reviewed paper

Von MS-COCO zur FGSV: Die Entwicklung und Validierung eines hierarchischen Klassifikationsschemas für die hochspezifische KI-gestützte Verkehrszählung in Deutschland

Jeroen Staab, Dorothee Stiller, Tobias Leichtle, Marlene Kühnl, Lena Huber, Stefan Gamperer, Klaus Martin, Robert Behringer, Helmuth Ammerl, Michael Wurm, Hannes Taubenböck

(Dr. Jeroen Staab, J.Staab Research (JSR), Bei der Schleifmühle 11, 85049 Ingolstadt, email@jstaab.de)
(Dr. Dorothee Stiller, German Aerospace Center (DLR), Münchener Straße 20, 82234 Wessling, dorothee.stiller@dlr.de)
(Dr. Tobias Leichtle, German Aerospace Center (DLR), Münchener Straße 20, 82234 Wessling, tobias.leichtle@dlr.de)
(Marlene Kühnl, Company for Remote Sensing and Environmental Research (SLU), Kohlsteiner Straße 5, 81243 Munich, marlene.kuehnl@slu-web.de)
(Lena Huber, OBERMEYER Infrastruktur GmbH & Co. KG, Hansastraße 40, 80686 München, lena.huber@obermeyer-group.com)
(Stefan Gamperer, Ingenieurbüro Behringer & Partner mdb, Luitpoldallee 32, 84453 Mühldorf a. Inn, stefan.gamperer@ib-behringer.de)
(Dr. Klaus Martin, Company for Remote Sensing and Environmental Research (SLU), Kohlsteiner Straße 5, 81243 Munich, klaus.martin@slu-web.de)
(Robert Behringer, Ingenieurbüro Behringer & Partner mdb, Luitpoldallee 32, 84453 Mühldorf a. Inn, r.behringer@ib-behringer.de)
(Helmuth Ammerl, OBERMEYER Infrastruktur GmbH & Co. KG, Hansastraße 40, 80686 München, helmuth.ammerl@obermeyer-group.com)
(Dr. Michael Wurm, German Aerospace Center (DLR), Münchener Straße 20, 82234 Wessling, michael.wurm@dlr.de)
(Prof. Dr. Hannes Taubenböck, German Aerospace Center (DLR), Münchener Straße 20, 82234 Wessling;Institute for Geography and Geology, Julius-Maximilians-Universität Würzburg, 97074 Würzburg, hannes.taubenboeck@dlr.de)

1 ABSTRACT

Die effektive und nachhaltige Steuerung urbaner Verkehrssysteme erfordert präzise Daten über eine zunehmend diverse Gruppe von Verkehrsteilnehmenden. Maschinelles Sehen (Computer Vision) bietet hierbei ein enormes Potenzial, um in der Stadt- und Verkehrsplanung die notwendigen hochfrequenten Daten für ein evidenzbasiertes Management bereitzustellen. Doch während viele KI-Lösungen überwiegend im wissenschaftlich-theoretischen Kontextentwickelt und erprobtwerden, bleibt deren praktische Anwendbarkeit häufig ungeprüft.

Anhand zweier Praxisbeispiele aus Süddeutschland beleuchten wir den interdisziplinären Raum zwischen modernen KI-Objektdetektoren einerseits, und den hiesig geltenden Standards zur Verkehrserhebung andererseits. Dies betrifft die Notwendigkeit einer hochgradig spezifizierten Klassifizierung von Verkehrsteilnehmenden gemäß den Richtlinien der Forschungsgesellschaft für Straßen- und Verkehrswesen (FGSV), Herausforderungen bei der Erfassung durch Wettereinflüsse, sowie strenge Datenschutzanforderungen. Die methodische Grundlage bildeteine manuell erhobene Referenzdatenbank mit über 17.000 annotierten Verkehrsteilnehmendenund vortrainierte YOLOv8-Modelle (You Only Look Once). Mit mehreren experimentellen Konfigurationen wurde untersucht, wie gut die Modelle den spezifischen Anforderungen der Planungspraxis gerecht werden. Im Speziellen wurde ein hierarchisches Klassifikationsschema entwickelt, das von internationalen Standards (MS-COCO, Common Objects in Context) über FGSV-Standards (Forschungsgesellschaft für Straßen- und Verkehrswesen) bis hin zu projektspezifischen Anforderungen reicht.

Die Untersuchung zeigt, dass die Detektionsgenauigkeit (mAP@50) stark von der semantischen Tiefe der Zielvariablen abhängig ist. Während auf der generischen Ebene mit lediglich 7 allgemein gehaltenen Klassen (bspw. Auto, Fahrrad, Person)eine sehr hohe Präzision von 86,3% erzielt wurde, sinkt dieser Wert bei der Anwendung der für die deutsche Planungspraxis notwendigen FGSV-Standards auf 78,0% und bei hochspezifischen Projektanforderungen (19 Klassen – von LastenfahrradfahrendePerson bis zu landwirtschaftlichem Spezialfahrzeug) auf 64,7%. Analog lässt sich auch der Einfluss von Bildqualität auf die Detektionsgenauigkeit quantifizieren. In der besten Konfiguration liegt die durchschnittliche Detektionsgenauigkeit bei 87,8%, wobei größere Objekte wie PKWs, LKWs und Busse mit bis zu 97,8% Genauigkeit detektiert wurden. Kleinere Objekte wie Personen, Motorräder und filigrane Fahrräder wurden jedoch nur in 86,8%, 83,2% und 67,5% Prozent der Fälle korrekt detektiert.

Im Vergleich zur manuellen Auswertung lassen sich mit den getesteten KI-Modellen deutlich größere Datenmengen automatisiert und mit robuster Genauigkeit verarbeiten. Damit werden sowohl großflächigere, als auch langfristigere Verkehrserhebungen möglich. Die erforderliche semantische Tiefe zwingt jedoch zur domänenspezifischen Optimierung. Gleichzeitig steigen mit der Automatisierung die Anforderungen an die Qualität der Datenerfassung allgemein. Valide Planungsgrundlagen hängen also zukünftig noch mehr von der interdisziplinären Schnittstelle von Technologie und Praxis ab.

