Deutsches Zentrum für
Schienenverkehrsforschung beim
Eisenbahn-Bundesamt

Berichte
des Deutschen Zentrums
für Schienenverkehrsforschung

Bericht 61 (2024)

Automatisierte digitale Bestands-erfassung gleisnaher Infrastruktur aus Befliegungsdaten

Schlussbericht

Berichte des Deutschen Zentrums
für Schienenverkehrsforschung, Bericht 61 (2024)
Projektnummer 2021-35-U-1202

Automatisierte digitale Bestandserfassung gleisnaher Infrastruktur aus Befliegungsdaten

von

Adrian Loy, Hanna Behnke, Jennifer Hahn, Miha Garafolj, Paula Feike, Ziyad Sheebaelhamd
Merantix Momentum, Berlin

im Auftrag des Deutschen Zentrums für Schienenverkehrsforschung beim Eisenbahn-Bundesamt

Impressum

HERAUSGEBER
Deutsches Zentrum für Schienenverkehrsforschung beim Eisenbahn-Bundesamt

August-Bebel-Straße 10
01219 Dresden

www.dzsf.bund.de

DURCHFÜHRUNG DER STUDIE
Merantix Momentum
Max-Urich-Straße 3, AI Campus
13355 Berlin

ABSCHLUSS DER STUDIE
Mai 2024

REDAKTION
Deutsches Zentrum für Schienenverkehrsforschung beim Eisenbahn-Bundesamt
Dr. Katharina Fricke, Dr. Sonja Szymczak; Fachbereiche Mobilität und Gesellschaft/ Klimaschutz, Umwelt und Nachhaltigkeit

BILDNACHWEIS
Ausschnitt aus DOP-Aufnahme: © GeoBasis-DE/BKG (2023)

PUBLIKATION ALS PDF
https://www.dzsf.bund.de/Forschungsergebnisse/Forschungsberichte

ISSN 2629-7973

doi: 10.48755/dzsf.240014.01

Dresden, November 2024

Inhaltsverzeichnis

Abkürzungsverzeichnis

ADAM	Adaptive Moment Estimation
AP	Average Precision
API	Application Programming Interface (Programmierschnittstelle)
BYOL	Bootstrap Your Own Latent
CNN	Convolutional Neural Network
DETR	Detection Transformer
DGM	Digitales Geländemodell
DINO	Distributed Instance-wise Recognition
DOP	Digitales Orthophoto
DOM	Digitales Oberflächenmodell
DZSF	Deutsches Zentrum für Schienenverkehrsforschung, Dresden
EBA	Eisenbahn-Bundesamt, Bonn
ESA	European Space Agency
GKE	Google Kubernetes Engine
IoU	Intersection over Union
mAP	Mean Average Precision
MAE	Masked Auto-Encoding
ML	Machine Learning (Maschinelles Lernen)
OSM	OpenStreetMaps
PAK	Projektbegleitender Arbeitskreis
ReLU	Rectified Linear Unit
SAM	Segment Anything-Modell
SGD	Stochastic Gradient Descent
SSW	Schallschutzwände
STEGO	Self-supervised Transformer with Energy-based Graph Optimization
ViT	Vision Transformer-Architektur

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Eisenbahninfrastruktur nimmt eine zentrale Stellung im Verkehrssystem ein, indem sie einen effizienten Transport von Gütern und Passagieren ermöglicht und Städte und Gebiete miteinander verbindet. Präzise und aktuelle Informationen über Infrastrukturelemente entlang der Schienenwege, wie Schallschutzwände (SSW), Zugangspunkte, Böschungen und Überführungen, sind für verschiedene Anwendungen wie die Lärmkartierung, die Infrastrukturwartung und das Notfallmanagement unabdingbar.

Jedoch stellt die genaue Erfassung solch gleisnaher Infrastrukturobjekte sowie die einheitliche Aufbereitung der entsprechenden Informationen essenzielle Herausforderungen für die genannten Anwendungsfelder dar. Bestehende digitale Datensätze, welche Informationen über Eisenbahninfrastrukturelemente beinhalten, weisen einige Einschränkungen wie etwa Unstimmigkeiten in der Datenqualität, Unvollständigkeit und veraltete Informationen auf.

Zur Bewältigung dieser Herausforderungen wurde in dem Projekt „Automatisierte digitale Bestandserfassung gleisnaher Infrastruktur aus Befliegungsdaten"[1] ein innovativer Prozessierungs-Workflow ausgearbeitet. Dieser Workflow soll die automatisierte Erfassung und Analyse von gleisnahen Infrastrukturobjekten, am Beispiel von SSW, mittels Fernerkundungsdaten und Methoden aus dem Bereich des maschinellen Lernens (ML) unterstützen.

Die Entwicklung dieses Prozessierungs-Workflows erforderte im ersten Schritt eine sorgfältige Anforderungsanalyse und Recherche zu aktuellen und anwendbaren Technologien. Die Grundlage und das erste Ziel im Rahmen des Projekts bestand darin, mit öffentlich zugänglichen Bilddaten eine erforderliche Datengrundlage zu erarbeiten, welche Informationen über die Position und Spezifikationen von SSW liefert. Aufbauend darauf wurde der Prozessierungs-Workflow und eine technische Implementierung basierend auf ML-Methoden entwickelt. Abschließend wurde sowohl die technische Funktionalität als auch die Möglichkeit zur Verwendung der entwickelten Lösung für den spezifischen Anwendungsfall der Lärmkartierung evaluiert und kritisch diskutiert.

Der folgende Bericht ist entsprechend dieser Arbeitspakete und Projektphasen aufgegliedert. Beginnend mit der Anforderungsanalyse, beschreibt Kapitel 2, welche möglichen Infrastrukturelemente aus Fernerkundungsdaten gewonnen werden können. Hierbei stand insbesondere die Lärmkartierung im Fokus. Dabei werden im ersten Teil in Kapitel 2.1 die erarbeiteten Anforderungen für die Extraktion von Infrastrukturelementen aus Fernerkundungsdaten entlang der Schienenstrecken basierend auf Interviews mit Nutzenden und Workshopergebnissen beschrieben. Im Anschluss ist die Literaturrecherche zu Methodiken der Extraktion von Infrastrukturelementen aus Fernerkundungsdaten sowie relevanten Techniken aus ähnlichen Anwendungsfeldern in Kapitel 3 illustriert. Unter Berücksichtigung der identifizierten Anforderungen sowie technologischen Möglichkeiten und Risiken werden in Kapitel 4 das zu Beginn definierte Vorgehen sowie Maßnahmen zur Evaluation und Qualitätssicherung dargelegt. Dies bildete auf der einen Seite eine Grundlage zum Verständnis, inwiefern Fernerkundungsdaten für die Infrastrukturbewertung genutzt werden können, und auf der anderen Seite wertvolle Erkenntnisse für die Auswahl der geeigneten Methoden.

In der darauffolgenden Phase wurde eine eingehende Analyse und Aufbereitung der Datengrundlage, wie in Kapitel 5 beschrieben, durchgeführt. Im Detail werden hier die Schritte der Datenexploration in Kapitel 5.1, die Datenfusion in Kapitel 5.2, sowie die abschließende Aufbereitung des Trainings- und Testdatensatzes in Kapitel 5.3 dargelegt. Dabei werden zunächst die notwendigen Vorverarbeitungsschritte für das

[1] Projekt-Webseite:
https://www.dzsf.bund.de/SharedDocs/Standardartikel/DZSF/Projekte/Projekt_130_Bestandserfassung_gleisnahe_infrastruktur.html

Trainieren des Modells sowie die Methodik zur Verbesserung der Datenqualität und zur Erleichterung eines effektiven Lernens in Kapitel beschrieben.

Im Anschluss an die Exploration und Erarbeitung der Datengrundlage stand die Entwicklung und Bewertung der technischen Lösung unter Verwendung eines ML-Modells, wobei die ML-Lösung speziell auf die Erkennung von SSW aus Befliegungsdaten zugeschnitten wurde. Entsprechend wird in Kapitel 6 ein umfassendes Verständnis der angewandten Methodik, des Entscheidungsprozesses und entsprechenden Ergebnissen der ML-Lösung im Rahmen des Projektes geschaffen. In Kapitel 6.1 und 6.2 werden die Hintergründe zur Auswahl der ausgewählten Modellarchitektur erläutert und die Faktoren, die zur Entscheidungsfindung beigetragen haben, dargelegt. Darüber hinaus werden die Ergebnisse der effektivsten Modell-Version anhand entsprechender Metriken in Kapitel 6.3 illustriert. Hier werden detaillierte Ergebnisse der Evaluation sowie Einblicke in die Genauigkeit und Robustheit der ML-Lösung erläutert. Zusätzlich zur Entwicklung des Prozessierungs-Workflows wird in Kapitel 7 die Integration der Lösung in das bestehende Software-System QGIS (QGIS Association, 2024) sowie deren Funktionalität und Nutzung beschrieben.

In der abschließenden Diskussion werden Herausforderungen und zukünftige Möglichkeiten, die sich innerhalb der unterschiedlichen Arbeitspakete gezeigt haben, genauer beleuchtet. Im Zuge dessen wird das Ergebnis des Projektes im Hinblick auf Funktionalität, Integration und Anwendbarkeit diskutiert sowie zukünftige Möglichkeiten zur Weiterentwicklung und zum Aufbau auf das bisherige Projekt in Kapitel 8.2 und 8.3 aufgezeigt.

2 Anforderungsanalyse

2.1 Methodik der Anforderungsanalyse

Um die Bedürfnisse der involvierten Stakeholder zu erfassen, wurde als Teil der ersten Projektphase eine mehrstufige Anforderungsanalyse durchgeführt (siehe Abbildung 1). Eine erste Bedarfsanalyse und Voruntersuchung mit Mitarbeitenden des Referats 53 des Eisenbahn-Bundesamtes (EBA) vor Beginn des Projektes diente der Gewinnung erster Einblicke in den Anwendungsfall der Lärmkartierung. Als Teil eines zweiten semistrukturierten Leitfadeninterviews mit zwei Mitarbeitenden des EBA-Referats 53 wurden zu Beginn der ersten Projektphase die Herausforderungen und mögliche Ansätze zur Unterstützung der Lärmkartierung näher untersucht. Neben der Lärmkartierung (Hauptfokus des Projekts) gibt es weitere Möglichkeiten, die Projektergebnisse wertstiftend einzusetzen. Diese wurden in zwei Workshops mit sieben Mitgliedern des projektbegleitenden Arbeitskreises (PAK) diskutiert. Der PAK setzt sich aus Fachpersonen und Vertretenden verschiedener Unternehmen und Instituten zusammen, darunter staatliche Einrichtungen, private Unternehmen, spezialisierte Forschungszentren und Behörden, deren Arbeit die Erfassung und Nutzung von Infrastrukturobjekten im Umfeld des Verkehrsträgers Schiene für die Lärmkartierung und andere behördliche Aufgaben umfasst. Die Workshops wurden mithilfe von Miro-Boards durchgeführt, um eine gemeinsame Diskussion über Anwendungsfälle und Priorisierung von Infrastrukturelementen zu ermöglichen und zu dokumentieren.

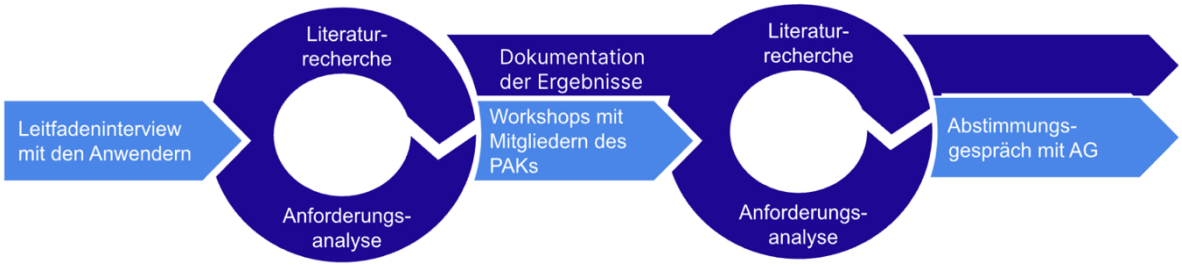


Abbildung 1: Vorgehen in der ersten Projektphase beginnend mit Leitfadeninterviews mit den Anwendern gefolgt von einer iterativen Literaturrecherche und Anforderungsanalyse sowie parallele Workshops mit entsprechenden Stakeholdern und Dokumentation der Ergebnisse (eigene Abbildung)

2.2 Anwendungsfall: Lärmkartierung an Schienenwegen

Das Hauptziel der Lärmkartierung an Schienenwegen besteht darin, Lärmkarten für das öffentliche Schienennetz Deutschlands gemäß den Vorschriften der Umgebungslärmrichtlinie (2002/49/EG) (EUR-Lex, 2022) zu erstellen. Diese Lärmkarten sollen die Umweltlärmexposition entlang der Eisenbahnstrecken bewerten und visualisieren sowie der Öffentlichkeit Informationen über die Lärmbelastung liefern. Die Lärmkartierung basiert auf der Modellierung der Lärmausbreitung und liefert Ergebnisse, die als Grundlage für Lärmminderungsplanung dienen. Die Berichterstattung der Ergebnisse an die EU-Kommission ist ein weiterer wichtiger Aspekt. Die EU-Umgebungslärmrichtlinie sieht einen rundenbasierten Überarbeitungszyklus vor, wobei jeder Kartierungszyklus fünf Jahre dauert (Eisenbahn-Bundesamt, 2023b). Die Ergebnisse der Lärmkartierung für die Ermittlung und Erstellung der Belastung durch Umgebungslärm, das

Informieren der Öffentlichkeit und der Lärmaktionsplanung können über den interaktiven Kartendienst auf dem GeoPortal.EBA (Eisenbahn-Bundesamt, 2023a) abgerufen werden.

Verschiedene Interessengruppen sind in den Prozess der Lärmkartierung involviert: Das EBA ist für die Durchführung der Lärmkartierung und die Bereitstellung der Ergebnisse für Schienenwege von Eisenbahnen des Bundes verantwortlich. Die EU-Kommission ist der Empfänger der gemeldeten Ergebnisse. Die Öffentlichkeit ist eine wichtige Interessengruppe, da sie über die Umweltlärmexposition informiert werden muss. Weitere Rollen spielen Organisationen wie z. B. die DB InfraGO, welche Daten zu SSW und anderen Infrastrukturelementen entlang der Schienenwege bereitstellen. Zentrale Bundes- und Landesbehörden sowie Kommunen tragen ebenfalls Daten zu SSW und anderen relevanten Infrastrukturelementen bei.

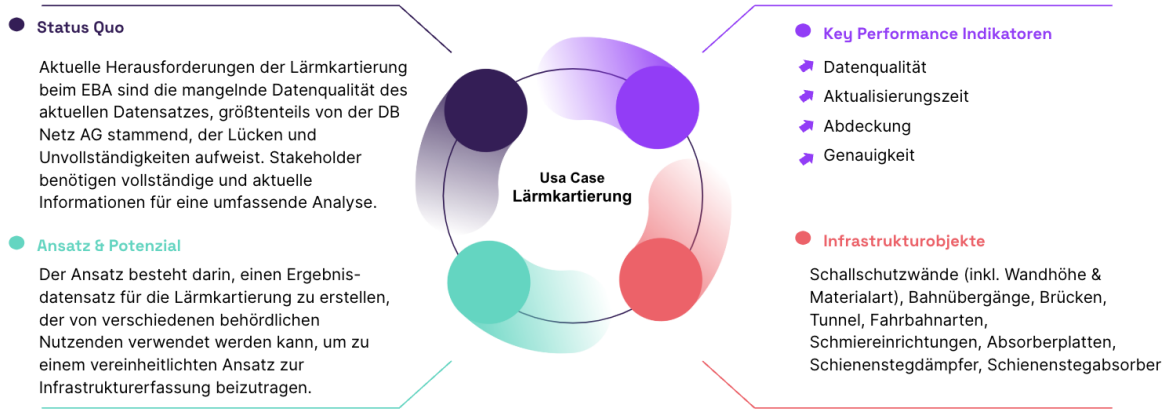


Abbildung 2: Anforderungsanalyse des Anwendungsfalls der Lärmkartierung an Schienenwegen. In der Abbildung ist der erarbeitete Status Quo, der geplante Ansatz und das Potenzial, die Key-Performance-Indikatoren sowie relevante Infrastrukturobjekte beschrieben (eigene Abbildung)

Das EBA steht derzeit vor mehreren Herausforderungen in Bezug auf die Lärmkartierung an Schienenwegen. Ein zentraler Aspekt dabei ist die Datenqualität: Der aktuelle, aus verschiedenen Quellen zusammengefügte Datensatz von SSW weist Lücken und Unvollständigkeiten auf, die die Gesamtqualität der Daten beeinträchtigen (siehe Abbildung 2).

Das EBA und die weiteren PAK-Mitglieder streben an, die Datenqualität, die Aktualisierungshäufigkeit, die Abdeckung und die Genauigkeit der Lärmkartierung zu verbessern und gleichzeitig detaillierte Informationen über Infrastrukturobjekte, insbesondere SSW, zu erhalten. Dies soll eine umfassende Analyse ermöglichen und als zuverlässige Grundlage für die Lärmkartierung dienen. Die Verbesserung der Datenqualität beinhaltet die Behebung von Lücken und Unvollständigkeiten im aktuellen Datensatz. Insbesondere würde ein neuer Datensatz ermöglichen, einen Abgleich gegen die gelieferten Daten vornehmen zu können, um anhand einer zweiten Quelle die Vollständigkeit der Daten prüfen zu können. Gleichzeitig wird angestrebt, die Aktualisierungshäufigkeit zu erhöhen, um schnell auf Infrastrukturänderungen reagieren zu können. Das EBA-Referat 53 erhält jährlich aktualisierte Daten der DB InfraGO AG und anderen datenführenden Stellen, dennoch kommt es vor, dass z. B. neu errichtete SSW noch nicht in die Daten aufgenommen wurden. Die Abdeckung der Lärmkartierung konzentriert sich auf relevante Bereiche beidseitig entlang der Schienenwege entsprechend des vorgegebenen Kartierungskorridors, um eine umfassende Bewertung der Lärmexposition zu gewährleisten. Das EBA sowie die PAK-Mitglieder haben außerdem Interesse an der Verbesserung der Informationsgenauigkeit, um präzisere Analysen und Modellierungen durchzuführen.

2.3 Weitere Anwendungsfälle

Auch wenn die Lärmkartierung primärer Fokus des Projekts ist, konnten im Rahmen der Anforderungsanalyse eine Reihe von weiteren möglichen Anwendungen identifiziert werden, die von der Erfassung gleisnaher Infrastrukturobjekte profitieren können. Diese Anwendungsfälle umfassen verschiedene Bereiche wie z. B. Notfall- und Katastrophenmanagement, erneuerbare Energieerzeugung, Landmarken-Navigation und Vegetationsmanagement. Durch die Nutzung der automatisierten digitalen Bestandsaufnahme mit hochauflösenden Satelliten- und Befliegungsdaten sollen wertvolle Informationen gewonnen werden, die als Grundlage für zielgerichtete Maßnahmen und Planung in diesen Bereichen dienen. Die Anwendungsfälle werden nachfolgend erläutert.

Das Notfallmanagement ist von entscheidender Bedeutung, um die wirksame Bereitstellung von Rettungsteams und Hilfsorganisationen in kritischen Situationen zu unterstützen und gleichzeitig die Reaktionszeit zu minimieren. Ein Aspekt des Notfallmanagements besteht darin, den Zugang für Rettungsteams, wie z. B. dem Technischen Hilfswerk (THW) und der Feuerwehr, zum Ereignisort entlang der Schienenwege schnell zu ermöglichen. Genaue und aktuelle Informationen über den Standort und die Eigenschaften von Infrastrukturelementen wie Brücken, SSW, Unterführungen und wichtigen Zugangspunkten sind dabei von entscheidender Bedeutung. Indem die Notfallhilfe mit solch präzisen Informationen versorgt wird, kann sie ihre Einsätze besser planen, potenzielle Hindernisse oder Gefahren proaktiv identifizieren und diese schnell und sicher überwinden.

Das Katastrophenmanagement befasst sich mit groß angelegten Notfällen oder Katastrophen, die eine koordinierte Zusammenarbeit mehrerer Behörden und Organisationen erfordern. In solchen Szenarien ist ein umfassendes Verständnis der Infrastruktur und der verfügbaren Ressourcen von entscheidender Bedeutung. Indem Informationen über Infrastrukturelemente wie Brücken, Bahnstrecken, Bahnübergänge und andere relevante Punkte entlang der Schienenwege erfasst und bereitgestellt werden, können Behörden schnell und effektiv auf Katastrophen reagieren. Dieses Wissen umfasst den genauen Standort, die Zugänglichkeit und mögliche Einschränkungen der genannten Objekte, wodurch Entscheidungsträger fundierte Entscheidungen treffen und eine koordinierte Einsatzplanung durchführen können.

Im Bereich der erneuerbaren Energie bieten die Identifizierung und Erfassung von gleisnahen Infrastrukturelementen Potenzial zur Förderung nachhaltiger Praktiken. Ein interessanter Anwendungsfall besteht darin, potenzielle Gebiete zu identifizieren, die für die Installation von Photovoltaikanlagen in der Bahnumgebung geeignet sind. Genaue Informationen über vorhandene Infrastrukturelemente sind für die Bestimmung der Machbarkeit und der Rentabilität der Erzeugung erneuerbarer Energien entlang der Bahnstrecken von entscheidender Bedeutung. Durch die Ermittlung potenzieller Standorte für Photovoltaikanlagen wird die Nutzung nachhaltiger Energielösungen weiter vorangetrieben.

Die Landmarken-Navigation nutzt Infrastrukturelemente wie Brücken oder Bahnübergänge, um eine hochgenaue Eigenlokalisation von Schienenfahrzeugen zu ermöglichen. Diese Anwendung hat besondere Bedeutung für z. B. Messzüge, die präzise Messungen entlang der Bahngleise durchführen. Dabei werden die identifizierten Infrastrukturelemente als markante Punkte genutzt, was eine genaue Positionsbestimmung ermöglicht und bei wichtigen Aufgaben wie der Streckenwartung und Leistungsbewertung hilft. Zudem können dadurch autonome Fahrzeuge sicher und zuverlässig auf den Straßen navigieren, was die Verkehrssicherheit erhöht, und die Weiterentwicklung des automatisierten Fahrens vorantreibt.

Vegetationsmanagement entlang der Bahngleise ist entscheidend für einen sicheren und zuverlässigen Betrieb. Detaillierte Kenntnisse über den Standort und den Zustand von Infrastrukturelementen wie Böschungen sind bei der Planung und Umsetzung wirksamer Maßnahmen zur Vegetationskontrolle unerlässlich. Durch die Erfassung und Nutzung dieser Informationen können Betreibende und Behörden umfassende Strategien zur Überwachung und Instandhaltung der Vegetation entlang der Bahnstrecken ent-

wickeln, um potenzielle Risiken zu minimieren und einen störungsfreien Betrieb zu gewährleisten. Zusätzlich ermöglicht die gezielte Detektion von potenziellen Standorten wie Böschungen und von Vegetation selbst die Identifizierung von Stellen, an denen Vegetation eine Gefährdung für den Bahnbetrieb darstellen kann, beispielsweise durch die Verdeckung von Signalen oder das Eindringen in das Lichtraumprofil.

Zusammenfassend wird deutlich, dass der Einsatz automatisierter digitaler Bestandsaufnahmemethoden wertvolle Erkenntnisse über verschiedene Infrastrukturelemente schafft, von denen nicht nur die Lärmkartierung profitiert.

2.4 Infrastrukturobjekte

Im folgenden Abschnitt werden verschiedene Arten von Infrastrukturobjekten vorgestellt, deren automatisierte Identifikation für die zuvor vorgestellten Anwendungsfälle von Relevanz ist. Außerdem werden die Nutzeranforderungen an die Genauigkeit der Erfassung je Infrastrukturobjekt dokumentiert und mit Hinblick auf mögliche technische Herausforderungen kommentiert. Ziel war es, eine Priorisierung der Anforderungen zu ermöglichen, um das Projektvorgehen im Rahmen der technischen Möglichkeiten an die spezifischen Ziele der beteiligten Interessengruppen anzupassen.

Infrastrukturobjekt: Schallschutzwände

SSW beeinflussen direkt die Ausbreitung von Emissionen entlang von Schienenwegen und spielen eine entscheidende Rolle bei der Lärmkartierung. Als integrale Bestandteile der Bahninfrastruktur beeinflussen SSW maßgeblich die Lärmausbreitung und sollten in Kartierungsmaßnahmen einbezogen werden, um präzise und verlässliche Ergebnisse zu erzielen. Durch die Erfassung von SSW in Lärmkartierungsübungen können verschiedene Lärmminderungsmaßnahmen geplant und bewertet werden. Dies ermöglicht den Behörden, die Wirksamkeit bestehender SSW zu beurteilen, potenzielle Lücken oder Mängel zu identifizieren und bei Bedarf zusätzliche Schutzmaßnahmen zu planen. Darüber hinaus bietet die Einbeziehung von SSW in die Lärmkartierung den Interessengruppen einen umfassenden Überblick über die Lärmsituation entlang der Bahngleise, was fundierte Entscheidungen in Bezug auf Lärmschutz und die Planung zukünftiger Infrastrukturprojekte erleichtert. Der Ist-Zustand, Ziel-Zustand sowie Herausforderungen bei der Erfassung der SSW sind in Tabelle 1 zusammengefasst und werden nachfolgend erläutert. Der jeweilige Ziel-Zustand wurde im Rahmen des PAK durch die Mitglieder definiert.

Die **Aktualisierungshäufigkeit** für SSW in der Lärmkartierung entspricht dem fünf-Jahres-Zyklus, der in der EU-Umgebungslärmrichtlinie (EUR-Lex, 2022) festgelegt ist. Dies bedeutet, dass die Datenlieferung der DB InfraGO AG, den zentralen Landesbehörden und den Kommunen bezüglich SSW diesem fünf-Jahres-Rhythmus folgt. Aufgrund des großen Sanierungsbedarfs und fortlaufenden Bauvorhaben von SSW (BMDV, 2024), besteht die Notwendigkeit, kürzere Aktualisierungszyklen für SSW zu etablieren.

Eine hohe **Genauigkeit** für den Standort von SSW ist von entscheidender Bedeutung, mit einer Toleranz von ≤ ±0,3 m in der Position. Dabei soll der Standort sowohl für hohe als auch für oft schwerer erkennbare, niedrige SSW vorzugsweise mit einer Positionsabweichung von ≤ ±0,3 m bestimmt werden. Es werden zusätzliche Informationen benötigt, um die Genauigkeit der räumlichen Daten zu beurteilen. Hierzu wird die mögliche Anwendung von Fernerkundungstechniken überprüft. Hier werden unter anderem Digitale Orthofotos (DOP) in Betracht gezogen und analysiert, die eine begrenzte Auflösung mit sich bringen. Alle SSW entlang der Bahngleise oder innerhalb von 1 bis 200 m von der Gleisachse sollten nach Möglichkeit erfasst werden.

Tabelle 1: Anforderungsanalyse für das Infrastrukturobjekt „Schallschutzwände“ bei der Lärmkartierung mit jeweiligem Ist- und Zielzustand sowie Herausforderungen der beschriebenen Merkmale

Merkmale	Aktueller Status	Angestrebter Zielzustand	Herausforderungen
Aktualisierungs-frequenz	Fünfjahresintervalle	Häufigere Aktualisierungen	Abhängig von verfügbaren Datenbeständen
Genauigkeit	≤ ±0,3 m	≤ ±0,3 m	Begrenzte Auflösung von Fernerkundungsdaten, z. B. DOP20-Luftbildern mit 0,2 m
Datenqualität	Ungenau und unvollständig	Verbesserte Datenqualität	Fehlende oder falsche Lärmschutzwanddaten
Höhe	Genauigkeit von 0,1 m	Genauigkeit von 0,1 m	Schwierigkeiten aufgrund einer begrenzten Auflösung von Höhendaten, z. B. von 1 m in DOM
Material	Schwierig zu bestimmen ohne Datenbankinformationen	Identifizierung des Lärmschutzwandmaterials	Herausforderungen bei der Unterscheidung von Materialien aus öffentlich verfügbaren Datenquellen

Die Anforderungen an die **Erfassung der Höhe** von SSW sind präzise Höhenmessungen mit einer Genauigkeit von 0,1 m. Zur Erfassung der Höhe ist jedoch die Auflösung der zur Verfügung stehenden Daten zu beachten, da sich dies auf die Anzahl der möglichen Höhenpunkte pro Längeneinheit auswirken kann. Im Rahmen der Anforderungsanalyse wurde die Ableitung von Höheninformationen der SSW aus digitalen Oberflächenmodellen (DOM) betrachtet. Eine mögliche Methode zur Bewältigung dieser Herausforderung besteht darin, benachbarte Höhenwerte zu mitteln, wenn der Höhenunterschied zwischen ihnen weniger als 0,5 m beträgt. Anschließend könnten Abschnitte gebildet werden, die als Lärmkartierungswandabschnitte bezeichnet werden und eine Höhenabweichung von etwa 0,5 m im Vergleich zum benachbarten Lärmkartierungswandabschnitt aufweisen. Diese Option würde eine kontinuierliche und inkrementelle Höhenvariation entlang der SSW ermöglichen. Alternativ könnte eine andere Methode angewendet werden, bei der eine einheitliche Durchschnittshöhe für die gesamte SSW festgelegt wird. In diesem Fall würden die maximalen und minimalen Höhenwerte bestimmt und angegeben werden. Beide Optionen bieten Möglichkeiten, Höheninformationen von SSW in die Analyse einzubeziehen. Die Wahl des Ansatzes hängt von der technischen Machbarkeit und den spezifischen Anforderungen des Projekts ab und zielt darauf ab, aussagekräftige Darstellungen der Höhe von SSW für praktische Analysen und Bewertungen zu erhalten.

