In der Rubrik **[FORSCHUNG.KOMPAKT]** stellen wir aktuelle Forschungsergebnisse übersichtlich und frei zugänglich dar, die eine wissenschaftliche Begutachtung (»Peer Review«) durchlaufen haben. Die Originalfassung dieses Artikels wurde veröffentlicht unter **Sielemann, A., Loercher, L., Schumacher, M., Wolf, S., Roschani, M., Ziehn, J., and Beyerer, J. (2024). Synset Signset Germany: A Synthetic Dataset for German Traffic Sign Recognition. In 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC).**, und kann unter der nebenstehenden Adresse abgerufen werden.

https://synset.de/datasets/synset-signset-ger/

[FORSCHUNG.KOMPAKT]

Anne Sielemann, Stefan Wolf, Masoud Roschani, Jens Ziehn

Synthetische Daten, die die Realität ersetzen können ↗

Daten sind der Rohstoff, aus dem Verfahren der künstlichen Intelligenz (KI) erzeugt werden. Über sie werden mit maschinellem Lernen KI-Modelle trainiert, die unterschiedlichste Aufgaben lösen – zum Beispiel die Erkennung von Objekten in Bildern oder die Steuerung von Fahrzeugen. Aber sie sind gleichzeitig ein knapper Rohstoff: Die Erhebung und Aufbereitung von realen Daten ist teuer, langwierig und aufwendig, insbesondere die Annotation der zugehörigen korrekten KI-Ausgabe. Eine Lösung sollen synthetische Daten darstellen, also Daten, die künstlich erzeugt sind, in der Regel über Simulation oder selbst über KI-Modelle. Doch für viele Anwendungen müssen sie zunächst beweisen, dass sie realistisch genug sind, um Realdaten ersetzen zu können. Wie das funktionieren kann, zeigen wir an zwei Beispielen.

Die Bedeutung künstlich erzeugter Daten hat in den letzten Jahren erheblich an Bedeutung gewonnen, aus vielfältigen Gründen. Moderne Verfahren des maschinellen Lernens (ML) und der künstlichen Intelligenz (KI) erfordern immer größere Umfänge an hochwertigen Daten, oft Rohdaten (beispielsweise Bilder) einschließlich sogenannter Annotationen (beispielsweise die im Bild sichtbaren Objekte, ihre Positionen und Umrisse, und Ähnliches).

Für diese fordert der europäische »AI-Act« (dt.: »Verordnung über künstliche Intelligenz«, Europäische Union, 2024), dass »Trainings-, Validierungs- und Testdatensätze, einschließlich [Annotationen], […] relevant, hinreichend repräsentativ und so weit wie möglich fehlerfrei und vollständig sein« sollen.

Dieser Anspruch ist in der Praxis oft nur schwer zu erfüllen: Je komplexer die KI-Anwendung, umso mehr Daten sind erforderlich, und umso schwieriger ist es, Annotationen bereitzustellen, die in einem vertretbaren Sinne »fehlerfrei« sind. Ein oft zitiertes Beispiel ist der »Cityscapes«-Datensatz, einer der ersten Datensätze für automatisiertes Fahren, mit denen KI trainiert wurde, in Kamerabildern Fahrzeuge, Fußgänger und befahrbare Straßen zu erkennen (s. Cordts u. a., 2016).

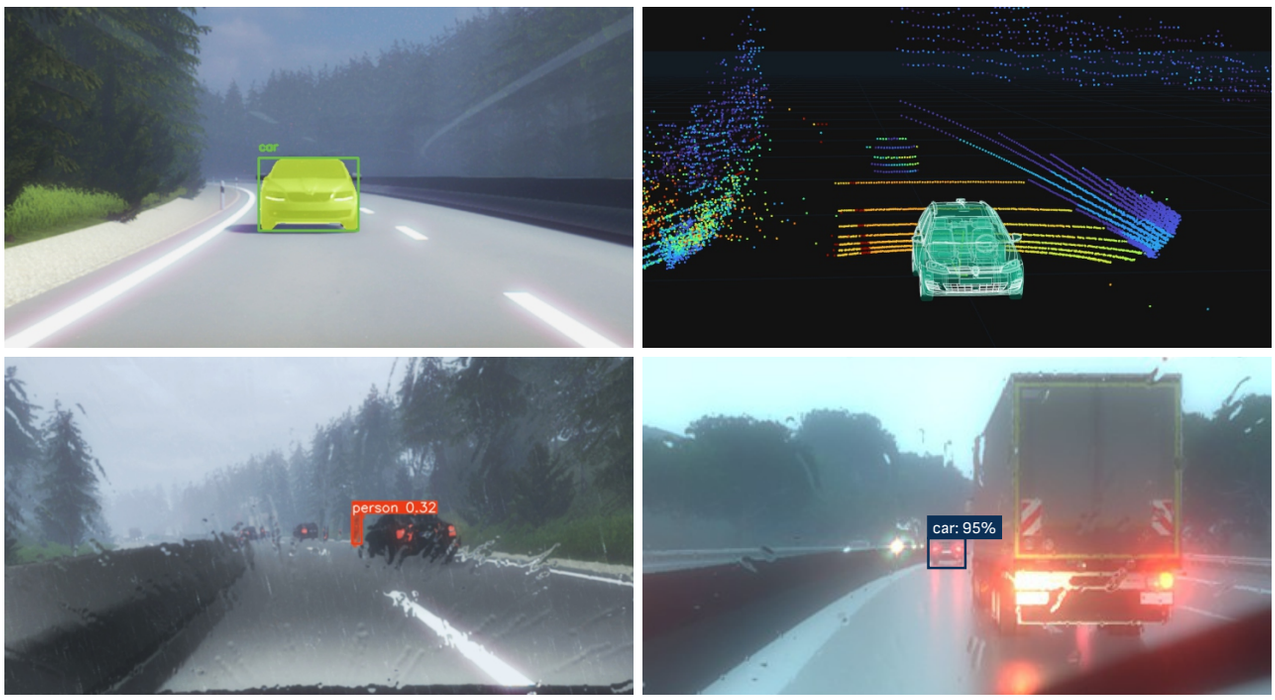

Abb. 1: Synthetische Daten unterschiedlicher Sensoren, wie Kamera und Laserscanner, ermöglichen die Analyse von KI unter schwierigen Bedingungen wie Blendungen oder Regen. (Simuliert in OCTAS®, Abb. oben aus Luettner u. a., 2025).