Keywords: Verkehrszählung, Kamera, MS-COCO, FGSV, Verkehrsplanung

2 EINLEITUNG

Die zuverlässige Erfassung von Verkehrsströmen bildet eine zentrale Grundlage für Verkehrs-, Umwelt- und Stadtplanung (Matthias et al., 2020; Schulthoff et al., 2022). Über klassische, aggregierte Verkehrsvolumina hinaus, besteht dabei zunehmend der Bedarf, die Vielfalt heutiger Verkehrsteilnehmender – darunter Fuß- und Radverkehr, motorisierter Individualverkehr, Lieferverkehre sowie Sonderfahrzeuge – differenziert und in hoher zeitlicher Auflösung abzubilden (Stiller et al., 2026). Vor diesem Hintergrund haben sich in den letzten Jahren Verfahren des maschinellen Sehens als vielversprechender Ansatz etabliert. Mit Künstlicher Intelligenz (KI) ausgestattet, ermöglichen kamerabasierte Systeme eine hochfrequente, flächendeckende und prinzipiell automatisierbare Erfassung von Verkehrsteilnehmenden (Al-qaness et al., 2021). Insbesondere YOLO-basierte Objektdetektoren haben sich aufgrund ihrer Echtzeitfähigkeit in zahlreichen verkehrsnahen Anwendungen bewährt (Redmon et al., 2016; Zhang et al., 2022).

Gleichzeitig weisen mehrere Studien (Galich et al., 2025; Lin et al., 2014; Molchanov et al., 2017; Stiller et al., 2026; Xia et al., 2024) darauf hin, dass generische, vortrainierte Modelle unter realen Einsatzbedingungen häufig an ihre Grenzen stoßen, etwa bei kleinen oder teilverdeckten Objekten, variierenden Licht- und Wetterbedingungen sowie bei der Übertragung internationaler Objektklassen auf nationale, planungsrelevante Kategorisierungen, wie dieder Forschungsgesellschaft für Straßen- und Verkehrswesen (FGSV). Hinzu kommen datenschutzrechtliche Anforderungen, die insbesondere in Deutschland eine Reduktion der Bildauflösung oder Anonymisierungsmaßnahmen erforderlich machen und damit die visuelle Detailtiefe weiter einschränken (Stiller et al., 2026).

Um zu evaluieren, wie zuverlässig KI-Systeme unter diesen Rahmenbedingungen für Verkehrszählungen in Deutschland eingesetzt werden können, adressiert diese Studie die identifizierten Einschränkungen in einem systematischen, praxisnahen Untersuchungsdesign. Anhand zweier realer Anwendungsfälle aus Süddeutschland wird schrittweise analysiert, (i) wie leistungsfähig ein generisches, vortrainiertes YOLO Modell „Out-of-the-Box“ ist und welchen Einfluss unterschiedliche Aufnahmebedingungen wie Beleuchtung und Witterung auf die Erkennungsleistung haben; (ii) wie sich eine zunehmende semantische Differenzierung der Verkehrsteilnehmenden, von internationalen Standardklassen (MS COCO) über FGSV-nahe Kategorien bis hin zu hochspezifischen, praxisnahen Klassifikationen, auf die erreichbare Modellgüte auswirkt; sowie (iii) inwieweit ein für die Verkehrszählung optimiertes KI-Modell mit etablierten manuellen Zählungen unter realen Praxisbedingungen vergleichbar ist. Durch diese strukturierte Variation zentraler Einflussfaktoren leistet die Arbeit einen Beitrag zur Einordnung der technischen Möglichkeiten und Grenzen KI-gestützter Verkehrszählung und zeigt auf, unter welchen Bedingungen ein belastbarer Mehrwert für die Planungspraxis entsteht.

Im folgenden Kapitel wird das methodische Setup der Untersuchung beschrieben, einschließlich der zugrunde liegenden Praxisanforderungen, der Referenzdatenbasis sowie des experimentellen Designs. Anschließend werden die Ergebnisse der systematischen Evaluation gezeigtund im Hinblick auf die identifizierten Einflussfaktoren diskutiert, bevor abschließend die zentralen Schlussfolgerungen für den Einsatz KI-gestützter Verkehrszählung in der Planungspraxis gezogen werden.

3 METHODISCHER RAHMEN

Für eine praxisnahe Evaluation basiert die vorliegende Untersuchung auf zwei realen Verkehrsplanungsprojekten aus der Metropolregion München. Beide Projekte zielten auf die Quantifizierung von Verkehrsströmen und speziell der Differenzierung von Verkehrsteilnehmenden im Kontext infrastruktureller Planungs- und Bewertungsprozesse ab.

3.1 Hierarchisches Klassifikationsschema

Unterschiedliche Verkehrsteilnehmende stellen jeweils spezifische Anforderungen an Planung, Bemessung und Bewertung von Verkehrsinfrastruktur, da sie sich hinsichtlich Flächenbedarf, Betriebsverhalten, Sicherheitsanforderungen und Umweltwirkungen systematisch unterscheiden (Schulthoff et al., 2022).

Demgegenüber stehen praktische Einschränkungen der KI-gestützten Kameraauswertung. Zum einen müssen, grundsätzlich, für jede unterschiedene Klasse an Verkehrsteilnehmenden repräsentative Daten für

das Lernverfahren vorliegen. Zum anderen kann eine ungleiche Repräsentation einzelner Verkehrsteilnehmerim Datensatz jedoch zu systematischen statistischen Verzerrungen führen. Mit zunehmender Klassifikationsdifferenzierung steigt damit der Bedarf an umfangreichen und ausgewogenen Referenzdaten pro Klasse. Gleichzeitig nimmt die erreichbare Modellgüte – die Treffsicherheit – typischerweise ab. In diesem Spannungsfeld haben wir ein dreistufiges, hierarchisches Klassifikationsschema entwickelt (Abbildung 1). Der Vergleich der Modellgüte über die drei Ebenen zeigt, wie robust die Modelle gegenüber praxisnaher Klassifikationskomplexität sind.

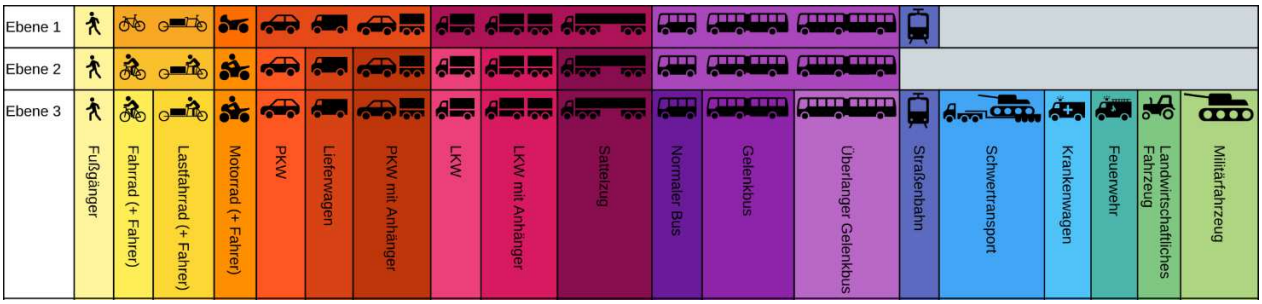

Abb. 1: Hierarchisches Klassifikationsschema. Über drei Ebenen verteilt werden Verkehrsteilnehmende zunehmend spezifischer differenziert. Farbliche Nuancen im Hintergrund unterstreichen die semantischen Unterschiede.

