglich im Training ungesehener Gesten auf dem Testdatensatz. Aus diesem Grund wurde eine weitere Modifikation der Trainingsdaten vorgenommen. In Datensatz B wurde zusätzlich eine Untermenge 0^* der Aktionsgesten der Restklasse für das Training entfernt. 0^* beinhaltet die Videosamples der Klassen *A1–A18* sowie *A20* im *NTU RGB+D 120* Datensatz. Das Training mit Datensatz B wird weiterhin als Training *mit Rückhaltung* bezeichnet. Die Modifikationen des Testdatensatzes C entsprechen den Modifikationen von Datensatz A und somit beinhaltet dieser Datenproben einer Menge von Klassen 0^* , die in Datensatz B nicht enthalten sind.

Für die Datensätze A bis C wurden jeweils zweimal Merkmale generiert, einmal 17 und einmal 65 zweidimensionale Koordinaten von Gelenkpunkten inklusive deren Erkennungswahrscheinlichkeit pro Person und Frame. Die Merkmale beschreiben dabei die Knotenpunkte der Graphen, die das Modell zur räumlich-zeitlichen Faltung verwendet. Insgesamt wurden zwei Graphtopologien in separaten Modellen erstellt, die sich in der

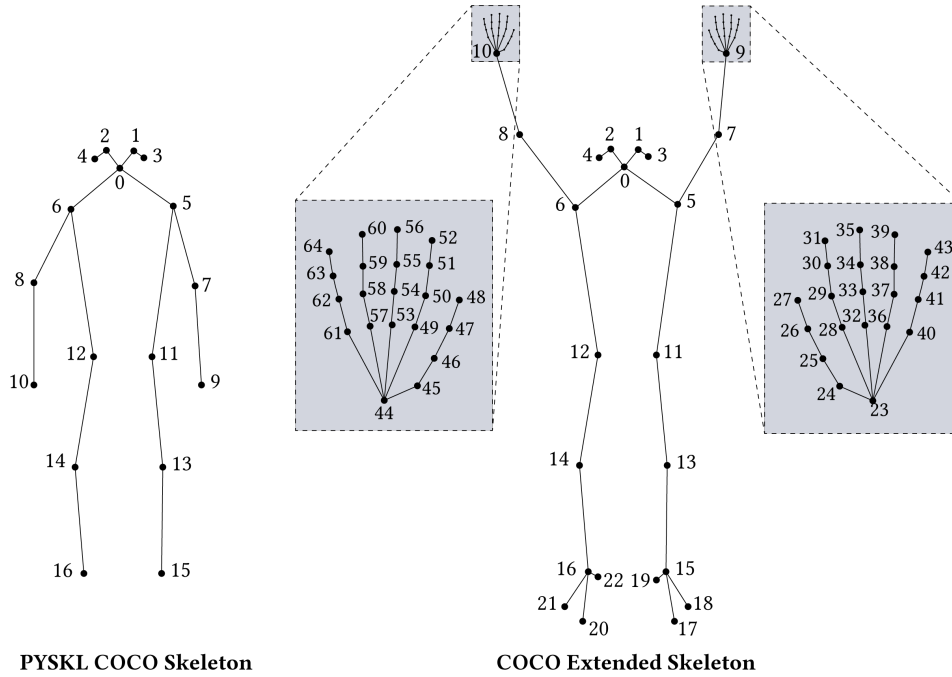


Abb. 3: Räumliche Komponenten der in dieser Arbeit untersuchten Graphtopologien sowie Indizierung der Knotenpunkte.