Der Cityscapes-Datensatz enthält 5 000 Kamerabilder, in denen pixelgenau per Hand annotiert wurde, welche Objektklasse darin zu sehen ist. Die Autoren geben den Annotationsaufwand je Bild, einschließlich Qualitätsprüfung, mit anderthalb Stunden an. Während diese Maßstäbe sicherlich nicht mehr repräsentativ sind, unterstreichen sie doch, dass die Aufwände hochwertiger Annotationen ein erheblicher Faktor der KI-Entwicklung sind.

Auf der Suche nach Sicherheitsgrenzen von KI

Das gilt insbesondere für Daten, die nicht zum *Training* einer KI-Funktion dienen sollen, sondern für den *Test* ihrer korrekten und robusten Funktion: Während viele KI-Funktionen, insbesondere sogenannte »Foundation Models«, die auf immensen Datenmengen trainiert wurden, bereits heute eine sehr hohe Leistungsfähigkeit erbringen, ist es sehr viel aufwendiger, nachzuweisen, dass keine kritischen Schwächen oder Restrisiken verbleiben.

Beispielsweise definiert die Sicherheitsnorm ISO 26262 (»Straßenfahrzeuge – Funktionale Sicherheit«) sogenannte »Automotive Safety Integrity Level« (ASIL). Diese stellen benötigte Sicherheitsniveaus für unterschiedliche Fahrzeugkomponenten dar. Auf dem höchsten Sicherheitsniveau ASIL-D, auf dem beispielsweise die Fehlauslösung von Airbags oder Lenkeingriffen während der Fahrt abgesichert werden, muss nachgewiesen werden, dass höchstens ein Fehler in 10^8 Betriebsstunden auftreten kann. Übertragen auf komplexe KI-Funktionen bedeutet das, dass auch sehr seltene Szenarien zuverlässig abgedeckt sein müssen.

Eine oft zitierte Beispielrechnung von 2015 (Wachenfeld und Winner, 2015) kommt zu dem Ergebnis, dass »6,62 Milliarden Testkilometer auf der Autobahn absolviert werden [müssten]«, um statistisch nachzuweisen, dass ein Autobahnpilot sicherer fährt als der Mensch. Auch wenn diese Rechenbeispiele schon von den Autoren selbst als theoretisch bezeichnet wurden, und mittlerweile über zehn Jahre alt sind, besteht das Grundproblem weiter: Je leistungsfähiger die KI, umso seltener treten Fehler auf, und umso schwieriger sind sie zu finden. Auch entstehen herausfordernde Situationen für KI meist nicht durch einzelne Faktoren (beispielsweise grüne Kleidung), sondern durch eine komplexe Kombination (Kleidung, Hintergrund, Umgebungslicht, Bewegungen).

→ **Kontakt**

Anne Sielemann
Fraunhofer IOSB

Fraunhoferstr. 1
76131 Karlsruhe

Mail: anne.sielemann
@iosb.fraunhofer.de

Web: synset.de

Solche Szenarien in der Realität zu finden, bedeutet erhebliche Aufwände und potenziell auch erhebliche Risiken, denn sie treten oft in Szenarien und Umgebungsbedingungen ein, die auch für menschliche Sicherheitsfahrer schwierig zu beurteilen und beherrschen sind. Zudem bleibt herausfordernd, zu argumentieren, welche Fälle erfolgreich abgedeckt sind, und wo eventuell »blinde Flecken« verbleiben.

Auf der ISO 26262 aufbauend beschreibt deshalb die ISO 21448 (»Straßenfahrzeuge — Sicherheit der Sollfunktion«, s. Schnieder und Hosse, 2019) die Herausforderung sogenannter »unbekannter unsicherer Szenarien«, also Fälle, in denen die KI ein falsches Ergebnis liefert, ohne dass es Entwicklern oder Zulassungsstellen bekannt wäre. Hier verbergen sich erhebliche Risiken für den Betrieb sicherheitskritischer KI-Funktionen.

Künstliche Daten

Hier haben Simulationen ein erhebliches Potenzial: Mit ihrer Hilfe lassen sich im virtuellen Raum »künstliche Daten« erzeugen, die gezielt definierten Parametern entsprechen. Diese Daten können in großer Zahl erzeugt werden, ohne Erhebungen im Realverkehr erforderlich zu machen. Insbesondere für seltene und risikoreiche Szenarien lassen sich auf diesem Weg große Anzahlen mit geringem Aufwand erzeugen. Noch größere Effizienzvorteile haben synthetische Daten in Bezug auf Annotationen. Während in realen Daten erst mit erheblichem, oft manuellem Aufwand sichtbare Objekte und weitere Informationen annotiert werden müssen, erlauben Simulationen, automatisch und meist mit vernachlässigbarem Mehraufwand entsprechende Informationen auszuleiten.

Darüber hinaus ermöglichen Simulationen, gezielt Grenzfälle abzudecken, wie überbelichtete oder verwackelte Bilder, Regen und Nebel, Verschmutzungen, oder Sensoreigenschaften wie Linsenverzeichnungen oder Rauschen, und auf diese Weise technische Parameter zu quantifizieren, ab denen die benötigte Leistung der KI-Modelle unter kritische Grenzen fällt. Durch die Verbindung mit detaillierten und praktisch fehlerfreien Annotationen lassen sich nicht nur Fehler der KI messen, sondern bereits frühere Auffälligkeiten in ihrer Leistung erkennen, wie ungenauer erkannte Objektumrisse oder Änderungen in den für die Klassifikation betrachteten Bildbereichen.

→ **Wie trainiert und testet man KI richtig?**
KI-Verfahren — konkreter: moderne Verfahren des maschinellen Lernens — nutzen Daten um den größten Teil ihrer Funktion zu »erlernen«. Nur ein kleiner Teil ihrer Fähigkeiten entsteht durch gezielt entwickelte Funktionen, das Allermeiste wird über Daten definiert.

Im sogenannten **»überwachten Lernen«** umfassen Trainingsdaten Datenpunkte, die jeweils aus zwei Teilen bestehen: Einer Eingabe (bspw. das Bild eines Verkehrsschilds) und einer zugehörigen korrekten Ausgabe (bspw. die Klasse »Tempo 30«). Im Training bekommt die KI die Eingabe präsentiert und muss eine Ausgabe erzeugen. Weicht diese von der korrekten Ausgabe ab, bekommt sie »Feedback«, um ihr gelerntes Modell zu verbessern.