Die Identifizierung des Materials, aus dem die SSW besteht, ist für die **Bestimmung der Schallabsorptionskoeffizienten** von großer Bedeutung. Jedoch ist die Ableitung basierend auf den visuellen Eigenschaften einer SSW oft eine Herausforderung. Es kann angenommen werden, dass transparente Wände oder solche, die ausschließlich aus Beton bestehen, schallreflektierende Eigenschaften aufweisen. Bei einer Metallstruktur mit eingebettetem schallabsorbierendem Material kann jedoch das Schallabsorptionsverhalten des Metalls nicht automatisch abgeleitet werden, und die schallabsorbierenden Eigenschaften des eingebetteten Materials können visuell nicht bestimmt werden. Für aussagekräftige Analysen sind spezifische Informationen über die SSW erforderlich. Dazu zählen auch Materialspezifikationen wie Beton, Beton mit Absorptionsbeton, Beton mit Mauerwerk, Beton mit Verkleidung, Beton mit Kunststoff, Beton mit Glas, Beton mit Leichtmetall, Mauerwerk, Stahl, Aluminium, Aluminium mit Glas, Stahl und Alumi-

nium, Kunststoff, Acryl, Weichholz, Hartholz, Holz (unbestimmt), Gummi, Glas, Stein (nicht schallabsorbierender Aufbau), Stein (schallabsorbierender Aufbau), z. B. Gabionenwand aus gestapelten Steinen mit Erde, sandgefülltem Zwischenraum und anderen Materialtypen. Derzeit spielen die Materialspezifikationen nur dann eine Rolle für das EBA, wenn SSW keine Informationen zur Schallabsorption enthalten. In solchen Fällen unterscheidet das EBA zwischen schallreflektierenden SSW (bestehend aus transparenten Elementen wie Glas) und stark absorbierenden SSW (bestehend aus nicht-transparenten Elementen, die mit schallabsorbierenden Materialien beschichtet sind). Jeder Typ von Lärmschutzwand wird basierend auf den jeweiligen Materialeigenschaften mit einem Schallabsorptionsspektrum versehen. Die Erfassung von Absorptionskoeffizienten und die Klassifizierung von SSW auf der Grundlage ihrer Materialeigenschaften sind entscheidend für eine effektive Lärmkartierung und -bewertung.

Weitere Infrastrukturobjekte

Neben den SSW sind verschiedene weitere gleisnahe Infrastrukturobjekte für die zuvor beschriebenen Anwendungsfälle von Relevanz. Dazu gehören Bahnüberführungen, Bahnunterführungen, Gleis- arten, Absorptionsplatten, Schienenstegdämpfer, Schienenstegabsorber und Evakuierungspunkte. Die Integration dieser Objekte in einen maschinellen Lernansatz bietet ein großes Potenzial für die beschriebenen Anwendungsfälle.

Die durch maschinelles Lernen unterstützte Kartierung dieser Infrastrukturelemente birgt jedoch auch Herausforderungen in Bezug auf Datenverfügbarkeit und -qualität. Da ML-Methoden Muster und Prognosen aus Daten erlernen, ist der Erfolg der Klassifizierung pro Infrastrukturobjekt maßgeblich von der verfügbaren Datengrundlage abhängig. Entsprechend müssen Datenquellen, Auflösung und Datenintegration angemessen berücksichtigt werden, um die Wirksamkeit des Kartierungsprozesses zu gewährleisten.

Zusammenfassung der Anforderungsanalyse

Die Anforderungsanalyse zeigt die Relevanz der digitalen Erfassung von gleisnaher Infrastruktur für verschiedene Anwendungsfälle auf. Die Analyse verdeutlicht, dass aufgrund der Vielfalt an Datenquellen und Inhalten sowie verzögerten Datenerfassungen und längeren Berichtszyklen bei reellen Infrastrukturobjekten zurzeit eine flächendeckende, einheitliche und aktuelle Erfassung der Informationen nicht zu jedem Zeitpunkt gewährleistet werden kann. Vor diesem Hintergrund wurde das Ziel des Projekts definiert, die Mitarbeitenden des EBA und weitere potenzielle Nutzende technologiegetrieben dabei zu unterstützen, den manuellen Aufwand der Kartographierung und Nachdigitalisierung von fehlenden Objekten zu reduzieren und eine verbesserte Genauigkeit bei der Informationsgewinnung zu erzielen.

Der Schwerpunkt wurde in Abstimmung mit dem projektbegleitenden Arbeitskreis auf die Ableitung von Informationen für die Lärmkartierung durch das EBA gelegt und die Auswahl der zugrundeliegenden ML-Methode anhand der Anforderungen des EBA priorisiert. Dies beinhaltet die Identifizierung von SSW, idealerweise einschließlich ihrer Höhe und materiellen Eigenschaften. Es ist jedoch wichtig, bestimmte Einschränkungen durch die verfügbaren Daten anzuerkennen, sodass die in Tabelle 2 zusammengefassten funktionalen und nicht-funktionalen Systemanforderungen definiert wurden.

Tabelle 2: Zusammenfassung der definierten funktionalen und nicht-funktionalen Systemanforderungen

Funktionale Anforderungen
☐ **Objekterkennung und -klassifizierung:** Das System muss Infrastrukturobjekte wie SSW identifizieren und klassifizieren können.
☐ **Lokalisierung und Geometrieschätzung:** Es sollen die Position und die geometrischen Eigenschaften der erkannten Infrastrukturobjekte geschätzt werden.
☐ **Genauigkeitsbewertung:** Die Genauigkeitsbewertung von SSW im Umkreis von 200 m entlang der Gleise hat Priorität, wobei die Herausforderung der Materialerkennung berücksichtigt werden muss.
☐ **Inferenz:** Die Inferenzzeit ist nicht der limitierende Faktor, aber unter Berücksichtigung der Rechenbeschränkungen sollte das Modell möglichst auf ganz Deutschland angewendet werden können.
☐ **Benutzeroberfläche:** Den Nutzenden soll Zugriff auf den Prozessierungs-Workflow und die erkannten und lokalisierten Infrastrukturobjekte über ein Plug-In für die Open-Source-Software QGIS ermöglicht werden.
☐ **Anpassungsfähigkeit:** Das System muss in der Lage sein, neuere, aktualisierte und möglicherweise erweiterte Versionen der ausgewählten Datensätze zu verarbeiten, solange sich deren Struktur nicht verändert hat.
Nicht-funktionale Anforderungen
☐ **Ausführungszeit:** Der Verarbeitungsworkflow, der neuere Versionen der verwendeten Datenquellen verwendet, muss auf einem Server ohne Grafikkarte, aber mit ausreichenden Rechenressourcen (RAM, CPUs, Festplattenspeicher), oder einem durchschnittlichen persönlichen Laptop ausgeführt werden können.
☐ **Sicherheit und Datenschutz**: Es wird sichergestellt, dass die Datenverarbeitungs- und Datenhandhabungsmethoden den relevanten Datenschutzbestimmungen und -vorschriften entsprechen.
☐ **Benutzerfreundlichkeit:** Das QGIS-Plug-In muss benutzerfreundlich sein und lediglich eine minimale Schulung für Nutzende erfordern, um effektiv damit arbeiten zu können.
☐ **Dokumentation:** Eine umfassende Dokumentation muss zur Verfügung gestellt werden, einschließlich Benutzerhandbüchern und technischer Dokumentation, um die Übernahme, Einrichtung und Nutzung der entwickelten Methodik durch relevante Interessengruppen zu unterstützen.

Aufbauend auf der Anforderungsanalyse konzentriert sich das folgende Kapitel auf die Datenverfügbarkeit und vorhandene Literatur zum maschinellen Lernen sowie verwandte Projekte. Es werden die mit der Datenerfassung verbundenen Herausforderungen und Chancen untersucht, Fusionstechniken zur Verbesserung der Kartierungsgenauigkeit erläutert, relevante ML-Literatur überprüft und die vorgeschlagene ML-Methodik zur Identifizierung von gleisnaher Infrastruktur präsentiert.

3 Literaturrecherche

3.1 Verfügbare Datenquellen

Ein erster Schritt bei der Entwicklung einer datenbasierten Fernerkundungslösung besteht darin, die verfügbaren Datensätze zu erkunden, bestehende Einschränkungen zu verstehen und enthaltene Informationen zu bewerten. In Tabelle 3 werden die betrachteten Hauptdatenquellen aufgeführt und deren möglicher Beitrag zur endgültigen Lösung hervorgehoben.

Zu den möglichen Datenquellen zählen unter anderem DOP, DOM und digitale Geländemodelle (DGM) sowie unterschiedliche Satellitendaten, die in Betracht gezogen werden. Satellitendaten bieten dabei einen hohen Aktualisierungszyklus, was für die Bestimmung von SSW von hoher Relevanz sein kann. Die Kombination von DOP und DOM sollte dabei die grundlegenden Informationen bereitstellen, eine multispektrale Darstellung der Szene anhand der DOPs und Höhenkarten basierend auf dem DOM. Dadurch sollte eine ML-Lösung ermöglicht werden, welche relevante Objekte anhand der multispektralen Darstellung segmentiert und die Höhenkarten abfragt. Die anderen Datenquellen, wie Geodaten der DB InfraGO AG sowie OpenStreetMap Foundation, sollten in erster Linie zur Bereitstellung von Annotationen für Beispiele von Infrastrukturelementen dienen. Zusätzlich sollen DGM verwendet werden, um in Kombination mit dem DOM die relative Höhe zum Boden, zu berechnen.

Darüber hinaus war es zur Gewährleistung der Zuverlässigkeit der Szeneninformationen, d. h. aller bereitgestellter Informationen für einen bestimmten Patch, ob visuelles RGB, multispektral oder anderweitig, wesentlich, dass in den Bildern die gewünschten Infrastrukturobjekte von Spezialisten mit Fachkenntnissen in der Interpretation von Fernerkundungsbildern gekennzeichnet werden. Im Rahmen dieses Projekts war die primäre Quelle für die Kennzeichnung des Datensatzes die von der DB InfraGO AG bereitgestellte Datenbank. Die Datenbank liefert Geometrien und Attributinformationen relevanter Infrastrukturelemente in der Nähe der Bahngleise. Diese Geometrien dienten als gekennzeichnete Masken, die mit den entsprechenden Patches im fusionierten gerasterten Datensatz abgeglichen werden können. Ein Patch ist definiert als ein abgegrenzter Bereich eines Datensatzes, der als eigenständige Einheit betrachtet wird und für verschiedene Analysezwecke verwendet werden kann.

Tabelle 3: Mögliche für Behörden zugängliche und öffentliche Datensätze, deren Quelle, Beschreibung, Auflösung und Aktualisierungszyklen

Datenbezeichnung	Quelle	Beschreibung	Auflösung	Zugänglichkeit	Aktualisierungszyklen
Digitales Oberflächenmodell (DOM)	Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland	Oberflächenhöhendaten für ganz Deutschland	1 m räumliche Auflösung in der Lage, 1 cm Auflösung in der Höhe und eine Kachelgröße von 5 km x 5 km	Behördliche Nutzungsvereinbarung mit BKG	Es ist ein Aktualitätszyklus von <= 3 Jahren definiert (AdV, 2021a)
Digitales Geländemodell (DGM)	Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland	Formen und Höhen der Geländeoberfläche für ganz Deutschland	10 m räumliche Auflösung pro Pixel und eine Patchgröße von 20 km x 20 km	Behördliche Nutzungsvereinbarung mit BKG	Es ist eine Grundaktualität von 10 Jahren definiert (Aktualitätsstand: 2021) (AdV, 2021b)
Digitale Orthophotos (DOP)	Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland	Zerrungsfreie und georeferenzierte Luftbilder für ganz Deutschland	Räumliche Auflösung von 400 cm^2 und eine Kachelgröße von 1 km x 1 km	Behördliche Nutzungsvereinbarung mit BKG	Es ist ein laufender Aktualitätszyklus von <= 3 Jahren definiert (AdV, 2021c).
Geodaten	OpenStreetMap Foundation	Markierte Objekte mit ihren Koordinaten in (Breitengrad, Längengrad)	-	Freier Zugang	Es ist ein laufender Aktualitätszyklus definiert.
Geodaten	DB InfraGO AG	3D-Form von relevanten Infrastrukturen zusammen mit dem Material	-	Zugriff über DZSF/EBA	Es ist ein laufender Aktualitätszyklus definiert. Das EBA erhält aktualisierte Daten für jede neue Runde der Lärmkartierung (Aktualitätsdefizit <= 5 Jahren).

MillionAID Dataset	„On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances and MillionAID.” (Long et al., 2020)	1 Million Instanzen für die Klassifizierung von Fernerkundungsszenen mit 51 semantischen Szenenkategorien.	Variable Auflösung	Freier Zugang	-
Functional Map of the World (fMoW) Dataset	"Functional map of the world." (Christie et al., 2018)	1 Million multispektrale Satellitenbilder aus über 200 Ländern mit 63 semantischen Kategorien	Variable Auflösung	Freier Zugang	-
Copernicus Satellitendaten	European Space Agency, 2024	Hochauflösende Radar- oder Bilddaten von polar-umlaufenden Satelliten.	Variable Auflösung	Freier Zugang	Häufiger Aktualisierungszyklus definiert, z. B. Sentinel-2: 5 Tage.
Geodaten	Kommunen/ zentrale Landesstellen	Berichte von Freiwilligen über Infrastrukturobjekte	-	Zugriff über DZSF/EBA	Aktualisierungszyklen unbekannt.

3.2 Relevante Literatur

Die vorliegende ML-Herausforderung kann als semantische Segmentierungsaufgabe formuliert werden. Die semantische Segmentierung ist ein grundlegendes Problem des Forschungsbereichs Computer Vision. Hierbei wird jedem Pixel in einem Bild ein aussagekräftiges Label zugewiesen. Sie spielt eine wichtige Rolle in verschiedenen Anwendungen wie autonomes Fahren, Szenenverständnis und medizinische Bildanalyse. In den letzten zehn Jahren hat die semantische Segmentierung signifikante Fortschritte verzeichnet, die durch die Verfügbarkeit von umfangreichen annotierten Datensätzen, leistungsstarken Deep-Learning-Architekturen und erhöhten Rechenressourcen vorangetrieben wurden. Diese Literaturübersicht soll einen umfassenden Überblick über moderne Techniken, Herausforderungen und jüngste Fortschritte in der semantischen Segmentierung geben, wobei ein spezieller Fokus auf ihrer Relevanz in Anwendung mit Datensätzen mit wenigen Annotationen liegt.

Frühe Arbeiten im Bereich der semantischen Segmentierung haben stark auf Convolutional Neural Networks (CNNs) (Long et al., 2020; Wang et al.2019; Chen et al., 2016; Long et al., 2014) zurückgegriffen. CNNs sind im Bereich Computer Vision weit verbreitet, da sie eine starke induktive Annahme für die Bildmodellierung bieten, nämlich Lokalität, Translationsäquivalenz und Translationsinvarianz. Ihr Erfolg bei dichten Vorhersageaufgaben wie semantischer Segmentierung wird ihrer Struktur zugeschrieben, welche sich in mehrere Schichten (Layer) unterteilt. Diese liefert hierarchische Repräsentationen, bei denen frühe Schichten lokale Muster wie z. B. Kanten ausgeben und tiefere Schichten globale semantische Repräsentationen der Szene liefern. CNNs sind zudem dateneffizient, da sie in der Lage sind, bedeutungsvolle Merkmale aus kleineren Datenmengen zu extrahieren. CNNs und Neural Networks (NNs) generell sind in der Regel so aufgebaut, dass sie aus einem sogenannten Backbone und spezialisierten, aufgabenspezifischen Schichten, die darauf aufbauen, bestehen. Das Backbone ist dabei der Kern des Modells, in dem die meisten Berechnungen stattfinden. Oft werden Backbones während eines Trainingsprozesses nicht weiter trainiert, aber mit den Gewichten eines vorherigen Trainingsprozesses initialisiert, man spricht dabei von Transfer Learning. Für die Initialisierung sind dann nur eine geringere Anzahl von Annotationen notwendig, da die Backbones jedoch bereits trainiert sind.

ResNets (Ronneberger et al., 2015) sind eine häufige Wahl für Backbones von auf CNN basierenden Segmentierungsnetzwerken. Dies liegt an ihrer Fähigkeit, mit der Tiefe des Netzwerks gut zu skalieren und weiterhin trainierbar zu sein. Somit wird eine Vorabtrainierung (Pre-Training) auf großen Datensätzen ermöglicht und hochwertige Merkmale für die nachgeschaltete Segmentierungsaufgabe bereitgestellt. CNNs können jedoch Probleme haben, globale Beziehungen in einem Bild zu modellieren, da die verwendete Rechenoperation lokal ist und ihr rezeptives Feld, also die Größe des Bildausschnitts, welches zur Ausgabe der Operation beiträgt, dadurch eingeschränkt ist. In CNNs kann das rezeptive Feld i. d. R. nur mit der Tiefe des Netzwerkes wachsen, was mit einem Verlust an Auflösung durch Pooling-Schichten einhergeht. In den Pooling-Schichten werden die stärksten Merkmale herausgefiltert und die schwächeren Merkmale entsprechend verworfen. Das reduziert die Datenmenge deutlich und die Verarbeitung wird dadurch beschleunigt (Long et al., 2014). Dies beschränkt die Fähigkeit von CNNs, Informationen zwischen verschiedenen Bereichen des Bildes auszutauschen, was ihre Fähigkeit zur Durchführung dichter Vorhersagen mit hoher Auflösung einschränkt (He et al., 2015).

In jüngster Zeit wurde der Aufmerksamkeitsmechanismus (Attention Mechanism), (Ranftl et al., 2021), der ursprünglich im Zusammenhang mit maschineller Übersetzung eingeführt wurde, zur Lösung von Computer Vision-Problemen angepasst und die Vision Transformer-Architektur (ViT) entwickelt (Vaswani et al., 2017). In einem ViT wird ein Bild in nicht überlappende Patches aufgeteilt, und jedem Patch wird ein Vektor zugewiesen, abhängig von den Pixeln, die sich im Patch befinden. Dieser Prozess wird Tokenisierung genannt. Der Tokenisierungsprozess ist der einzige Teil in einem ViT, bei dem ein Auflösungsverlust auftritt, abhängig von den gewählten Patchgrößen. Trotz der Tokenisierung kann man die Informationen in einem ViT erhalten, indem man den Informationsaggregationsprozess über die Pixel in

einem Patch sorgfältig gestaltet (He et al., 2015). Ab diesem Punkt behält die ViT-Architektur bei zunehmender Tiefe eine konstante Auflösung bei, während sie in globalen Aufmerksamkeitsschichten (Attention Layers) Beziehungen zwischen allen Patches modelliert. Es wurde gezeigt, dass die auf der ViT basierende Architektur zu höher auflösenden semantischen Masken führt, insbesondere bei großen Datensätzen (He et al., 2015). Die ViT-Architektur übertrifft die Performanz von CNNs vor allem dann, wenn viele Daten verfügbar sind, da sie weniger induktive Annahmen haben und damit flexibler in Bezug auf die Erfassung von Informationen sind. Der Nachteil der Transformer-Architektur besteht jedoch darin, dass globale Aufmerksamkeitsschichten eine quadratische Rechenkomplexität mit der Menge der Eingabetoken aufweisen. Dies macht es rechnerisch unlösbar, hochauflösende Bilder in großen Mengen zu verarbeiten, was für Satelliten- und Befliegungsdaten problematisch ist. Diese Einschränkung hat zu spezialisierten Architekturen (Dosovitskiy et al., 2020a) (Liu et al., 2021) geführt, die dieses Problem durch begrenzte Aufmerksamkeitsschichten angehen und den Umfang der zu bearbeitenden Pixel begrenzen. Diese spezialisierten Architekturen nutzen kurze Attention Spans[2], die die rechnerische und speicherbezogene Komplexität nur linear mit der Länge der Tokens wachsen lassen, während gleichzeitig eine globale Attention Mixing für relationales Modellieren zwischen Token über große Distanzen gewährleistet wird. Die ViT-Architektur hat sich in semantischen Segmentierungsaufgaben (He et al., 2015) (Li et al., 2022; Cheng et al., 2021; Xie et al., 2021) als erfolgreich erwiesen und kann aufgrund der Vielzahl effizienter Implementierungen des Attention Mechanism hochauflösende Fernerkundungsbilder verarbeiten (AlMarzouqi und Saoud, 2022; Wang et al., 2021b; Yamazaki et al., 2023; Li et al., 2020).

Die begrenzte Verfügbarkeit von Annotationsdaten (Label) ist ein häufiges Problem bei realen Datensätzen, die oft aus einer hohen Anzahl an nicht-annotierten Daten bestehen. Dieses Problem ist bei semantischen Segmentierungsaufgaben noch verstärkt, da zur Markierung eine pixelgenaue Annotation erforderlich ist. Insbesondere bei Fernerkundungsdaten erfordert dies viel Domänenwissen sowie erheblichen Zeit- und Ressourcenaufwand. Es stellt sich die Frage, wie riesige Mengen an Daten ohne Annotationen genutzt werden können, um leistungsstarke Modelle mithilfe einer begrenzten Anzahl gekennzeichneter Beispielobjekte abzuleiten. In letzter Zeit wurde das kontrastive Lernen (Contrastive Learning) genutzt (Wang, et al., 2021a), welches zu dem Gebiet der selbstüberwachten Trainingsmethoden (Self-supervised Learning) zählt. Hierbei werden nicht-annotierte Bilder manipuliert (beispielsweise gespiegelt, Farbwerte verändert), um Paare zu bilden, deren Repräsentationen ähnlicher sein sollten als die Repräsentationen von unterschiedlichen Bildern. Obwohl kontrastives Lernen sehr effektiv ist, ist es schwierig den Trainingsprozess zu stabilisieren. Hard-Negative-Sample-Mining (Chen et al., 2023) und ähnliche Techniken sind erforderlich, um zu verhindern, dass das Netzwerk in einen Modus von konstanten Repräsentationen verfällt.

Ein alternativer Ansatz ist Bootstrap Your Own Latent (BYOL) (Robinson et al., 2020), bei dem keine negativen Samples erforderlich sind und das Netzwerk darauf abzielt, ein Bild und seine augmentierte Ansicht abzugleichen. Dafür werden zwei Netzwerke verwendet, ein Zielsystem und ein Online-Netzwerk, wobei das Online-Netzwerk das Zielsystem in einer Student-Teacher-Methode destilliert und dessen Ausgabe als Trainingssignal verwendet wird. Aus diesem Prinzip sind sehr leistungsstarke Modelle hervorgegangen, wie z. B. Distributed Instance-wise Recognition (DINO) (Grill et al., 2020; Caron et al., 2021, Oquab et al., 2023a), welche leistungsstarke Repräsentationen für Bilder erzeugt, die für nachgelagerte Aufgaben feinabgestimmt (fine-tuned) werden können. Ein Nachteil der mit BYOL trainierten Modelle ist jedoch, dass sie zwar besonders für die Klassifikation von ganzen Bildern geeignet sind, aber keine unterteilbaren Objektrepräsentationen aufweisen, da dies während des Trainings nicht gefördert wird. Intuitiv ist nicht zu erwarten, dass während des Trainings nur auf der Ebene der Bildrepräsentation objektspezifische Repräsentationen erzeugt werden können.

[2] Im Kontext des maschinellen Lernens bezieht sich Attention Span darauf, wie gut ein Modell in der Lage ist, sich selektiv auf relevante Informationen in einer gegebenen Eingabesequenz zu konzentrieren, wodurch es komplexe Abhängigkeiten besser verarbeiten kann und gute Leistung bei anspruchsvollen Aufgaben mit komplexen Daten zeigt.

Eine Verbesserung gegenüber mit BYOL trainierten Modellen ist die Methode des Object Discovery and Representation Networks (ODIN)[3], welches von Hénaff et al. (2022) eingeführt wurde. In ODIN wird ein selbst-überwachtes Trainingssignal auf Objektebene mithilfe unterschiedlicher Bildaugmentierungen verwendet. Auch Hénaff et al. (2022) haben diese Idee für die Entwicklung von Self-supervised Transformer with Energy-based Graph Optimization (STEGO)[4] genutzt, um bei vortrainierten Modellen, die während des Trainings keine objektspezifische Trennung erzwangen, die Objektseparabilität durchzusetzen, um Feature-Korrespondenzen[5] explizit zu verfeinern und zu vergrößern. Ein Problem von auf BYOL basierenden Methoden besteht darin, dass sie aufgrund ihres hohen Speicher-, Rechen- und Datenbedarfs besonders aufwändig zu trainieren sind. Eine Alternative, die aus der natürlichen Sprachverarbeitung übernommen wurde, ist das Masked-Auto-Encoding (MAE) (Hamilton et al., 2022), bei dem das Netzwerk etwa 75 % des Bildes maskiert und versucht, es aus den verbleibenden 25 % wiederherzustellen. Diese Methode schafft ein Gleichgewicht zwischen der Überwachung auf Bild- und Objektebene, indem auf der Ebene des Bild-Patches gearbeitet wird.

Trotz des vielversprechenden selbstüberwachten Lernens leistungsstarker gelernter Repräsentationen ohne Annotationen ist das Trainieren eines Segmentierungsmodells ein äußerst daten- und ressourcenintensiver Prozess. In jüngster Zeit hat sich der Fokus der Forschung für ML von der Entwicklung kleiner spezialisierter Netzwerke zu größeren und vielseitigeren Modellen hin entwickelt, die sich durch ihre starken Repräsentationen besonders gut als Backbone für eine Vielzahl an Aufgaben eignen. Sie werden auch oft als Foundation-Modelle bezeichnet, wobei der Begriff noch nicht klar abgegrenzt ist. Solche Foundation-Modelle können als Grundgerüst für verschiedene nachgelagerte Aufgaben dienen, ohne dass ihre Repräsentationen stark verändert werden müssen. Foundation-Modelle für leistungsstarke Repräsentationen wurden für szenenunabhängige RGB-Bilder (Grill et al., 2020; Caron et al., 2021; Hamilton et al., 2022), für RGB-Luft- und Satellitenbilder (He et al., 2021; Cha et al., 2023; Sun et al., 2022) sowie für multispektrale Luftbilder (Wang et al., 2022) entwickelt. Diese Modelle werden entweder mit MAE oder BYOL trainiert, jedoch nicht explizit für Segmentierungsaufgaben. In jüngster Zeit wurden das Segment Anything-Modell (SAM) (Ke et al., 2023) und seine hochauflösendere Variante (Kirillov et al., 2023) als Foundation-Modelle für die allgemeine Segmentierung eingeführt. SAM gibt bei einer Eingabe in Form von Text, Pixelposition oder einer Bounding Box drei plausible hochwertige Masken des angewiesenen Objekts in der Szene aus. Diese eine-zu-viele-Beziehung zwischen der Anweisung und den Masken ermöglicht es dem Modell, mit Mehrdeutigkeit umzugehen, da eine Anfrage mehreren Benutzerintentionen entsprechen kann: Die Abfrage nach dem Objekt, zu dem ein einzelner Punkt gehört, könnte beispielsweise als Antwort ein T-Shirt, den Menschen, der das T-Shirt trägt, oder das Auto, in dem der Mensch sitzt, umfassen.

Foundation-Modelle, die sich darauf konzentrieren, leistungsstarke Repräsentationen aus Luft- und Satellitenbildern zu extrahieren, könnten theoretisch mit einem Segmentierungskopf (z. B. UPerNet (Ke et al., 2023)) feinabgestimmt werden. Ein Vorteil wäre, dass das Modell auf die Verteilung der Eingabedaten abgestimmt wäre. Dies würde jedoch voraussichtlich Millionen von gekennzeichneten Bildern erfordern, um die Leistung von SAM zu erreichen.

SAMs Backbone hingegen basiert auf einem mit MAE trainierten ViT mit fensterbasierter Aufmerksamkeit und globalen Mischschichten (Mixing Layers). Dieses Backbone wurde für eine Segmentierungsaufgabe feinabgestimmt, bei der die Merkmale des Backbones durch einen Maskendekodierer (Masked Decoder) geleitet werden, der eine Modifikation der Architektur des in Detection Transformer[6] (DETR) verwendeten Bounding Box Decoder (deutsch: Dekodierer für Begrenzungsrahmen) darstellt. Im Masked Decoder werden die Anfrage und die Repräsentationen fusioniert, und es werden vier Ausgabemasken (sortiert

[3] Deutsch: Objekterkennung- und Repräsentationsnetzwerke

[4] Deutsch: Selbstüberwachter Transformator mit energiebasierter Graphenoptimierung

[5] Im Kontext von Computer Vision bezieht sich Feature-Korrespondenz auf das Zuordnen von charakteristischen Punkten oder Regionen in mehreren Bildern, die demselben realen Objekt oder Szenario entsprechen.