räumlichen Ebene durch die Anzahl an Knotenpunkten unterscheiden und in Abb. 3 schematisch dargestellt sind. Die erste Topologie entspricht derjenigen von Duan et al. [Du22], welche auf Gelenkpunkten des *COCO*-Datensatzes [Li14] basiert. Hierfür wurden Merkmale genutzt, die in der Implementierung von Duan et al. [Du22] zur Verfügung gestellt sind. Die zweite Topologie verwendet eine Teilmenge derjenigen Gelenkpunkte, die in der *COCO-Wholebody* Annotation [Ji20] definiert sind. Um diese Gelenkpunkte zu erhalten, wurden für den gesamten Datensatz zweidimensionale *WholeBody* Gelenkpunkte mithilfe der vortrainierten *MMPose* [MM20] Modelle *Faster R-CNN R50* [Re15] zur Objekterkennung sowie *HRNet w48* [Su19] zur Posenschätzung aus den RGB-Videodaten extrahiert. Anschließend wurden Samples mit mangelhaften Posenschätzungen automatisiert aus dem Trainingssatz entfernt. Da die Entfernung der Probanden zu den Kameras im Datensatz zwischen 2 und 4.5 Metern beträgt [Li19], gewährleistet dies eine ausreichende Auflösung, um Gelenkpunkte der Extremitäten zu generieren. Die resultierenden Gelenkpunkte wurden auf räumlicher Ebene ähnlich des *PYSKL COCO Skeleton* miteinander verbunden, jedoch inklusive Daten für Hände und Füße. Der finale Graph wird im Folgenden als *PYSKL Extended Skeleton* bezeichnet und ist auf räumlicher Ebene in Abbildung Abb. 3 rechts dargestellt.

4.2 Training

Um die Funktionalität der *Dustbin*-Methode gegenüber im Training ungesehenen Klassen quantitativ evaluieren zu können, wurde zunächst ein *ST-GCN++* Modell in *PYSKL COCO Skeleton* sowie in *COCO Extended Skeleton* Topologie auf Datensatz B 100 Epochen trainiert³. Zusätzlich dazu wurden in beiden Topologien *ST-GCN++* Modelle auf Datensatz A trainiert⁴, mit dem Ziel einer höheren Generalisierung. Diese Modelle sollen später für Feldversuche genutzt werden. Deren Training bestand insgesamt aus 500 Epochen. Alle Modelle wurden “from scratch“ trainiert. In Anlehnung an die Implementierung von Duan et al. [Du22] diente als Grundlage der Validierung aller Trainingsdurchläufe der verwendete Testdatensatz C.

Im Training wurde *Stochastic Gradient Descent* mit *Nesterov Momentum* zur Optimierung genutzt sowie die Lernrate während des Trainings durch *Cosine Annealing* über alle Epochen angepasst, mit einer minimalen Lernrate von null. Als Verlustfunktion diente *Cross-Entropy Loss*. Da beide Varianten der Trainingsdatensätze ohne weitere Maßnahmen im Training zu einer Überrepräsentation der Samples der Restklasse gegenüber Samples der anderen trainierten Klassen führen würden, wurde im Dateneinzug eine automatisierte Balancierung implementiert. Diese berechnet mithilfe der normierten Häufigkeit der Datenproben im Datensatz eine Auswahlwahrscheinlichkeit pro Klasse. Anhand dieser Wahrscheinlichkeit werden pro Epoche Samples aus den jeweiligen Klassen gezogen. Da die Gesten der Restklasse als eine gemeinsame Klasse betrachtet werden, wird in diesem Fall zufällig aus Klasse null anhand der Auswahlwahrscheinlichkeit der Restklasse gezogen. Dabei wird nicht zwischen Gesten innerhalb der Restklasse unterschieden.

5 Numerische Resultate

Weiterhin werden die Ergebnisse des Experiments besprochen. Für die Evaluation aller Modelle wurden jeweils die Gewichte der Epoche mit der höchsten Top-1 Accuracy auf den Validierungsdaten genutzt. Für die Testdaten wird kein Sampling im Dateneinzug anhand der Klassenhäufigkeit verwendet.

5.1 Evaluation der mit Rückhaltung trainierten Modelle

Zunächst werden zur Bewertung der *Dustbin*-Methode die mit Rückhaltung trainierten Modelle untersucht. Die Ergebnisse sind in Tab. 2 gezeigt. Eine Top-1 Accuracy von 1.0 auf im Training ungesehenen Klassen, hier gekennzeichnet durch 0*, würde bedeuten, dass alle Samples dieser Menge korrekt der Restklasse 0 zugeordnet wurden.

³ Siehe Abschnitt 5.1 für die Evaluation.

⁴ Siehe Abschnitt 5.2 für die Evaluation.

Klassen IDs	PYSKL COCO Skeleton			COCO Extended Skeleton		
	0'	1-24	0*	0'	1-24	0*
Top-1 Accuracy	0.9252	0.8203	0.8263	0.9580	0.8635	0.8745

Tab. 2: *Cross-Subject* Evaluation der mit Rückhaltung trainierten Modelle auf dem Testdatensatz C. Gezeigt wird für beide Graphtopologien jeweils die gemittelte Top-1 Accuracy der explizit trainierten Klassen eins bis vierundzwanzig (1–24), im Training der Restklasse zugeordneten Klassen (0') sowie der im Training ungesehenen Klassen (gekennzeichnet durch 0*).