Ein häufiges Risiko ist, dass das KI-Modell im Training kein übertragbares Modell erlernt, sondern lediglich die Trainingseingaben und die zugehörigen korrekten Ausgaben auswendig lernt. Man spricht hier von **»Overfitting«**. Unter anderem deshalb testet man das Trainingsergebnis in der Praxis quasi niemals auf den Trainingsdaten, sondern auf einem separaten Testdatensatz. Dieser sollte aus derselben Quelle stammen wie der Trainingsdatensatz, aber keinen übereinstimmenden Datenpunkt enthalten. Meist entstehen diese Datensätze, indem man einen großen Datensatz erzeugt, und ihn dann (zum Beispiel im Verhältnis 70% : 30%) in Trainings- und Testdaten aufteilt. Man spricht deswegen auch von **Trainings- und Test-Splits**.

Gute Trainings- und Testdaten müssen **»i.i.d.«** sein: Independent and identically distributed — unabhängige Datensätze, die aber derselben stochastischen Verteilung folgen müssen. Auf diese Weise sind die »Prüfungsfragen« gleichschwer wie die »Übungsfragen«, aber Auswendiglernen hilft nicht weiter. Sind die Trainings- und Testdaten tatsächlich »i.i.d.«, dann ist ein KI-Verfahren genau dann gut auf dem Testdatensatz, wenn es im Training tatsächlich das Prinzip korrekt gelernt und »verstanden« hat.

Entsprechend gilt für den Beitrag: Die Formulierung »Wir trainieren und testen auf demselben Datensatz« meint, dass jeweils die entsprechenden Trainings- und Testsplits genutzt wurden. Oft verfügt ein Datensatz bereits über einen empfohlenen Split. Buchstäblich auf denselben Datenpunkten zu trainieren und zu testen ist im maschinellen Lernen fast immer nur eins: tabu.

Direktvergleich mit Realdatensätzen

Um Eignung künstlicher Daten für Training und Test von KI zu bewerten, muss insbesondere quantitativ bewertet werden, inwieweit diese Daten in der Lage sind, Realdaten zu ersetzen. Da ein solcher Vergleich in umfangreichen Anwendungsfällen komplex ist, wurden KI-Anwendungen in der Mobilität herangezogen, die einen relativ klar abgegrenzten Fokus haben, und es ermöglichen, »synthetische Zwillinge« von realen Datensätzen zu erzeugen, die in statis-

Abb. 2: Vergleich von Bildern des realen CompCars-Datensatzes (oben) mit ähnlichen Bildern des synthetischen Synset®-Boulevard-Datensatzes

Abb. 3: Beispielbilder einschließlich Annotationsbild im Synset®-Boulevard-Datensatz.

tischen wesentlichen Eigenschaften übereinstimmen sollen, ausdrücklich ohne zu versuchen, konkrete Einzelbilder zu reproduzieren. Ziel ist zu prüfen, ob sich ein konkreter realer Datensatz durch ein künstliches Erzeugungsprinzip ersetzen lässt, aus dem sich prinzipiell beliebig viele Daten zufällig erzeugen lassen.

Diese Beispielanwendungen sind »Vehicle Make and Model Recognition« (VMMR) einerseits, und Verkehrszeichenerkennung andererseits. Die Zielstellung bei VMMR ist es, basierend auf einem Bild eines Fahrzeugs auf dessen Marke und Modell zu schließen, sowie potenziell weitere Informationen wie Baujahr oder Farbe. Der Anwendungsfall ist verbreitet in Verkehrsanalyse und polizeilicher Fahndung, gleichzeitig gibt es nur eine sehr begrenzte Anzahl an öffentlichen Datensätzen. Die Zielstellung bei Verkehrszeichenerkennung ist, wie der Name vermuten lässt, in einem Bildausschnitt, der ein Verkehrszeichen zeigt, die Bedeutung des Verkehrszeichens zu klassifizieren. In beiden Anwendungen ist die Annahme, dass es eine vorab bekannte (oft: sehr große) Anzahl entsprechender »Klassen« gibt, die (wieder-)erkannt werden muss. Trainings- und Testdaten umfassen daher Beispielbilder, geordnet nach den sichtbaren Objektklassen.

Vehicle Make and Model Recognition (VMMR)

Für den Anwendungsfall VMMR wird der CompCars-Datensatz von Yang, Luo u. a. (2015)[1] als Ausgangspunkt genutzt, da es sich hierbei um den größten öffentlich verfügbaren Datensatz handelt, der Fahrzeuge in frontaler Verkehrskamera-Perspektive zeigt. Der basierend darauf erzeugte synthetische Zwilling wurde unter dem Namen »Synset® Boulevard« publiziert in Sielemann, Wolf u. a. (2024)[2], jeweils in Abb. 2.

Synset® Boulevard wurde in der Simulation OCTAS®[3] erzeugt. Dazu wurden kommerzielle 3D-Modelle von Fahrzeugen aufbereitet, um physikalisch-basiertes Rendering zu ermöglichen und insbesondere die Scheinwerferfunktion abzubilden. Diese Fahrzeuge wurden auf einer vereinfachten mehrspurigen Straße platziert, deren Texturen mit Fahrbahnmarkierungen prozedural zufällig erzeugt wurden. Die Fahrzeuge wurden dann mit zufällig gezogenen Lichtbedingungen, Positionen und Scheinwerferzuständen aus ebenfalls zufällig gezogenen Kamerapositionen abgebildet, wobei Bilder mit dem Cycles-Renderer[4] berechnet wurden. Anschließend wurden Kameraeffekte wie Linsenreflexe und Überstrahlungen simuliert. Zudem liegt jedem Bild ein Annotationsbild bei (s. Abb. 3), das den Umriss des Fahrzeugs im Bild sowie die Lage und den Typ von sichtbaren Fahrbahnmarkierungen angibt.