[6] Detection Transformer Architektur, eine Transformer-Architektur, die ein ViT als Grundlage verwendet.

nach ihrer Intersection over Union[7] (IoU) erzeugt. Diese eine-zu-viele-Beziehung zwischen der Anweisung und den Masken ermöglicht es dem Modell, gegenüber Mehrdeutigkeit robust zu sein. Es ist möglich, SAM mit einem Rastergitter aus Anfragepunkten auf Luftbildern anzuwenden und redundante Masken durch Nicht-Maxima-Unterdrückung zu beseitigen. Dies liefert eine Menge Segmentierungsmasken. SAM weist jedoch den Ausgabemasken keine semantischen Labels zu. Daher ist eine zusätzliche Trainingsebene erforderlich, die SAM zur semantischen Zuordnung hinzugefügt werden muss, um die Information zu erhalten, um welche Art von Objekten es sich bei einer ausgegebenen Segmentierungsmaske handelt. Ein weiteres Problem besteht darin, dass SAM nicht auf Fernerkundungsdaten trainiert wurde. Diese Daten sind daher nicht zwangsläufig Teil der Verteilung der Trainingsdaten, in diesem Fall würde man von Out-of-Distribution-Daten sprechen. Daher muss die Leistung auf den für das Projekt relevanten Daten sorgfältig bewertet werden.

Eine Möglichkeit, um das Out-of-Distribution-Problem anzugehen, ist, SAM auf einer Teilmenge annotierter Fernerkundungsdaten feinabzustimmen. Dies ist nicht nur rechenintensiv und erfordert eine große Menge an fein granular gelabelten Daten, sondern ist auch anfällig für katastrophales Vergessen, wenn in einer Schicht erstellte Verbindungen in der nächsten wieder überschrieben werden (Xiao et al., 2018; Carion et al., 2020). Ein Versuch, die Rechen- und Datenanforderungen zu bewältigen, könnte der Einsatz von Low-Rank-Adaptern[8] (Kirkpatrick et al., 2017) sein. Durch die Feinabstimmung von SAM in dieser Weise kann jedoch keine Integration von multispektralen Kanälen ermöglicht werden, da der ViT-Backbone nur RGB-Kanäle unterstützt.

Eine weitere relevante Adapterarchitektur für dichte Vorhersagen mit Hilfe von SAM oder anderen Foundation-Modellen wurde in Hu et al. (2021) eingeführt, die aus drei Hauptkomponenten besteht: einem räumlichen Vorab-Modul (Spatial Prior Module), einem räumlichen Merkmalsinjektor (Spatial Feature Injector) und einem räumlichen Merkmalsextraktor (Spatial Feature Extractor). Das Spatial Prior Module enthält ein Backbone, mit dem domänenspezifische Merkmale extrahiert werden können. Im Kontext dieses Projekts könnte dies ein auf Fernerkundungsbildern trainiertes Foundation-Modell sein. Der Spatial Feature Injector nimmt die aus dem ViT extrahierten Merkmale an einem bestimmten Block und führt eine Cross-Attention[9] mit den Adaptertokens durch, wobei die Adaptertokens als Schlüssel und Werte dienen und die Merkmale des ViT als Abfrage dienen. Dadurch werden die Informationen des ViT-Backbones des Foundation-Modells und des Adapters fusioniert und es wird ein Merkmal gleicher Größe wie das ViT-Merkmal zurückgegeben. Diese neuen Merkmale werden dann durch den nächsten Transformer-Block im ViT-Backbone weiterverarbeitet. Der Spatial Feature Extractor nimmt dann die verarbeiteten Merkmale aus dem ViT und verwendet sie als Schlüssel und Wert, während die Adaptertokens als Abfrage dienen. In dieser Schicht werden die Tokens des Adapters mit den Merkmalen des ViT aktualisiert. Dadurch wird der Zustand des Adapters erweitert. Mit dieser Konfiguration kann erfolgreich Information in und aus dem vortrainierten Modell weitergeleitet werden, was die Integration von multi-spektralen Bildinformationen in Foundation-Modellen wie SAM oder anderen Modellen ermöglichen könnte. Die drei Hauptkomponenten – Spatial Prior Module, Spatial Feature Injector und Spatial Feature Extractor – sind anpassungsfähig und ermöglichen die Integration von domänenspezifischen Merkmalen aus verschiedenen vortrainierten Modellen. Auch andere einfachere Adaptertechniken haben sich als erfolgreich erwiesen (Chen et al., 2022; Chen et al., 2023), bei denen die Autoren einen ViT-Backbone angepasst haben, um eine Feinobjektsegmentierung durchzuführen. Das Prinzip ist jedoch das gleiche.

[7] Intersection over Union (Schnitt über Vereinigung), eine Metrik in der Computer Vision, die die Überlappung zwischen den vorhergesagten und den tatsächlichen Begrenzungsrahmen oder Segmentierungsmasken bewertet.

[8] Low Rank Adapter sind Komponenten, die einem großen Netzwerk hinzugefügt werden können und effizientes fine-tuning ermöglichen.

[9] Im Kontext des maschinellen Lernens bezieht sich Cross Attention auf einen Aufmerksamkeitsmechanismus, der es einem Modell ermöglicht, sich gleichzeitig selektiv auf verschiedene Teile von zwei oder mehr Eingabesequenzen zu konzentrieren, wodurch der Informationsaustausch erleichtert wird und die Leistung bei verschiedenen Aufgaben verbessert wird.

3.3 Relevante öffentliche Projekte

Die steigende Nachfrage nach effizienten und nachhaltigen Transportsystemen hat zu einer Vielzahl von Projekten geführt, die sich auf die Extraktion von gleisnahen Infrastrukturobjekten konzentrieren. Diese Projekte zielen darauf ab, die Sicherheit des Schienenverkehrs zu verbessern, die Vegetationspflege zu optimieren und Veränderungen in der Landnutzung zu überwachen, um nur einige Ziele zu nennen. Dieses Kapitel gibt einen Überblick über aktuelle Projekte in diesem Bereich und hebt die verwendeten ML-Methoden, Innovationen und Herausforderungen hervor.

Übersicht der Projekte. In den letzten Jahren wurden mehrere Projekte zur Extraktion von Infrastrukturobjekten aus Satelliten- und Befliegungsdaten im Schienenumfeld durchgeführt, darunter:

- "Ableitung des Baumbestandes entlang des deutschen Schienennetzes" (Frick et al., 2021),
- "SENSchiene – Satellitengestützte Erfassung von Flächeneigenschaften und Nutzungsveränderungen im Umfeld des Verkehrsträgers Schiene" (Preußler et al., 2024),
- "Anforderungskatalog für eine webbasierte Plattform zur Bereitstellung, Darstellung und Analyse von Geodaten – mHUB-B" (Hasberg et al., 2021),
- "safe.trAIn: Sichere Künstliche Intelligenz am Beispiel fahrerloser Regionalzug" (Siemens AG, 2023),
- „Grün an der Bahn – Wie die DB Bäume und Sträucher an ihren Strecken pflegt “ (Deutsche Bahn AG, 2023),
- „Computer Vision – Transformationsprozess in eine digitale und effizientere Bankwelt“ (DB E.C.O. Group, 2022)

Eingesetzte ML-Methoden. Die Projekte setzen unterschiedliche ML-Methoden ein, um ihre Ziele zu erreichen. Beispielsweise verwendet das Projekt "Ableitung des Baumbestandes entlang des deutschen Schienennetzes" eine Wasserscheidentransformation kombiniert mit Entscheidungsbäumen zur Bewertung des Risikos von Baumstürzen auf Bahngleise. Das im Projekt entwickelte GIS-Tool kann Berechnungen für verschiedene Bereiche entlang des deutschen Schienennetzes durchführen und berücksichtigt dabei Faktoren wie Baumhöhe, Abstand zur Infrastruktur, Windgeschwindigkeit und Baumtyp.

"safe.trAIn" entwickelt eine Methodik zur Integration von KI und Sicherheitsaspekten in fahrerlosen Zügen, während das Projekt "Grün an der Bahn" sich auf den Einsatz von Satellitendaten und KI zur Erfassung der Vegetation konzentriert. Das Digitalisierungsprojekt "Computer Vision – Transformationsprozess in eine digitale und effizientere Bahnwelt" zielt darauf ab, eine Lösung zur visuellen Objekterkennung entlang der Gleise zu entwickeln.

Forschungsbeitrag der Projekte. Die Projekte zeigen bereits bemerkenswerte Innovationen und Fortschritte auf. Z. B. hat das Projekt "Ableitung des Baumbestandes entlang des deutschen Schienennetzes" ein GIS-Tool zur Einzelbaumdetektion entwickelt. Das Projekt "SENSchiene" nutzt frei verfügbare Satellitendaten zur automatischen Erfassung von Gebiets- und Streckenmerkmalen im Schienenbereich. "mHUB-B" hat eine gemeinsame Dateninfrastruktur für die Verkehrspolitikbewertung eingerichtet, während "safe.trAIn" Sicherheit und Transparenz von KI-Algorithmen in autonomen Zügen verbessert.

Die Projekte "Grün an der Bahn" und "Computer Vision – Transformationsprozess in eine digitale und effizientere Bahnwelt" demonstrieren die Kraft der Kombination von Satellitentechnologie und KI zur Verringerung der Störanfälligkeit des Bahnbetriebs, indem sie präzise Informationen über die Vegetation liefern und eine umfassende Überwachung der Bahninfrastruktur ermöglichen.

Zusammenfassend haben die genannten Projekte bedeutende Fortschritte und Innovationen in der automatisierten digitalen Extraktion von Infrastrukturobjekten entlang von Bahngleisen geleistet und sind für

dieses Forschungsvorhaben auf vielfältige Weise von Relevanz. So setzt ein Großteil der genannten Projekte erfolgreich auf den Einsatz von Befliegungs- und Satellitendaten und zeigt, dass diese durchaus geeignet sind, um gleisnahe Infrastruktur zu erkennen. Von traditionellen Algorithmen der visuellen Bildverarbeitung bis zu neuesten technologischen ML-Ansätzen kommen diverse Technologien zum Einsatz. Das Projekt „safe.trAIn" z. B. evaluiert und entwickelt moderne ML-Methoden zur Analyse gleisnaher Infrastruktur, u. a. bei wenigen bzw. gänzlich fehlenden Annotationen. Einige Projekte stellen neben einem Forschungsbericht auch einsetzbare GIS-Tools zur Verfügung, um den Einsatz und die Reproduzierbarkeit der Ergebnisse zu vereinfachen.

Dennoch konzentrieren sich diese Projekte in erster Linie auf einen spezifischen Teilbereich, wie die ausschließliche Entwicklung einer ML-Methodik. Im Gegensatz dazu war das Ziel dieses Projekts, eine ganzheitliche Lösung zu entwickeln. Diese sollte von der Aufbereitung der Daten, über den Einsatz moderner ML-Modelle, bis hin zur Bereitstellung eines einsetzbaren Plug-Ins reichen. Dabei sollen aus vorangegangenen Projekten erarbeitete Ansätze kombiniert, auf die Erkennung von SSW übertragen und darüberhinausgehend erweitert werden.

4 Fazit zur Anforderungsanalyse und Literaturrecherche

Das folgende Kapitel beschreibt die initial vorgesehene Methodik bezüglich der ML-Lösung und eine mögliche Evaluierung dieser. Darüber hinaus werden die Datengrundlage und die allgemeine Systemarchitektur der geplanten Lösung dargelegt. Die folgenden Annahmen und der entsprechende Plan resultieren dabei als Ergebnis aus der Anforderungsanalyse und Literaturrecherche im Rahmen des ersten Arbeitspakets. Eine detaillierte Beschreibung der Umsetzung und entsprechenden Herausforderungen sowie Anpassungen im Zuge dessen werden in den folgenden Kapiteln 5 bis 7 beschrieben.

4.1 Vorgeschlagene ML-Methodik

Das vorherige Kapitel verdeutlicht, dass es viele verschiedene Ansätze für die ML-Lösung in diesem Projekt gibt. In den letzten Jahren wurde eine Vielzahl an Modellen veröffentlicht und zugänglich gemacht, welche auf großen Datensätzen selbstüberwacht trainiert wurden. Deswegen sollte im weiteren Vorgehen auf ein leistungsstarkes, bereits trainiertes Foundation-Modell aufgebaut werden. Da Foundation-Modelle häufig auf der Transformer-Architektur basieren und auf riesigen Datenmengen vortrainiert sind, ist die Verwendung dieser Architektur aufgrund ihrer überlegenen Leistung zu bevorzugen. Die verfügbaren Foundation-Modelle liefern bereits nützliche Repräsentationen, welchen einen Lösungsansatz für das vorliegende Problem liefern. Deshalb wurde nicht geplant, ein Modell von Grund auf mit einer selbstüberwachten Methodik wie z. B. ODIN zu trainieren. Falls nur wenige Datenpunkte verfügbar gewesen wären, wäre die CNN-Architektur aufgrund ihrer Dateneffizienz die bessere Wahl.

Auf der Grundlage der bereits vortrainierten Transformer-basierten Foundation-Modelle wurde beabsichtigt, eine annotations-effiziente Feinanpassung oder ein Zero-Shot-Prompting[10] durchzuführen. In Tabelle 4 sind Beispiele für relevante zugrunde liegende Modelle (Backbones) aufgeführt, die im Rahmen der Modellentwicklung evaluiert werden sollten (siehe Kapitel 6.1).

Tabelle 4: Beispiele für relevante zugrunde liegende Modelle (Backbones) und entsprechende Lizenzierung

Modell	**Beschreibung**	**Lizenzierung**
Segment Anything Model	Foundation-Modell für die allgemeine RGB-Bildsegmentierung	Zulässige Verwendung (Apache License, Version 2.0)
DINO-trained ViT	Foundation-Modell für allgemeine RGB-Bildrepräsentationen	Zulässige Verwendung (Apache License, Version 2.0)
Remote Sensing MAE	Foundation model für RGB-Luftbildrepräsentationen	Zulässige Verwendung (MIT License)

Kombinationen der bereitgestellten Architekturen sind durch den Einsatz von Adaptern möglich. Zunächst sollten die Zero-Shot-Fähigkeiten der Backbones bewertet und bei Bedarf gegebenenfalls nachtrainiert (fine-tuning) werden.

[10] Zero-Shot-Prompting ist eine Methode, bei der das Modell ohne vorheriges Training für eine bestimmte Aufgaben abgefragt wird, indem eine sorgfältig konstruierte Anfrage spezifiziert wird.

Alle Modelle weisen jedoch den Nachteil auf, dass sie den segmentierten Objekten keine semantische Bedeutung zuweisen. Um dieses Problem zu adressieren, sollen ähnlichkeitsbasierte Suchtechniken anhand weniger Exemplare der relevanten semantischen Klassen bewertet werden. Eine weitere Möglichkeit besteht in einer Feinanpassung (fine-tune Training) der letzten Schichten des Modells. Um jedoch ein maschinelles Modell trainieren zu können, muss ein Zugang zu Annotations- daten für die relevanten Objekte in der Szene vorhanden sein. Da die Backbones jedoch bereits trainiert sind, ist nur eine geringe Anzahl an Annotationen notwendig. An dieser Stelle wird angenommen, dass der Datensatz der DB InfraGO AG zu Infrastrukturobjekten im Schienenumfeld (siehe Tabelle 3) hierzu ausreichend ist. Sollte diese Hypothese nicht zutreffen, sollten gegebenenfalls Daten von OpenStreetMaps (OSM) oder dem Geo-Portal für allgemeine Infrastrukturobjekte eingearbeitet werden. Annotationen können in Form von Polygonmasken oder Ankerpunkten und -linien vorliegen, falls Polygonmasken nicht realisierbar sind. Da die Auflösung der DOP-Daten höher ist als bei Satellitendaten, war nicht davon auszugehen, dass die Hinzunahme von Satellitendaten zu den DOM- und DOP-Daten einen Gewinn bei der Erkennung von Infrastrukturobjekten bringt. Es ist jedoch zu beachten, dass Satellitendaten, wie in Kapitel 3.1 aufgezeigt, in der Regel häufiger aktualisiert werden als die DOP- und DOM-Daten. Dadurch sollten potenziell häufigere Aktualisierungszyklen genutzt werden können, was eine verbesserte Erkennung von Veränderungen im Laufe der Zeit ermöglichen würde.

4.2 Evaluation und Qualitätssicherung

Zur Sicherstellung einer hohen Qualität und Genauigkeit bei der Verwendung und Kombination mehrerer Datenquellen sowie insbesondere im Zuge der ML-Modellentwicklung ist ein umfassender Evaluations- und Qualitätssicherungsprozess notwendig. Die Qualität des Daten-Fusion-Algorithmus spielt eine wesentliche Rolle für den Erfolg des nachgelagerten Machine-Learning-Modells. Daher sollte zur Beurteilung der Korrektheit des Daten-Fusion-Algorithmus die geometrische Genauigkeit stichprobenartig überprüft werden. Die geometrische Genauigkeit misst die Positionsgenauigkeit der Ausrichtung zwischen den DOM und den DOP. Eine geringe Stichprobenanzahl ist hierbei ausreichend, da bei Datensätzen mit korrekten Georeferenzdaten davon auszugehen ist, dass alle Koordinatenpunkte von einem Fehler betroffen wären. Bei der Stichprobenkontrolle wird eine geringe Menge an Annotationsobjekten ausgewählt, bei denen sowohl die Korrektheit der GPS-Position als auch das Vorhandensein des Objektes zum Zeitpunkt der Aufnahme der Orthophotos bekannt ist. Für diese Objekte soll manuell verifiziert werden, dass das Objekt visuell im fusionierten Datensatz an der erwarteten Stelle vorhanden ist und dass das Höhenprofil mit der visuellen Information im Einklang ist (siehe Kapitel 5.2).

Im Zusammenhang mit der Bewertung des Instanz-Segmentierungsmodells soll ein umfassender Ansatz angewendet werden, bei dem eine Kombination von Bewertungsmetriken zum Einsatz kommt (siehe Kapitel 6.3). Eine solche Metrik ist IoU, welche die Überlappung zwischen vorhergesagten Segmentierungsmasken und Ground-Truth-Masken für einzelne Objektinstanzen misst. Diese Metrik soll eine lokale Bewertung der Segmentierungsgenauigkeit sowie der Fähigkeit des Modells, Objekte genau abzugrenzen ermöglichen. Eine weitere wertvolle Metrik ist die Average Precision (AP), die sowohl die Lokalisationsgenauigkeit als auch die Segmentierungsqualität berücksichtigt. Durch die Berechnung von Präzisions- (Precision)[11] und Wiederkennungswerten (Recall)[12] bei mehreren IoU-Schwellenwerten kann die Leistung des Modells bei verschiedenen Überlappungsstufen bewertet werden. Diese Analyse soll Erkenntnisse über die Fähigkeit des Modells liefern, Objektinstanzen zu identifizieren – unter Berücksichtigung des Trade-offs zwischen Präzision und Wiedererkennung. Um eine umfassende Bewertung zu erhalten, kann die Mean Average Precision (mAP, Mittelwert der durchschnittlichen Genauigkeit) berechnet werden, in-

[11] Precision: Verhältnis von richtig positiven Instanzen zur Gesamtzahl der vorhergesagten positiven Instanzen (Saito und Rehmsmeier, 2015)

[12] Recall: Verhältnis der wahrhaft positiven Vorhersagen zur Gesamtzahl der tatsächlich positiven Instanzen (Saito und Rehmsmeier, 2015)

dem die AP-Werte über mehrere IoU-Schwellenwerte gemittelt werden. Diese Metrik bietet eine Gesamtbewertung der Leistung des Modells in Aufgaben zur Instanzsegmentierung und berücksichtigt Variationen in Objektgrößen, -formen und -komplexitäten. mAP liefert wertvolle Erkenntnisse über die allgemeine Leistung des Modells in verschiedenen Szenarien und bietet einen repräsentativen Wert, der dessen Effektivität zusammenfasst. Zusammenfassend soll die Kombination von IoU, AP und mAP als Bewertungsmetriken eine umfassende und detaillierte Analyse der Leistung eines Instanz-Segmentierungsmodells bieten. Diese Metriken bewerten sowohl die lokale Segmentierungsgenauigkeit als auch die Gesamtwirksamkeit und berücksichtigen verschiedene Überlappungsstufen, um fundierte Entscheidungen für die Optimierung und Verbesserung des Modells zu ermöglichen.

4.3 Angestrebte Systemarchitektur

Für den Arbeitsfluss der Nutzenden des Systems ist insbesondere das QGIS-Plugin, beschrieben in Kapitel 7, relevant. Dies erlaubt den Nutzenden das Segmentierungsmodell auszurufen und mit den detektierten Infrastrukturobjekten zu interagieren und trägt somit zur Anwendungsfreundlichkeit und Weiternutzung z. B. für die Lärmkartierung bei.

Da Inferenz, also die Vorhersage (in diesem Fall die Objektdetektion) mit Hilfe von trainierten Modellen sehr rechenintensiv ist und die Hauptdatenquellen (DOP20 und DOM1) nur in niedriger zeitlicher Frequenz aktualisiert werden (Zielwert: alle 36 Monate), kann das Ergebnis der Referenz für die relevanten Gebiete (200 m Distanz zu Eisenbahngleisen) vorberechnet werden. Dies verhindert unnötige redundante Berechnungen im Vergleich zu einem System, bei dem die Inferenz innerhalb des QGIS-Plugins durchgeführt wird. Es ist somit aus Umweltaspekten und Nutzersicht zu bevorzugen. Abbildung 3 veranschaulicht das System der Inferenz-Pipeline, welches bei einem Update der Eingangsdaten (DOP) einmalig angewendet werden muss, um den Datensatz der detektierten Infrastrukturobjekte zu aktualisieren. Dieser Datensatz von detektierten Infrastrukturobjekten im Schienenumfeld kann dann innerhalb des QGIS-Plugins, beschrieben in Kapitel 7, visualisiert und verwendet werden.

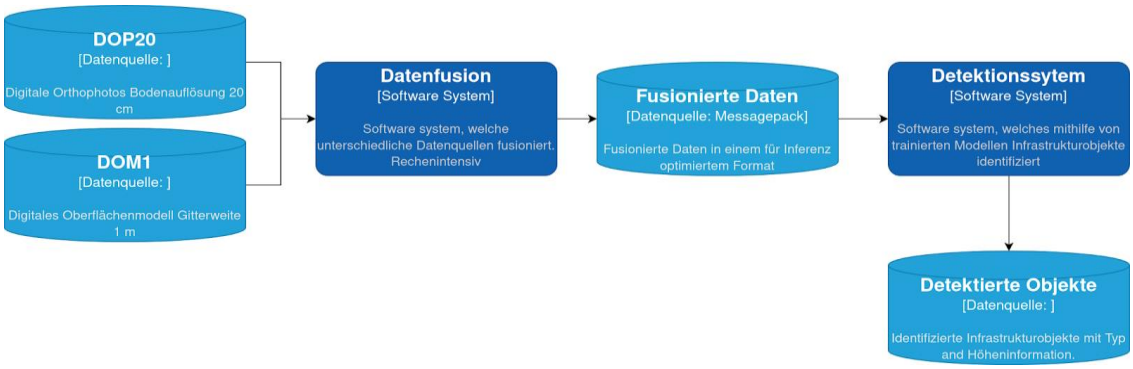


Abbildung 3: Angestrebte Systemarchitektur der Inferenz Pipeline von der Bilddateneingabe über die Fusionierung der Daten hin zur Identifizierung von Infrastrukturobjekten (eigene Abbildung)

Zusammenfassend wurde in dem vorangegangenen Kapitel dargestellt, dass die automatisierte, digitale Bestandserfassung gleisnaher Infrastruktur aus Fernerkundungsdaten eine entscheidende Rolle für verschiedene Anwendungsfälle spielt und durch den Einsatz von ML-Methoden unterstützt werden kann. Nach Abschluss der Dokumentation der erarbeiteten Anforderungen und einem möglichen weiteren Vorgehen werden im Folgenden, als Teil der zweiten Projektphase, die erforderlichen Schritte zur Schaffung der Datengrundlage ausführlich erläutert. Dieser Prozess umfasst zunächst die Datenexploration, wobei die verfügbaren Datenquellen analysiert werden. Anschließend erfolgt die Erklärung der Datenfusion, gefolgt von der initialen Aufbereitung des Trainings- und des Testdatensatzes.

5 Datengrundlage

5.1 Datenexploration

Um einen umfassenden Überblick über die verfügbaren Daten zu erhalten, wurde eine Datenexploration durchgeführt. Zunächst wurde die Eignung von Satellitendaten als Informationsquelle bewertet. Dabei wurden verschiedene Datenquellen in Betracht gezogen, insbesondere solche, die eine hohe temporale Auflösung haben und aus dem Erdbeobachtungsprogramm Copernicus (European Space Agency, 2024) stammen. Im Rahmen dieser Untersuchung wurden die folgenden Missionen genauer betrachtet:

- **Sentinel 1**: Diese Mission stellt Radarbilder mit Auflösungen von 10 bis 40 m zur Verfügung.
- **Sentinel 2**: Die Sentinel-2-Mission bietet hochauflösende optische Bilder mit einer Auflösung von 10 m für sichtbare und nahinfrarote Daten sowie einer Auflösung von 20 m für Daten im Bereich des Rotrandes und des kurzwelligen Infrarots.

Nach einer gründlichen Analyse wurde festgestellt, dass aus der dargestellten Auswahl die Sentinel-2-Mission die am besten geeignete Datenquelle für das Projekt darstellt. Die hohe räumliche Auflösung und Verfügbarkeit von sichtbaren optischen Bildern stimmen mit den Anforderungen der nachgelagerten maschinellen Lernaufgaben überein, was es ermöglichen könnte, vortrainierte Modelle effektiv zu nutzen.

Um die höchstmögliche Bildqualität für das Projekt zu gewährleisten, wurden gezielt Daten aus der Sentinel-2-Mission für den Zeitraum vom 12. Juni 2023 bis zum 13. Juni 2023 abgerufen. Es ist festzuhalten, dass die abgeleiteten Schlussfolgerungen auch für zu anderen Zeitpunkten aufgenommene Sentinel-2-Daten gelten, da die Limitation in der inhärenten Datenaufbereitung und den spezifischen Anforderungen des Projekts begründet liegt. Abbildung 4 veranschaulicht beispielhaft die abgerufenen Bilder und umfasst einen Ausschnitt von 200 m in der Umgebung der SSW. Die Abbildung zeigt dabei, dass einzelne Infrastrukturelemente aufgrund der Auflösung schwer erkennbar und nicht eindeutig unterscheidbar sind.

Abbildung 4: Beispiele von Sentinel-2-Daten (Sentinel-2: Copernicus Sentinel Daten (2023), verarbeitet von der European Space Agency (ESA))

In Abbildung 5 wird im direkten Vergleich der DOP und Satellitendaten anschaulich deutlich, dass das betrachtete Infrastrukturelement dieses Projektes, die SSW, aufgrund der unzureichenden räumlichen Auflösung in den Sentinel-2-Daten nicht identifizierbar ist. Darüber hinaus treten häufig Verdeckungen auf, die durch ungünstige Wetterbedingungen (z. B. Wolken) verursacht werden. Obwohl die Sentinel-2-Daten eine hohe Aktualisierungsrate aufweisen und zeitlich hoch aufgelöst Informationen liefern können,

liefern sie keine relevanten Erkenntnisse für das Projektziel und wurden daher in der weiteren Analyse nicht mehr berücksichtigt.

Für das Projekt sollten stattdessen Daten von DOP, DOM und DGM zur Erfassung von SSW verwendet werden. Die DOP-Daten bestehen aus zerrungsfreien und georeferenzierten Luftbildern mit einer räumlichen Auflösung von 20 cm und einer Kachelgröße von 1 km x 1 km. Das DGM weist eine räumliche Auflösung von 10 m pro Pixel und eine Kachelgröße von 20 km x 20 km auf, während das DOM mit Oberflächenhöhendaten eine räumliche Auflösung von 1 m pro Pixel und eine Kachelgröße von 5 km x 5 km aufweist.

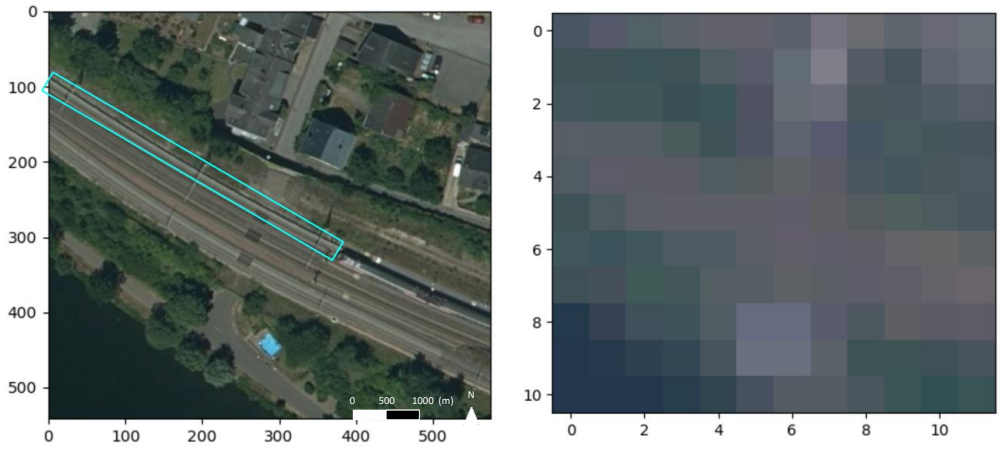


Abbildung 5: Beispielaufnahme von SSW aus dem DOP-Datensatz (links) mit Markierung der SSW sowie eine Aufnahme aus der Sentinel-2-Mission (rechts) (DOP: © GeoBasis-DE / BKG (2023); Sentinel-2: Copernicus Sentinel Daten (2023), verarbeitet von der ESA)

Im nächsten Schritt erfolgten die Analyse und Bewertung der Daten zu Infrastrukturobjekten im Schienenumfeld, die von der DB InfraGO AG stammen. Bei der Untersuchung der SSW-Daten stellte sich heraus, dass insgesamt 26.855 Labels verfügbar sind. Diese Anzahl an Labels ist ausreichend für das Training und die Evaluierung der in Kapitel 3.2 vorgestellten maschinellen Lernalgorithmen im Kontext überwachter und semi-überwachter Modelle. Die Fülle der verfügbaren Annotationen in diesem Datensatz machte die Inanspruchnahme zusätzlicher Datenquellen entsprechend überflüssig.