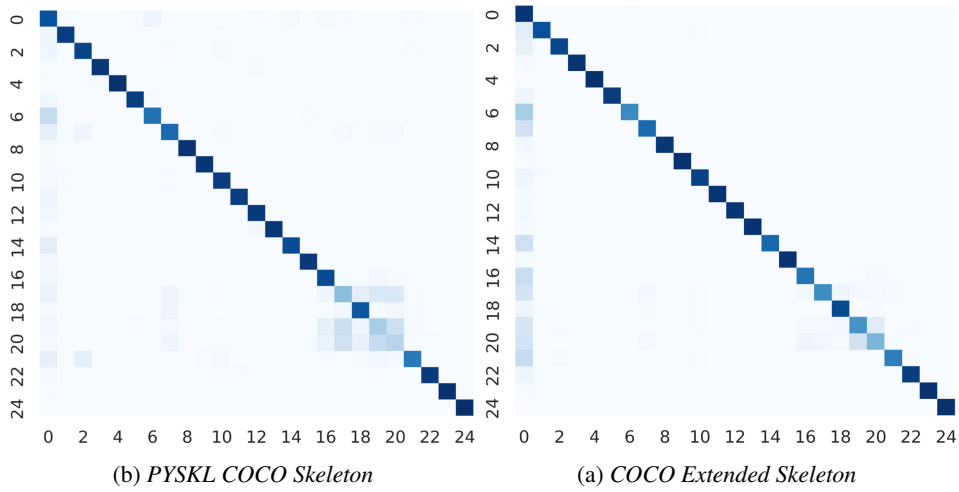


Abb. 4: Zeilenweise normalisierte Konfusionsmatrizen der mit Datensatz A trainierten *ST-GCN++* Modelle in beiden Topologien. Dunkelblau entspricht einem Wert von 1.0. Wahre Label IDs sind auf der y-Achse abgebildet.

Insgesamt ergaben die gemittelten Top-1 Accuracies über allen explizit trainierten Klassen; Samples der Menge 0* sowie den im Training zurückgehaltenen Klassen in der *COCO Extended Skeleton* Topologie höhere Werte als in der *PYSKL COCO Skeleton* Topologie. Ohne einen spezifischen Schwellwert der Erkennungswahrscheinlichkeit für die Restklasse gegenüber explizit trainierter Klassen erreichte die Restklasse in beiden Topologien eine höhere Top-1 Accuracy. In über 82 % der Fällen für die *PYSKL COCO Skeleton* Topologie und in über 87 % für die *COCO Extended Topologie* wurden im Training ungesehene Klassen korrekt der Restklasse zugeordnet.

5.2 Evaluation der ohne Rückhaltung trainierten Modelle

Die Ergebnisse der *Cross-Subject* Evaluierung der mit Datensatz B trainierten Modelle werden in Tab. 3 gezeigt. Die zugehörigen Konfusionsmatrizen sind in Abb. 4 dargestellt.

Klasse	Evaluationsergebnisse	
	ST-GCN++	ST-GCN++
	PYSKL COCO Skeleton	COCO Extended Skeleton
0 Restklasse	0.8698	0.9740
1 Brille abnehmen	0.9416	0.8768
2 Winken mit einer Hand	0.9234	0.9167
3 Hüpfen auf einem Bein	0.9673	0.9819
4 Aufspringen	0.9928	1.0000
5 Telefonieren	0.9455	0.9529
6 Handy/ Tablet bedienen	0.7455	0.6486
7 Mit Finger zeigen	0.7790	0.7790
8 Kopf nicken	0.9819	0.9709
9 Kopf schütteln	0.9782	0.9928
10 Salut	0.9457	0.9348
11 Arme kreuzen („Stopp“)	0.9420	0.9783
12 Taumeln	0.9601	0.9746
13 Fallen	0.9709	0.9855
14 Auf Person zeigen	0.8913	0.7862
15 Ball prellen	0.9632	0.9826
16 „Leise“-Zeichen	0.8970	0.7361
17 Daumen hoch	0.4278	0.6441
18 Daumen runter	0.8487	0.9045
19 „OK“-Zeichen	0.3496	0.6181
20 „Victory“-Zeichen	0.2974	0.4618
21 Faust schütteln	0.7153	0.6997
22 Beide Hände oben	0.9547	0.9461
23 Armkreisen	0.9913	0.9948
24 Armschwingen	0.9983	0.9948
Top-1 Accuracy	0.8598	0.9496
Top-5 Accuracy	0.9931	0.9946
Balanced Accuracy	0.8511	0.8694