Der reale CompCars-Datensatz (genauer gesagt dessen Teildatensatz für Verkehrskameras) umfasst 50 000 Bilder über 281 unterschiedliche Fahrzeugmarken, -modelle und -baujahre, die bis 2014 in China erhoben wurden. Die Anzahl der Fahrzeuge je Klasse variiert, es sind ferner eine Reihe an bekannten menschlichen Fehlern in der Annotation enthalten (vgl. Sánchez, Parra u. a. (2021)); ferner ist zu berücksichtigen, dass asiatische Fahrzeuge überwiegen, die in Europa mitunter seltener oder in optisch leicht anderen Varianten anzutreffen sind.

Der synthetische Synset®-Boulevard-Datensatz umfasst 32 400 Bilder (jeweils noch einmal für vier unterschiedliche Kamerasimulationen) über 162

1 mmlab.ie.cuhk.edu.hk/datasets/comp_cars/

2 synset.de/datasets/synset-blvd/
3 octas.org
4 cycles-renderer.org

Tab. 1: Die wichtigsten öffentlich verfügbaren Datensätze für VMMR in Verkehrskameras.

Datensatz und Publikation	Anzahl Bilder	Anzahl Klassen	real/synth.	Perspektive
CompCars, Yang, Luo u. a. (2015)	50 000	281	real	frontal
BoxCars21k, Sochor, Herout und Havel (2016)	63 750	148	real	gemischt
BoxCars116k, Sochor, Špaňhel und Herout (2018)	116 286	693	real	gemischt
Synset® Boulevard, Sielemann, Wolf u. a. (2024)	32 400	162	synth.	frontal

Tab. 2: Untersuchung von Trainingsvorgängen eines ConvNeXt-Small-Modells, trainiert auf Synset® Boulevard und getestet auf dem realen CompCars-Datensatz, zunächst für alle übereinstimmenden Modelle, dann eingegrenzt auf die Teile, in denen das Modelljahr (»Jhr.«) oder sogar der genaue Facelift (»Fcl.«) übereinstimmen. Bei maximaler Übereinstimmung erreichen sowohl synthetische wie auch echte Trainingsdaten ein Testergebnis von 100%.

Jhr.	Fcl.	F_1 SS Blvd.	F_1 CompCars	Anz. Modelle
✗	✗	50,3	97,4	51
✓	✗	93,5	100,0	21
✓	✓	100,0	100,0	12

Klassen, die jeweils durchgängig in 200 Bildern abgedeckt sind. Synset® Boulevard fokussiert darauf, insbesondere auch aktuellere und europäische Marken (neuestes Baujahr 2022) abzudecken. Zudem beinhaltet Synset® Boulevard eine stärkere Variation von Lichtverhältnissen und Kamerawinkeln als der CompCars-Datensatz.

Die Datensätze unterscheiden sich ferner dadurch, dass CompCars in derselben Klasse zum Teil unterschiedliche Baujahre und Facelifts kombiniert, wohingegen Synset® Boulevard innerhalb einer Klasse strikt eine einzige Fahrzeugvariante enthält.

Um zu messen, wie gut die synthetischen Daten reale Daten ersetzen können, wurden neuronale Netze auf den synthetischen Synset®-Daten trainiert und auf den realen CompCars-Daten getestet. Das dabei erzielte Ergebnis wurde verglichen mit der Leistung, die neuronale Netze erzielen, wenn sie direkt auf dem realen CompCars-Datensatz trainiert werden. Die Ergebnisse zeigt Tab. 2 für ein KI-Modell vom Typ ConvNeXt-Small Liu, Mao, Wu u. a., 2022. Da CompCars keine jahres- und Facelift-genauen Klassen beinhaltet, wurde neben dem Vergleich aller übereinstimmenden Modelle der CompCars-Datensatz ferner eingegrenzt auf Teile, in denen das Modelljahr (»Jhr.«) oder sogar der genaue Facelift (»Fcl.«) übereinstimmen.

Angegeben ist der F_1-*Score* bei Training auf dem jeweiligen Datensatz und Test auf CompCars, als gängiges Maß in der Bewertung der Klassifikationsleistung von maschinellen Lernverfahren, wobei ein F_1-Score von 100% eine fehlerfreie Klassifikation anzeigt. Es zeigt sich, dass das Trainingsergebnis auf synthetischen Daten zunächst wesentlich schlechter ist als auf den Realdaten, wenn man Klassen in CompCars aufnimmt, die eine große Mischung von Jahren und Facelifts beinhalten. Schränkt man die Daten jedoch ein, sodass auch CompCars nur präzise Jahre bzw. Facelifts beinhaltet, steigt die Ergebnisqualität erheblich. Bei maximaler Vergleichbarkeit verbleiben lediglich 12 übereinstimmende Modelle — für diese führt ein Training auf rein synthetischen Daten aber zu fehlerfreien Testergebnissen auf Realdaten, ebenso wie wenn das KI-Modell direkt auf Realdaten trainiert wird.

Ein differenzierterer Vergleich ist anhand der verfügbaren Daten in diesem Anwendungsfall nicht möglich, da sowohl die realen wie auch die synthetischen Daten die maximal mögliche Leistung von 100% erreichen.

Verkehrszeichenerkennung

Für den Anwendungsfall der Verkehrszeichenerkennung dient der reale deutsche GTSRB-Datensatz (»German Traffic Sign Recognition Benchmark«, veröffentlicht in Stallkamp u. a., 2011)[5] als Ausgangspunkt, der 51 882 Bilder über 43 deutsche Verkehrsschildklassen umfasst. Im Gegensatz zum VMMR-Anwendungsfall gibt es für Verkehrsschildklassifikation deutlich mehr öffentlich verfügbare Datensätze (s. Tab. 3), darunter auch der CATERED-Datensatz (veröffentlicht in Siniosoglou, Sarigiannidis u. a., 2021), der bereits einen synthetischen Zwilling von GTSRB darstellt. Für die Untersuchung erzeugen wir einen weiteren Datensatz, Synset® Signset Germany (veröffentlicht in Sielemann, Loercher u. a., 2024)[6].

Der synthetische CATERED-Datensatz umfasst 94 478 Bilder über dieselben 43 Klassen wie GTSRB, während Synset® Signset 105 500 enthält, die sich aber über insgesamt 211 Klassen verteilen (mit jeweils 500 Bildern pro Klasse); die ursprünglichen 43 Klassen aus GTSRB sind vollständig als Teildatensatz enthalten (s. Abb. 4).

5 benchmark.ini.rub.de
6 synset.de/datasets/synset-signset-ger/

Abb. 4: Überblick über die Klassen in Synset® Signset Germany. Die erste Zeile zu 43 Zeichen entspricht den Klassen im realen GTSRB-Datensatz.