- Ebene 1: Der für internationale Computer-Vision Benchmarks gängige Datensatz MS-COCO (Lin et al., 2014) unterscheidet sieben für diesen Kontext relevante Objektklassen (z.B.person,bicycle,car,truck). Auf eine Übersetzung ins Deutsche wird bewusst verzichtet, um taxonomische Unterschiede nicht zu verlieren. Speziell die Klasse truckinkludiertsowohl schwere LKW, als auchleichtere Lieferwägen und Fahrzeuge mit offener Ladefläche (Pickup-Trucks).
- Ebene 2: Die in Deutschland für nationaleVerkehrsplanung relevanten Empfehlungen für Verkehrserhebungen (Forschungsgesellschaft für Straßen- und Verkehrswesen, 2012)schreiben dahingegen eine etwas feinere Differenzierung der Klassen vor. Hier werden motorisierte Fahrzeuge in achtthematische Grundklassen eingeteilt. Diese werden sowohl um Fahrradfahrende (im Folgenden Radler) und Fußgänger ergänzt. Insgesamt entspricht Ebene 2 damit zehn Objektklassen.
- Ebene 3:Diese Ebene erweitert Ebene 2 um zusätzliche, für individuelle Planungsfragestellungen relevante Unterkategorien. Dazu zählen unter anderem LKW mit und ohne Anhänger, Sattelzüge, Gelenk- und Standardbusse, ausgewählte Sonder- und Einsatzfahrzeuge, sowie Lastenradler. Insgesamt entspricht Ebene 3 damit neunzehn Klassen.

3.2 Praxisnahe Datengrundlage

Die Datenerhebung im Gelände erfolgte mittels temporär installierter Kameras an insgesamt 34 Standorten mit 51 einzelnen Erhebungen und einer Gesamtdauer von rund 1.200 Stunden Videomaterial. Die Kameras waren zwar in konstantenHöhen (5 Meter), jedoch in unterschiedlichen Abständen zur Zähllinie montiert, wodurch stark variierende Objektgrößen, perspektivische Verzerrungen und gegenseitige Verdeckungen auftraten. Alle Aufnahmen liegen in einer einheitlichen räumlichen Auflösung von480 × 720 Pixeln vor. Bis auf drei Ausnahmen wurden statt RGB- nur Graustufeninformationen aufgezeichnet.

Zusätzlich decken die Aufnahmen ein breites Spektrum an Umweltbedingungen ab, einschließlich wechselnder Beleuchtungssituationen, Witterungseinflüsse und tageszeitlicher Effekte. Diese Rahmenbedingungen stellen zentrale Herausforderungen für die automatisierte Objekterkennung dar, bilden jedoch zugleich die Realität verkehrsplanerischer Erhebungen ab und sind daher integraler Bestandteil des Untersuchungsdesigns. Für die weitere Analyse wurden die stündlich aufgeteilten Videosequenzen entsprechend in drei Beleuchtungskategorien unterschieden:sonnig (direkte Sonneneinstrahlung mit ausgeprägten Schattenwürfen), bedeckt (diffuse, gleichmäßige Ausleuchtung) unddunkel (Dämmerung und Nachtaufnahmen mit künstlicher Beleuchtung). Diese sind in Abbildung 2 beispielhaft dargestellt.

Abb. 2: Drei beispielhafte Aufnahmen illustrieren die unterschiedlichen Bildkompositionen an Standorten. Außerdem zeigen sie unterschiedliche Beleuchtungssituationen: a) sonnig, b) bedeckt und c) dunkel.

3.3 Annotation Referenzdatenbank

Um die Güte eines automatisierten Ansatzes evaluieren zu können, braucht es annotierte Referenzdaten. In Abwägung von Arbeitsumfang und technischen Anforderungen wurden aus der insgesamt verfügbaren Datengrundlage eine stratifizierte Stichprobe gezogen. Zunächst wurde die Grundgesamtheit an Videodaten auf komplementäre Verkehrssituationen reduziert und jeweils stundenweise Ausschnitte zu unterschiedlichen Aufnahmezeitpunkten und Beleuchtungskategorienpro Standort ausgewählt. Die so strukturierte Referenzdatenbank erlaubt eine differenzierte Analyse der kameragestützten Verkehrszählung unter variierenden Praxisbedingungen.

Um Redundanzen durch signalbedingt wartenden Verkehr und eine hohe Diversität der Trainingsdaten sicherzustellen, wurden aus den Videos in einem festen zeitlichen Intervall Einzelbilder (Frames) extrahiert. Gleichzeitig bestand der Anspruch, auch seltene, aber planerisch relevante Verkehrsteilnehmende ausreichend häufig abzubilden. In ausgewählten Szenen, die insbesondere Spezialfahrzeuge, LKW mit Anhänger oder Lastenradler zeigten, wurde die Abtastrate gezielt erhöht. Insgesamt umfasst die annotierte Referenzdatenbank 3.924 Einzelbilder.

In jedem der ausgewählten Einzelbilder müssen alle erkennbaren Verkehrsteilnehmenden einzeln markiert und einer semantischen Klasse zugeordnet werden. Die räumliche Lokalisierung der Objekte erfolgte mittels rechteckiger Bildausschnitte (Bounding Boxes), die den jeweiligen Verkehrsteilnehmenden eindeutig im Bild verorten. Die Annotation wurde in einem teilautomatischen, mehrstufigen Prozess durchgeführt, bei dem qualifizierte wissenschaftliche Hilfskräfte durch ein vortrainiertes Objekterkennungsmodell der YOLOv8-Familie unterstützt wurden. Die automatisch erzeugten Bounding Boxes wurden anschließend manuell überprüft, bei Bedarf korrigiert und um fehlende Objekte ergänzt. In diesem manuellen Arbeitsschritt erfolgte auch die Zuordnung zu den definierten semantischen Kategorien bis zur feinsten Klassifikationsebene 3.