Tab. 3: Ergebnisse der untersuchten Konfigurationen von *ST-GCN++* in *Cross-Subject* Evaluation auf Datensatz C. Die Modelle wurden auf Datensatz A trainiert. Die Accuracies einzelner Klassen werden für die jeweilige Konfiguration sowie die Top-1 Accuracy, Top-5 Accuracy und Balanced Accuracy über allen Klassen gezeigt. Der größte Wert pro Zeile ist jeweils hervorgehoben.

Insgesamt erreichten beide Modelle auf den Testdaten hohe Accuracies, die *COCO Extended Skeleton* Topologie sogar über allen Klassen eine durchschnittliche Top-1 Accuracy von über 94 %. Die Accuracies der Restklassen liegen in beiden Modellen über der gemittelten Top-1 Accuracy. Größere Konfusionen mit der Restklasse wurden beispielsweise in Klasse sechs beobachtet. Schwierigkeiten bei der Erkennung bereiteten insbesondere die Klassen 16 bis 20, deren Aktionen sich weitestgehend durch Hand- bzw. Fingerstellungen definieren. Wie Abb. 4 zeigt, entstanden in diesem Bereich besonders in der *PYSKL COCO Skeleton*

Topologie, welche keine Gelenkinformationen der Hände beinhaltet, einige Verwechslungen. Die Accuracies dieser Klassen fallen in *COCO Extended Skeleton* Topologie jedoch ebenfalls geringer aus als die gemittelte Top-1 Accuracy. Die besten Ergebnisse erzielten in beiden Modellen die Klassen vier, 22 sowie 23.

6 Diskussion

Insgesamt lieferte *ST-GCN++* als Modell zur Gestenerkennung vielversprechende Accuracies auf den Testdaten in allen untersuchten Konfigurationen. Wie aufgrund zusätzlicher Gelenkinformationen erwartet, konnte im *COCO Extended Skeleton* Layout gegenüber *PYSKL COCO Skeleton* die Erkennung von Handgesten verbessert werden. Hierzu gehörten die Aktionen *Daumen hoch*, *Daumen runter*, „OK“-Zeichen und „Victory“-Zeichen. Auch bezüglich der Genauigkeit der Restklasse übertraf *COCO Extended Skeleton* die Topologie *PYSKL COCO Skeleton* mit weniger Keypoints und lieferte damit insgesamt eine höhere Top-1 Accuracy. Mit einer Genauigkeit von über 99% bei beiden Topologien sind besonders die Gesten *Aufspringen*, *Armkreisen* und *Armschwingen* positiv hervorzuheben. Diese Aktionen zeichnen sich verglichen mit anderen Gesten durch einen hohen Bewegungsgrad aus, wobei sich die Bewegung zudem auf beide Arme erstreckt. Betrachtet man andere Klassen mit einem hohen Bewegungsgrad wie *Hüpfen auf einem Bein* oder *Ball prellen*, so erreichten diese ebenfalls vergleichsweise hohe Genauigkeiten. Demgegenüber erzielten Gesten geringerer Bewegungen wie *Handy/Tablet bedienen* oder *Mit Finger zeigen* geringere Werte. Die Gesten *Kopf nicken* und *Kopf schütteln* konnten in beiden Topologien mit mehr als 97% Accuracy größtenteils korrekt klassifiziert werden.