Tab. 3: Die wichtigsten öffentlich verfügbaren Datensätze für Verkehrszeichenerkennung (nach Datum sortiert).

Datensatz und Publikation	Anzahl Bilder	Anzahl Klassen	⌀ Bilder/Klasse	real/synth.	Region
MASTIF, Šegvić u. a. (2010)	6 428	94	68	real	HRV
MASTIF, Šegvić u. a. (2010)	5 215	86	61	real	HRV
Stereopolis, Belaroussi u. a. (2010)	251	10	25	real	FRA
MASTIF, Šegvić u. a. (2010)	1 473	51	29	real	HRV
STS (set 1&2), Larsson und Felsberg (2011)	6 652	19	350	real	SWE
GTSRB, Stallkamp u. a. (2011)	51 882	43	1 207	real	DEU
LISA, Mogelmose, Trivedi und Moeslund (2012)	7 855	49	160	real	USA
BTSC, Mathias u. a. (2013)	7 125	62	115	real	BEL
TT100K, Zhu u. a. (2016)	30 000	221	136	real	CHN
CURE-TSR, Temel u. a. (2017)	2 206 106	14	157 579	mixed	BEL
TSRD, Huang (2018)	6 164	58	106	real	CHN
European DS, Serna und Ruichek (2018)	82 476	164	503	real	EUR
DFG, Tabernik und Skočaj (2019)	17 598	200	88	real	SVN
voll annot. MTSD, Ertler u. a. (2020)	257 541	400	644	real	weltweit
teilw. annot. MTSD, Ertler u. a. (2020)	96 613	400	242	real	weltweit
CATERED Siniosoglou, Sarigiannidis u. a. (2021)	94 478	43	2 197	synth.	DEU
Synset® Signset, Sielemann, Loercher u. a. (2024)	105 500	211	500	synth.	DEU

Die Erzeugung des Synset®-Signset-Germany-Datensatzes ähnelt der Erzeugung von Synset®-Boulevard, weicht aber in wesentlichen Punkten davon ab, wie in Abb. 6 gezeigt. Auch Synset® Signset wird in OCTAS® über physikalisch-basiertes Rendering erzeugt; allerdings wurden die Schilder-Klassen nicht über kommerzielle 3D-Modelle erzeugt, sondern basierend auf den offenen Daten der Wikipedia.[7]

Die dortigen »idealen« Bilder müssen anschließend an das reale Erscheinungsbild im Verkehr angepasst werden, wozu Verfärbungen, Abnutzungen und Vandalismus (insbesondere das Anbringen von Aufklebern) gehört. Zu diesem Zweck wurden beim Tiefbauamt in Karlsruhe über 200 ausgemusterte Verkehrsschilder abfotografiert. Für jedes Foto eines entsprechend verschmutzten Schildes wurde teilautomatisiert eine idealisierte Fassung wiederhergestellt, wie das Schild in etwa auf Wikipedia aussehen würde. Zudem wurde eine Maske mit Schmutz- und Kratzerpositionen erzeugt. Basierend darauf wurde ein generatives KI-Modell trainiert (das *Generative Adversarial Network* Pix2Pix, Isola u. a., 2017), das die umgekehrte Transformation erlernte: Die Umwandlung eines idealen Bildes von Wikipedia, zusammen mit einer Schadensmaske, hin zu einem entsprechend verschmutzen und zerkratzten Schild, das anschließend als realitätsnahe Textur für das Rendering genutzt werden konnte. Um unabhängig von der Größe und Form der einzelnen Schilder zu sein, wurde das generative KI-Modell nicht auf gesamten Schildern trainiert, sondern auf 256×256 Pixeln großen Ausschnitten.

In der Datensatzerzeugung wird dies eingesetzt, indem zuerst prozedural (das heißt, über ein klassisches Rechenmodell) eine Schadensmaske erzeugt wird, die mit dem idealen Wikipedia-Bild verbunden wird. Anschließend übersetzt das generative KI-Modell dies zu einer Schildtextur, wodurch jedes Bild im Datensatz ein anderes Muster aus Kratzern und Verschmutzungen aufweist. Die Haupt-Schilderklassen werden zufällig mit weiteren, jeweils passenden Zusatzschildern verbunden (beispielsweise der Einschränkung der Geschwindigkeitsbegrenzung auf Nässe), und auf vertikalen oder horizontalen Masten platziert.

Anschließend werden diese 3D-Szenen wieder, wie in Synset® Boulevard, zu physikalisch-basierten Kamerabildern umgewandelt, einerseits über Raytracing mit dem Cycles-Renderer, andererseits über

7 de.wikipedia.org/wiki/Bildtafel_der_Verkehrszeichen_in_der_Bundesrepublik_Deutschland_seit_2017

Abb. 5: Vergleich von Bildern des realen GTSRB-Datensatzes (oben) mit ähnlichen Bildern des synthetischen CATERED-Datensatzes (mitte), sowie des neu erzeugten synthetischen Synset®-Boulevard-Datensatzes (unten).

Rasterisierung mit dem Ogre3D-Renderer.[8] Im Vergleich zu Synset® Boulevard werden in Synset® Signset noch weitere Abbildungsartefakte aufgenommen, darunter Verwacklungen und chromatische Aberrationen, da diese im Originaldatensatz GTSRB deutlich ausgeprägter zu beobachten waren. Alle Parameter, wie Lichtverhältnisse, Kamerawinkel oder Verwacklungen werden über Wahrscheinlichkeitsverteilungen gezogen, sodass der gesamte Datensatz explizit beschriebenen stochastischen Eigenschaften folgt. Die gewählten Parameter je Einzelbild werden in Annotationsdateien bereitgestellt.

Die Ergebnisse beim Training eines KI-Modells sind in Tab. 4 dargestellt. Die erste und wenig überraschende Erkenntnis aus der Tabelle ist, dass das neuronale Netzwerk immer die besten Ergebnisse erzielt, wenn es auf einem Datensatz getestet wird, auf dem es auch trainiert wurde. Das Training mit dem synthetischen Synset®-Signset-Datensatz erzielt jedoch durchgängig die zweitbesten Ergebnisse.