Ein besonderer methodischer Aspekt ergibt sich außerdem bei Verkehrsteilnehmenden, die – wie bspw. Radler – visuell aus mehreren Objekten bestehen. Diese werden nur in Ebene 1 als separate Entitäten (z. B. person und bicycle) gehandhabt. Um diese semantische Mehrdeutigkeit aufzulösen, wurde das hierarchische Klassifikationsschema um eine logische Fusionsregel ergänzt. Während des Annotationsprozesses werden die Entitäten noch einzeln erfasst. Während der anschließenden Überführung wird geprüft, ob sich die Bounding Boxes von Person und Zweirad räumlich überlagern und die Entitäten für die Ebene 2 und 3 zusammengeführt.

4 EXPERIMENTELLES EVALUATIONSDESIGN

Ziel dieser Studie ist es, die Leistungsfähigkeit einer kameragestützten, vollautomatisierten Verkehrszählung unter dem Einfluss realer Faktoren aufzuzeigen. Die experimentelle Validierung folgt einem kontrollierten Design, bei dem jeweils ein Einflussfaktor variiert und alle übrigen Parameter konstant gehalten werden. Auf diese Weise werden der Einfluss der Aufnahmebedingungen, der semantischen Klassifikation und der Modellarchitektur getrennt untersucht. Die gewonnenen Erkenntnisse werden anschließend zusammengeführt, um eine konsistente Modellkonfiguration für den Vergleich mit der Planungspraxis abzuleiten.

4.1 Variable Aufnahmebedingungen

Optische Systeme sind in der Planungspraxis zwangsläufig mit wechselnden Licht- und Sichtverhältnissen konfrontiert, die sich der direkten Kontrolle entziehen und die visuelle Erkennbarkeit von Verkehrsteilnehmenden erheblich beeinflussen können (Staab et al., 2021; Xia et al., 2024; Zhang et al., 2022).Um in einem ersten Benchmark diesen Einfluss auf kameragestützte Ansätze isoliert zu analysieren, wird ein auf die semantische Ebene 1 vortrainiertesYOLOv8neingesetzt. Modellarchitektur und semantische Klassifikation werden also konstant gehalten, sodass beobachtete Leistungsunterschiede ausschließlich auf die Aufnahmebedingungen zurückzuführen sind. Als Stratum dienen die bereits in der Datengrundlage systematisch erfassten Beleuchtungskategorien (vgl. Abschnitt 3.2).

Zur Bewertung der Modellleistung werden True Positives (korrekt erkannte Objekte), False Positives (fälschlich erkannte Objekte) und False Negatives (nicht erkannte Objekte) betrachtet. Auf dieser Basis beschreibt Precision die Zuverlässigkeit der Detektionen, während Recall die Vollständigkeit der Erkennung quantifiziert. Der F1-Score kombiniert beide Maße als harmonisches Mittel und erlaubt so eine ausgewogene Bewertung der Modellleistung. Die Metrik hat sich insbesondere bei unausgeglichenen Klassenverteilungen als robuste Kenngröße etabliert und wird häufig zur Evaluation von Objekterkennungsmodellen verwendet (Powers, 2011; Sokolova & Lapalme, 2009).

Um die erzielten Werte einordnen zu können, wurde außerdem über alle Beleuchtungsszenarien hinweg ein allgemeines Benchmark berechnet.

4.2 Diversifikation der Klassifikation

Da die verschiedenen Verkehrsteilnehmenden jeweils individuelle Anforderungen an die Planung stellen, ist es gerade in heterogenen, urban geprägten Gebieten wichtig, diese differenziert zu berücksichtigen. Ziel des zweiten Experiments ist es, den Einfluss zunehmender semantischer Differenzierung auf die erreichbare Detektionsleistung zu quantifizieren. Hierbei wird über die drei Varianten des hierarchischen Klassifikationsschemas (Ebene 1–3) iteriert, während die anderen Parameter konstant gehalten werden: Als Datengrundlage werden ausschließlich ausreichend belichteteFrames (sonnig und bedeckt, insgesamt 3.845 Bilder) herangezogen und nach dieser Variable in 50% Trainings-, 30% Validierungs- und 20% Testdaten stratifiziert aufgeteilt. Die Versuche wurden je mit identischen Trainingsparametern über 100 Epochen und einer Batchgröße von 8 trainiert. Weitere Hyperparamter, wie die Lernrate, wurden nicht manuell angepasst und entsprechen damit den Standardeinstellungen der verwendeten YOLOv8-Implementierung. Maßnahmen zum expliziten Ausgleich von Klassenungleichgewichten (z.B. Oversampling oder Klassengewichte) wurden nicht eingesetzt, um die Effekte semantischer Differenzierung unter realitätsnahen Datenverteilungen abzubilden.

Die Modellgüte wird mittels mAP@50 bewertet. Diese mean Average Precision bei einer Intersection of Union von > 0,5 ist ein integriertes Maß dafür, wie gut die Verkehrsteilnehmende sowohl korrekt detektiert und klassifiziert, als auch im Bild lokalisiert werden. Die Auswertung erfolgte als Mittelwert über die jeweils betrachteten Klassen, sodass die Ergebnisse über die drei Ebenen direkt vergleichbar sind.

Ergänzend werden jeweils Konfusionsmatrizen ausgegeben. Sie stellendie manuelle Klassifikation der automatisierten Objektklasse gegenüber. Dadurch lässt sich nachvollziehen, zwischen welchen Verkehrsteilnehmenden systematische Verwechslungen auftreten.

4.3 Vergleich von Architekturen unterschiedlicher Komplexität

Technisch besteht die Möglichkeit, die Leistungsgrenzen der automatisierten Verkehrserkennung durch komplexere Modellarchitekturen zu verschieben. Im Fokus des dritten Experiments steht die Frage, ob der Einsatz größerer Netzwerkarchitekturen in Kombination mit längeren Trainingsphasen eine robustere Klassifikation ermöglicht. Zu diesem Zweck werden zwei Modellvarianten derselben Architektur- und Modellgeneration miteinander verglichen: das kompakte und für Echtzeitanwendungen optimierte Modell YOLO v8n mit rund 3 Millionen trainierbaren Parametern sowie das mit etwa 68 Millionen Parametern deutlich komplexere Modell YOLOv8x. Die Trainingsdauer wird architekturspezifisch gewählt, um eine stabile Konvergenz der Modelle zu gewährleisten. Datengrundlage und semantisches Klassifikationsschema (Ebene 1) werdenüber beide Trainingsläufe hinweg konstant gehalten.

Wie in den vorherigen Experimenten wird die Modellleistung mittels mAP@50 bewertet. Ergänzend werden klassenweise Leistungswerte berechnet, um architekturspezifische Effekte differenzierter analysieren zu können. Das Delta ebendieser Werte zeigt etwaige Leistungssteigerungen an.