Bezüglich der *Dustbin*-Methode wurden über allen Konfigurationen im Test hohe Accuracies für die Restklasse in *Cross-Subject* Evaluation erreicht. Trotz einer vergleichsweise geringen Anzahl Trainingsepochen erzielten die Modelle mit Rückhaltung auf den im Training vorenthaltenen Aktionen über 82% Accuracy, in *COCO Extended Skeleton* Topologie sogar über 87%. Die Ergebnisse der Evaluation sprechen demnach für die hier untersuchte *Dustbin*-Methode zum Umgang mit ungesesehenen Aktionen. Ein weiteres Argument für diese Methodik ist deren einfache Implementierung, verglichen mit auf Wahlen basierenden Frameworks [Ro20]. Es besteht zudem die Möglichkeit, dass sich die Erkennung durch eine größere Auswahl an Klassen der Menge 0^* noch verbessern ließe. Hierzu könnten ergänzende Datensätze genutzt werden, um eine höhere Variation im Training der Restklasse zu erreichen. Wie Abb. 4 zeigt, bestehen bei einigen Klassen geringe Konfusionen gegenüber der Restklasse, wobei die Anpassung eines Schwellenwertes für die Erkennungswahrscheinlichkeit der Restklasse diese Unterscheidung möglicherweise noch verbessern könnte.

Sowohl der Einfluss des Abstandes von Personen zur Kamera als auch die Bewegung der Kamera wurden in dieser Arbeit nicht quantitativ evaluiert. Zudem ist zu beachten, dass sich die verwendeten Videosamples von den Lichtverhältnissen und dem Versuchsaufbau stark von Videodaten im Straßenverkehr unterscheiden, wo zudem kontinuierliche Videodaten



Abb. 5: Typische Probleme der Posenschätzungen bei qualitativen Versuchen im Straßenverkehr bei 3 Meter (a) sowie 5 Meter (b) Entfernung zur Kamera. Im Fall (a) konnte die Geste „Victory“-Zeichen im *COCO Extended Skeleton* nicht mit ausreichender Wahrscheinlichkeit zugeordnet werden. Trotz Fehler in der Posenschätzung wurde im Fall (b) die Geste *Arme kreuzen* („Stopp“) korrekt erkannt.

vorliegen. Der Mangel an öffentlich verfügbaren Datensätzen erschwert jedoch eine quantitative Evaluation. Aus diesem Grund wurden mit den mit Datensatz A trainierten Modellen qualitative Versuche auf kontinuierlichen Videodaten stationärer Kameras im Straßenverkehr durchgeführt. Probanden wurden dabei in unterschiedlichen Entfernungen gefilmt. Zu jedem Frame t wurden die Ergebnisse der Posenschätzung in einer Warteschlange der Größe⁵ $T = 59$ gespeichert und für jedes t die letzten T Frames für die Gestenerkennung von *ST-GCN++* verwendet. Die *Dustbin*-Methode konnte auch hier nicht explizit trainierte Gesten zuverlässig zuordnen. In den Ergebnissen wurde außerdem eine Abhängigkeit zwischen der Erkennbarkeit von Gesten auf größere Entfernungen und dem Grad der Bewegung des gesamten Körpers beobachtet. Besonders gut erkennbare Klassen auf eine Distanz bis zu 20 Metern waren unter anderem die Aktionen *Aufspringen* und *Hüpfen auf einem Bein*. Handgesten konnten in beiden Topologien selbst auf kurze Distanzen unter herrschenden Lichtverhältnissen nicht korrekt vorhergesagt werden, wobei bereits in der Gestenerkennung größere Fehler auftraten. Beispiele hierfür werden in Abb. 5 gezeigt. Dennoch schien der erweiterte Graph in der qualitativen Analyse Vorteile für die Gestenerkennung von Passanten zu bringen, insbesondere bei fehlenden Gelenkpunkten im nahen Sichtbereich. Eine Kombination beider Graphtopologien in verschiedenen Entfernungen wäre demnach zukünftig möglich.

7 Zusammenfassung und Ausblick

Unter dem Aspekt der Interaktion zwischen Passanten und automatisierten Fahrzeugen wurden in dieser Veröffentlichung gut erkennbare Aktionsklassen herausgestellt sowie Aussagen über die Vor- und Nachteile unterschiedlich detaillierter Graphtopologien aufgezeigt. Dabei wurden sowohl im Training als auch in der Evaluation Gegebenheiten einer *OSR*

⁵ Dies entspricht dem Median der explizit trainierten Klassen der Trainingsdatensätze.

berücksichtigt und ein verfügbarer Datensatz entsprechend dieser Anforderungen angepasst. Hierfür wurden zwei Methodiken vorgestellt und verfolgt.