In einer detaillierten Analyse der zur Verfügung gestellten Daten zu SSW wurde beispielsweise die Längenverteilung der einzelnen SSW untersucht, um die Pixelabdeckung der jeweiligen Label zu ermitteln. Die Untersuchung ergab eine minimale Länge von 1,2 m und eine maximale Länge von 12,9 km, wobei der Median bei 40 m liegt. Abbildung 6 veranschaulicht diese Verteilung im logarithmischen Maßstab.

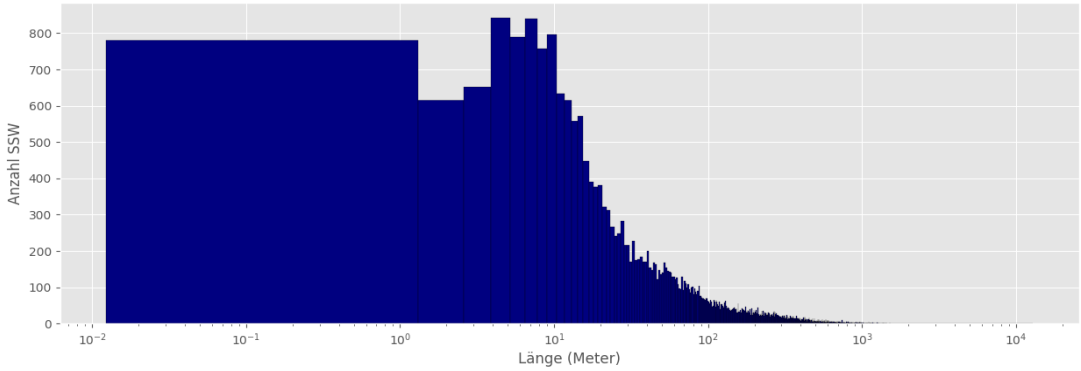


Abbildung 6: Histogramm der Längenverteilung der SSW

5.2 Datenfusion und räumliche Auswahl

Dieses Kapitel beschreibt die Schritte zur Kombination der verschiedene Datenquellen, um für die Erfassung der untersuchten Objekte relevante Informationen zusammenzutragen (siehe Abbildung 7). Durch die Fusion der Datensätze (DOP, DOM und DGM) kombiniert die resultierende zusammengesetzte Darstellung die detaillierten visuellen Informationen der Orthophotos mit den räumlichen Höhendaten des DOM und sollte eine umfassendere und aussagekräftigere Darstellung des untersuchten Gebiets erzeugen.

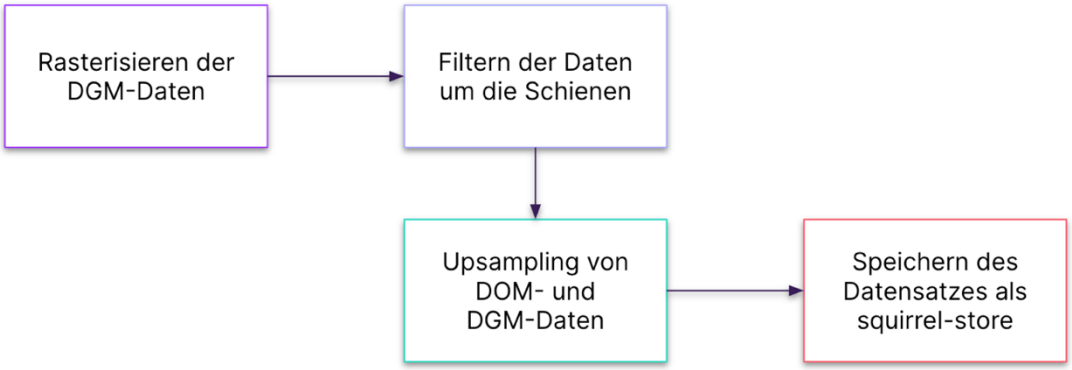


Abbildung 7: Methodik zur Fusion der DGM- und DOM- Daten (eigene Abbildung)

Die Rasterisierung des DGM, d. h. die Umwandlung der .xyz-Dateien in .tif-Dateien, wurde durchgeführt, um ein einheitliches Application Programming Interface (API)[13] für die Abfrage aller Daten innerhalb eines bestimmten geografischen Begrenzungsrahmen zu etablieren. Dieser Schritt sollte eine einheitliche Handhabung von DOP, DOM und DGM ermöglichen.

Nach der Rasterisierung der DGM-Daten wurde für jede Datenquelle eine API entwickelt, die eine effiziente Abfrage einer definierten Bounding Box (deutsch: Begrenzungsrahmen) ermöglicht. Es wurden bundesweit alle 200 m x 200 m Bounding Boxen ausgewählt, in denen sich Gleise befinden. Bounding Boxen, in denen sich keine Schiene befinden, wurden aus dem Datensatz entfernt. Der Schienenverkehrsnetzdatensatz wurde aus dem Open Data Portal der Deutschen Bahn AG bezogen (Deutsche Bahn AG, 2022).

Nachdem die Daten für die Abfrage nahe den Bahnschienen ausgewählt wurden, wurden die DOM- und DGM-Daten auf die Bodenauflösung der DOP-Daten hochskaliert, was einer Auflösung von 20 cm pro Pixel entspricht. Dieser Schritt war von zentraler Bedeutung, um eine 1:1-Zuordnung zwischen jedem Pixel in den DOP und den Höhendaten sicherzustellen. Für das Hochskalieren der schlechter aufgelösten Datenquellen wurde ein Schema des "nearest neighbor"-Upsamplings verwendet. Beim "nearest neighbor"-Upsampling handelt es sich um eine Methode zur Verbesserung der Auflösung von Rasterdaten, bei der der Wert des nächstgelegenen Nachbarpixels für jedes verkleinerte Pixel übernommen wird.

In Abbildung 8 ist ein Beispiel der fusionierten Daten dargestellt. Hierbei wurde ein normalisiertes DOM (nDOM) erzeugt, indem die DOM- und DGM-Daten subtrahiert wurden, um die absolute Höhe jedes Pixels im DOP-Bild zu bestimmen. Weitere Beispiele sind in Abbildung 32 im Anhang zu finden.

[13] Deutsch: Programmierschnittstelle

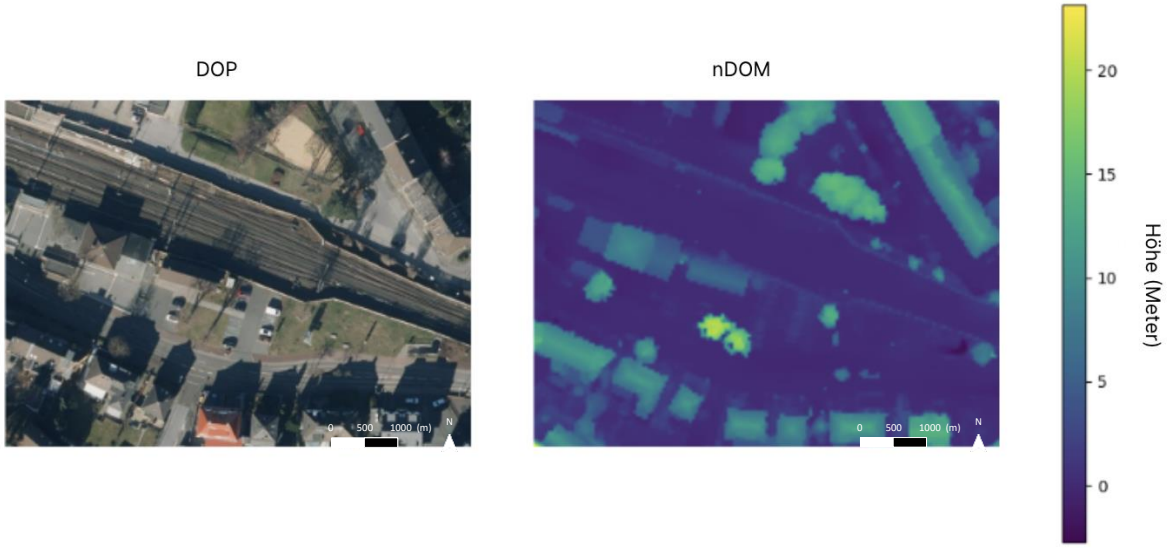


Abbildung 8: Beispiel zur Erstellung des nDOMs nach der Datenfusion von DOM, DGM und DOP (Geobasisdaten: © GeoBasis-DE / BKG (2023))

5.3 Aufarbeitung der Trainings- und Testdatensätze

Im vorliegenden Abschnitt wird die methodische Vorgehensweise zur Aufbereitung und Aufteilung des Datensatzes in Trainings-, und Testdatensatz für die Analyse erörtert. Dieser Schritt ist von Bedeutung, um eine unabhängige Validierung und Bewertung der ausgewählten Modelle sicherzustellen. Das ML-Modell, beschrieben in Kapitel 6.1 und 6.2 wird mit dem Trainingsdatensatz trainiert und das beste Modell basierend auf dem Validierungsdatensatz ausgewählt. Anschließend wird das beste Modell mit dem Testdatensatz unter anderem auf Präzision und Leistung überprüft (siehe Kapitel 6.3). Zusätzlich zur Bearbeitung des Testdatensatzes, um eine hohe Qualität der Annotationen sicherzustellen, werden die notwendigen Vorverarbeitungsschritte für das Trainieren des Modells, sowie die Methodik zur Verbesserung der Datenqualität und zur Erleichterung eines effektiven Lernens beschrieben.

Zunächst wurde die Verteilung der DOP-Aufnahmen im Hinblick auf die generellen Aufnahmezeitpunkte sowie eine bundesweite Repräsentation innerhalb der Testdaten analysiert. Die gesamte Verteilung der Aufnahmezeitpunkte aller verwendeten DOP-Aufnahmen (Training- und Testdatensatz) ist in Abbildung 9 veranschaulicht. Hier ist zu sehen, dass der größte Anteil der DOP-Aufnahmen in den Jahren 2020 und 2021 entstanden ist. Als Grundlage für den Trainings- und Validierungsdatensatz wurde dieser ursprüngliche Datensatz von 365.075 DOP-Aufnahmen der Größe 5.000 x 5.000 Pixel zu Kacheln der Größe 1.000 x 1.000 Pixel vorverarbeitet. Darüber hinaus wurden nicht benötigte Bereiche, welche weit von Gleisumgebungen entfernt liegen, entfernt, wodurch 307.000 der DOP-Kacheln verblieben.

Zusätzlich zu den 365.075 DOP-Aufnahmen wurde ein Testdatensatz vorbereitet. Dieser Testdatensatz umfasst insgesamt 400 zufällig ausgewählte DOP-Aufnahmen, welche mindestens eine Instanz einer SSW enthalten. Im Folgenden bezeichnen die Bilddaten des Trainingsdatensatzes entsprechend vorverarbeitete DOP-Kacheln, während der Testdatensatz weiterhin aus den ursprünglichen DOP-Aufnahmen besteht. Zur Erstellung des Testdatensatzes wurde das Zufallsverfahren verwendet, um eine angemessene Bandbreite an Aufnahmezeitpunkten und eine ausgeglichene räumliche Verteilung auf verschiedene Regionen in Deutschland sicherzustellen. Die spezifische Verteilung der SSW ausschließlich im Testdatensatz über die verschiedenen Bundesländer wird in Abbildung 10 veranschaulicht. Die Verteilung orientiert sich entsprechend an der Anzahl der Aufnahmen für jedes Bundesland, wobei die unterschiedliche Größe der Bundesländer in Betracht gezogen und repräsentiert wurde.

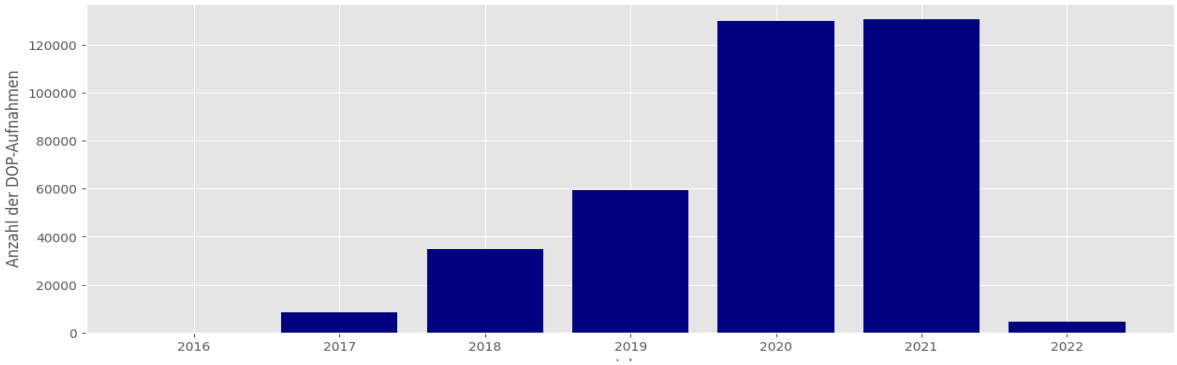


Abbildung 9: Verteilung der Aufnahmezeitpunkte aller verwendeten DOP-Aufnahmen pro Jahr

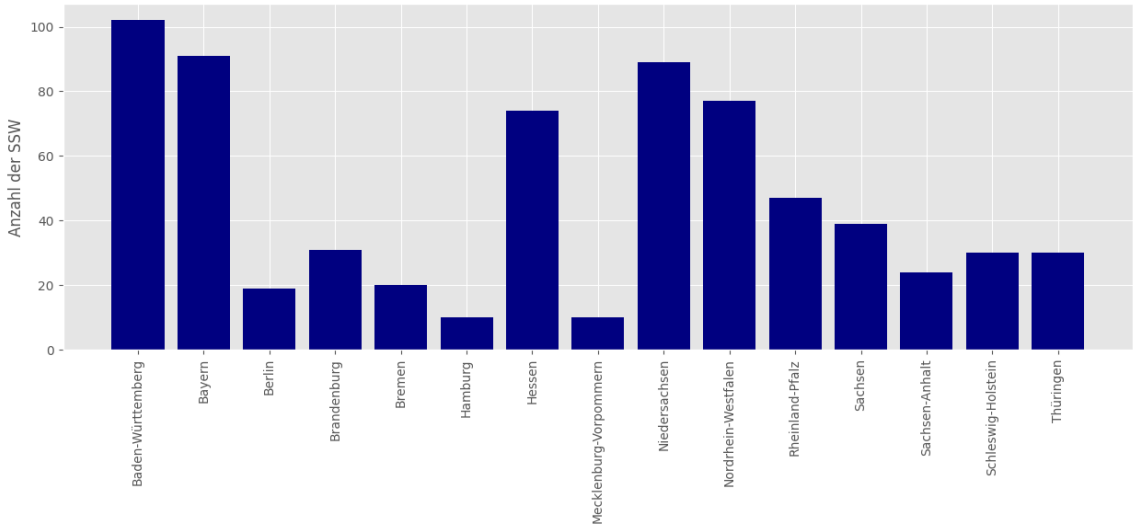


Abbildung 10: Verteilung der SSW-Anzahl über alle Bundesländer hinweg innerhalb des Testdatensatzes

Um die Qualität und Zuverlässigkeit der Annotationsdaten abgeleitet aus den Infrastrukturdaten der DB InfraGO AG zu erhöhen, wurden diese Testdaten von verschiedenen Daten-Annotierenden überarbeitet. Jede und jeder Annotierende erhielt eine spezifische Teilmenge des Datensatzes zur Bearbeitung. Die Hauptaufgabe der Annotierenden bestand darin, die bereits vorhandenen Labels gründlich zu überprüfen, zu verfeinern und bei Bedarf notwendige Ergänzungen oder Streichungen vorzunehmen, um die Genauigkeit der Übereinstimmung der Annotation mit SSW auf das höchstmögliche Niveau zu bringen. Dies schaffte eine robuste und qualitativ hochwertige Grundlage zur späteren Evaluation der Modelle.

Dieser manuelle Annotationsprozess wurde in einer cloudbasierten QGIS-Umgebung durchgeführt. Diese Plattform ermöglichte es jeder und jedem Annotierenden, effizient an der ihm oder ihr zugewiesenen Teilmenge der Daten zu arbeiten. Die Bereitstellung der Annotationsumgebung erfolgte auf der Google Kubernetes Engine (GKE). Die einzelnen Benutzenden stellten eine Verbindung zu einer QGIS-Desktop-Anwendung her, die auf einem Kubernetes-Pod gehostet wurde, und verbanden sich über das VNC-Protokoll von ihren lokalen Rechnern aus (siehe Abbildung 11). Es ist wichtig zu beachten, dass jede benutzende Person über eine isolierte QGIS-Umgebung verfügte, wobei lediglich die PostGIS-Datenbank gemeinsam genutzt wurde, in der alle Annotationen der DB InfraGO AG gespeichert waren. Abbildung 12 veranschaulicht beispielhaft die Ansicht von Annotierenden während der Konfiguration.

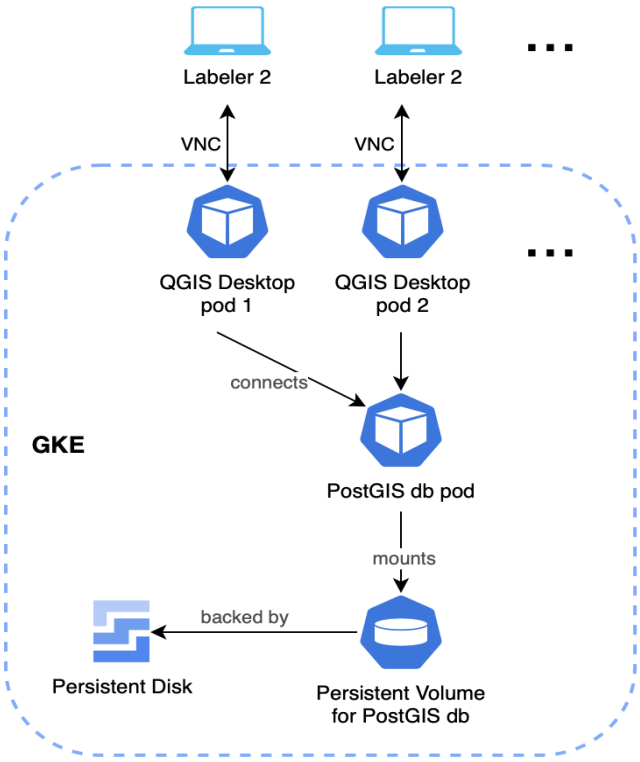


Abbildung 11: Aufbau des Google Kubernetes Engine Aufbaus zur Annotation und die Überprüfung der Annotationen/annotierten Datensätze (eigene Abbildung)

In der QGIS-Umgebung wurden die ursprünglichen annotierten Daten der DB InfraGO AG angezeigt und von den einzelnen Annotierenden mithilfe der in QGIS integrierten Funktionen angepasst. Alle Annotationen der SSW waren in Form von Linien angezeigt, die der Mittellinie einer SSW folgten. Der Testdatensatz beinhaltet insgesamt 2.518 SSW, wobei im Verlauf des Annotations-Prozesses insgesamt 157 neue Annotationen hinzugefügt, 17 Annotationen gelöscht und 545 Annotationen modifiziert wurden. Während des Prozesses fiel auf, dass die meisten ursprünglichen Annotationen der DB InfraGO AG eine geringe räumliche Abweichung von der Position der SSW in den DOP-Bildern aufwiesen. Diese Abweichungen werden vermutlich durch eine ungenaue Ausrichtung der Aufnahmewinkel verursacht, was zu einem konstanten Versatz führte. In einigen Fällen wurde bei einer Annotation keine SSW gefunden. Hinsichtlich der fehlenden oder zusätzlichen Annotationsdaten war dies auf die Diskrepanz zwischen dem Zeitpunkt der Bildaufnahme und dem Zeitpunkt, zu dem die Annotationsdaten von der DB InfraGO AG erstellt wurden, zurückzuführen und wurde daher erwartet.

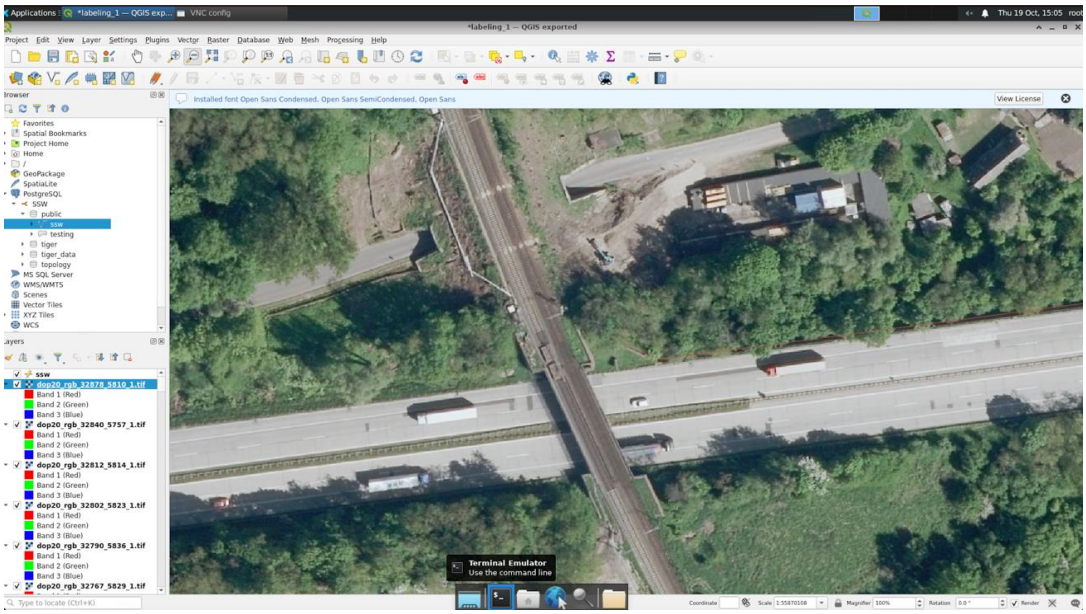


Abbildung 12: Beispielansicht der QGIS-Umgebung zur Annotation (Markierung) und Überprüfung der Annotations-Daten (Quelle: eigene Aufnahme der QGIS-Oberfläche; DOP: © GeoBasis-DE/BKG (2023))

Es ist wichtig zu beachten, dass die Annotationen des Trainingsdatensatzes unverändert in ihrem Rohformat verwendet werden sollten. Sowohl der Trainings- als auch der Testdatensatz wurden im Messagepack-Format gespeichert, um ein effizientes Laden aus der Cloud zu ermöglichen und das Speichern redundanter Daten zu verhindern. Der Datensatz ist in ein eigenes Paket namens „Squirrel“ (Squirrel Developer Team, 2022), einer Python-Bibliothek, verpackt, welche speziell zur kollaborativen Nutzung von Datensätzen entwickelt wurde und die effiziente Verwaltung der serialisierten Daten ermöglicht. Serialisiert bedeutet, dass die Daten in einem maschinenlesbaren Format gespeichert werden, um eine schnelle Verarbeitung zu ermöglichen.

Im Anschluss an die Bearbeitung und Verbesserung des Testdatensatzes mussten sowohl die zu trainierenden (307.000 DOP-Kacheln) – als auch die Test-Daten (DOP-Aufnahmen) entsprechend der spezifischen Aufgabenstellung des maschinellen Lernens zusätzlich aufbereitet werden. Bei der Erkennung von SSW handelt es sich im Hinblick auf ML, wie in Kapitel 3.2 beschrieben, um ein semantisches Segmentierungsproblem, weshalb die Annotationen, das heißt die Markierungen von SSW, welche identifiziert werden sollen, in einem binären Maskenformat vorliegen müssen. Allerdings wurden die entsprechenden Beschriftungen der SSW in Form einer Liniengeometrie im Koordinatensystem jedes Bildes gespeichert. Obwohl alle Informationen in den Daten erhalten bleiben, kann ein solches Format nicht trainiert und verarbeitet werden und muss daher angepasst werden.

Aus diesem Grund wurde die binäre Maske, welche durch die Liniengeometrie erzeugt wird, mithilfe einer Dilatation erweitert, um eine breitere Segmentierungs-Maske zu erhalten. Die Dilatation ist eine Operation, die häufig in der Bildverarbeitung angewendet wird, insbesondere bei Aufgaben wie der Objekterkennung und -segmentierung. Hierbei handelt es sich um eine morphologische Operation, welche die Grenzen von Objekten in einem Bild erweitert. In diesem speziellen Fall wurde diese Dilatations-Operation durchgeführt, da die Annotationen der SSW ungenau oder unzuverlässig sind. Die Ungenauigkeiten im Datensatz treten in erster Linie in Form von Verschiebungen der Liniengeometrie gegenüber den tatsächlichen Wänden im Bild auf. Entsprechend wurden die Linien im Raster der DOP-Kacheln abgebildet und um eine Kernelgröße von 15 Pixeln erweitert. Dabei konnte anhand von stichprobenartig ausgewählten Bilddaten aus dem DOP-Datensatz visuell bestätigt werden, dass diese Veränderung bereits ausreichend ist, um die meisten Wände innerhalb der generierten Masken in dem Datensatz zu erfassen.

Die folgende Abbildung 13 stellt ein exemplarisches Beispiel für die Dilatation in einem DOP-Bildausschnitt dar.

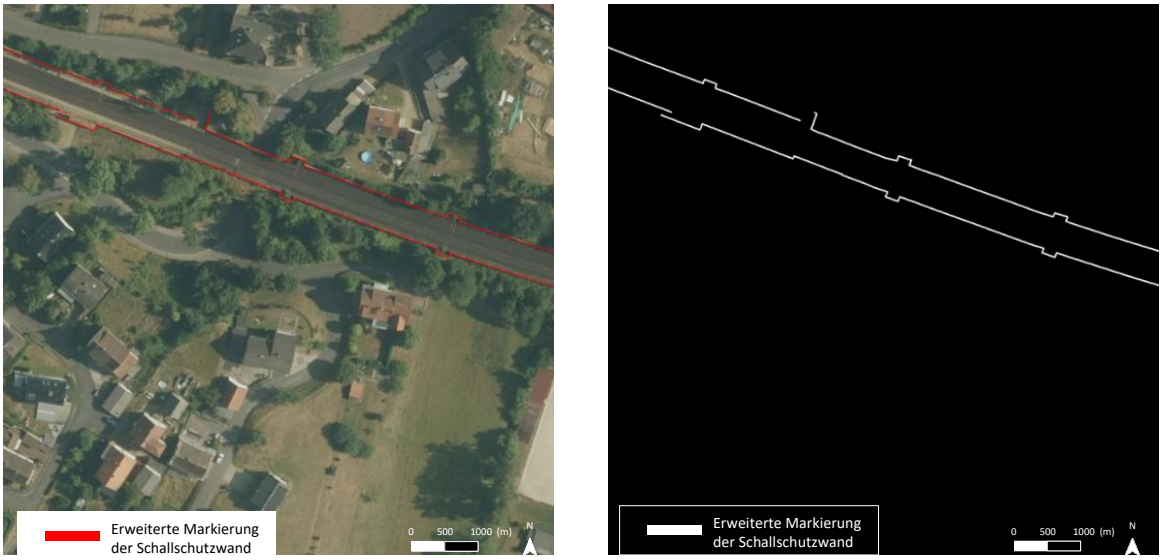


Abbildung 13: Beispiel für die Dilatation der als SSW markierten Pixel in einem DOP-Bildausschnitt. Auf der linken Seite ist der Original-Bildausschnitt mit entsprechend erweiterter Markierung der SSW in rot und auf der rechten Seite lediglich die erweiterte Markierung dargestellt (DOP: © GeoBasis-DE/BKG (2023))

An dieser Stelle gilt es außerdem anzumerken, dass die Trainings- und Validierungsdaten explizit keine DOM- und DGM-Daten beinhalten. Obwohl die Qualität des Abgleichs zwischen DOP-, DOM- und DGM-Daten sich im Zuge der Datenevaluierung (siehe Kapitel 5.1) als vielversprechend erwies, wurde die weitere Verwendung der Höhendaten im Zuge der im Folgenden beschriebenen Modell- und Ergebnisevaluation ausgeschlossen. Aufgrund der geringeren Auflösung wurden bei der Verwendung der Höhendaten (DOM und DGM) deutlich mehr Bereiche fälschlicherweise als SSW klassifiziert, sodass die Verwendung des kombinierten Datensatzes schlechtere Ergebnisse erzielte. Entsprechend wurden die DOM- und DGM-Daten für den Trainingsprozess ausgeschlossen, da diese nicht zur besseren Klassifizierung beigetragen haben.

Vor dem eigentlichen Trainingsvorgang wurde zudem der Anteil der DOP-Kacheln, welche SSW beinhalten, im Vergleich zu Bilddaten ohne SSW verglichen. Hierbei wurden von den 307.000 verbleibenden Bildern nach dem Vorverarbeitungsprozess lediglich 12.102 Bilddaten mit SSW identifiziert, was 3,9 % der Daten entspricht.