4.4 Beispielhafte Anwendung

Zur exemplarischen Prüfung der praktischen Anwendbarkeit und Übertragbarkeit des entwickelten KI-Verfahrens wurde ein Vergleich mit einem etablierten manuellen Erhebungsverfahren durchgeführt. Zum Einsatz kommt das zuvor trainierte YOLOv8x-Modell. Als Referenz dient eine in der Planungspraxis übliche Verkehrszählung, bei der geschulte Mitarbeitende vier fünfzehnminütige Videos manuell auswerteten. Dabei werden alle Fahrzeuge im fließenden Verkehr gezählt.

Da in der vorliegenden Anwendung keine explizite Objektverfolgung eingesetzt wird, basiert die KI-gestützte Auswertung auf der Detektion von Verkehrsteilnehmenden in einzelnen Bildframes. Um Mehrfacherfassungen zu reduzieren und eine zeitlich konsistente Aggregation zu ermöglichen, werden aus den vier Videosequenzen in einem festen Intervall von zehn Sekunden Einzelbilder extrahiert. Die Detektionsergebnisse werden anschließend zeitlich aggregiert und auf dieselben Erhebungsintervalle bezogen wie die manuelle Zählung. Zur vergleichenden Bewertung beider Verfahren wird der zeitliche Verlauf der aggregierten Zählwerte analysiert und statistisch gegenübergestellt, wobei ein Korrelationskoeffizient als Maß für die Übereinstimmung der zeitlichen Dynamik herangezogen wird.

5 ERGEBNISSE

In diesem Kapitel werden die Ergebnisse der experimentellen Evaluation der kameragestützten Verkehrszählung dargestellt. Ausgangspunkt bilden die manuell geprüften Referenzdaten, die die empirische Grundlage für alle nachfolgenden Analysen darstellen. Auf dieser Basis werden zunächst Umfang und Struktur der annotierten Verkehrsteilnehmenden beschrieben, bevor die Ergebnisse der drei Experimente zur Robustheit gegenüber Aufnahmebedingungen, zur semantischen Differenzierung und zur Modellarchitektur vorgestellt werden.

5.1 Hochspezifische Verkehrszählung

Die Referenzdatenbank umfasst insgesamt 17.172 manuell geprüfte Annotationen (s. Tabelle 1). Die Klassenverteilung erinnert stark an reale Verkehrsverhältnisse und ist eben nicht gleichverteilt.Personenkraftwagen stellen mit 70,4 % den größten Anteil der Annotationen, gefolgt von Lieferwagen (10,4 %) sowie Fußgänger (10,2 %). Zusammengenommen entfallen damit rund 91 % aller Annotationen auf diese drei dominanten Klassen.Die verbleibenden Klassen weisen teils sehr geringe Fallzahlen auf. Die aus verkehrsplanerischer Sicht ebenfalls relevantenKrankenwägen, überlange Busse, Schwertransporte und Militärfahrzeuge konnten im verfügbaren Videomaterial nicht erfolgreich identifiziert werden und sind folglich nicht Bestandteil der Referenzdatenbank.

Semantische Klasse	#Annotationen	Semantische Klasse (Fortsetzung)	#Annotationen
PKW	12084	Motorrad	94
Fussgänger	1759	Gelenkbus	84
Lieferwagen	1780	Lastfahrrad	59
LKW	485	PKW mit Anhänger	47
Fahrrad	351	Landwirt. Fahrzeug	11
NormalerBus	196	Strassenbahn	5
LKWmitAnhänger	113	Feuerwehr	1
Sattelzug	103	… Übrige	0

Tab. 1: Verteilung der manuell geprüften Annotationen nach semantischen Klassen in der Referenzdatenbank

Die in der Praxis inhärente Ungleichverteilung an Objektkategorien (engl. class imbalance) ist methodisch besonders relevant, da sie die Modellleistung klassenabhängig beeinflusst und die Interpretation aggregierter Gütemaße erschwert (Japkowicz & Stephen, 2002). Im Kontext des hierentwickelten, hierarchischen Klassifikationsschemas ist dieser Effekt proportional zur semantischen Tiefe der jeweils betrachteten Ebene.

5.2 Benchmark Out-of-the-box-Modell

Um gleich zu Beginn der Studie die Einsatzfähigkeit eines frei verfügbaren, auf Ebene 1 vortrainierten Objekterkennungsmodells einordnen zu können, wurdenvier einfache Benchmarks durchgeführt.Zum einen im Vergleich zwischen allen Annotationen und über alle 3.924 Bilder und zum anderen eine Evaluation, die für die drei Beleuchtungskategorien sonnig, bedeckt und dunkelseparat durchgeführt wurde.

Über alle Beleuchtungsszenarien hinweg wurde ein gewichteter F1-Score von 0,48 erreicht.Dieser Wert ergibt sich aus einer Präzision von 0,79 und einer Sensitivität von 0,35.Das bedeutet, dass ein Großteil der vom Modell detektierten Objekte korrekt klassifiziert wurde, gleichzeitig jedoch nur rund 35 % der tatsächlich vorhandenen Objekte erkannt wurden, während etwa 65 % unentdeckt blieben.Die vergleichsweise hohe Precision weist darauf hin, dass das Modell zurückhaltend detektiert und Fehlklassifikationen weitgehend vermeidet, jedoch auf Kosten einer systematischen Untererfassung. Im Detail wird speziell die Klasse carmit einem F1-score von 0,54 überdurchschnittlich gut erkannt. Für andere Verkehrsteilnehmer, insbesondere person, bus und truck, fällt die Erkennungsleistung deutlich geringer aus (F1-Scores zwischen 0,17 und 0,26). Seltene Klassen wie bicycle, motorcycle und train wurden nahezu nicht erkannt.

Unter guten Lichtverhältnissen (sonnig) erreicht das Modell hingegen einen etwas höheren F1-Score von 0,51. Bei moderaten Beleuchtungsbedingungen (bedeckt) sinkt die Erkennungsleistung wieder auf 0,48. Deutlich stärker fällt der Leistungsabfall bei dunklen Aufnahmebedingungen aus, bei denen lediglich ein F1-Score von 0,30 erzielt wird.

5.3 Zielkonflikt semantischer Differenzierung

In der methodischen Praxis ist es üblich, Modelle auf ihre zukünftigen Anwendungen hin anzupassen. Dabei werden die vortrainierten Netzwerke auf die spezifischen Eingangsdaten und Aufgabenstellungen hin angepasst. Allerdingsverdeutlichen die Ergebnisse einen grundlegenden Zielkonflikt der kameragestützten Verkehrserkennung. Bei konstanter Modellarchitektur (YOLOv8n), identischer Datengrundlage und unveränderten Trainingsparametern sinkt die Modellgüte mit zunehmender semantischer Tiefe deutlich. Wie die Abweichungen entlang der Diagonalen anzeigen, nehmen Verwechslungen zu (Abbildung 3).