Das Modell *ST-GCN++* lieferte sowohl auf den Testdaten als auch auf kontinuierlichen Videodaten bei Ganzkörperposen vielversprechende Ergebnisse. Die Erkennung von Handgesten konnte auf den Testdaten durch die Nutzung einer erweiterten Graphtopologie mit 65 Gelenkpunkten zwar verbessert werden, Ergebnisse auf Testdaten sowie qualitative Tests der untersuchten Modelle zeigen jedoch Probleme bei der Nutzung von Handgesten für den Anwendungsfall auf. Besonders gute Ergebnisse wurden bei Gesten erreicht, welche einen hohen Bewegungsgrad sowie eine Bewegung mehrerer Körperpartien beinhalteten. Hierzu gehörten die Gesten *Aufspringen*, *Armkreisen* und *Armschwingen*. Zur Lösung des *OSR* Problems wurde die *Dustbin*-Methode untersucht, welche auf im Training ungesehenen Aktionsklassen eine Accuracy von über 82% und im erweiterten Layout über 87% erreichte. In einem Training ohne Rückhaltung ungesehener Aktionsklassen der Restklasse erreichte die erweiterte Topologie auf trainierten Aktionsgesten eine Top-1 Accuracy von fast 95% sowie über 97% bezüglich der Restklasse.

Die Ergebnisse dieser Arbeit lassen auch auf zukünftige Herausforderungen schließen. Wie Untersuchungen von Brand; Schmitz [BM23] zeigen, variieren im Straßenverkehr genutzte Gesten bislang stark. Die vorliegende Arbeit belegt, dass auch die Erkennbarkeit einzelner Aktionen stark divergieren kann. Daher wird die Vermutung gestützt, dass Passanten zukünftig ein Aktionsalphabet zur Interaktion mit Verkehrsteilnehmern mit Automatisierungsfunktionen erlernen müssen. Um eine bestmögliche Selektion von Aktionsklassen zu gewährleisten, können zukünftige Forschungen an die hier präsentierten Ergebnisse anknüpfen.

Literatur

- [BIG18] Busto, P. P.; Iqbal, A.; Gall, J.: Open set domain adaptation for image and action recognition. *IEEE transactions on pattern analysis and machine intelligence* 42/2, S. 413–429, 2018.
- [BM23] Brand, T.; M., S.: Analysis of Pedestrian Gestures in a Virtual Traffic Environment, Paper presented at the 1st Würtual Reality XR Meeting, Würzburg, 11.-13.04.2023. 2023.
- [BYK21] Bao, W.; Yu, Q.; Kong, Y.: Evidential deep learning for open set action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. S. 13349–13358, 2021.
- [CVL11] Crowley-Koch, B. J.; Van Houten, R.; Lim, E.: Effects of pedestrian prompts on motorist yielding at crosswalks. *Journal of applied behavior analysis* 44/1, 2011.

- [Du22] Duan, H.; Wang, J.; Chen, K.; Lin, D.: PYSKL: Towards Good Practices for Skeleton Action Recognition. In: Proceedings of the 30th ACM International Conference on Multimedia. 2022.
- [Gi14] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
- [He20] He, J.; Zhang, C.; He, X.; Dong, R.: Visual Recognition of traffic police gestures with convolutional pose machine and handcrafted features. *Neurocomputing* 390/, 2020.
- [Ji20] Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P.: Whole-Body Human Pose Estimation in the Wild. In: Proceedings of the European Conference on Computer Vision (ECCV). 2020.
- [KW16] Kipf, T. N.; Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907/*, 2016.
- [Li14] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. L.: Microsoft COCO: Common Objects in Context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V* 13. Springer, 2014.
- [Li19] Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; Kot, A. C.: NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE transactions on pattern analysis and machine intelligence* 42/10, 2019.
- [MAA16] Mueid, R. M.; Ahmed, C.; Ahad, M. A. R.: Pedestrian activity classification using patterns of motion and histogram of oriented gradient. *Journal on Multimodal User Interfaces* 10/, 2016.
- [MB24] Mercedes-Benz Australia/Pacific Pty Ltd: The Mercedes-Benz F 015 Luxury in Motion, URL: <https://www.mercedes-benz.com.au/passengercars/campaigns/mercedes-benz-f-015.html>, Stand: 12.03.2024.
- [Mi21] Mishra, A.; Kim, J.; Cha, J.; Kim, D.; Kim, S.: Authorized Traffic Controller Hand Gesture Recognition for Situation-Aware Autonomous Driving. *Sensors* 21/23, 2021.
- [MM20] MMPose Contributors, 2020, URL: <https://github.com/open-mmlab/mmpose>, Stand: 25.04.2023.
- [Na19] Najamuddin, A.: The Meaning of Gesture in Social Cultural Context. *El-Tsaqafah: Jurnal Jurusan PBA* 18/1, 2019.
- [NM15] Nissan Motor Co., Ltd.: Nissan IDS Concept: Nissan’s vision for the future of EVs and autonomous driving, 2015, URL: <https://usa.nissannews.com/en-US/releases/nissan-ids-concept-nissan-s-vision-for-the-future-of-evs-and-autonomous-driving?selectedTabId=releases>, Stand: 12.03.2024.