Schlechter schneidet der synthetische CATERED-Datensatz ab: Während dieser als Trainingsdatensatz beim Testen mit dem realen GTSRB-Datensatz nur eine Genauigkeit von 76,4% erreicht, erzielt Synset® Signset 98,5% (während GTSRB auf sich selbst trainiert 99,9% erreicht).

Doch nicht nur der Realismusgrad lässt sich so abschätzen, auch die Frage, welcher Datensatz herausfordernd genug ist, um ein robustes KI-Modell zu trainieren. Während Training mit Synset® eine Genauigkeit von rund 98% beim Test GTSRB liefert, trifft das Gegenteil nicht zu: Ein auf GTRSB trainiertes Netz erreicht eine Genauigkeit von unter 90%, wenn es mit Synset® getestet wird. Dies deutet darauf hin, dass Synset® tatsächlich der schwierigere Datensatz ist und wohl eine stärkere Generalisierung fördert, ohne dass die Anwendbarkeit auf das einfachere GTSRB nennenswert beeinträchtigt wird.

Tab. 4: Klassifikationsgenauigkeit von ConvNeXt-Small-Modellen, jeweils auf einem der drei Datensätzen trainiert, und anschließend auf allen drei Datensätzen für Verkehrszeichenerkennung getestet.

Training ↓ Test →	S. Signset	GTSRB	CATERED
Synset® Signset	99,5%	98,3%	84,4%
GTSRB	89,4%	99,9%	77,1%
CATERED	50,0%	76,4%	86,1%

Und obwohl CATERED erkennbar stark von GTSRB und Synset® Signset abweicht, kann ein Netz mithilfe von Synset® Signset trainiert werden, um zu eine Genauigkeit von ca. 84% auf CATERED zu erreichen — wohingegen das Training auf dem realen GTSRB-Datensatz nur eine Genauigkeit von ca. 77% auf CATERED erreicht. Auch das unterstreicht, dass Synset® Signset ein robustes Trainingsergebnis und gute Generalisierungsfähigkeit erreicht.

Mehr als nur ein Ersatz für Realdaten

Die Ergebnisse zeigen, dass der erreichte Realismusgrad der synthetischen Daten es ermöglicht, reale Daten adäquat zu ersetzen, was die Voraussetzung für alle weiteren Verwendungen synthetischer Daten ist. Wie eingangs angedeutet, erlauben synthetische Daten aber nicht nur das kostengünstige Ersetzen von Realdaten — sie können auch Untersuchungen ermöglichen, die mit Realdaten nahezu unmöglich wären.

Beispielhaft wurde ein weiterer Datensatz erzeugt, der dasselbe Erzeugungsprinzip wie Synset® Signset Germany nutzt, aber gezielte einzelne Variationen beinhaltet (veröffentlicht in Sielemann, Barner u. a., 2025).[9] Dazu wurde der Datensatz auf 82 Klassen eingeschränkt, die nach Form (rund, dreieckig, rechteckig, sonstige) und nach üblichem Auftretensort (innerorts, außerorts, gemischt) gruppiert sind. Basierend darauf wurden sechs Datensätze über je 90 200 Bilder erzeugt (541 200 insgesamt), nach den folgenden Prinzipien:

8 ogre3d.org

9 synset.de/datasets/synset-signset-ger/background-effect

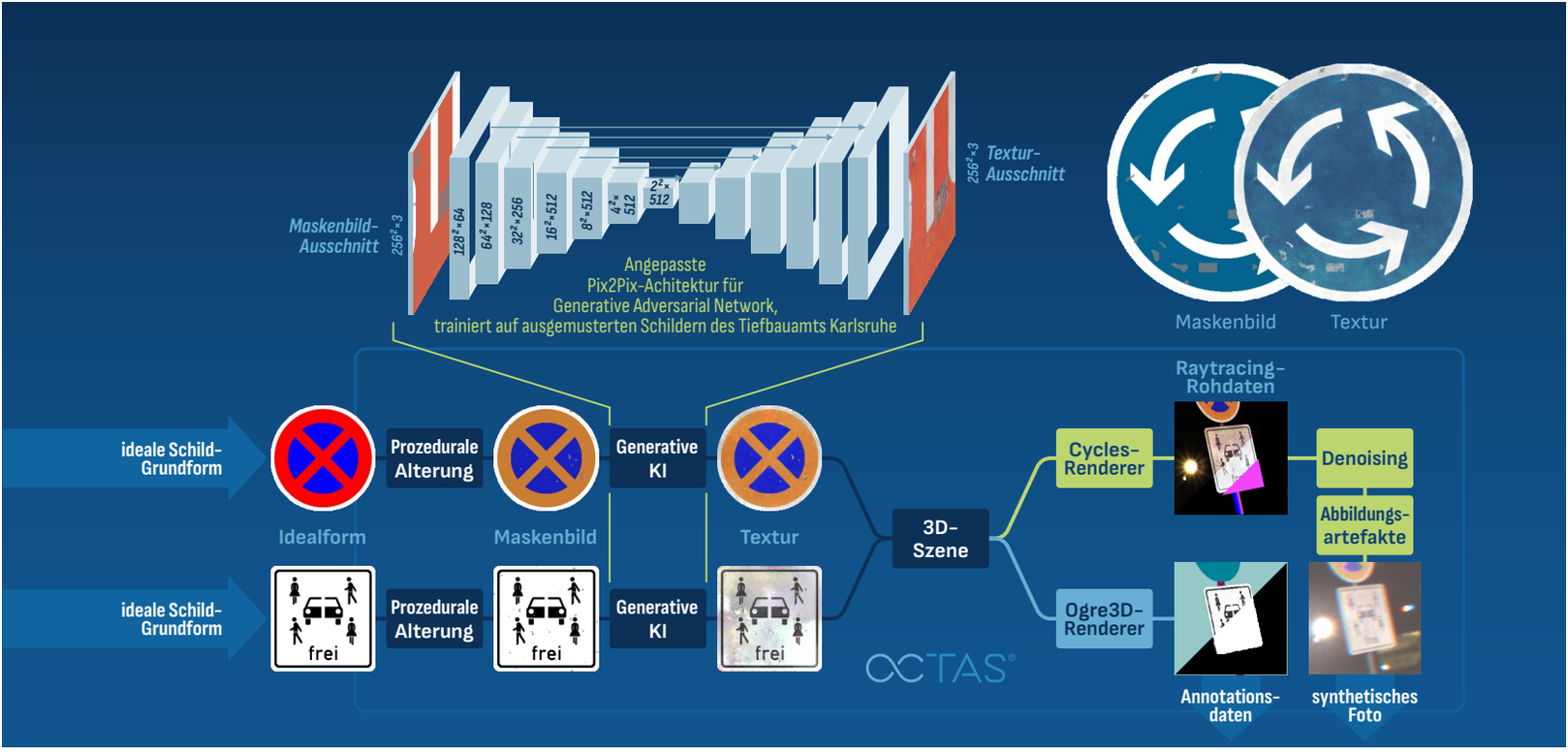

Abb. 6: Überblick über die Erzeugungspipeline für Verkehrsschildbilder, die in der KAMO-Simulationsplattform OCTAS® aufgebaut wurde. Die Verarbeitungskette verbindet generative KI mit analytischen, physikalisch basierten Modellen – KI, um komplexe und chaotische Effekte wie die Alterung und Verschmutzung von Verkehrsschildern abzubilden, physikalisch-basierte Modelle hingegen, um klar beschreibbare optische Phänomene darzustellen, wie Überbelichtung, Brechungs- und Beugungseffekte oder Rauschen. Beide Ansätze sind kontrollierbar und auf realen Daten basierend umgesetzt, sodass für gezielt für realitätsnahe Szenarien Daten erzeugt werden können.

- Jeder Datensatz enthält exakt dieselbe Abfolge von Schildern
- Drei Datensätze (die Hälfte) stellen jedes Schild in seiner üblichen Umgebung dar: Innerorts-Schilder vor urbanen Umgebungen, Außerorts-Schilder vor ländlichen Umgebungen. Die anderen drei Datensätze variieren die Umgebungen völlig zufällig, ohne Bezug zum Schildinhalt.
- Die jeweiligen drei Datensätze wiederum sind exakt übereinstimmend bis auf den Betrachtungswinkel des Schildes: Ein Datensatz enthält die Schilder strikt frontal und mittig, ein Datensatz variiert den Winkel leicht, und ein Datensatz variiert den Winkel extrem.

So entstehen Datensätze, die sich gezielt in genau einem Merkmal unterscheiden, und ansonsten identisch sind. In Sielemann, Barner u. a., 2025 wird dies genutzt, um eine gängige Annahme aus dem Bereich der erklärbaren KI (XAI, »explainable AI«) zu widerlegen. Ein gängiger Ansatz der XAI besteht darin, für ein neuronales Netz zu berechnen, welche Bildbereiche einen positiven (bestärkenden) oder negativen (abschwächenden) Einfluss auf die korrekte Klassifikation haben. Dazu dient beispielsweise die Berechnung sogenannter Kernel-SHAP-Werte (Lundberg, 2017), bei denen Bildausschnitte zwischen -1 und +1 bewertet werden, je nachdem, ob sie eher *gegen* oder *für* die erkannte Klasse sprechen.

Unter der Annahme, dass ein robustes, gut trainiertes neuronales Netz nur auf das Schild selbst »schaut«, um es zu klassifizieren, und den Hintergrund ignoriert, wird ein sogenannter »Pixel-Ratio« berechnet: Das Verhältnis zwischen positiven Aufmerksamkeiten innerhalb des Objektumrisses (des Schildes) und den gesamten positiven Aufmerksamkeiten (einschließlich des Hintergrunds). Ist dieses Verhältnis groß, hat das Netzwerk bei der Klassifikation vor allem auf den Schildinhalt geachtet. Ist es klein, hat das Netzwerk stark den Hintergrund berücksichtigt. Der Annahme zufolge ist ein hoher Pixel-Ratio also ein Zeichen für ein gesundes, gut funktionierendes neuronales Netz; ein niedriger Pixel-Ratio hingegen ein Anzeichen, dass das Netz nicht wirklich »verstanden« hat, was die Aufgabe ist.

Die Untersuchungen in Sielemann, Barner u. a., 2025 allerdings widerlegen diese gängige Interpretation von Kernel-SHAP und Pixel-Ratio. Die Publikation zeigt, dass neuronale Netze unterschiedliche Gründe haben können, auch auf den Hintergrund zu achten. Bei starken Kameravariationen nimmt die Aufmerksamkeit auf den Hintergrund zu, mutmaßlich um zwischen Hintergrund und Schild zu unterscheiden und so den eigentlichen Schildinhalt überhaupt zu finden. Ebenso nimmt die Aufmerksamkeit zu, wenn der Hintergrund tatsächlich Nutzinformation enthält. Ein verwackeltes Schild, das Tempo 30 oder Tempo 80 zeigen könnte, kann potenziell eingegrenzt werden danach, ob es sich in einer städtischen oder in einer Überland-Umgebung befindet. In beiden Fällen lässt die Klassifikationsgüte trotz sinkendem Pixel-Ratio nicht nach.

Abb. 7: Synthetische Daten erlauben gezielte Variationen, die in Realität unmöglich sind: Die Abbildung zeigt für zwei Verkehrsschildklassen jeweils das erste Bild der sechs Teildatensätze im Synset®-Signset-Variationsdatensatz. Das identische Schild mit identischen Kratzern erscheint in drei verschiedenen Winkeln und in zwei verschiedenen Umgebungen, und erlaubt so gezielte Analysen, wie Kontext das Training von neuronalen Netzen beeinflusst.

Das zeigt, dass die Interpretation des Pixel-Ratio für erklärbare KI zumindest komplizierter ist, als bisweilen angenommen. Ein solch gezielter Test einzelner Einflussfaktoren wäre mit Realdaten kaum durchführbar gewesen, zudem bringen synthetische Daten für den Anwendungsfall noch weitere nützliche Eigenschaften ein: Die Annotation, wo sich das Schild im Bild befindet, kann automatisch mit erzeugt werden, und durch das stochastisch-basierte Erzeugungsprinzip lässt sich mathematisch definieren, welche Korrelationen zwischen Trainings- und Testdatensatz überhaupt existieren können. So können neuronale Netze nicht von Korrelationen zwischen Trainings- und Testdaten profitieren, die nicht ausdrücklich bezweckt sind. Das ist ein wichtiger Baustein für die Belastbarkeit der Ergebnisse.