Für die generische Klassifikationsebene (Ebene 1 mit sieben Klassen) wird eine hohe Detektionsleistung mit einem mAP@50 von 0,863 erreicht. Die Konfusionsmatrix (Abbildung 3a) weist in diesem Fall eine stark ausgeprägte Diagonale auf, was auf eine robuste und weitgehend fehlerfreie Trennung der siebenVerkehrsteilnehmerkategorien hinweist. Fehlklassifikationen treten nur vereinzelt auf und betreffen überwiegend Randfälle.

Mit der Umstellung auf eine FGSV-nahe Klassifikation (Ebene 2, elf Klassen) reduziert sich die Detektionsleistung auf einen mAP@50 von 0,780. Parallel dazu verändert sich die Struktur der Fehlklassifikationen. Die Konfusionsmatrix (Abbildung 3b) zeigt vermehrt systematische Verwechslungen zwischen semantisch und visuell ähnlichen Klassen, etwa zwischen Fahrzeugen mit und ohne Anhänger oder zwischen unterschiedlichen Nutzfahrzeugtypen. Trotz dieses Leistungsrückgangs bleibt die Mehrzahl der Zuordnungen korrekt, was auf eine grundsätzlich stabile Modellleistung auch bei erhöhter semantischer Komplexität hinweist.

Die ausführlichste Semantik (Ebene 3) führt zu einem weiteren, deutlich ausgeprägten Leistungsabfall. Der mAP@50fällt auf 0,647. Die Konfusionsmatrix (Abbildung 3c) ist im Vergleich nochmals stärker aufgefächert. Im Detail ist erkennbar, dass die Fehlzuordnungen zwischen strukturell ähnlichen Klassen auftreten. Besonders deutlich ist das im Falle des einen Feuerwehrfahrzeuges, welches durch die KI fälschlicherweise als LKW detektiert wurde.

5.4 Auswirkungen der Modellarchitektur

Im Gelände soll eine kameragestützte Verkehrszählung möglichst alle Fahrzeuge erfassen. Vor diesem Hintergrund stellt sich die Frage, ob sich die beschriebenen Genauigkeitsverluste technisch durch den Einsatz leistungsfähigerer Modellarchitekturen abmildern lassen.

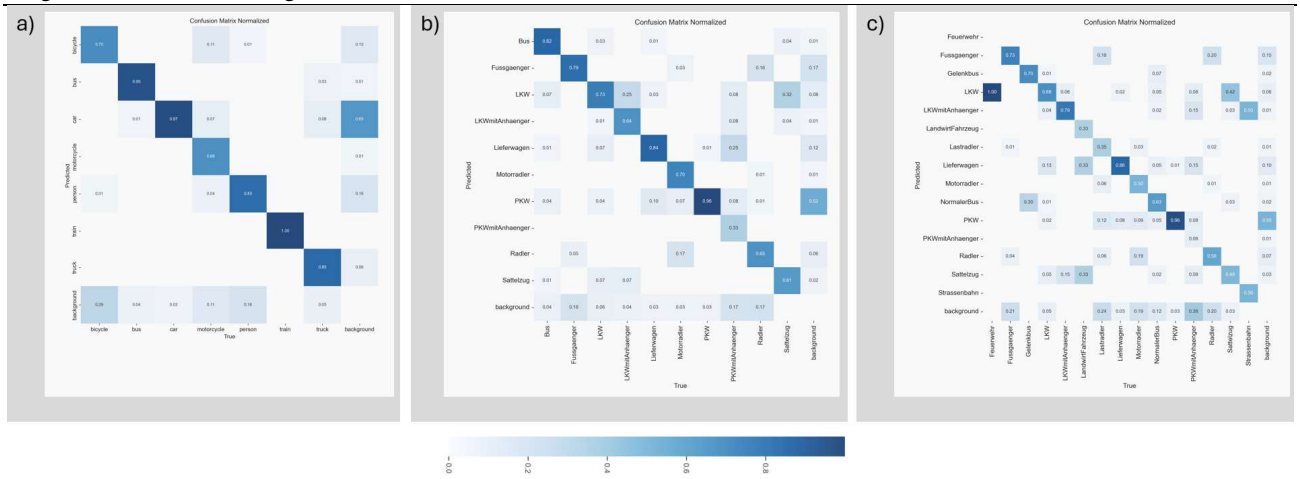

Abb. 3: Normalisierte Konfusionsmatrizen für drei semantische Klassifikationsebenen: a) generische Klassifikation (Ebene 1), b) FGSV-nahe Klassifikation (Ebene 2) und c) projektspezifische Klassifikation (Ebene 3). Die Diagonale zeigt korrekte Zuordnungen, Einträge außerhalb der Diagonale kennzeichnen systematische Fehlklassifikationen zwischen Klassen. Die Farbskala repräsentiert spalten-normalisierte Anteile im Wertebereich von 0 bis 1.

Im Vergleich zwischen YOLOv8n und YOLOv8x zeigt Tabelle 2, dass eine größere Modellarchitektur durchaus zu einer leichten, systematischen Verbesserung der aggregierten Detektionsleistung führt. Der mittlere mAP@50 steigt von 0,863 auf 0,878, was einem Zuwachs von +0,015 entspricht.Eine differenzierte Betrachtung der klassenweisen Ergebnisse zeigt, dass insbesondere kleinere oder strukturell komplexere Objekte wie motorcycle, bicycle und person von der höheren Modellkapazität profitieren, während für größere und visuell klar abgegrenzte Klassen wie bus, car und truck keine konsistenten Verbesserungen zu beobachten sind. In einzelnen Fällen fällt die Detektionsleistung des größeren Modells sogar leicht geringer aus.

Semantische Klasse	***mAP@50***		**Δ**
	YOLO v8n	***YOLO v8x***	
train	0.995	0.995	±0.000
car	0.983	0.978	−0.005
bus	0.943	0.928	−0.015
truck	0.912	0.870	−0.042
person	0.853	0.868	+0.015
motorcycle	0.713	0.832	+0.119
bicycle	0.643	0.675	+0.032
arithmetische Mittel	0.863	0.878	+0.015

Tab. 2: Klassenweise Detektionsleistung für zwei unterschiedlich komplexe Modellarchitekturen bei identischer Datengrundlage und generischer Klassifikation (Ebene 1). Angegeben sind die mittleren Genauigkeiten pro Klasse sowie die Differenz (v8x – v8n).