Für den Trainingsprozess ist solch eine unverhältnismäßige Datengrundlage problematisch, da in den meisten Fällen die Wand als wichtigstes Trainingselement fehlt. Das Modell würde in diesem Fall in erster Linie lernen und vorhersagen, dass alle Objekte zum Hintergrund gehören und dass es in 96,1 % der Bilder richtigliegen würde. Entsprechend wurde an dieser Stelle entschieden, für den weiteren Trainingsprozess lediglich die 12.102 Bilddaten, welche eine SSW enthalten, als Trainings- und Validierungsdatensatz zu verwenden. Dabei wurden 80 % der Daten für das Training und 20 % zur Validierung verwendet. Die Reduzierung des Datensatzes ermöglicht dennoch ein erfolgreiches Training des Modells, da ein Großteil der Bildausschnitte der verwendeten Bilddaten keine Wände enthält und diese daher als Negativbeispiel dienen können.

6 Technische Lösung: Modellentwicklung und Evaluation

Anschließend an die detaillierte Erläuterung der Datengrundlage wird im folgenden Kapitel ein umfassendes Verständnis der angewandten Methodik, des Entscheidungsprozesses und entsprechenden Ergebnissen der ML-Lösung im Rahmen des Projektes geschaffen. Dabei werden die Hintergründe zur Auswahl der ausgewählten Modellarchitektur erläutert und die Faktoren, die zur Entscheidungsfindung beigetragen haben, dargelegt. Darüber hinaus werden die Ergebnisse der effektivsten Modell-Version anhand entsprechender Metriken illustriert. Hier werden detaillierte Ergebnisse der Evaluation sowie Einblicke in die Genauigkeit und Robustheit der ML-Lösung gegeben.

6.1 Backbone Auswahl

Wie in Kapitel 3.2 beschrieben, besteht ein klassischer Ansatz im Bereich des maschinellen Lernens darin, auf großen Datenmengen vortrainierte Modelle als vorgefertigte Architektur, dem sogenannten Backbone, zu nutzen. Anschließend werden nur wenige Modell-Schichten auf einem spezifisch konfigurierten Backbone-Netzwerk, welches unverändert bleibt und somit als “eingefroren” gilt, trainiert, um die spezifische Aufgabe zu lösen. Da das Modell-Backbone die entscheidende Grundlage für das weitere Training bildet, ist eine gezielte Auswahl und Evaluation notwendig. Dafür wurde im Rahmen des Projekts eine neuartige Bewertungsmethode für verschiedene ViT-Backbones entwickelt, um festzustellen, welche Backbone-Modelle auch ohne Training und annotierte Trainingsdaten Merkmale erzeugen, welche die SSW am besten von der Umgebung separieren können.

Um die Qualität und Leistungsfähigkeit eines Backbones zu evaluieren, wurde eine repräsentative DOP-Kachel, welche eine klar identifizierbare Wand im RGB-Farbraum beinhaltet, ausgewählt. Die SSW wurde dabei manuell mit Punkten, die innerhalb der DOP-Kachel klar auf eine SSW fallen, markiert. Abbildung 14 zeigt die beiden verwendeten und manuell markierten DOP-Kacheln.

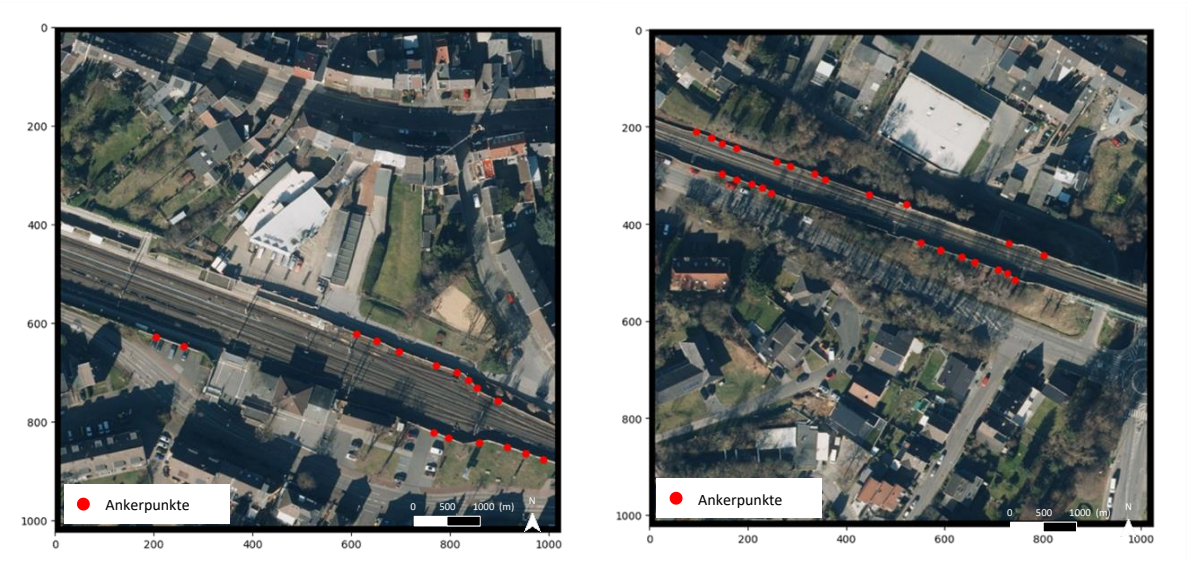


Abbildung 14: Repräsentative DOP-Kacheln mit manuell identifizierten und markierten Ankerpunkten in Rot auf vorhandenen SSW als Grundlage zur Evaluation der Backbone-Modelle (DOP: © GeoBasis-DE/BKG (2023))

Solch ein manuell mit einzelnen, als SSW identifizierten Punkten annotiertes Bild wurde als Ankerbild definiert, wobei die annotierten Punkte als Ankerpunkte gelten. Das Ankerbild wurde dann mit dem ViT-Backbone der Wahl kodiert, sodass die Merkmale der Ankerpunkte, sogenannte Anker-Merkmale, gespeichert wurden. Diese Anker-Merkmale stellen die wahren Daten, die sogenannte Ground Truth, der SSW aus der Perspektive des Modells dar.

Um die Anker-Merkmale qualitativ zu bewerten, wurde anschließend ein neues „Test"-Bild kodiert und eine pixelweise Kosinusähnlichkeit (Cosine Similarity) mit jedem der Ankerpunkte berechnet. Die Kosinusähnlichkeit misst die Ähnlichkeit zwischen zwei Vektoren in einem n-dimensionalen Raum. Dazu wird der Kosinus des Winkels zwischen den Vektoren, das heißt das Skalarprodukt der Vektoren dividiert durch das Produkt ihrer Längen berechnet (Han et al., 2012). Der Wert der Kosinusähnlichkeit gibt hier bei jedem Pixel an, wie nahe die Darstellung dieses Pixels nach Ansicht des Modells an der Darstellung der Anker-Merkmale liegt. Dabei wurde der durchschnittliche Wert (Kosinusähnlichkeit) für alle Anker-Merkmale berechnet, um zu einer robusten qualitativen Bewertung zu gelangen, die unabhängig der zusätzlichen verwendbaren Annotationen der Test- und Trainingsdaten ist. Das Ergebnis dieser Berechnung ist das normalisierte innere Produkt im Bereich (0, 1), wobei 1 eine perfekte Ähnlichkeit und 0 keine Ähnlichkeit darstellt. Die Kosinusähnlichkeitskarte der „Test"-Kachel ist in Abbildung 15 (links) dargestellt.

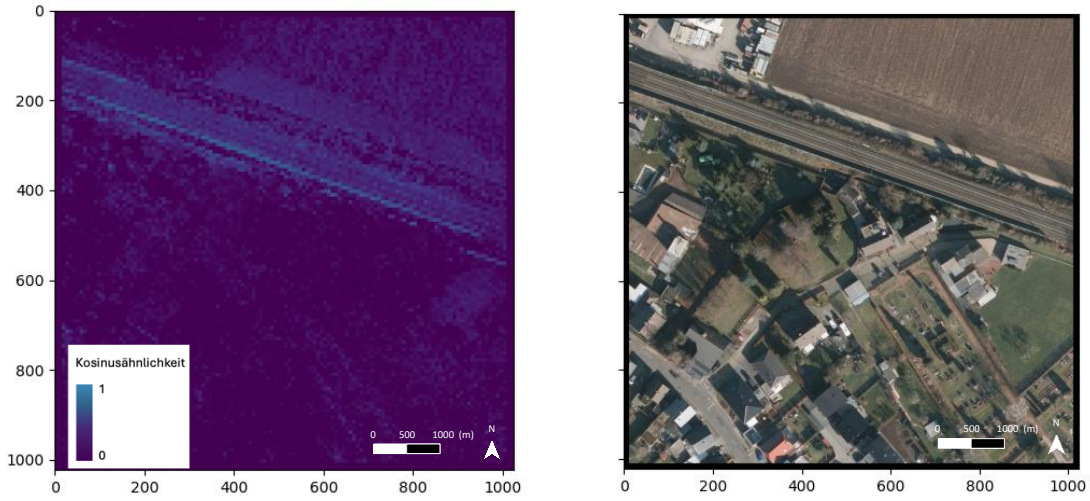


Abbildung 15: Ausgewählte „Test"-DOP-Kachel (rechts) zur Bewertung der Backbone-Modelle sowie eine Kosinusähnlichkeitskarte (links) für die berechnete „Test"-Kachel (DOP: © GeoBasis-DE/BKG (2023))

Eine Kosinusähnlichkeitskarte ist eine spezielle Art von Aktivierungskarte. Aktivierungskarten sind visuelle Darstellungen der Ausgabe von Neuronen innerhalb eines neuronalen Netzwerks, erzeugt durch die Anwendung von erlernten Filtern auf Eingabedaten (Ayyar et al., 2023). Aktivierungskarten visualisieren und ermöglichen Nachvollziehbarkeit, welche Merkmale das Netz aus den Eingangsdaten zu extrahieren gelernt hat. Sie helfen dabei, das Verhalten des Netzwerks zu interpretieren und zu verstehen, welche Art von Informationen in den verschiedenen Schichten erfasst werden. Eine Kosinusähnlichkeitskarte visualisiert die Ähnlichkeit zwischen Eingabevektoren und den Gewichtsvektoren einer bestimmten Schicht eines neuronalen Netzwerks. Die im Folgenden aufgezeigten Aktivierungskarten beziehen sich ebenfalls auf die Kosinusähnlichkeit. Eine höhere Aktivierung (d. h. hohe Kosinusähnlichkeitswerte; hellere Darstellung) in der Karte deutet darauf hin, dass die Merkmale in den Eingabedaten und den gelernten Gewichtsvektoren ähnlicher sind, während eine niedrigere Aktivierung (d. h. niedrige Kosinusähnlichkeitswerte; dunklere Darstellung) auf geringere Ähnlichkeit hinweist. Mithilfe dieses Vergleichs und der Kosinusähnlichkeitskarte wurden bei der Analyse folgende Fragestellungen beurteilt:

1. Existiert eine SSW?
2. Ist die Kosinusähnlichkeit in der Region hoch, falls eine SSW existiert? Falls ja, ist die SSW auf der Kosinusähnlichkeitskarte gut lokalisiert?
3. Ist die Kosinusähnlichkeit in der Region niedrig, falls keine SSW vorhanden ist?

Die hier beschriebene Evaluationsmethode wurde für die in Tabelle 5 dargestellten Modelle umgesetzt. Dabei basieren alle ausgewählten Backbone-Modelle auf der zuvor erwähnten und Kapitel 3.2 erläuterten ViT-Architektur. SAM ViT Basis (Kirillov et al., 2023b) und SAM ViT Groß (Kirillov et al., 2023c) erzeugen Objektmasken aus Eingaben wie Punkten oder Boxen und können zur Erzeugung von Masken für alle Objekte in einem Bild verwendet werden. Das Modell ist so konzipiert und trainiert, dass es durch die Technik von Zero-Shot-Lernen auf neue Bildverteilungen und Aufgaben übertragen werden kann. Zero-Shot-Lernen bedeutet, dass das Modell eine Aufgabe lösen kann, ohne dass es spezifisch dafür trainiert wurde oder direkte Trainingsdaten dazu erhalten hat. Stattdessen werden Informationen und Attribute aus ähnlichen Aufgaben, für die das Modell bereits trainiert wurde, verwendet, welche als Brücke zu neuen Informationen dienen. In Bezug auf SAM hat das Modell im Rahmen des bisherigen Trainings bereits ein breites Verständnis oder ein allgemeines Konzept dafür entwickelt, was ein Objekt ausmacht. Es hat gelernt, die Eigenschaften und Merkmale zu erkennen und zu verstehen, durch welche Merkmale verschiedene Objekte definiert werden. Das Modell besitzt die Fähigkeit, dieses Wissen zu verallgemeinern und Vorhersagen für Aufgaben oder Objekte zu treffen, die es während des Trainings nicht gesehen hat. Die Modelle SAM ViT Basis und SAM ViT Groß unterscheiden sich dabei lediglich in der Anzahl der Schichten und Parameter (siehe Tabelle 5).

DINO V2 Klein (Oquab et al., 2024b), DINO V2 Basis (Oquab et al., 2024c) und DINO V2 Groß (Oquab et al., 2024d) sind Varianten des DINO-Ansatzes, welcher ebenfalls auf der Idee des selbstüberwachten Lernens (siehe Kapitel 3.2) beruht. DINO basiert auf Instanz-Diskriminierung und kontrastivem Lernen. Kernidee des Modells ist es, visuelle Repräsentationen zu erlernen, indem zwischen verschiedenen Instanzen desselben Objekts oder derselben Umgebung unterschieden wird, ohne definierte Klassenbezeichnungen zu verwenden (Instanz-Diskriminierung). Dies wird durch die Gegenüberstellung der Merkmale verschiedener Ansichten desselben Bildes erreicht. Dazu nutzt DINO kontrastives Lernen, um das Modell zu trainieren. Ziel ist es, die Repräsentationen verschiedener Ansichten desselben Bildes (positive Paare) einander anzunähern und gleichzeitig die Repräsentationen verschiedener Bilder (negative Paare) im Lernprozess voneinander zu trennen. Auch hier variieren die Größen der Modelle in Bezug auf die Anzahl der Schichten und Parameter (siehe Tabelle 5).

vit_base_patch8_224.dino (Dosovitskiy et al., 2020b) und vit_small_patch8_224.dino (Dosovitskiy et al., 2020c) gehören gleichermaßen zur ViT-Familie und sind speziell für die Verarbeitung von Bildern mit einer Patch-Größe von 8 x 8 und einer Eingangsgröße von 224 x 224 Pixeln ausgelegt. Die Unterschiede liegen hier in der Größe des Modells und der Kapazität zur Merkmalsextraktion. Das Basismodell (vit_base) ist größer und komplexer, was sich in der Anzahl der trainierbaren Parameter (siehe Tabelle 5), dem Ausmaß der Rechenkomplexität sowie der Anzahl der möglichen Aktivierungen in dem neuronalen Netz widerspiegelt. Entsprechend hat das Modell mit mehr Parametern eine höhere Kapazität zur Modellierung komplexer Daten und Erkennung detaillierterer Merkmale, z. B. eines Objektes, wodurch sich aber auch die Rechenkomplexität erhöht.

Die Patchgröße (Patch Size), angegeben in Tabelle 5, bezieht sich in der Bildverarbeitung auf die Abmessungen eines typischerweise quadratischen Ausschnitts aus einem Bild. Diese Patches werden als Eingabe für die Algorithmen zur Bildklassifizierung oder Objekterkennung verwendet. wobei die optimale Patchgröße von der spezifischen Aufgabe, dem Datensatz und den Merkmalen der zu analysierenden Daten abhängt. Bei der Segmentierung von SSW ist es wichtig, eine Patchgröße zu wählen, die sowohl die notwendigen Details erfasst als auch effizient verarbeitbar ist. Kleinere Patches, wie 8 x 8, können feine Details besser erfassen, während größere Patches, wie 14 x1 4 oder 16 x 16, besser für größere Strukturen geeignet sein können. Die Patchgröße ist durch das gewählte Modell vorgegeben und eine optimale Patch-Größe somit auch von weiteren Modell-Parametern abhängig.

Tabelle 5: Die evaluierten Backbone-Modelle und die jeweiligen Parameter, sowie die verarbeitete Patchgröße

Modell Name	**Anzahl der Parameter**	**Patchgröße (in Pixel)**
SAM ViT Basis	91 M	16 x 16
SAM ViT Groß	308 M	16 x 16
DINO V2 Klein	22.1 M	14 x 14
DINO V2 Basis	86.6 M	14 x 14
DINO V2 Groß	300 M	14 x 14
vit_small_patch8_224.dino	21.7 M	8 x 8
vit_base_patch8_224.dino	85.8 M	8 x 8

Im Rahmen der Evaluation der Backbone-Modelle wurden die Ankermerkmale aus den in Abbildung 14 dargestellten DOP-Kacheln zur Analyse von jeweils zwanzig Test-DOP-Kacheln verwendet. Diese Analyse hatte die Beantwortung folgender Fragen zum Ziel:

1. Sind größere Backbone Modelle besser?
2. Ist DINOv2 besser als DINOv1?
3. Ist SAM besser als DINO V2?

Um festzustellen, ob größere Modelle besser für die Aufgabenstellung geeignet sind, wurden diese zwanzig DOP-Kacheln und entsprechend erzeugte Aktivierungskarten untersucht. Im Folgenden sowie im Anhang sind Beispiele von dieser Auswahl illustriert. Die Ergebnisse zeigten, dass größere Modelle in der Tat bessere Leistungen erbringen.

Das Beispiel in Abbildung 17 zeigt die mit DINOv2-Modellen unterschiedlicher Größe erstellten Aktivierungskarten für die in Abbildung 16 illustrierte DOP-Kachel.

Abbildung 16: Rohdatenbild einer DOP-Kachel (DOP: © GeoBasis-DE / BKG (2023))

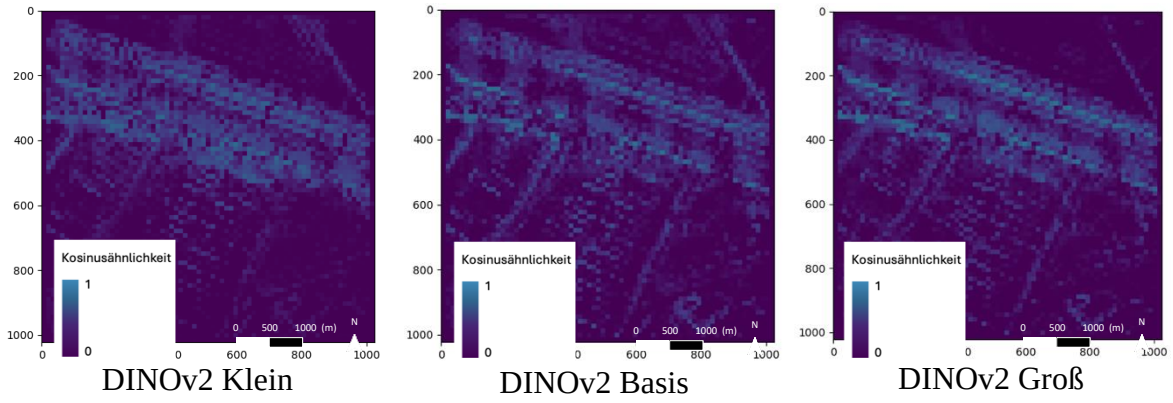


Abbildung 17: Aktivierungskarten zur Erkennung von SSW mit drei unterschiedlichen Modellgrößen von DINOv2: DINOv2 Klein (links), Basis (mittig) und Groß (rechts). Die Aktivierungskarten zeigen eine deutlichere Präzision für das große DINOv2-Modell (DOP: © GeoBasis-DE/BKG (2023))

Die in Abbildung 17 gewählten Aktivierungskarten wurden zur Visualisierung der unterschiedlich erzeugten Ausgaben bzw. Aktivierungen der DINOv2-Modelle ausgewählt und repräsentieren den Unterschied in der Präzision der Modelle entlang verschiedener SSW. Dabei ist in Abbildung 17 deutlich zu erkennen, dass die erzielten Ergebnisse von DINOv2 mit zunehmender Größe des Modells weniger Rauschen und Ungenauigkeiten aufweisen und entsprechend besser mit den tatsächlichen Wänden übereinstimmen. Jedoch ist dabei die Verbesserung vom Basismodell zum großen Modell viel geringer als die signifikante Verbesserung, die beim Übergang vom kleinen Modell zum Basismodell zu beobachten ist. Dieses Muster konnte für alle in Tabelle 5 genannten Modelle basierend auf den ausgewählten DOP-Kacheln beobachtet werden. Weitere Beispiele dazu sind im Anhang zu finden.

Um anschließend festzustellen, ob das DINOv2-Modell eine bessere Leistung als das vorherige DINOv1-Modell erzielen kann, wurde die Qualität der Kosinusähnlichkeitskarten beider Modelle bei einer festen Modellgröße anhand aller ausgewählten Kacheln verglichen. In der folgenden Abbildung 18 sind die Kosinusähnlichkeitskarten der Basisversionen beider Modelle dargestellt.

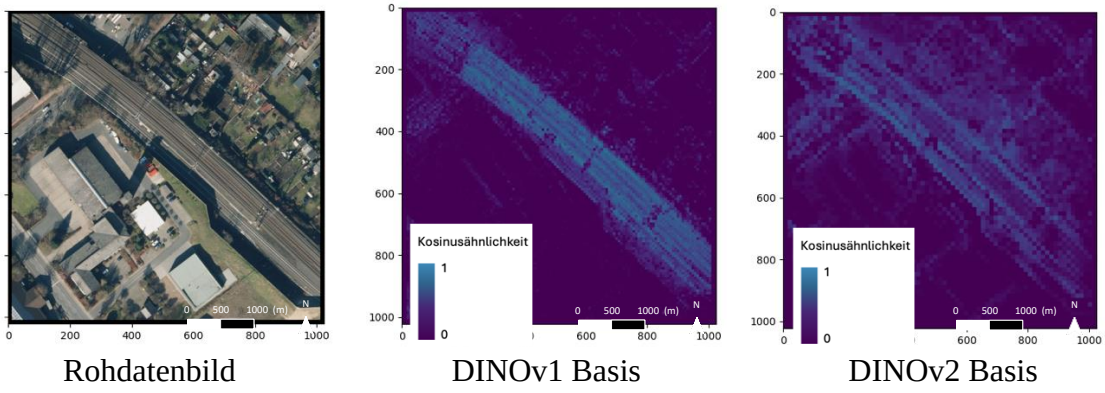


Abbildung 18: Beispiele einer DOP-Kachel als Rohdatenbild (links) sowie zwei unterschiedliche Aktivierungskarten, erzeugt durch die Verwendung des DINOv1 Basismodells (mittig) im Vergleich zu DINOv2 Basismodell (rechts) zum Vergleich der beiden DINO Backbone Modell-Versionen (DOP: © GeoBasis-DE/ BKG (2023))

In Abbildung 18 ist zu erkennen, dass das DINOv1-Modell eine höher aufgelöste Aktivierungskarte (deutlich hellere Farbmarkierung) erzeugt, was vor allem auf die geringere Patchgröße zurückzuführen ist, die bei der Tokenisierung (siehe Kapitel 3.2) des Bildes verwendet wurde. Nichtsdestotrotz wird bei diesem

Modell die Semantik der Wand mit der Umgebung bzw. dem gesamten Gleisbereich verwechselt. Auf der anderen Seite ist die DINOv2-Aktivierungskarte weniger verrauscht und stellt die SSW klarer dar. Daraus konnte die Annahme geschlossen werden, dass DINOv2 Merkmale von linearen Elementen und Strukturen besser erfasst und diese im Vergleich zu DINOv1 besser vom Hintergrund trennt. Weitere Beispiele dazu finden sich ebenfalls im Anhang.

Abschließend war es wichtig festzustellen, wie das SAM-Backbone, unter Verwendung der gleichen ausgewählten DOP-Kacheln, im Vergleich zu den von DINO trainierten Backbones abschneidet. Da SAM ein Basismodell ist, das speziell für die Aufgabenstellung der Bildsegmentierung trainiert wurde, war hier grundsätzlich zu erwarten, dass die Kosinusähnlichkeitskarten von höherer Qualität sind als bei einem für allgemeine Zwecke trainierten Backbone.

Dennoch zeigt Abbildung 19 deutlich, dass SAM dazu neigt, eine hohe Kosinusähnlichkeitsaktivierung im gesamten Gleisbereich zwischen zwei Wänden auszugeben. Hier zeigt der gesamte Gleisbereich eine helle Farbmarkierung, repräsentativ für die starke Aktivierung, wohingegen DINOv2 Basis eine klarere Präzision beim Erkennen der SSW-Linie aufzeigt.

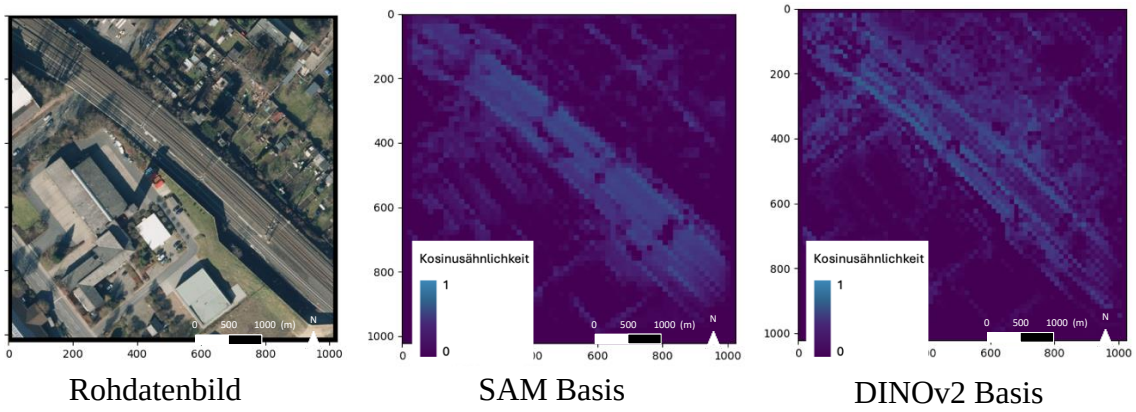


Abbildung 19: Rohdatenbild (links) sowie die jeweiligen Aktivierungskarten von SAM-Basis (mittig) und DINOv2 Basis (rechts) im Vergleich, wobei die SAM-Aktivierungskarte deutlich stärkeres Rauschen aufzeigt (DOP: © GeoBasis-DE/BKG (2023))

Diese Ergebnisse lassen sich darauf zurückführen, dass SAM darauf trainiert wurde, größere anstelle von granularen Objekten in einer Region zu segmentieren. Daher entspricht die Repräsentation der Wand dem nächstgelegenen Objekt in der Umgebung, in diesem Fall dem Gleisbereich.

Diese Hypothese wurde validiert, indem SAM rund um die Bereiche von SSW in den Bilddaten zur Segmentierung der SSW aufgerufen wurde. Die ausgegebenen Segmentierungs-Masken wurden analysiert. Ein Beispiel DOP-Kachelausschnitt ist in Abbildung 20 zu sehen.

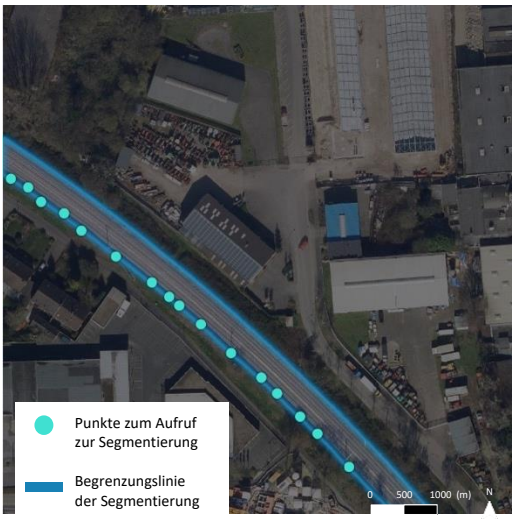


Abbildung 20: Darstellung der mit blauen Punkten markierten Stellen entlang der SSW, an denen SAM zur Segmentierung aufgerufen wurde sowie das Segmentierungsergebnis, dem gesamten Gleisbereich innerhalb der blau markierten Begrenzungslinie (DOP: © GeoBasis-DE/BKG (2023))

Abbildung 20 zeigt zum einen manuell in hellblau markierte Punkte, welche die Stellen entlang der SSW illustrieren, an denen das Modell zur Segmentierung aufgerufen wurde. Gleichermaßen zeigt die Abbildung die Ausgabe im Hinblick auf die Segmentierung. Hier gibt das Modell den gesamten Gleisbereich aus, d. h. den aufgehellten Bereich im Vergleich zur restlichen DOP-Kachel, eingeschlossen zwischen den blauen Linien. Im Zuge dieser zusätzlichen Validierung zeigte sich somit, dass die Ausgabemaske lediglich den zwischen den Wänden eingeschlossenen Gleisbereich anstelle der SSW abbildet.

Zusammenfassend zeigte die qualitative Analyse, dass DINOv2 das am besten geeignete Backbone-Modell für die spezielle Aufgabenstellung der SSW-Segmentierung ist. Das Modell wies eine leistungsstarke Darstellung auf, ohne dass ein weiterer Trainingsprozess erforderlich war. Dies war ein starkes Indiz dafür, dass das Modell mit etwas Feinabstimmung der DINOv2-Merkmale die Aufgabe effizient und mit einer geringen Anzahl von Annotationen lösen kann.