Fazit

Die beschriebenen Ansätze zeigen nicht nur, dass simulierte Daten in der Lage sind, reale Daten zu ersetzen, sondern insbesondere auch, wie man ihre Eignung quantifizieren kann. Dabei ist das Trainieren auf synthetischen Daten und das Testen auf realen

Abb. 8: Der »Pixel-Ratio« errechnet sich durch die positiven Kernel-SHAP-Werte innerhalb der Schildfläche, geteilt durch die Summe aller positiven Kernel-SHAP-Werte im Bild. Er misst intuitiv, wie relevant der Schildinhalt für die Klassifikation war, im Vergleich zum Hintergrund.

Daten nur ein erster Schritt. Es zeigt sich, dass synthetische Daten auch weitere Zielvorgaben erfüllen können — insbesondere, anspruchsvollere Fälle abzudecken, und damit robustere KI zu trainieren.

Darüber hinaus bieten synthetische Daten aber auch eine Chance, nicht nur bessere KI-Modelle zu trainieren, sondern KI besser zu verstehen. Durch Simulationen lassen sich gezielt »Paralleluniversen« aufbauen, deren Eigenschaften vollständig bekannt sind, und wo es somit — im Gegensatz zur realen Welt — immer eine richtige und eine falsche Lösung oder Schlussfolgerung gibt. Über solche Datensätze lassen sich dann gezielt beispielsweise Methoden der erklärbaren KI untersuchen. ■

Literaturverweise

Belaroussi, Rachid u. a. (2010). »Road sign detection in images: A case study«. In: *2010 20th International Conference on Pattern Recognition*. IEEE.

Cordts, Marius u. a. (2016). »The Cityscapes Dataset for Semantic Urban Scene Understanding«. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ertler, Christian u. a. (2020). »Traffic Sign Detection and Classification around the World«. In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Europäische Union (Juli 2024). *Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, […] und (EU) 2020/1828 (Verordnung über künstliche Intelligenz)*.

Huang, LinLin (2018). *Chinese Traffic Sign Database*. https://nlpr.ia.ac.cn/pal/trafficdata/recognition.html.

Isola, Phillip u. a. (2017). »Image-to-image translation with conditional adversarial networks«. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, S. 1125–1134.

Larsson, Fredrik und Michael Felsberg (2011). »Using Fourier descriptors and spatial models for traffic sign recognition«. In: *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*. Springer, S. 238–249.

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu u. a. (Juni 2022). »A ConvNet for the 2020s«. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Feichtenhofer, Christoph and Darrell, Trevor and Xie, Saining.

Luettner, Florian u. a. (2025). »From Real-World Traffic Data to Relevant Critical Scenarios«. In: *Proceedings of the 2025 IEEE International Automated Vehicle Validation Conference (IAVVC)*. IEEE.

Lundberg, Scott (2017). »A unified approach to interpreting model predictions«. In: *arXiv preprint arXiv:1705.07874*.

Mathias, Markus u. a. (2013). »Traffic sign recognition—How far are we from the solution?« In: *The 2013 international joint conference on Neural networks (IJCNN)*. IEEE.

Mogelmose, Andreas, Mohan Manubhai Trivedi und Thomas B Moeslund (2012). »Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey«. In: *IEEE transactions on intelligent transportation systems* 13.4, S. 1484–1497.

Sánchez, Héctor Corrales, Noelia Hernández Parra u. a. (2021). »Are We Ready for Accurate and Unbiased Fine-Grained Vehicle Classification in Realistic Environments?« In: *IEEE Access* 9. Alonso, Ignacio Parra and Nebot, Eduardo and Fernández-Llorca, David, S. 116338–116355. DOI: 10.1109/ACCESS.2021.3104340.

Schnieder, Lars und René S Hosse (2019). *Leitfaden Safety of the Intended Functionality*. Springer.

Šegvić, S u. a. (2010). »A computer vision assisted geoinformation inventory for traffic infrastructure«. In: *13th international IEEE conference on intelligent transportation systems*. IEEE, S. 66–73.

Serna, Citlalli Gamez und Yassine Ruichek (2018). »Classification of traffic signs: The european dataset«. In: *IEEE Access* 6, S. 78136–78148.

Sielemann, Anne, Valentin Barner u. a. (2025). »Measuring the Effect of Background on Classification and Feature Importance in Deep Learning for AV Perception«. In: *Proceedings of the 2025 IEEE International Automated Vehicle Validation Conference (IAVVC)*. IEEE.

Sielemann, Anne, Lena Loercher u. a. (2024). »Synset Signset Germany: a Synthetic Dataset for German Traffic Sign Recognition«. In: *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, S. 3383–3390.

Sielemann, Anne, Stefan Wolf u. a. (2024). »Synset Boulevard: A Synthetic Image Dataset for VMMR«. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, S. 9146–9153.

Siniosoglou, Ilias, Panagiotis Sarigiannidis u. a. (2021). »Synthetic Traffic Signs Dataset for Traffic Sign Detection & Recognition In Distributed Smart Systems«. In: *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, S. 302–308.

Sochor, Jakub, Adam Herout und Jiri Havel (Juni 2016). »BoxCars: 3D Boxes as CNN Input for Improved Fine-Grained Vehicle Recognition«. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sochor, Jakub, Jakub Špaňhel und Adam Herout (2018). »Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance«. In: *IEEE transactions on intelligent transportation systems* 20.1, S. 97–108.

Stallkamp, Johannes u. a. (2011). »The German traffic sign recognition benchmark: a multi-class classification competition«. In: *The 2011 international joint conference on neural networks*. IEEE.

Tabernik, Domen und Danijel Skočaj (2019). »Deep Learning for Large-Scale Traffic-Sign Detection and Recognition«. In: *IEEE Transactions on Intelligent Transportation Systems*. ISSN: 1524-9050. DOI: 10.1109/TITS.2019.2913588.

Temel, D. u. a. (2017). »CURE-TSR: Challenging unreal and real environments for traffic sign recognition«. In: *Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning for Intelligent Transportation Systems*.

Wachenfeld, Walther und Hermann Winner (2015). »Die Freigabe des autonomen Fahrens«. In: *Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte*, S. 439–464.

Yang, Linjie, Ping Luo u. a. (Juni 2015). »A Large-Scale Car Dataset for Fine-Grained Categorization and Verification«. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Change Loy, Chen and Tang, Xiaoou.

Zhu, Zhe u. a. (Juni 2016). »Traffic-Sign Detection and Classification in the Wild«. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.