5.5 Vergleich zur Praxis

Das mit einer Erkennungsgenauigkeit von 87,8 % leistungsstärkste Modell wurde nachfolgend auf vier aufeinanderfolgende Videosequenzen angewendet. Unabhängig von Fahrtrichtung und Fahrzeugtyp ergab die manuelle Zählung insgesamt 200 Fahrzeuge, während die KI-basierte Auswertung 585 Detektionen lieferte. Trotz dieser deutlichen Abweichung auf absoluter Ebene zeigt der Korrelationskoeffizient von r = 0,82 eine hohe Übereinstimmung der relativen Verkehrsdynamik über die Zeit.

Der detaillierte Vergleich der Zeitreihen verdeutlicht, dass die Differenzen primär auf grundlegende konzeptionelle Unterschiede zwischen beiden Verfahren zurückzuführen sind. Ab etwa 5:00 Uhr morgens steigt die Anzahl der erfassten Fahrzeuge in beiden Ansätzen kontinuierlich bis 6:00 Uhr an. Während die Verläufe in den frühen Zeitintervallen noch nahezu parallel verlaufen, reißen die automatisierten Detektionen in der letzten Viertelstunde sprungartig nach oben aus. Dieser Effekt ist darauf zurückzuführen, dass die manuelle Zählung ausschließlich den fließenden Verkehr berücksichtigt, während die KI-basierte Methode auch ruhende sowie an Lichtsignalen wartende Verkehrsteilnehmende in konsekutiven Einzelbildern

mehrfach erfasst. Der beobachtete Unterschied reflektiert somit weniger eine Fehlleistung des Modells als vielmehr unterschiedliche, systemimanente Zähl- und Abgrenzungslogiken.

6 DISKUSSION

Die vorliegende Studie zeigt, dass kamerabasierte KI-Verfahren grundsätzlich in der Lage sind, Verkehrsdynamiken automatisiert und hochfrequent abzubilden. Die Studie zeigt aber auch, dass ihre Leistungsfähigkeit stark von Aufnahmebedingungen, semantischer Zielsetzung und Modellkonfiguration abhängt. Drei komplementäre Experimente verdeutlichen, dass weder generische Modelle noch hochkomplexe Architekturen per se schlüsselfertige Lösungen darstellen. Vielmehr entsteht Modellgüte aus dem Zusammenspiel von Datenqualität, semantischer Angemessenheit und technischer Auslegung. Diese Ergebnisse sind insofern erwartbar, als sie zentrale Annahmen der Computer-Vision-Literatur bestätigen (z.B. Benenson et al., 2014; Brunetti et al., 2018; Redmon et al., 2016), machen den Zielkonflikt jedoch erstmals explizit für verkehrsplanerische Anwendungen sichtbar.

Die Ergebnisse aus der händischen Annotation und des ersten Experiments bestätigen, dass Aufnahmebedingungen wie Kameraausrichtung, Witterung und Tagesgang einen maßgeblichen Einfluss auf die kameragestützte Verkehrszählung haben. Aber auch datenschutzbedingte Rahmenbedingungen limitieren mögliche Einsatzzwecke. Die in unserem Fall maximal reduzierte räumliche Auflösung und eingeschränkte Farbinformation führen zu einem geringen visuellen Signal, das selbst für annotierende Menschen eine eindeutige Klassifikation erschwert. Trotz relativ großer Stichprobe (Tabelle 1) konnte die maximal angestrebte semantische Tiefe (Ebene 3, vgl. Abbildung 1) in der Praxis nicht vollständig abgedeckt werden. Im Falle von militärischen und sonstigen Sonderfahrzeugen ist dies statistisch erwartbar. Im Falle von Krankenwägen jedoch wäre ebenfalls denkbar, dass diese und ihre Aufschriften schlicht nicht identifizierbar sind. Ebenso lässt sich die Länge eines Busses oder eines LKW-Gespanns weniger in der Frontalansicht, als erst im Laufe des Abbiegeprozesses im seitlichen Profil festhalten.

Außerdem stellt die in der Praxis übliche, stark unausgeglichene Klassenverteilung eine in den Computerwissenschaften bekannte Herausforderung für lernbasierte Objekterkennung dar (Lin et al., 2014; Molchanov et al., 2017). Ebendiese Effekte verstärkend offenbaren die Ergebnisse des zweiten Experiments den systematischen Zusammenhang zwischen semantischer Differenzierung und erreichbarer Modellgüte. Während auf generischen Ebenen hohe Detektionsleistungen erzielt werden, geht eine zunehmende semantische Auflösung mit einem kontinuierlichen Leistungsabfall einher. Dieser Befund ist konsistent mit der Literatur zu fein granularer Objekterkennung und tritt insbesondere dann auf, wenn visuelle Unterschiede zwischen Klassen gering sind (Stiller et al., 2026; Xia et al., 2024). Konkret machen die Konfusionsmatrizen (Abbildung 3) deutlich, dass es sich dabei nicht um zufällige Fehlklassifikationen handelt, sondernsich die Verwechslungen auf visuell und funktional benachbarte Klassenkonzentrieren. Vor diesem Hintergrund erscheint es sinnvoll, bestehende Empfehlungen zur Verkehrserhebung künftig nicht nur im Hinblick auf die zunehmende Diversifizierung der Verkehrsteilnehmenden weiterzuentwickeln, sondern auch deren visuelle und sensorische Trennbarkeit explizit zu berücksichtigen.

Die hohe zeitliche Korrelation der Zählverläufe in der beispielhaften Praxisanwendung zeigt, dass relative Verkehrsdynamiken zuverlässig abgebildet werden können. Mithilfe geeigneter Aggregations- und Kalibrierungsverfahren lassen sich KI-basierte Detektionen stochastisch in belastbare Verkehrskennzahlen überführen (Staab et al., 2021). Darüber hinaus erscheint die Integration objektbasierter Tracking-Ansätze als konsequenter nächster technologischer Schritt, um Mehrfachzählungen zu vermeiden und die Vergleichbarkeit mit etablierten Verfahren weiter zu erhöhen. Ein qualitativer Vergleich mit manuellen Verkehrszählungen verdeutlicht jedoch zugleich, dass vollautomatisierte KI-Lösungen nicht eins zu eins mit etablierten Erhebungsmethoden vergleichbar sind. Wenn menschliche Zählungen bewegte Fahrzeuge erfassen, erfolgt eine eindeutige Klassifizierung intuitiv und kontextabhängig.