An dieser Stelle gilt es jedoch zu erwähnen, dass die durchgeführte Analyse die anderen präsentierten Modelle keineswegs als ungeeignet einstuft. Vielmehr verdeutlichte die Analyse und Evaluierung, dass für die anderen Modelle weitere und längere Trainingsprozesse notwendig sind, um die Merkmale so weit zu verbessern, dass sie die Repräsentationskraft der DINOv2-Merkmale erreichen.

6.2 Segmentierungskopf (Segmentation Head)

Zur Bewältigung der speziellen Segmentierungsaufgabe wurde, wie im vorherigen Abschnitt erläutert, zunächst die Effizienz verschiedener Backbone-Modelle im Hinblick auf ihre Repräsentationskraft untersucht. In diesem Prozess wurde DINOv2-Modell als das vielversprechendste Backbone identifiziert. Darauf aufbauend wurde die Komplexität des finalen Modells mit kleinen iterativen Schritten erhöht, um ein vollständiges Segmentierungsmodell zu entwickeln.

Im Zuge dessen wurde zunächst eine lineare Sonde, wie in Abbildung 21 illustriert, als rudimentäres Modell eingesetzt. In Deep Learning[14] bezieht sich eine lineare Sonde (Linear Probe) auf einen einfachen linearen Klassifikator, der auf den vorher trainierten Merkmalen eines neuronalen Netzwerks, typischerweise eines CNN, trainiert wird. Der Begriff Sonde impliziert, dass dieser lineare Klassifikator verwendet

[14] Deep Learning: Teilmenge von ML, wobei neuronale Netze verwendet werden (LeCun, 2015)

wird, um die Repräsentationen zu untersuchen bzw. zu sondieren, die vom neuronalen Netzwerk gelernt wurden. Anstatt das gesamte neuronale Netzwerk fein abzustimmen, was rechenintensiv sein kann, konzentriert sich eine lineare Sonde darauf, nur die Klassifikatorschicht zu trainieren, während der vorher trainierte Merkmalsextraktor (Backbone) eingefroren bleibt (Alain und Bengio, 2018).

Um die Leistung zu verbessern, wurden hier in der Modellarchitektur entsprechend systematisch zusätzliche Faltungsschichten integriert. Faltungsschichten (Convolutional Layer) sind eine grundlegende Komponente eines CNN. Diese Schicht führt Faltungsoperationen auf Eingabedaten durch, um Merkmale wie Kanten, Texturen und Strukturen zu extrahieren. Die Ergebnisse werden als Aktivierungskarten oder Feature Maps, wie bereits in Kapitel 6.1 illustriert, bezeichnet und dienen als Eingabe für nachfolgende Schichten im Netzwerk.

Auf jede dieser Faltungsschichten folgte eine Rectified Linear Unit (ReLU) nicht-lineare Aktivierungsfunktion, welche die Erhaltung wertvoller Informationen in der Nähe der Bildränder durch Polsterungen bzw. Erweiterung, sogenanntes Zero-Padding, sicherstellte. Bei der Anwendung von Zero-Padding werden zusätzliche Nullen um den Rand eines Bildes, also schwarze Pixel, hinzugefügt. ReLU ist eine nichtlineare Aktivierungsfunktion, die häufig in tiefen neuronalen Netzen verwendet wird und dazu beiträgt, das Problem des Verschwindens der berechneten Gradienten zu mildern. Die Aktivierungsfunktion behält für positive Eingaben, das heißt Eingaben größer Null bei Merkmalen, welche den Eingangsdaten ähnlicher sind, den entsprechenden Wert bei, während negative Eingaben, das heißt Eingaben kleiner Null, auf null gesetzt werden. Durch das Entfernen der negativen Komponenten der Aktivierung können neuronale Netze besser trainiert werden, da sie eine effektivere, d. h. der Aufgabe entsprechend gezieltere, Signalübertragung ermöglichen (PyTorch, 2023b).

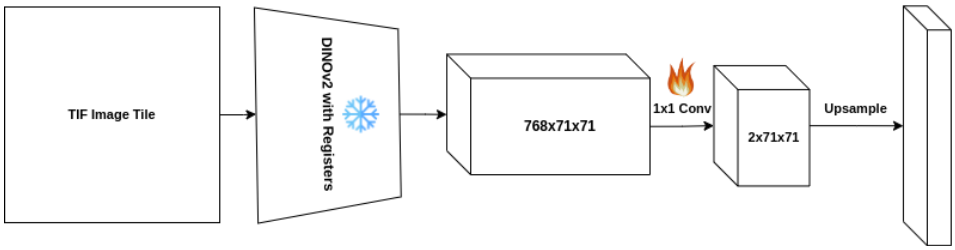


Abbildung 21: Architektur der initial verwendeten linearen Sonde: Von der Eingabe der TIF-Bildkachel auf der linken Seite durch das ViT-Backbone, gefolgt von zwei Faltungsschichten

Dieser Segmentierungskopf, der nach den Faltungsschichten eingesetzt wurde, verwendet am Ende bilineare Interpolation (Upsampling, siehe Abbildung 21), um die Logits an die Bildgröße anzupassen. Logits, auch Wahrscheinlichkeits-Logits genannt, sind die Rohausgaben eines Modells, bevor sie in Wahrscheinlichkeiten umgewandelt werden. Sie repräsentieren die unskalierten Werte der letzten Schichten eines neuronalen Netzwerks und dienen als Grundlage für die Berechnung von Wahrscheinlichkeiten für verschiedene Klassen. Die abschließende Aktivierungsfunktion wird dann auf die Logits angewendet, um Wahrscheinlichkeiten zu generieren, die die Zuordnung zu den verschiedenen Klassen darstellen (Lowe et al., 2022).

In diesem Fall wurde als Aktivierungsfunktion die Sigmoid-Funktion (Nwankpa et al., 2018) angewendet, um die Wahrscheinlichkeit des Vorkommens einer Wand an jedem Pixel vorherzusagen. Die Sigmoid-Funktion, auch als logistische Funktion bezeichnet, ist eine nichtlineare Aktivierungsfunktion, die häufig in neuronalen Netzen verwendet wird. Sie transformiert alle Eingabewerte in einen Wertebereich zwischen 0 und 1, was sie besonders nützlich für binäre Klassifikationsprobleme macht. Somit kann sie die Wahrscheinlichkeit für das Vorliegen einer bestimmten Klasse, in diesem Fall einer vorhandenen SSW, ausgeben.

Das Backbone-Modell wurde wie zuvor beschrieben eingefroren, während der Segmentierungskopf anhand der Dice-Verlustfunktion granular abgestimmt wurde. Die Dice-Verlustfunktion basiert auf dem Dice-Koeffizienten, der die Ähnlichkeit zwischen zwei Mengen misst. In der Bildsegmentierung wird der Dice-Koeffizient verwendet, um die Ähnlichkeit zwischen dem vorhergesagten Segment und dem tatsächlichen Segment zu bewerten. Entsprechend zielt die Dice Loss-Funktion darauf ab, den Dice-Koeffizienten zu maximieren, indem sie die Differenz zwischen dem vorhergesagten und dem tatsächlichen Segment minimiert. Sie ist definiert als das Komplement des Dice-Koeffizienten, was bedeutet, dass ein niedriger Dice Loss einen höheren Dice-Koeffizienten und damit eine bessere Segmentierung anzeigt (Sudre et al., 2017).

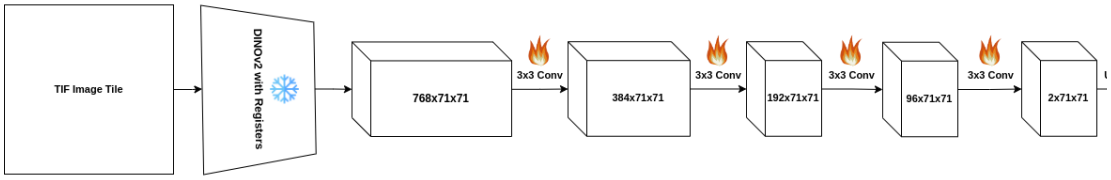


Abbildung 22: Finale Modell-Architektur: Die zuvor dargestellte lineare Sonde wurde um drei weitere Faltungsschichten mit jeweils der Hälfte der Parameter

Die Modelle wurden mit unterschiedlicher Modellkomplexität trainiert und die Ergebnisse des Validierungsdatensatzes im Verlauf des Trainings überwacht, um die beste Modellkonfiguration auszuwählen. Für das Training und die Validierung der Modelle diente der in Kapitel 5.3 erläuterte Trainings- und Validierungsdatensatz von 12.102 DOP-Kacheln.

Für die Bewertung der Durchläufe wurde die beschriebene Dice-Verlustfunktion betrachtet und das Modell mit dem niedrigsten Validierungsverlust (Dice Loss) ausgewählt. Dabei wurden unterschiedliche Parameter in Bezug auf die Tiefe des Netzwerks variiert und getestet.

Die Anzahl der verwendeten Kernel entspricht der Anzahl der Kanäle in der Modellausgabe, wobei in der letzten Schicht zwei Kanäle angestrebt wurden, um die Wahrscheinlichkeits-Logits zu erhalten. Ein Kernel, auch als Filter bekannt, ist eine kleine Matrix, die in neuronalen Netzwerken, insbesondere in CNNs, verwendet wird. Diese Kernel führen Faltungsoperationen auf die Eingabedaten aus, indem sie über diese gleiten, um spezifische Merkmale wie Kanten, Texturen oder Muster zu erkennen. Die Gewichte in der Kernel-Matrix werden während des Trainings angepasst, um die Merkmalsdetektion zu optimieren (Long et al., 2014). Zum Erlangen der Wahrscheinlichkeits-Logits müssen die extrahierten Merkmale in dem finalen Kernel in eine Form verfeinert und verdichtet werden, die zur Berechnung der Wahrscheinlichkeits-Logits verwendet werden kann. Diese Wahrscheinlichkeits-Logits können dann in Wahrscheinlichkeiten für die endgültige Klassifizierung umgewandelt werden. Um die Wahrscheinlichkeiten für beide Klassen (SSW oder nicht SSW) zu erhalten, werden somit zwei finale Kanäle benötigt. Entsprechend wurde die Anzahl der Kernel für jede Tiefe halbiert, um einen gleichmäßigen Übergang zu erzielen.

Die Faltungsschichten starten mit 768, da dies die Ausgabe aus dem DINOv2 Modell ist und werden nach dem Standardvorgehen in ML jeweils halbiert. Entsprechend begannen die darauf aufbauenden Experimente mit 384 3 x 3 Faltungs-Kernel (Convolutional Kernel) und in der letzten Schicht wurden, wie beschrieben, zwei Faltungs-Kernel verwendet. Die Modelle mit zusätzlicher Tiefe der Faltungsschichten wurden mit jeweils einer Liste als [384, 192] und [384, 192, 96] definiert. Dies bedeutet, dass 384 3 x 3 Faltungs-Kernel in der ersten Schicht angewandt wurden, anschließend 192 in der zweiten Schicht, und für ein weiteres Modell 96 in der dritten Schicht, jeweils gefolgt von zwei abschließenden 3 x 3-Faltungen.

Bei den Experimenten wurden somit, wie beschrieben, folgende Architekturen mit unterschiedlicher Tiefe verglichen, welche jeweils mit zwei Faltungs-Kerneln abschließen:

- Lineare Sonde (Linear Probe)
- 384 3 x 3 Faltungsschichten
- [384, 192] 3 x 3 Faltungsschichten
- [384, 192, 96] 3 x 3 Faltungsschichten

Die Ergebnisse der Experimente zeigten hier, dass der Segmentierungskopf, der drei aufeinanderfolgende Faltungsschichten (Convolutional Layers) mit der Anzahl [384, 192, 96] von verwendeten Kerneln in der jeweiligen Schicht die beste Leistung erzielt. Die finale Architektur ist entsprechend in Abbildung 22 dargestellt. Abbildung 23 illustriert den Validierungsverlust, welcher die Entscheidungsgrundlage für das beste Modell bildete, über die Trainings-Epochen hinweg und verdeutlicht die Unterschiede zwischen den verschiedenen Modell-Architekturen. Eine Trainings-Epoche bezieht sich hier auf den Durchlauf durch den gesamten Trainingsdatensatz des Modells bzw. des neuronalen Netzwerks während des Trainingsprozesses. Während einer Epoche werden alle Trainingsdaten einmal dem Netzwerk präsentiert und das Netzwerk aktualisiert seine Gewichte entsprechend den berechneten Fehlern, um die Leistung zu verbessern (Afaq und Rao, 2020).

Darüber hinaus bildet Abbildung 24 den durchschnittlichen Jaccard-Index für jede Trainings-Epoche als zusätzliches Qualitätsmaß und zur weiteren Illustration ab. Der Jaccard-Index ist ein Maß für die Ähnlichkeit zwischen zwei Mengen und beläuft sich auf Werte zwischen 0 und 1, wobei 0 keine Ähnlichkeit und 1 eine perfekte Ähnlichkeit zwischen Mengen repräsentiert (Chandra und Bhattacharya, 2022).

Für jede Architektur wurde das Modell mit fünf verschiedenen zufälligen Seeds, das heißt Startwerten für die Initialisierung des Netzwerks, trainiert. In Abbildung 23 sind entsprechend die Mittelwerte zusammen mit den Konfidenzintervallen dargestellt.

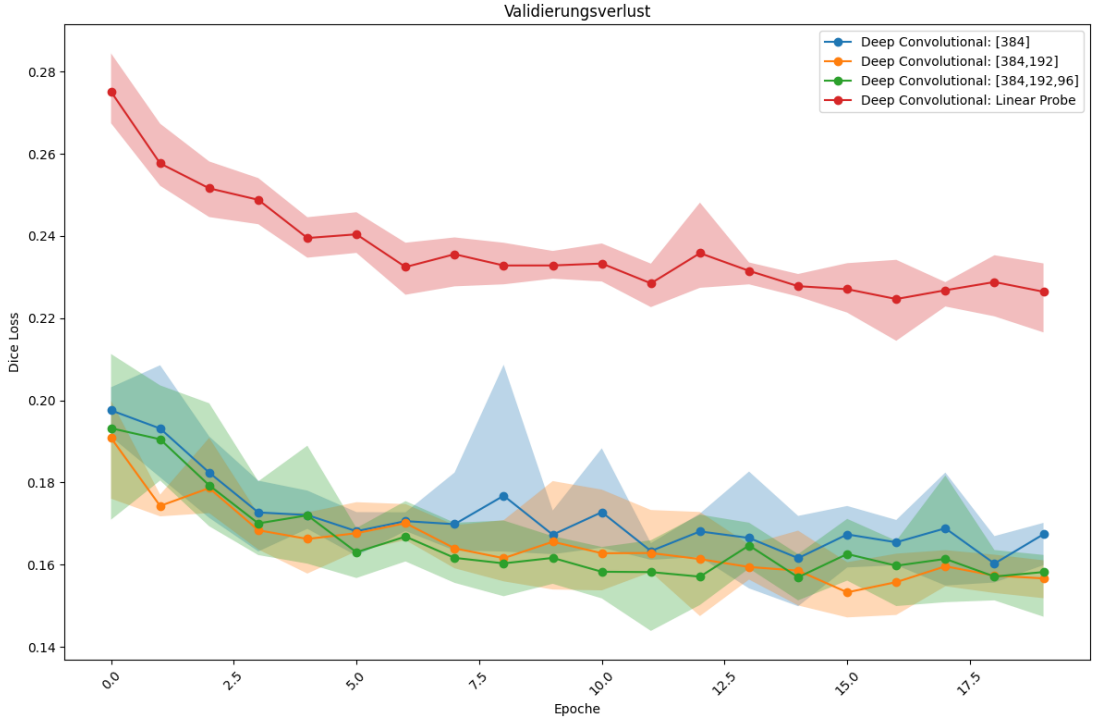


Abbildung 23: Vergleich der der Validierungsverluste gemessen am Dice Loss von unterschiedlichen Architekturen (Linear Probe sowie Architekturen mit zusätzlichen Faltungsschichten) über Trainingsepochen hinweg

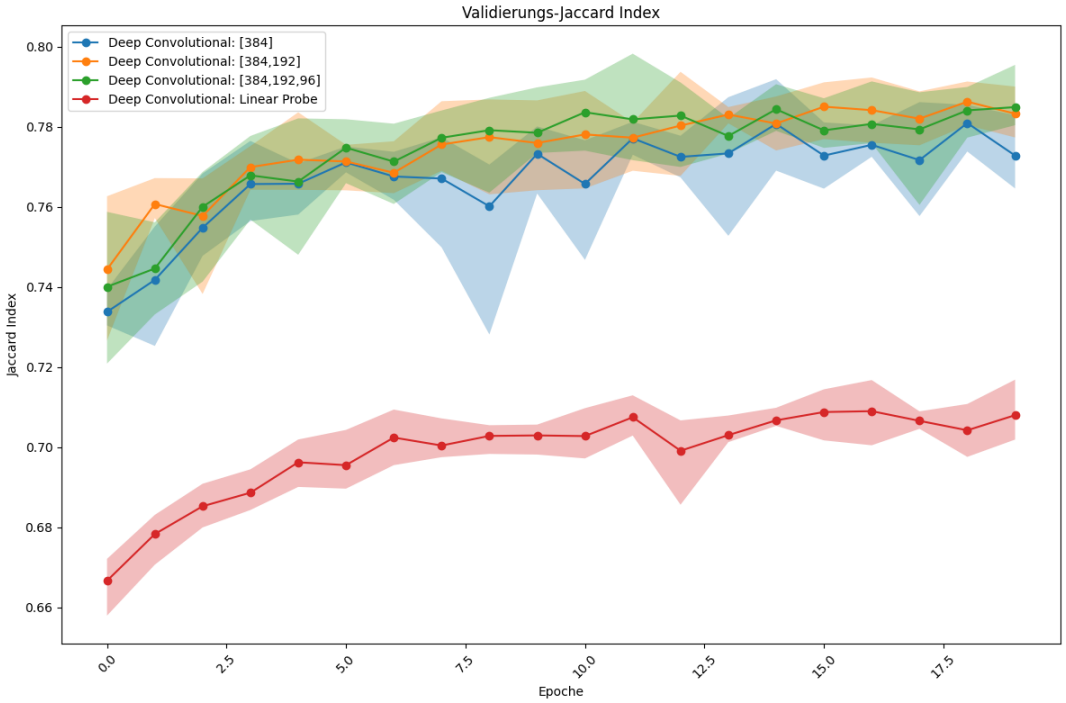


Abbildung 24: Vergleich des Jaccard-Index von unterschiedlichen Architekturen (Linear Probe sowie Architekturen mit zusätzlichen Faltungsschichten) über Trainingsepochen hinweg

Für ein tieferes Verständnis des Trainingsprozesses, einschließlich spezifischer Hyperparameter, sind in der folgenden Tabelle 6 alle relevanten Hyperparameter, die für das Training des beschriebenen Modells verwendet und manuell getestet wurden, dargestellt.

Die **Batch-Größe** stellt die Anzahl der Bilder (DOP-Kacheln) in einem Vorwärts-Durchlauf des neuronalen Netzes während des Trainings dar. Die hier verwendete Größe (8) ist bereits die maximale Batch-Größe, die in den verwendeten Arbeitsspeicher passt. Das Modell wurde mit geringeren Batch-Größen getestet, wobei jedoch eine schlechtere Leistung aufgrund von hoher Varianz der Modellgradienten festgestellt wurde. Beim Training eines neuronalen Netzwerks werden die Gradienten verwendet, um die Gewichtungen und Tendenz (Bias) des Modells iterativ anzupassen und die Leistung entsprechend zu verbessern. Ein stabiler und konsistenter Gradient hilft dem Modell, effizienter zu lernen und schneller zu konvergieren, was hier durch eine größere Batch-Größe erreicht wurde.

Für den zum Trainieren des Modells verwendete Optimierungsalgorithmus, auch kurz **Optimierer** genannt, wurden Stochastic Gradient Descent (SGD) und Adaptive Moment Estimation (ADAM) getestet (Ruder, 2017). SGD ist ein einfacher Optimierungsalgorithmus, der den Gradienten der Verlustfunktion für jeden Datenpunkt im Datensatz berechnet und die Modellparameter basierend darauf aktualisiert. Dabei wird eine konstante Lernrate verwendet und es gibt keine speziellen Anpassungen für unterschiedliche Parameter. ADAM ist im Vergleich zu SGD ein fortschrittlicher Optimierungsalgorithmus, der adaptive Lernraten für jeden Modellparameter berechnet und bewegte Durchschnittswerte der vergangenen Gradienten speichert. Dadurch kann ADAM sich an verschiedene Daten und Probleme anpassen und bietet oft eine schnellere und stabilere Konvergenz im Vergleich zu SGD (Ruder, 2017). Im Zuge der Evaluation des optimalen Modells für die Segmentierungsaufgabe hat sich gezeigt, dass SGD nicht stabil genug war und somit ADAM final ausgewählt wurde.

Die **Lernrate** bezeichnet die Schrittgröße, mit der die Modellgewichte in Abhängigkeit von einem Batch aktualisiert werden. Wenn die Lernrate zu hoch ist, werden die Modellgewichte aggressiver, das heißt öfter, aktualisiert und das Training divergiert. Wenn die Lernrate zu klein ist, wird das Training jedoch

langsamer. Die Lernrate wurde auf einer logarithmischen Skala über die Werte innerhalb der Liste [0,01, 0,001, 0,0001, 0,00001] variiert und derjenige mit der besten Leistung (0,0001) ermittelt.

Wie zuvor beschrieben, definieren die **Epochen** die Anzahl der Male, die das Modell den gesamten Datensatz sieht. Hier wurde zunächst ein hoher Wert festgelegt und nachfolgend das beste Modell in der Epoche mit dem geringsten Verlust gewählt.

Die Größe der **Faltungsschichten** (Hidden Layer Sizes) beschreibt die zuvor im Rahmen der Architektur definierte Tiefe der Faltungsschichten und ist in Abbildung 22 illustriert. **Convolution Stride** bezieht sich auf die Schrittgröße, mit der ein Filter oder Kernel über das Eingangsbild gleitet. Der Stride bestimmt, um wie viele Pixel sich der Filter in jeder Richtung bewegt, wenn er über das Bild gefaltet wird. Ein Stride von 1 bedeutet, dass der Filter Pixel für Pixel über das Bild bewegt wird, wodurch eine dichtere Abtastung des Bildes erfolgt. Das **Basismodell** beschreibt das eingefrorene ViT-Backbone-Modell.

Tabelle 6: Hyperparameter und entsprechend ausgewählte Werte des ausgewählten Segmentierungs-Modells nach manuellen Experimenten

Parameter	**Wert**
Batch Größe	8
Optimierer (Optimizer)	Adam
Verlustfunktion (Loss Function)	Dice Loss
Lernrate (Learning Rate)	0,0001
Epochen	20
Faltungsschichten (Hidden Layer Sizes)	[384,196,96]
Convolution Stride	1
Basismodell	vit_base_patch14_dinov2.lvd142m

Anschließend an die detaillierte Beschreibung und den Hergang der Modellarchitektur werden im Folgenden die Ergebnisse und Evaluation in Bezug auf den spezifischen Anwendungsfall der Segmentierung von SSW beschrieben.

6.3 Evaluation

Mithilfe des entwickelten Modells wurde für die DOP, welche den Bereich von 200 m beiderseits der bundesweiten Gleisstrecken abdecken, die Inferenz, d. h. die vorhergesagten SSW berechnet. Bei stichprobenartiger Überprüfung werden die in den DOP mit dem menschlichen Auge gut identifizierbaren SSW gut durch das Modell erkannt (s.). Zusätzlich werden jedoch auch Strukturen, welche eine ähnliche Geometrie besitzen und einen ähnlichen Schattenwurf verursachen als SSW markiert. Bei visuellen Unterbrechungen der SSW werden diese, auch wenn sie durchgehend sind, in mehrere Objekte unterteilt.

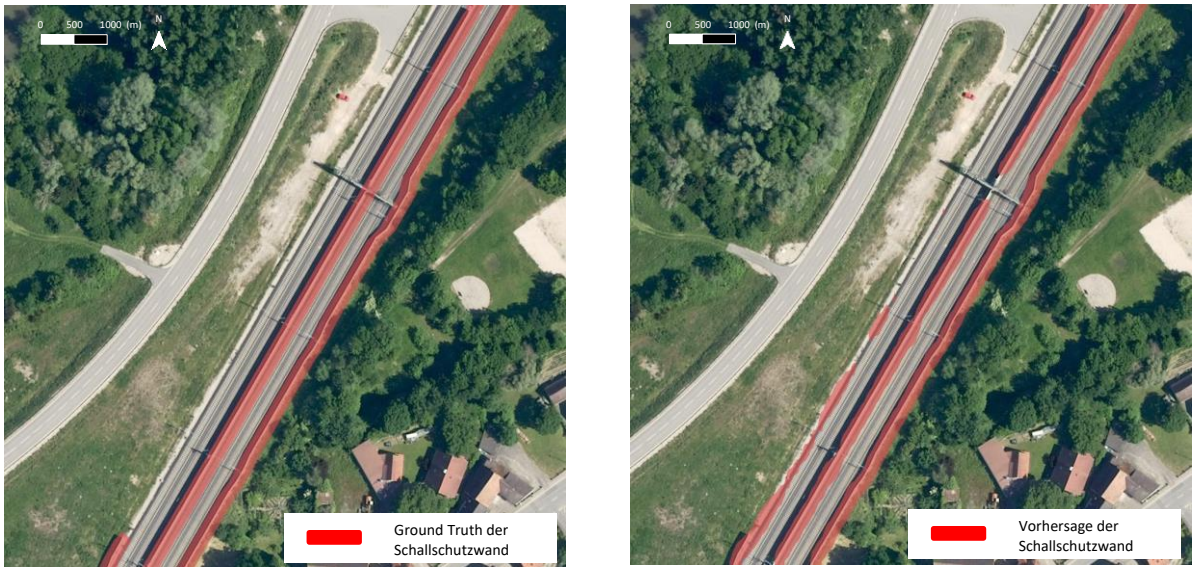


Abbildung 25: Beispiel für die Ground Truth (links) und Vorhersage (rechts) von SSW in einer DOP-Beispielkachel. Die in der Ground Truth vorhandenen SSW werden gut identifiziert, jedoch werden auch zusätzliche, ähnliche Strukturen als SSW erkannt. Eine durchgehende SSW wird aufgrund von einer quer verlaufenden Struktur unterteilt.

Zur objektiven Analyse der genannten Beobachtungen und Evaluierung des entwickelten ML-Modells, welches den geringsten Validierungsverlust erzielte, wurde der manuell annotierte und qualifizierte Datensatz von 400 DOP-Aufnahmen, wie in Kapitel 5.3 beschrieben, verwendet und vorverarbeitet.

Die Metriken der Modell(ergebnis)qualität, die für den manuell annotierten Testdatensatz zur Beurteilung des Segmentierungsmodells ermittelt und evaluiert wurden, sind in Tabelle 7 aufgeführt. Hierbei wurden die für die Segmentierungsaufgabe relevanten Metriken Präzision (Precision)[15], Wiedererkennung (Recall)[16], der F1-Wert[17] sowie der Jaccard Index betrachtet. Darüber hinaus wurden die richtig identifizierten positiven Instanzen, d. h. richtig identifizierte SSW, (True Positives) von insgesamt 2518 SSW in den Testdaten, sowie die falsch identifizierten Instanzen und falsch-negative Instanzen (False Negatives), d. h. SSW, die nicht als solche erkannt wurden, analysiert. Die Tabelle 7 illustriert die Ergebnisse der unterschiedlichen Modell-Architekturen in Bezug auf die Anzahl der Faltungsschichten, welche in Kapitel 6.2 beschrieben wurden, als Mittelwert basierend auf dem gesamten Testdatensatz sowie die jeweilige Standardabweichung. Die angegebenen Metriken und zusätzliche Ergebnisse der unterschiedlichen Modellarchitekturen unterstreichen an dieser Stelle die zuvor getroffene Auswahl des Modells, welches aus drei aufeinanderfolgende Faltungsschichten (Convolutional Layers) mit der Anzahl [384, 192, 96] von verwendeten Kerneln in der jeweiligen Schicht besteht.