-
- [PI21] Projekt INITIATIVE: Initiative Projekt – Intelligente Mensch-Technik-Kommunikation im gemischten Verkehr, 2021, URL: <https://www.initiative-projekt.de>, Stand: 21.03.2024.
 - [PIG21] Projekt INITIATIVE: Gesamtziel des Vorhabens, 2021, URL: <https://www.initiative-projekt.de/gesamtziel-des-vorhabens/>, Stand: 21.03.2024.
 - [PIS21] Projekt INITIATIVE: Szenarien Katalog, 2021, URL: <https://www.initiative-projekt.de/szenarien-katalog/>, Stand: 21.03.2024.
 - [Po19] Pop, D. O.; Rogozan, A.; Chatelain, C.; Nashashibi, F.; Bensrhair, A.: Multi-Task Deep Learning for Pedestrian Detection, Action Recognition and Time to Cross Prediction. *IEEE Access* 7/, 2019.
 - [Re15] Ren, S.; He, K.; Girshick, R.; Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in neural information processing systems* 28/, 2015.
 - [Ro20] Roitberg, A.; Ma, C.; Haurilet, M.; Stiefelhagen, R.: Open Set Driver Activity Recognition. In: 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020.
 - [SA24] Semcon AB: The Smiling Car, URL: <https://semcon.com/smilingcar/>, Stand: 12.03.2024.
 - [Sc12] Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; Boult, T. E.: Toward Open Set Recognition. In: Bd. 35. 7, IEEE, 2012.
 - [Sh18] Shu, Y.; Shi, Y.; Wang, Y.; Zou, Y.; Yuan, Q.; Tian, Y.: Odn: Opening the deep network for open-set action recognition. In: 2018 IEEE international conference on multimedia and expo (ICME). IEEE, S. 1–6, 2018.
 - [St18] Stanciu, S. C.; Eby, D. W.; Molnar, L. J.; St. Louis, R. M.; Zanier, N.; Kostyniuk, L. P.: Pedestrians/Bicyclists and Autonomous Vehicles: How Will They Communicate? *Transportation research record* 2672/22, 2018.
 - [St20] Stroud, J.; Ross, D.; Sun, C.; Deng, J.; Sukthankar, R.: D3D: Distilled 3D Networks for Video Action Recognition. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020.
 - [Šu14] Šucha, M.: Road users' strategies and communication: driver-pedestrian interaction. *Transport Research Arena (TRA)* 1/, 2014.
 - [Su17] Sun, L.; Jia, K.; Chen, K.; Yeung, D.-Y.; Shi, B. E.; Savarese, S.: Lattice Long Short-Term Memory for Human Action Recognition. In: *Proceedings of the IEEE international conference on computer vision*. 2017.
 - [Su19] Sun, K.; Xiao, B.; Liu, D.; Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
 - [Wa16] Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: *European conference on computer vision*. Springer, 2016.

- [Zh21] Zha, X.; Zhu, W.; Xun, L.; Yang, S.; Liu, J.: Shifted Chunk Transformer for Spatio-Temporal Representational Learning. *Advances in Neural Information Processing Systems* 34/, 2021.
- [ZL17] Zhang, X.; LeCun, Y.: Universum Prescription: Regularization using Unlabeled Data. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Bd. 31. 1, 2017.
- [ZW14] Zhuang, X.; Wu, C.: Pedestrian gestures increase driver yielding at uncontrolled mid-block road crossings. *Accident Analysis & Prevention* 70/, 2014.