7 AUSBLICK

Verkehrs-, Umwelt- und Klimaziele erfordern zunehmend hochfrequente, differenzierte und flächendeckende Datengrundlagen, die mit klassischen Erhebungsmethoden nur noch eingeschränkt wirtschaftlich bereitzustellen sind (Schulthoff et al., 2022). Die vorliegende Studie zeigt, dass KI-gestützte, kamerabasierte Verkehrszählungen unter geeigneten Rahmenbedingungen eine valide und skalierbare Ergänzung bestehender Verfahren darstellen können. In drei komplementären Experimenten wurden zentrale

Einflussfaktoren systematisch untersucht, darunter Aufnahmebedingungen, semantische Klassifikationstiefe und Modellarchitektur.

Die Untersuchung zeigt erstens, dass generische Out-of-the-Box Modelle unter günstigen Aufnahmebedingungen eine solide Basisleistung erzielen, jedoch bei ungünstigen Lichtverhältnissen deutlich an Robustheit verlieren. Zweitens, konnte ein klarer Zielkonflikt zwischen semantischer Differenzierung und erreichbarer Detektionsgenauigkeit identifiziert werden. Mit zunehmender semantischer Tiefe steigt zwar die planerische Aussagekraft der Klassifikation, gleichzeitig sinkt jedoch die automatisiert erreichbare Modellgüte. Hierarchische Klassifikationsschemata erweisen sich als geeigneter Ansatz, um diesen Zielkonflikt explizit zu adressieren und Ergebnisse auf unterschiedlichen Aggregationsebenen belastbar zu interpretieren. Drittens, zeigt der Praxisvergleich, dass KI-basierte Detektionen trotz konzeptioneller Unterschiede eine hohe Übereinstimmung der zeitlichen Verkehrsdynamik mit manuellen Zählungen aufweisen und prinzipiell in belastbare Verkehrskennzahlen überführt werden können. Gleichzeitig macht die kritische Diskussion deutlich, dass maschinelles Sehen als eigenständiges Instrument mit spezifischen Stärken verstanden werden muss. Sein Mehrwert liegt weniger in der punktgenauen Reproduktion etablierter Erhebungen als in der kontinuierlichen, automatisierten Erfassung relativer Verkehrsdynamiken über größere räumliche und zeitliche Skalen hinweg.

Die alleinige Steigerung der technologischen Kapazität (Moore, 1965) übertragen auf hiesig stetig wachsende Modellgrößen führt jedenfalls zu messbaren, jedoch immernoch begrenzten Leistungsgewinnen. Zukünftige KI-Generationen werden bestimmte Defizite –so die Erwartungshaltung–kompensieren können, die Notwendigkeit für hochwertigeDaten und realistische semantische Klassifikationsschemata wird allerdings bleiben.

8 DANKSAGUNG

Diese Arbeit entstand im Rahmen des Projekts „OptiPlan“, das durch das Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) unter den Förderkennzeichen 033LK007A und 033LK007B gefördert wurde. Wir danken der Stadt München für ihre Unterstützung sowie den studentischen Hilfskräften Marina Trintz und Georg Starz für ihre engagierte Mitarbeit bei der Datenannotation.

9 REFERENZEN

Al-qaness, M. A. A., Abbasi, A. A., Fan, H., Ibrahim, R. A., Alsamhi, S. H., & Hawbani, A. (2021). An improved YOLO-based road traffic monitoring system. Computing, 103 (2), 211–230. https://doi.org/10.1007/s00607-020-00869-8

Benenson, R., Omran, M., Hosang, J., & Schiele, B. (2014). Ten Years of Pedestrian Detection, What Have We Learned? Computer Vision – ECCV 2014 Workshops, Lecture Notes in Computer Science, 613–627. https://doi.org/10.1007/978-3-319-16181-5_47

Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing, 300, 17–33. https://doi.org/10.1016/j.neucom.2018.01.092

Forschungsgesellschaft für Straßen- und Verkehrswesen (Hrsg.). (2012). Recommendations for Traffic Surveys – Empfehlungen für Verkehrserhebungen (EVE) (Ausg. 2012). FGSV-Verl.

Galich, A., Stiller, D., Wurm, M., & Taubenböck, H. (2025). AI-Based Counting of Traffic Participants: An Explorative Study Using Public Webcams. Future Transportation, 5 (3), 87. https://doi.org/10.3390/futuretransp5030087

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study1. Intelligent Data Analysis, 6 (5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Hrsg.), Computer Vision – ECCV 2014 (S. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Matthias, V., Bieser, J., Mocanu, T., Pregger, T., Quante, M., Ramacher, M. O. P., Seum, S., & Winkler, C. (2020). Modelling road transport emissions in Germany – Current day situation and scenarios for 2040. Transportation Research Part D: Transport and Environment, 87, 102536. https://doi.org/10.1016/j.trd.2020.102536

Molchanov, V. V., Vishnyakov, B. V., Vizilter, Y. V., Vishnyakova, O. V., & Knyaz, V. A. (2017). Pedestrian detection in video surveillance using fully convolutional YOLO neural network (J. Beyerer & F. Puente León, Hrsg.). https://doi.org/10.1117/12.2270326

Moore, G. E. (1965). Cramming more components onto integrated circuits. Electronics, 38 (8), 114–117.

Powers, D.M.W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788. http://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html

Schulthoff, M., Kaltschmitt, M., Balzer, C., Wilbrand, K., & Pomrehn, M. (2022). European road transport policy assessment: A case study for Germany. Environmental Sciences Europe, 34 (1), 92. https://doi.org/10.1186/s12302-022-00663-7

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45 (4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Staab, J., Udas, E., Mayer, M., Taubenböck, H., & Job, H. (2021). Comparing established visitor monitoring approaches with triggered trail camera images and machine learning based computer vision. Journal of Outdoor Recreation and Tourism, 35, 100387. https://doi.org/10.1016/j.jort.2021.100387

Stiller, D., Wurm, M., Staab, J., Stark, T., Starz, G., Rauh, J., Dech, S., & Taubenböck, H. (2026). Open webcam data for traffic monitoring: YOLOv8 detection of road users before and during COVID-19. Transportation Research Interdisciplinary Perspectives, 36, 101774. https://doi.org/10.1016/j.trip.2025.101774

Xia, W., Li, P., Huang, H., Li, Q., Yang, T., & Li, Z. (2024). TTD-YOLO: A Real-Time Traffic Target Detection Algorithm Based on YOLOV5. IEEE Access, 12, 66419–66431. https://doi.org/10.1109/ACCESS.2024.3394693

Zhang, Y., Guo, Z., Wu, J., Tian, Y., Tang, H., & Guo, X. (2022). Real-Time Vehicle Detection Based on Improved YOLO v5. Sustainability, 14 (19), Article 19. https://doi.org/10.3390/su141912274