[15] Präzision: Verhältnis von richtig positiven Instanzen zur Gesamtzahl der vorhergesagten positiven Instanzen (Saito und Rehmsmeier, 2015)

[16] Recall: Verhältnis der wahrhaft positiven Vorhersagen zur Gesamtzahl der tatsächlich positiven Instanzen (Saito und Rehmsmeier, 2015)

[17] Der F1-Wert ist das harmonische Mittel aus Precision und Recall. Der F1-Wert hilft, wenn man ein Gleichgewicht zwischen falsch-positiven und falsch-negativen Instanzen anstrebt (Saito und Rehmsmeier, 2015)

Tabelle 7: Ergebnisse der Evaluation unterschiedlicher Architekturen basierend auf dem Testdatensatz unter Anwendung der beschriebenen Metriken für die Segmentierungsaufgabe. Die Werte bezeichnen jeweils den Mittelwert berechnet basierend auf dem gesamten Testdatensatz sowie die jeweilige Standardabweichung

Metrik	**Deep Convolutional [384, 192, 96]**	**Deep Convolutional [384, 192]**	**Deep Convolutional [384]**	**Linear Probe**
Precision	0.31 ± 0.016	0.26 ± 0.023	0.20 ± 0.017	0.05 ± 0.0016
Recall	0.74 ± 0.001	0.74 ± 0.017	0.73 ± 0.015	0.632 ± 0.009
F1-Wert	0.43 ± 0.016	0.39 ± 0.022	0.325 ± 0.011	0.094 ± 0.0027
Jaccard Index	0.77 ± 0.002	0.77 ± 0.006	0.76 ± 0.004	0.68 ± 0.0021
True Positives	1267 ± 16	1268 ± 28	1246 ± 25	1076 ± 15
False Positives	2864 ± 220	3588 ± 493	4955 ± 545	20121 ± 943
False Negatives	447 ± 16	446 ± 29	464 ± 26	627 ± 15

Die in der Tabelle 7 genannten Metriken wurden bei einem IoU-Wert von 0,3 berechnet, welcher gewählt wurde, um die Leistungsfähigkeit des Modells bei der verwendeten Datengrundlage bestmöglich widerzuspiegeln. Der IoU-Wert gibt die Überlappung zwischen den vorhergesagten Begrenzungsrahmen und den Ground-Truth-Begrenzungsrahmen an. Er gilt als Schwellenwert im Rahmen der Evaluation und wird zur Ermittlung der richtigen Positiv-Ergebnisse (True Positives) im Vergleich zu falschen Positiv-Ergebnissen (False Positives) verwendet.

Abbildung 26 veranschaulicht die Überschneidung der Modell-Vorhersage (Prediction) mit der tatsächlichen SSW (Ground Truth) bei einem niedrigen IoU-Wert. Hier liegt der IoU-Wert unter 0,5 wobei die Wand sowohl in der Ground Truth als auch in der vorhergesagten Segmentierungsmaske liegt.

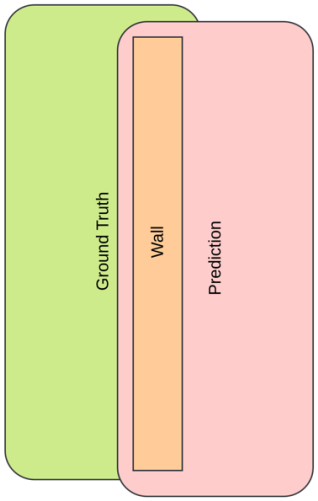


Abbildung 26: Darstellung einer Inferenz, welche einen niedrigen IoU-Wert aufzeigt, obwohl die Wand sowohl in der Ground Truth als auch der Vorhersage enthalten ist

In diesem Datensatz sind niedrige IoU-Werte unter anderem auf die Dilatation der Annotationen um 15 Pixel zurückzuführen, welche dazu führte, dass die Vorhersagen auch bei einer geringeren Überlappung zwischen der Vorhersage und der Ground Truth korrekt sind.

Die Abbildung 27 veranschaulicht den resultierenden Zusammenhang der Metriken Recall und Precision in Abhängigkeit des IoU-Wertes in einem der Experimente. Hierbei ist nochmals zu erkennen, dass eine Verringerung des IoU-Wertes sowohl die Genauigkeit (Precision) als auch die Wiedererkennung (Recall) erhöht.

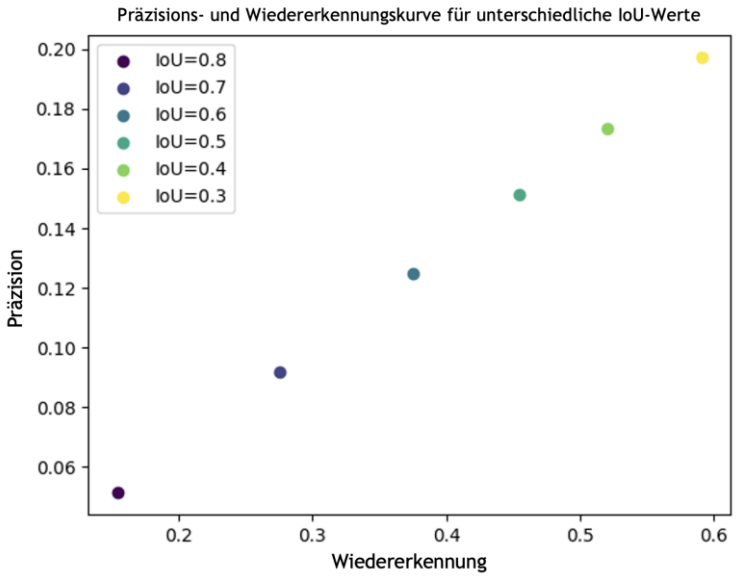


Abbildung 27: Zusammenhang von Recall und Precision in Abhängigkeit des gewählten IoU-Wertes. Precision und Recall steigen beide gleichermaßen bei niedrigerem IoU-Wert

Um die Leistung des Modells besser beurteilen zu können, wurden zudem kleine vorhergesagte Bereiche entfernt, die als Rauschen in der Berechnung angesehen werden können. Da diese kleinen Segmente zu falschen Leistungsindikatoren führen können, ist es wichtig, diese aus der Berechnung der Metriken herauszunehmen. Zur Entfernung dieser vorhergesagten “Inseln” wurde Median-Weichzeichnung (Median Blurring) angewandt sowie die optimale Kernelgröße quantitativ bestimmt, indem zwischen Präzision und Wiedererkennbarkeit abgewogen wurde.

Abbildung 28 illustriert diese Abwägung in Abhängigkeit der Größe des Median-Weichzeichnungs-Kernels. Die Ergebnisse zeigen, dass die Wahl einer Kernelgröße von 7 x 7 die Wiedererkennbarkeit (Recall) nicht reduziert und gleichzeitig die Präzision deutlich erhöht. Aus diesem Grund wurde ein 7 x 7-Kernel verwendet, um kleine Regionen herauszufiltern.

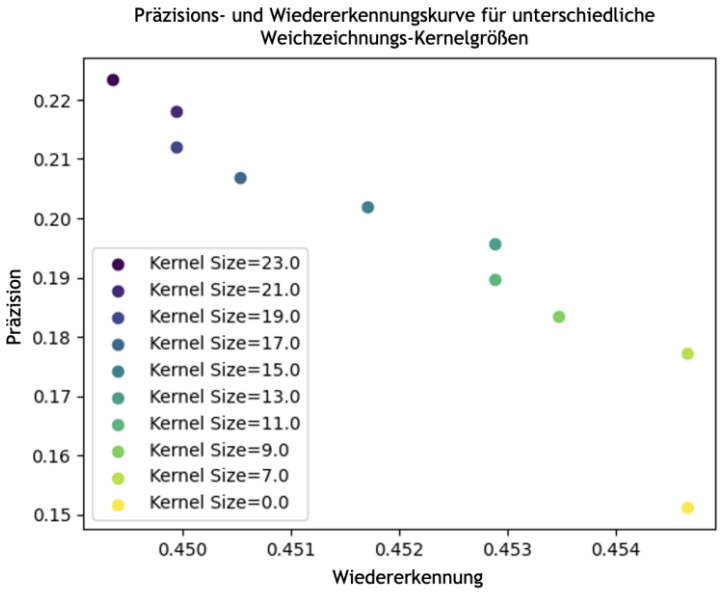


Abbildung 28: Korrelation zwischen Precision und Recall abhängig von unterschiedlich ausgewählten Kernel-Größen des Median-Weichzeichners von 0 bis 23, wobei Precision bei größerem Kernel steigt und der Recall entsprechend sinkt

Um verlässliche Metriken zu gewährleisten, muss sichergestellt werden, dass die richtig vorhergesagten Segmente in den Metriken gut repräsentiert werden. Ein Problem an dieser Stelle war jedoch, dass die vorhergesagten Masken zwar korrekt, aber gleichzeitig fragmentiert waren. Dieser Fall ist in Abbildung 29 illustriert.

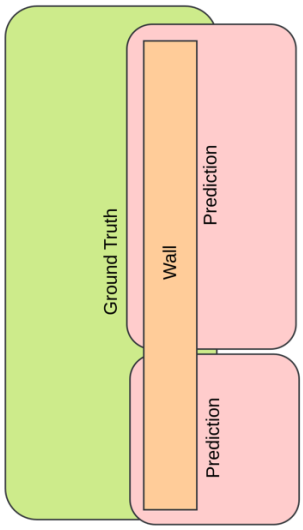


Abbildung 29: Darstellung einer fragmentierten Inferenz, wobei basierend auf der Ground Truth eine einzige SSW vorhanden ist, welche jedoch als zwei unabhängig vorhandene SSW von dem Modell segmentiert wurde

In solch einem Fall wird lediglich eine Vorhersage als richtiger Positiv-Wert (True Positive) gewertet (die Inferenz mit dem höheren IoU-Wert in der Ground Truth), während die andere als falsch positiv (False Positive) gewertet wird.

Dies kann zu niedrigen Jaccard-Index-Werten und niedrigen Wiederkennungswerten (Recall) führen, obwohl sich die vorhergesagten Masken gut mit den echten Wänden überschneiden. Um dies zu vermeiden,

wurden solch nebeneinanderliegende Inferenzen der gleichen Wand-Instanz zugeordnet, auch wenn sie eine bestimmte Anzahl von Pixeln voneinander entfernt sind. Zur gemeinsamen Zuordnung dieser nebeneinanderliegenden Inferenzen wurden Verbindungen zwischen Pixeln, welche n Schritte bzw. n-Hops[18] voneinander entfernt sind, hinzugefügt, indem die Segmentierungsmaske um einen Faktor von n Pixeln erweitert wurde. Um die Konnektivitäts-Parameter bestmöglich zu bestimmen, wurde der Zusammenhang zwischen Präzision und Recall analysiert, während die Konnektivitäts-Werte n variieren. Dieser Zusammenhang ist in Abbildung 30 dargestellt.

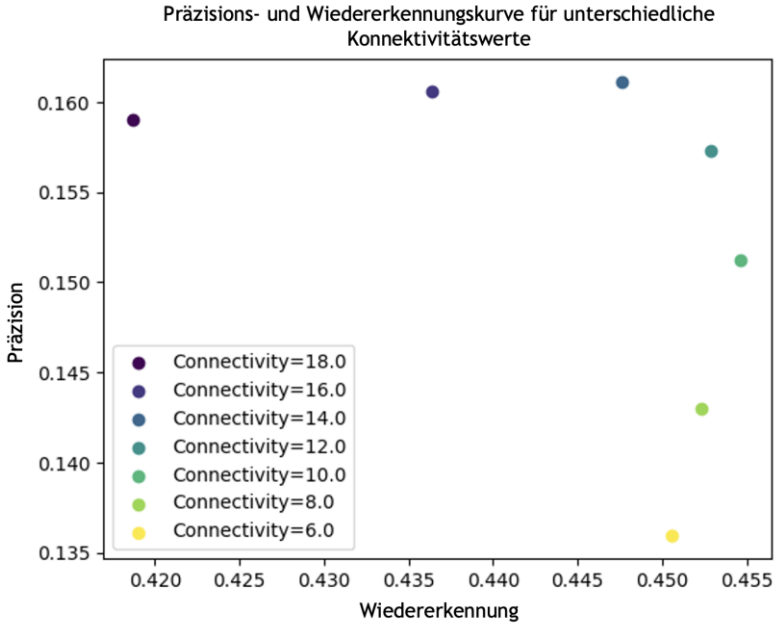


Abbildung 30: Korrelation zwischen Precision und Recall abhängig von unterschiedlichen Konnektivitäts-Werten n, wobei Recall mit einem sinkenden Konnektivitäts-Wert von 18.0 auf bis zu 10.0 steigt

In Abbildung 30 wird ersichtlich, dass ein Konnektivitäts-Wert n von 10.0 bis 12.0 optimal ist, wobei der finale Wert 10.0 gewählt wurde, um die Modell-Metriken basierend auf dem Testdatensatz zu ermitteln.

[18] Der n-Hop-Abstand misst den kürzesten Weg zwischen Pixeln unter Berücksichtigung von Bewegungen, die auf benachbarte Pixel beschränkt sind, wobei die Anzahl der erlaubten Schritte auf n begrenzt ist.

7 Integration – QGIS Plugin

Zusätzlich zur Vorverarbeitung der Daten und zum Aufbau sowie der Evaluierung des eigentlichen ML-Modells zur Segmentierung der SSW wurde eine Schnittstelle zur Verwendung des Modells mit der Open-Source-Software QGIS aufgebaut. Diese Schnittstelle und entsprechende Funktionalitäten sind im Folgenden beschrieben. Um die Interaktion mit dem Modell für Benutzende zu ermöglichen, wurde ein spezifisches Plug-In für QGIS, Deepness, verwendet.

Das Deepness QGIS-Plug-In vereint die gesamte Komplexität hinter einer einfachen Benutzeroberfläche, sodass Nutzende die Bilddaten mit dem bereitgestellten Segmentierungsmodell leicht verarbeiten können. Das Plug-In ist als Open-Source-Version verfügbar und wird aktiv weiterentwickelt, wodurch die stetige Aktualität und Sicherheit gewährleistet ist. Eine ausführliche Dokumentation zu dem Plug-In ist auf der offiziellen Deepness-Webseite zu finden (Aszkowski und Ptak, 2022).

7.1 Modell-Export

Im Anschluss an den Trainingsprozess des Modells, welches mithilfe der PyTorch-Bibliothek (PyTorch, 2023a) implementiert wurde, wurde die beste Version des Modells anhand der Dice Loss Bewertungs-Metrik ausgewählt. Wie in Tabelle 6 auf Seite 51 dargestellt, wurde für das finale Modell die Dice (Loss)-Verlustfunktion verwendet. Diese Verlustfunktion zielt, wie bereits in Kapitel 6.2 beschrieben, darauf ab, den Dice-Koeffizienten zu maximieren, indem sie die Differenz zwischen dem vorhergesagten und dem tatsächlichen Segment minimiert. Die Verwendung von Dice Loss als Verlustfunktion ermöglichte es somit, die Segmentierungsgenauigkeit während des Trainingsprozesses zu optimieren und sicherzustellen, dass das Modell robuste und präzise Segmentierungen erzeugt.

Abschließend wurde das Modell in einen ONNX (Open Neural Network Exchange) Computational Graphen exportiert, was eine standardisierte Modell-Repräsentation darstellt und in QGIS importiert werden kann.

7.2 Funktionalitäten

Mithilfe des Plug-Ins und des standardisierten Formats des Modells können die Funktionalitäten innerhalb von QGIS wie folgt genutzt werden:

Die Benutzenden laden zunächst die erforderlichen Bildebenen sowie die bereitgestellte onnx-Datei des Modells in QGIS in der Oberfläche des Plug-Ins hoch.

- Durch das Aufrufen der Funktion "Load default parameters" werden bereits alle folgenden erforderlichen Optionen und Parameter ausgefüllt. Der wichtigste Parameter für das Modell ist die Auflösung (Resolution). Das Modell wurde basierend auf Bildern mit einer Auflösung von 20 cm/px trainiert, sodass diese auch bei der Inferenz verwendet werden sollten.
- Nachdem alle Parameter angepasst wurden, wird der Inferenzprozess mit dem Klicken auf “Run” gestartet. Je nach CPU-Geschwindigkeit und Anzahl der Kerne dauert dieser Prozess etwa 3 – 10 Minuten für ein Bild der Größe 5.000 x 5.000 Pixel.

Um die Inferenz auf einem kleineren Bildbereich zu beschleunigen, kann in der Maske “Processed area mask” neben der gesamten Bildkachel („Entire Layer“) ebenfalls Sichtbarer Bereich (“Visible Layer”) angegeben werden. Dadurch wird das Modell nur auf dem Teil der Bildebene, welcher derzeit für die Benutzenden über die Benutzeroberfläche sichtbar ist, ausgeführt.

Auf diese Weise können Benutzende in den gewünschten Bereich hineinzoomen und die Berechnung der Segmentierung auf weniger als zehn Sekunden reduzieren. Bei dieser Annahme wird vorausgesetzt, dass der Bildbereich klein genug ist.

Bei der Auswahl des Inferenzbereichs gilt jedoch zu beachten, dass der ausgewählte Bildbereich zur Verarbeitung in Kacheln von 1000 x 1000 Pixel zerlegt wird, wodurch sich die Berechnung und Ausgabe des Modells leicht unterscheiden kann. Abhängig von dem ausgewählten Bildausschnitt in der Oberfläche weichen die Umrisse der Segmentierungen minimal ab, jedoch werden die gleichen SSW identifiziert und segmentiert. Zusätzlich betrachtet das Modell bei jeder Berechnung die gesamte zur Verfügung gestellte Umgebung, was bei unterschiedlicher Auswahl ebenfalls eine minimal andere Entscheidungsgrundlage für das Modell mit sich bringt.

Die Modellergebnisse und entsprechende Segmentierungs-Masken werden als neue Schicht in QGIS erstellt. Die Benutzenden können somit die Segmentierungs-Ergebnisse leicht auf Pixelebene nachvollziehen und im Detail analysieren, ob das Modell eine SSW an einer bestimmten Stelle inferenziert bzw. identifiziert hat.

Abbildung 31 zeigt die visuelle Oberfläche und entsprechend eingestellte Parameter in der Deepness Plug-In-Oberfläche innerhalb der QGIS-Software, nachdem die Bilddatei sowie das Modell geladen wurden.

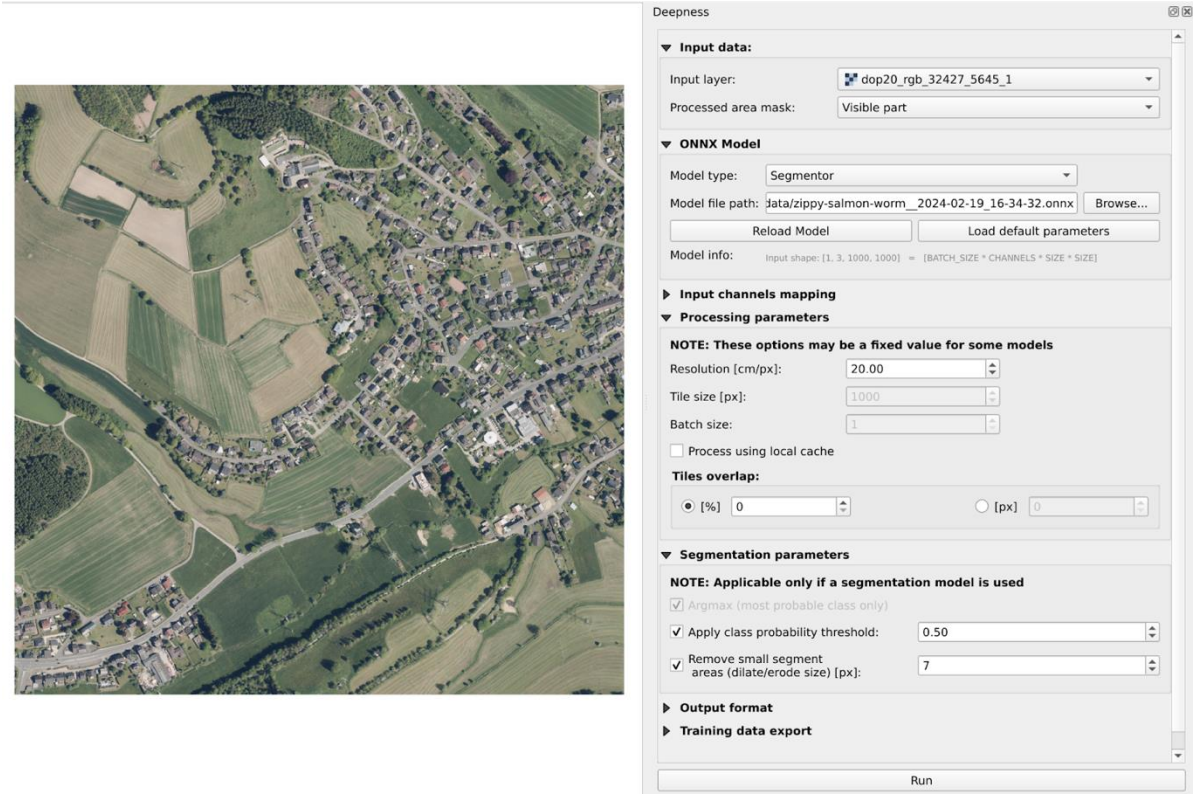


Abbildung 31: Oberfläche des Deepness-Plugins innerhalb von QGIS

8 Diskussion und Fazit

8.1 Diskussion der Ergebnisse

Während der Datenvorverarbeitung und des Aufbaus des ML-Modells wurden bereits wertvolle Erkenntnisse in Bezug auf die Möglichkeiten und Herausforderungen basierend auf den vorhandenen Bilddaten für das Projekt „Automatisierte digitale Bestandserfassung gleisnaher Infrastruktur aus Befliegungsdaten“ gewonnen. Das Verhältnis Objektgröße zu Bodenauflösung, sowie die grundsätzlich schwere Erkennbarkeit der SSW in den verwendeten Befliegungsdaten und die entsprechend erschwerte Annotation der Bilddaten, stellen eine Herausforderung für die Vorverarbeitung der Daten und die Inferenz dar. Durch die abweichenden Annotationen erforderte der Trainingsdatensatz zur Entwicklung des maschinellen Lernmodells weitere Anpassungen. Allgemein wurde der Trainingsdatensatz aufgrund von dem geringen Flächenanteil an SSW in den Bildern und der daraus resultierenden Unterrepräsentanz der Zielobjekte auf einen kleineren Teil der Bilddaten, welche SSW enthalten, reduziert. Diese Reduzierung des Trainingsdatensatzes hatte aufgrund großer Flächen ohne SSW jedoch keine Nachteile für das Training und die Validierung des Modells.

Zur Auswahl des geeigneten Backbone-Modells zur Extraktion der Merkmale wurde eine detaillierte qualitative Analyse unterschiedlicher Segmentierungs-Ergebnisse mithilfe von Aktivierungskarten durchgeführt. Im Zuge dessen wurde das DINOv2-Modell als bestgeeignetes Backbone-Modell identifiziert, worauf im nächsten Schritt der Segmentierungskopf zur spezifischen Segmentierung der Objekte aufbaut. An dieser Stelle wurden ebenfalls, basierend auf mehreren Experimenten, optimale Parameter für das ML-Modell sowie die Modell-Architektur mit drei aufeinanderfolgenden Faltungsschichten (Convolutional Layers) mit der Anzahl [384, 192, 96] von verwendeten Kerneln in der jeweiligen Schicht identifiziert und ausgewählt. Zusätzliche Schritte, wie die Filterung der Inferenzergebnisse zur Entfernung besonders kleiner Objekte, sowie die verbesserte Zuordnung nebeneinanderliegender Inferenzen konnten ebenfalls zur Verbesserung der Modell-Metriken beitragen. Ergebnisse zeigen, dass SSW mithilfe von ML bereits gut segmentiert und identifiziert werden können. In Abhängigkeit des gewählten IoU-Wertes und Parameter wurde ein Jaccard-Index von 0,77, sowie ein Präzisionswert (Precision) von 0,31 und Wiedererkennungswert (Recall) von 0,74 erreicht.

Bei der Verwendung und Angabe der Metriken ist jedoch zu beachten, dass die Evaluation der Ergebnisse basierend auf der pixelweisen Verarbeitung des ML-Modells erschwert wurde. Das Modell entscheidet bzw. segmentiert für jeden Pixel der eingegebenen Bilddaten, ob innerhalb des Pixels eine SSW existiert, sodass kein geometrisches Objekt ausgegeben wird. Durch diese pixelweise Segmentierung von SSW konnte entsprechend keine konkrete Anzahl an segmentierten SSW angegeben werden. Auch lässt sich durch äußert unterschiedliche Längenverhältnisse von SSW über den bundesweiten Raum keine durchschnittliche Länge und ein entsprechender Mittelwert aller segmentierten Pixel berechnen. Zudem verdeutlichte Abbildung 29, dass eine SSW oft so segmentiert wurde, als gäbe es mehrere kürzere SSW. Solche fragmentierten Segmentierungen visualisieren noch immer die korrekte Stelle der SSW in den Bilddaten, jedoch verfälscht dies die Evaluation der Segmentierung entsprechend. Gleichermaßen ist der angegebene IoU-Wert stark abhängig von der korrekten Überlagerung der Annotation und Segmentierung des Modells, wie in Abbildung 26 illustriert. Wenn die Annotationen und Segmentierungen deutlich abweichen, obwohl die gleiche SSW markiert wurde, kommt es dennoch zu falsch-positiven (False Positive) Ergebnissen und es wird ein geringer IoU-Wert ausgegeben.

Bei der Ergebnisauswertung wurde ebenfalls festgestellt, dass viele falsch-positive (False Positive) Ergebnisse in Form von kleinen „Inseln“ auftreten, d. h. sehr kleine Masken, welche aufgrund hoher Aktivierungen erscheinen. Dies lässt sich auf die Korrelation mit dem Auftreten von Mauern und einigen spezifischen Bereichen, wie z. B. der Schattengebung, in der Nähe der Bahnlinie zurückführen, was fehlerhafte Modell-

Vorhersagen provozieren kann. Um die Ausgabe von zu vielen kleinen Masken und isolierten Inseln zu verhindern bzw. zu vermindern, können Standardtechniken, wie dem Filtern von (zu) kleinen Flächen, eingesetzt werden. Um fragmentiert erkannte Abschnitte einer SSW als zusammenhängend bzw. durchlaufend zu erkennen, kann auf graphenbasierte Algorithmen zurückgegriffen werden. Da eine solche Lösung nicht im onnx-Datei-Format implementiert werden kann, hätte eine entsprechende Implementierung den Umfang und die Komplexität des Vorhabens maßgeblich erhöht und den Rahmen dieses Forschungsprojektes überschritten.

8.2 Zusammenfassung und Fazit

Die ersten Schritte und Resultate des Projekts bestanden in der Erfassung und Priorisierung des relevanten Anwendungsfalls sowie einer Anforderungsanalyse und entsprechenden Definition der funktionalen und nicht-funktionalen Anforderungen an das Projekt und Resultate (Kapitel 2). Darüber hinaus bildete eine umfassenden Literaturrecherche (Kapitel 3) die Grundlage für die technologische Implementierung des Prozessierungs-Workflows sowie die Gewinnung und Analyse der öffentlich zugänglichen Daten. Dadurch wurde ein umfassendes Verständnis für den Anwendungsfall der Lärmkartierung und die möglichen Daten als Grundlage zur Entwicklung der ML-Lösung (Kapitel 6) geschaffen. Gleichermaßen gilt an dieser Stelle anzumerken, dass der Fokus der Literaturrecherche in Bezug auf die ML-Entwicklung stark auf SAM als mögliches Backbone-Modell für das finale Modell lag. Im Zuge der schnellen Entwicklungsgeschwindigkeit im ML-Umfeld bildete die Recherche im Frühjahr 2023 nicht den tatsächlichen Forschungsstand zum Ende des Jahre 2023 ab. Somit wurde auch im Rahmen der Modellentwicklung weitere Literatur und Forschung, u. a. zu DINO als Backbone, beleuchtet und zur Entwicklung verwendet.

Aufbauend auf der ursprünglichen Anforderungsanalyse und der Recherche wurde als Grundlage und erstes Ziel ein bundesweiter Datensatz für das Training und die Erkennung von SSW für das ML-Modell erstellt. Für diesen Datensatz wurden mehrere Datenquellen wie Satellitendaten, DOM-, DGM und DOP-Daten verwendet und deren Nutzen für den spezifischen Anwendungsfall bewertet. Die Auswahl dieser Daten basierte insbesondere auf der Einschränkung und Anforderung zur Nutzung rein öffentlich zugänglicher sowie bundesweit vorhandener Daten.

Während dieser Datenuntersuchung und entsprechenden Datenvorverarbeitung wurde deutlich, dass die Verwendung von Open-Source-Satellitendaten für die Detektion von SSW in diesem Projekt aufgrund der geringen Auflösung ungeeignet ist. Die fehlende Detailgenauigkeit und Dimensionierung verhindern die präzise Erkennung von Objekten. Aus diesem Grund wurde im ersten Schritt eine Datengrundlage aus DOP, DGM und DOM geschaffen und, im Hinblick auf die Annotation, insbesondere auch auf die DOP konzentriert. Während des Annotations-Prozesses traten jedoch Herausforderungen bei der SSW-Erkennung auf. Diese Schwierigkeiten waren teilweise auf die geringe Auflösung der Aufnahmen, dichte Vegetation, welche Teile der SSW verdeckt, sowie die Verwechslungsgefahr mit anderen Objekten, wie Leitplanken, zurückzuführen. Im Zuge des Annotations-Prozesses waren insbesondere Schatten um den Gleisbereich hilfreich für Annotierende zur Identifizierung und Entscheidung von vorhandenen SSW. Darüber hinaus wurde festgestellt, dass die von der DB InfraGO AG bereitgestellten Annotationen (siehe Kapitel 5.3), obwohl quantitativ ausreichend, in Bezug auf die Lokalisierungsgenauigkeit bei der Überlagerung mit DOP Schwächen aufwiesen.

Die geringe Qualität sowie die schwierige Erkennbarkeit der SSW in den zugrundeliegenden Aufnahmen und die entsprechend erschwerte Annotation der Bilddaten stellten somit eine Herausforderung für die Vorverarbeitung der Daten dar. Durch die abweichenden Annotationen erforderte der Trainingsdatensatz zur Entwicklung des maschinellen Lernmodells weitere Anpassungen, welche in Kapitel 5.3 beschrieben wurden. Obwohl die Qualität des Abgleichs zwischen DOP-, DOM- und DGM-Daten sich im Zuge der Datenevaluierung als vielversprechend erwies, wurde die weitere Verwendung der Höhendaten (DOM) im Zuge der Modell- und Ergebnisevaluation ausgeschlossen. Aufgrund der geringeren Auflösung wurden bei der Verwendung der Höhendaten (DOM und DGM) deutlich mehr Bereiche fälschlicherweise als SSW

klassifiziert. Die Genauigkeit ist hierbei neben der Auflösung der Bilddaten insbesondere durch die beschriebene zeitliche Diskrepanz der Aufnahmen beeinträchtigt. Entsprechend wurden die DOM- und DGM-Daten für den Trainingsprozess ausgeschlossen, da diese nicht zur besseren Klassifizierung beigetragen haben.

Entsprechend konnten von allen zur Verfügung stehenden und betrachteten Daten lediglich die DOP-Datensätze zielführend für das ML-Modell im Rahmen des Prozessierungs-Workflows verwendet werden. Darüber hinaus wurde der Trainingsdatensatz aufgrund von nicht vorhandenen SSW in den Bildern und der daraus resultierenden Unterrepräsentanz der Zielobjekte auf 12.102 DOP-Kacheln, welche SSW enthalten, reduziert. Dies hatte aufgrund der Wahl des Backbone-Modells DINO v2 jedoch keine Nachteile für das Training und die Validierung des Modells.

Durch die beschriebene Selektion und Bereinigung der Daten konnte eine bundesweite Datengrundlage geschaffen werden, welche nicht nur eine einheitliche Informationsgrundlage darstellt, sondern auch für die Segmentierung weiterer Infrastrukturobjekte sowie für das Trainieren und Evaluieren zukünftiger ML-Modelle verwendet werden kann.

Unter Nutzung der geschaffenen Datenbasis konnte eine ML-Lösung basierend auf einem bestehenden, vortrainierten Backbone-Modelle mit weiterer Feinabstimmung (fine-tuning) implementiert werden, welche erfolgreich SSW mit Hilfe der DOP-Daten erkennt. Durch den erstellten Datensatz und die Implementierung der ML-Lösung wurden somit bundesweite Ergebnisse zu den berechneten Inferenzen, d. h. segmentierte SSW, erarbeitet und zur Verfügung gestellt.

Die Ergebnisse des Modells, illustriert in Kapitel 6.3, zeigen gute Werte auf Basis der Wiedererkennung (Recall), jedoch relativ niedrige Werte bei der Präzision (Precision). Dies bedeutet, dass das Modell entweder zu viele Vorhersagen zu SSW, die nicht vorhanden sind, macht oder dass die entsprechenden Vorhersagen fragmentiert sind. Eine solch fragmentierte Berechnung wurde beispielhaft in Abbildung 29 dargestellt.

Allgemein sind diese Ergebnisse durch die in Kapitel 5.3 beschriebenen ungenauen Annotationen der SSW, die Dilatation der Beschriftungen und die schwierige visuelle Repräsentation und Erkennbarkeit der SSW aus der Vogelperspektive begründet. Darüber hinaus lässt sich das bestehende Problem der Fragmentierung nicht gut in den Metriken abbilden. Hier wird meist nur eine von mehreren Inferenzen als „richtige“ SSW und damit entsprechend viele weitere Inferenzen als falsch erkannte SSW bewertet, obwohl mehrere erkannte Objekte zu einer einzigen SSW gehören. Entsprechend weist das Modell, basierend auf den relevanten Metriken, scheinbar schlechtere Ergebnisse auf als die tatsächlich gewonnenen visuellen Resultate in der Ausgabemaske.

Die Ergebnisse der ML-Modellierung und insbesondere der Experimente liefern bereits wertvolle Erkenntnisse für das Projekt. Dabei wurden neben den Herausforderungen, insbesondere in Bezug auf die Qualität und Komplexität der Datengrundlage, aber auch zahlreiche Möglichkeiten zur technologisch unterstützten SSW-Erkennung identifiziert. Es wurde gezeigt, dass die automatisierte Erkennung und Vorhersage von SSW mit Hilfe von ML-Methoden bereits erfolgreich möglich ist. Zu dieser Segmentierung von Infrastrukturobjekten basierend auf Befliegungsdaten haben in erster Linie innovative Modelle und Entwicklungen im Bereich Computer Vision beigetragen. Insbesondere in den letzten Monaten haben sich die frei verfügbaren (Open Source) Modelle, welche bereits Objekte ohne eine große Anzahl an annotierten Daten erkennen können, deutlich verbessert.

Durch die Verwendung des Modells mit Hilfe eines Plug-Ins innerhalb von QGIS können Nutzende die erarbeitete Lösung ohne Verwendung einer weiteren Software- oder Programmierkenntnissen zur Erkennung von SSW verwenden. Somit wurde eine nahtlose Integration des Prozessierungs-Workflows und Modells sichergestellt und das Modell kann zukünftig auch auf neue Input-Daten innerhalb der Software angewendet werden. Wie in Kapitel 7.2 beschrieben, lässt sich in der Oberfläche des Plug-Ins auch der zu

berechnende Inferenzbereich auswählen, sodass bei der Berechnung lediglich der gewünschte Ausschnitt (Area of Interest) ausgewählt werden kann. Über die Verwendung des Segmentierungsmodells hinaus bietet das Plug-In die Möglichkeit, Modelle für weitere Anwendungsfälle oder die Segmentierung anderer Infrastrukturobjekte zu verwenden. Das Identifizieren weiterer Infrastrukturobjekte bildet eine wichtige Grundlage, damit Anwendende in vollem Umfang von dieser Automatisierung profitieren können.

Da das Plug-In auf einer Open-Source-Lösung basiert, ist die stetige Weiterentwicklung und Aktualität dieser Schnittstelle zukünftig ebenfalls gegeben. Durch die Einbindung des Modells und insbesondere der bundesweiten Inferenz-Ergebnisse in QGIS können Nutzende außerdem zusätzliche Datenquellen als weitere Ebene (auf den Inferenzergebnissen) hinzufügen. Entsprechend können die beschriebenen Höhendaten (DOM), trotz Ausschluss im Rahmen des Trainings und der ML-Modellentwicklung, im Anschluss an die Inferenz als weitere Daten-Ebene in QGIS die Identifizierung von SSW untermauern. An dieser Stelle gilt es anzumerken, dass die Verfügbarkeit von Höhendaten mit höherer Auflösung wichtig ist, um das 3D-Modell der SSW vollständig zu erstellen und eine nachgelagerte Lärmkartierung zu ermöglichen.

In Anbetracht der erreichten Segmentierungs-Ergebnisse sowie deren Darstellung sind, neben dem Hinzufügen der Höhendaten, entsprechend weitere Anpassungen an dem möglichen Input sowie den Resultaten empfehlenswert, um die Überprüfung der Eingangsdaten für die Lärmkartierung und den entsprechenden Abgleich zu verbessern bzw. den Aufwand zu vermindern. Mögliche Ansätze seitens der Input-Daten sowie des Modells und Prozessierungs-Workflows werden im Folgenden abschließend erläutert.

8.3 Ausblick und Handlungsempfehlungen

Innerhalb des Forschungsprojektes wurde ein umfassender, einheitlicher Datensatz von Befliegungsdaten zur Ermittlung bzw. Segmentierung von SSW mit Hilfe des entwickelten Prozessierungs-Workflows und Modells erstellt. Die in Kapitel 6.3 vorgestellten Ergebnisse unterstreichen das Potential der entwickelten Lösung zur Segmentierung von SSW. Nutzende können die Implementierung mit Hilfe eines Plug-Ins innerhalb der QGIS-Oberfläche unterstützend zur Lärmkartierung verwenden, sowie die Standorte der SSW verifizieren.

Zur ausschließlichen Verwendung des Prozessierungs-Workflows und der Ergebnisse des ML-Modells zur Identifizierung der SSW sowie zur Lärmkartierung sind jedoch noch weitere Schritte und Anpassungen zur Optimierung der Lösung nötig. Trotz der guten Ergebnisse und der bereits bestehenden Möglichkeit, SSW basierend auf Befliegungsdaten zu analysieren, brachte die Verwendung von Fernerkundungsdaten für den speziellen Anwendungsfall der Lärmkartierung sowie die vorhandene Basis an annotierten Bilddaten gleichzeitig Limitierungen mit sich. Zur Verbesserung der Auflösung sowie Verringerung der zeitlichen Diskrepanz der verwendeten Aufnahmen ist es zukünftig empfehlenswert neben den öffentlich zugänglichen Daten auch weitere, aktuellere und räumlich höher aufgelöste Datenquellen, wie kostenpflichtige Satellitendaten oder zusätzliche Befliegungsdaten mit einer deutlich höheren Auflösung in Betracht zu ziehen. Bei der zukünftigen Auswahl und Beschaffung von Datenquellen sollte durchweg ein einheitlicher Zeitrahmen bei der Erstellung der Bilddaten im Fokus stehen, um die zeitliche Diskrepanz und differenzierte Informationsgrundlage zu verbessern.

Allgemein sind SSW in Luftaufnahmen deutlich schwerer zu erkennen und zu differenzieren als aus der Boden- und Gleisperspektive. Hier wäre zu evaluieren, inwiefern eine vollständige Erstellung eines Datensatzes mit Hilfe von Umgebungsaufnahmen im Zuge der Befahrung von Gleisen machbar ist. Mit umfassenden Bilddaten, die mehrere Perspektiven beinhalten, können ML-Modelle ein breites Spektrum an Informationen und Merkmalen zur Segmentierung von SSW erlernen und anwenden.

Bei der Verwendung der annotierten Daten der DB InfraGO AG wurden ebenfalls zahlreiche Ungenauigkeit der Annotationen, das heißt Beschriftungen und Lokalisierung der SSW, festgestellt. Die durchgeführten Experimente und die Evaluation unterschiedlicher Datenquellen sowie ML-Modelle (sichtbar in den Kosinusähnlichkeitskarten in den Abbildungen) zeigten deutlich, dass die Segmentierung der SSW mit höherer Genauigkeit durchgeführt werden kann, wenn qualitativ hochwertige und genauere Annotationen zur Verfügung stehen. Aus diesem Grund sollte, ähnlich den manuellen Annotationen (siehe Kapitel 5.3), der Fokus zukünftig ebenfalls auf der intensiven Verbesserung und Korrektur der Annotationen liegen, um die Qualität der Daten als Grundlage für das ML-Modell zu erhöhen. Zudem ist die Bereitstellung eines qualitativ hochwertigen Ground Truth-Datensatzes essenziell für die Weiterentwicklung des Modells und sollte entsprechend priorisiert werden. Ein Ground Truth-Datensatz beschreibt eine Sammlung von Daten, welche als Referenz für das Training und die Bewertung der ML-Modelle verwendet werden kann. Dieser Datensatz enthält die korrekten und genauen Annotationen für die Eingabebilder. An der Stelle gilt es anzumerken, dass die Qualität der Annotationen, d. h. vollständig, korrekt und präzise gesetzte Markierungen der SSW in den Bildaufnahmen, entscheidend sind. Insbesondere bei der Verwendung des beschriebenen Modells gilt, dass einige sehr gute Annotationen wichtiger sind als zahlreiche durchschnittlich annotierte Bilddaten, welche Fehler oder fehlende Annotationen beinhalten.

Im Hinblick auf die Datengrundlage gilt es ebenfalls anzumerken, dass die Erarbeitung einer bundesweit einheitlichen Bilddatengrundlage mit hoher Auflösung und qualitativ hochwertigen Annotationen maßgeblich zur weiteren Entwicklung des Modells beitragen kann.

In Bezug auf die Verbesserung des Modells kann in einem nächsten Schritt die Anpassung des Backbone-Modells (DINOv2) (Russell et al., 2024) untersucht werden, um selbstüberwachtes Lernen auf einer großen Menge von Daten durchzuführen, die keine SSW enthalten. Dies könnte die Leistungsfähigkeit des Modells erhöhen. Die Verwendung von Transformer-basierten Segmentierungsköpfen wie Mask2Former (Cheng et al., 2022) könnte ebenfalls die Qualität der Segmentierung verbessern, vorausgesetzt, sie werden mit einem qualitativ hochwertigen Datensatz trainiert. Ein Nachbearbeitungsschritt mit bedingten Zufallsfeldern könnte zudem eine konsistentere Segmentierungsmaske erzeugen, die weniger anfällig für Fragmentierung ist.

Über den Anwendungsfall der Lärmkartierung entlang von Schienenwegen hinaus lässt sich der erarbeitete Prozessierungs-Workflow und das ML-Modell bereits auf die Segmentierung von SSW in anderen Umgebungen, wie z. B. Bundesstraßen oder Autobahnen, anwenden. Obwohl der Fokus des Modells und der entsprechend ausgewählten Bildausschnitte zum Trainieren und Testen auf der Erkennung von SSW in der Umgebung von Gleisanlagen lag, baut das Modell auf einem Backbone-Modell auf, welches mit zahlreichen heterogenen Daten trainiert wurde. Die Datengrundlage für das bisherige Modell und den speziellen Anwendungsfall der Lärmkartierung wurde jedoch so abgestimmt und reduziert, dass sich die Bilddaten auf die Umgebung von Gleisanlagen beschränken. Somit ist es empfehlenswert, das Modell bzw. den Segmentation Head zukünftig nochmals mit annotierten Bilddaten von SSW entlang von Straßen zu trainieren, um das Modell für die spezielle Aufgabe zu verbessern. Anhand der diskutierten Ergebnisse und Segmentierungen des Modells ist allerdings festzustellen, dass das Modell bereits SSW entlang von Straßen erkennt und diese segmentieren kann.

Das Projekt „Automatisierte digitale Bestandserfassung gleisnaher Infrastruktur aus Befliegungsdaten“ schafft somit nicht nur einen entscheidenden Schritt zur Automatisierung der Lärmkartierung, sondern ebenfalls eine Grundlage für weitere Anwendungsfälle und die Segmentierung von Infrastrukturelementen rund um den Schienen- und Straßenverkehr.

Quellenverzeichnis

Afaq, S. and Rao, S., 2020. Significance Of Epochs On Training A Neural Network. *International Journal of Scientific & Technology Research,* 9(6).

Alain, G. und Bengio, Y., 2018. Understanding intermediate layers using linear classifier probes. *arXiv*. doi: https://arxiv.org/abs/1610.01644.

AlMarzouqi, H. und Saoud, L. S., 2022. Semantic Labeling of High Resolution Images Using EfficientUNets and Transformers. *arXiv*. doi: https://doi.org/10.48550/arXiv.2206.09731.

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 2024. Produkt- und Qualitätsstandard für Digitale Oberflächenmodelle (DOM). [Online], Verfügbar unter: https://www.adv-online.de/AdV-Produkte/Standards-und-Produktblaetter/Standards-der-Geotopographie/binarywriterservlet?imgUid=da14073e-de6b-1f71-96e7-436303dd7d12&uBasVariant=11111111-1111-1111-1111-111111111111 [Zugriff am 27.05.2024].

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 2024. Produkt- und Qualitätsstandard für Digitale Geländemodelle. [Online], Verfügbar unter: https://www.adv-online.de/AdV-Produkte/Standards-und-Produktblaetter/Standards-der-Geotopographie/binarywriterservlet?imgUid=2b14073e-de6b-1f71-96e7-436303dd7d12&uBasVariant=11111111-1111-1111-1111-111111111111 [Zugriff am 27.05.2024].

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 2024. Produkt- und Qualitätsstandard für Digitale Orthophotos. [Online], Verfügbar unter: https://www.adv-online.de/AdV-Produkte/Standards-und-Produktblaetter/Standards-der-Geotopographie/binarywriterservlet?imgUid=75419114-249e-4711-1fea-f5203b36c4c2&uBasVariant=11111111-1111-1111-1111-111111111111 [Zugriff am 27.05.2024].

Aszkowski, P. und Ptak, B., 2022. Home — QGIS: Deepness: Deep Neural Remote Sensing 0.6.3 documentation. [online] Verfügbar unter: https://qgis-plugin-deepness.readthedocs.io/en/latest/ [Zugriff am 24.05.2024].

Ayyar, M., Benois-Pineau, J. und Zemmari, A., 2023. Chapter 5 – A feature understanding method for explanation of image classification by convolutional neural networks. In: Explainable Deep Learning AI. s.l.:Academic Press, pp. 79 – 96.

Bundesministerium für Digitales und Verkehr (BMDV), 2024. Lärmvorsorge und Lärmsanierung an Schienenwegen. [online] Verfügbar unter: https://bmdv.bund.de/SharedDocs/DE/Artikel/E/schiene-laerm-umwelt-klimaschutz/laermvorsorge-und-laermsanierung.html, [Zugriff am 24.05.2024].

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. und Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. *arXiv*. doi: https://doi.org/10.48550/arXiv.2005.12872.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. und Joulin, A., 2021. Emerging Properties in Self-Supervised Vision Transformers. *arXiv*. doi: https://doi.org/10.48550/arXiv.2104.14294.

Cha, K., Seo, J. und Lee, T., 2023. A Billion-scale Foundation Model for Remote Sensing Images. *arXiv*. doi: https://doi.org/10.48550/arXiv.2304.05215.

Chandra, N. K. und Bhattacharya, S., 2022. Chapter 3 – Dependent Bayesian multiple hypothesis testing. Handbook of Statistics. s.l.:Elsevier, pp. 67 – 81.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. und Yuille, A. L., 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. arXiv. doi: https://doi.org/10.48550/arXiv.1606.00915.

Chen, T., Zhu, L., Ding, C., Cao, R., Wang, Y., Li, Z., Sun, L., Mao, P. und Zang, Y., 2023. SAM Fails to Segment Anything? – SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. *arXiv*. doi: https://doi.org/10.48550/arXiv.2304.09148.

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J. und Qiao, Y., 2022. Vision Transformer Adapter for Dense Predictions. *arXiv*. doi: https://doi.org/10.48550/arXiv.2205.08534.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. und Girdhar, R., 2022. Masked-attention Mask Transformer for Universal Image Segmentation. *arXiv*. doi: https://doi.org/10.48550/arXiv.2112.01527.

Cheng, B., Schwing, A. G. und Kirillov, A., 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. *arXiv*. doi: https://doi.org/10.48550/arXiv.2107.06278.

Deutsche Bahn AG, 2022. Schienennetz Deutsche Bahn. [online] Verfügbar unter: https://data.deutschebahn.com/dataset/geo-strecke.html, [Zugriff am 14.07.2023].

Deutsche Bahn AG, 2023. Grün an der Bahn – Wie die DB Bäume und Sträucher an ihren Strecken pflegt. [Online], Verfügbar unter: https://www.deutschebahn.com/de/presse/suche_Medienpakete/medienpaket_vegetationsmanagement-6854346, [Zugriff am 14.07.2023].

DB Engineering & Consulting (DB E.C.O. Group), 2022. „Computer Vision“ – Transformationsprozess in eine digitale und effizientere Bahnwelt. [Online] Verfügbar unter: https://db-eco.com/de/computer-vision-transformationsprozess-in-eine-digitale-und-effizientere-bahnwelt/, [Zugriff am 14.07.2023].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. und Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*. doi: https://doi.org/10.48550/arXiv.2010.11929

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. und Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. [Online], Verfügbar unter: https://huggingface.co/timm/vit_base_patch8_224.dino, [Zugriff am 27.05.2024].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. und Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. [Online], Verfügbar unter: https://huggingface.co/timm/vit_small_patch8_224.dino, [Zugriff am 27.05.2024].

Eisenbahn-Bundesamt, 2023. GeoPortal.EBA. [Online], Verfügbar unter: https://geoportal.eisenbahn-bundesamt.de/, [Zugriff am 04.08.2023].

Eisenbahn-Bundesamt, 2023. Lärm an Schienenwegen: Lärmkartierung. [Online], Verfügbar unter: https://www.eba.bund.de/DE/Themen/Laerm_an_Schienenwegen/Laermkartierung/laermkartierung_node.html, [Zugriff am 04.08.2023].

EUR-Lex, 2022. Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise. [Online], Verfügbar unter: https://eur-lex.europa.eu/eli/dir/2002/49/2020-03-25, [Zugriff am 04.08.2023].

European Space Agency, 2024. Copernicus Programme. [online] Verfügbar unter: https://sentiwiki.copernicus.eu/web/copernicus-programme. [Zugriff am 21.05.2024].

Frick, A., Stöckigt, B. und Wagner, K., 2021. Ableitung des Baumbestandes entlang des deutschen Schienennetzes, Dresden: Deutschen Zentrums für Schienenverkehrsforschung beim Eisenbahn-Bundesamt.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R. und Valko, M., 2020. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv*. doi: https://doi.org/10.48550/arXiv.2006.07733.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N. und Freeman, W. T., 2022. Unsupervised Semantic Segmentation by Distilling Feature Correspondences. *arXiv*. doi: https://doi.org/10.48550/arXiv.2203.08414.

Han, J., Kamber, M. und Pei, J., 2012. 2 – Getting to Know Your Data. In: Data Mining (Third Edition). s.l.:Morgan Kaufmann, pp. 39 – 82.

Hasberg, A., Hofmann, K., Bott, F. und Hoffmeister, D., 2021. Anforderungskatalog für eine webbasierte Plattform zur Bereitstellung, Darstellung und Analyse von Geodaten – mHUB-B (mFUND), Dresden: Deutschen Zentrums für Schienenverkehrsforschung beim Eisenbahn-Bundesamt.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. und Girshick, R., 2021. Masked Autoencoders Are Scalable Vision Learners. *arXiv*. doi: https://doi.org/10.48550/arXiv.2111.06377.

He, K., Zhang, X., Ren, S. und Sun, J., 2015. Deep Residual Learning for Image Recognition. *arXiv*. doi: https://doi.org/10.48550/arXiv.1512.03385.

Hénaff, O. J., Koppula, S., Shelhamer, E., Zoran, D., Jaegle, A., Zisserman, A., Carreira, J. und Arandjelović, R., 2022. Object discovery and representation networks. *arXiv*. doi: https://doi.org/10.48550/arXiv.2203.08777.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., und Chen, W., 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv*. doi: https://doi.org/10.48550/arXiv.2106.09685.

Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.-W., Tang, C.-K. und Yu, F., 2023. Segment Anything in High Quality.*arXiv*. Doi: https://doi.org/10.48550/arXiv.2306.01567.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D. und Hadsell, R., 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, Band 114, p. 3521 – 3526. Doi: https://doi.org/10.48550/arXiv.1612.00796.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P. und Girshick, R., 2023. Segment Anything. *arXiv*. Doi: https://doi.org/10.48550/arXiv.2304.02643.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P. und Girshick, R., 2023. Segment Anything. [Online], Verfügbar unter: https://huggingface.co/facebook/sam-vit-base#model-details, [Zugriff am 27.05.2024].

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P. und Girshick, R., 2023. Segment Anything. [Online], Verfügbar unter: https://huggingface.co/facebook/sam-vit-large, [Zugriff am 27.05.2024].

LeCun, Y., Bengio, Y. und Hinton, G., 2015. Deep learning. Nature, Band 521, p. 436 – 444.

Li, R., Zheng, S., Duan, C., Su, J. und Zhang, C., 2020. Multi-stage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images. *arXiv*. doi: https://doi.org/10.48550/arXiv.2011.14302.

Li, Y., Mao, H., Girshick, R. und H, K., 2022. Exploring Plain Vision Transformer Backbones for Object Detection. *arXiv*. doi: https://doi.org/10.48550/arXiv.2203.16527.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. und Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv*. doi: https://doi.org/10.48550/arXiv.2103.14030.

Long, J., Shelhamer, E. und Darrel, T., 2014. Fully Convolutional Networks for Semantic Segmentation. *arXiv*. doi: https://doi.org/10.48550/arXiv.1411.4038.

Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M. Y., Zhu, X. X., Zhang, L. und Li, D., 2020. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances and Million-AID. *arXiv*. doi: https://doi.org/10.48550/arXiv.2006.12485.

Nwankpa, C., Ijomah, W., Gachagan, A. und Marshall, S., 2018. Activation Functions: Comparison of trends in Practice and Research for Deep Learning. *arXiv*. doi: https://doi.org/10.48550/arXiv.1811.03378.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H. und Maira, J., 2023. DINOv2: Learning Robust Visual Features without Supervision. *arXiv*. doi: https://doi.org/10.48550/arXiv.2304.07193.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Maira, J., 2023. DINOv2: Learning Robust Visual Features without Supervision. [Online], Verfügbar unter: https://huggingface.co/facebook/dinov2-small, [Zugriff am 27.05.2024].

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Maira, J., 2023. DINOv2: Learning Robust Visual Features without Supervision. [Online], Verfügbar unter: https://huggingface.co/facebook/dinov2-base, [Zugriff am 27.05.2024].

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Maira, J., 2023. DINOv2: Learning Robust Visual

Features without Supervision. [Online], Verfügbar unter: https://huggingface.co/facebook/dinov2-large, [Zugriff am 27.05.2024].

Preußler, V., Fricke, K., Bott, F., Schulz, C., Kleinschmit, B. und Tintrup, G., 2024. SENSchiene -Satelliten-gestützte Erfassung von Flächeneigenschaften und Nutzungsveränderungen im Umfeld des Verkehrs-trägers Schiene. [online] Verfügbar unter: https://www.dzsf.bund.de/SharedDocs/Down-loads/DZSF/Veroeffentlichungen/Fachveroeffentlichungen/2024/2024-05_Tagungs-band_dgpf.pdf?__blob=publicationFile&amp, [Zugriff am 24.05.2024].

PyTorch, 2023. PyTorch. [online] Pytorch.org. Verfügbar unter: https://pytorch.org/, [Zugriff am 18.03.2024].

PyTorch, 2023. PyTorch - ReLU. [Online], Verfügbar unter: https://pytorch.org/docs/stable/gener-ated/torch.nn.ReLU.html, [Zugriff am 18.03.2024].

QGIS Association, 2024. QGIS Geographic Information System. [Online] Verfügbar unter: https://www.qgis.org, [Zugriff am 25.04.2024].

Robinson, J., Chuang, C.-Y., Sra, S. und Jegelka, S., 2020. Contrastive Learning with Hard Negative Sam-ples. *arXiv*. doi: https://doi.org/10.48550/arXiv.2010.04592.

Ranftl, R., Bochkovskiy, A. und Koltu, V., 2021. Vision Transformers for Dense Prediction. *arXiv*. doi: https://doi.org/10.48550/arXiv.2103.13413.

Ronneberger, O., Fischer, P. und Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*. doi: https://doi.org/10.48550/arXiv.1505.04597.

Ruder, S., 2017. An overview of gradient descent optimization algorithms. Dublin. *arXiv*. doi: https://doi.org/10.48550/arXiv.1609.04747.

Saito, T. und Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One, 10(3). doi: https://doi.org/10.1371/journal.pone.0118432.

Siemens AG, 2023. Siemens Aktiengesellschaft, safe trAIn - Mobilität neu denken - Sichere KI am Bei-spiel fahrerloser Regionalzug. [Online], Verfügbar unter: https://safetrain-projekt.de/, [Zugriff am 14.07.2023].

Squirrel Developer Team, 2024. Squirrel: A Python library that enables ML teams to share, load, and transform data in a collaborative, flexible, and efficient way. [online] GitHub. Verfügbar unter: https://github.com/merantix-momentum/squirrel-core, [Zugriff am 24.05.2024].

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. und Cardoso, C. M. J., 2017. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In: Lecture Notes in Computer Science. s.l.:Springer International Publishing, p. 240 – 248. doi: https://doi.org/10.1007/978-3-319-67558-9_28.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., He, Q., Yang, G., Wang, R., Lu, J. und Fu, K., 2022. RingMo: A Remote Sensing Foundation Model with Masked Image Modeling. IEEE Transactions on Geoscience and Remote Sensing, Band 61. doi: https://doi.org/10.1109/TGRS.2022.3194732.

Lowe, S.C., Earle, R., d'Eon, J., Trappenberg, T. and Oore, S., 2022. Logical Activation Functions: Logit-space equivalents of Probabilistic Boolean Operators. *NeurIPS 2022*. Verfügbar unter: https://openreview.net/forum?id=m6HNNpQO8dc [Zugriff am 27.05.2024].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. und Polosukhin, I., 2017. Attention Is All You Need. *arXiv*. doi: https://doi.org/10.48550/arXiv.1706.03762.

Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D. und Zhang, L., 2022. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Mode. *arXiv*. doi: https://doi.org/10.48550/arXiv.2208.03987.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W. und Xiao, B., 2019. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv*. doi: https://doi.org/10.48550/arXiv.1908.07919.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X. und Atkinson, P. M., 2021. UNetFormer: A UNet-like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. arXiv. doi: https://doi.org/10.48550/arXiv.2109.08937.

Wang, L., Li, R., Duan, C., Zhang, C., Meng, X. und Fang, S., 2021. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. arXiv. doi: https://doi.org/10.48550/arXiv.2104.12137.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y. und Sun, J., 2018. Unified Perceptual Parsing for Scene Understanding. *arXiv*. doi: https://doi.org/10.48550/arXiv.1807.10221.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M. und Luo, P., 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv*. doi: https://doi.org/10.48550/arXiv.2105.15203.

Yamazaki, K., Hanyu, T., Tran, M., Garcia, A., Tran, A., McCann, R., Liao, H., Rainwater, C., Adkins, M., Molthan, A., Cothren, J. und Le, N., 2023. AerialFormer: Multi-resolution Transformer for Aerial Image Segmentation. *arXiv*. doi: https://doi.org/10.48550/arXiv.2306.06842.

Anhänge

Beispiele Datenfusion

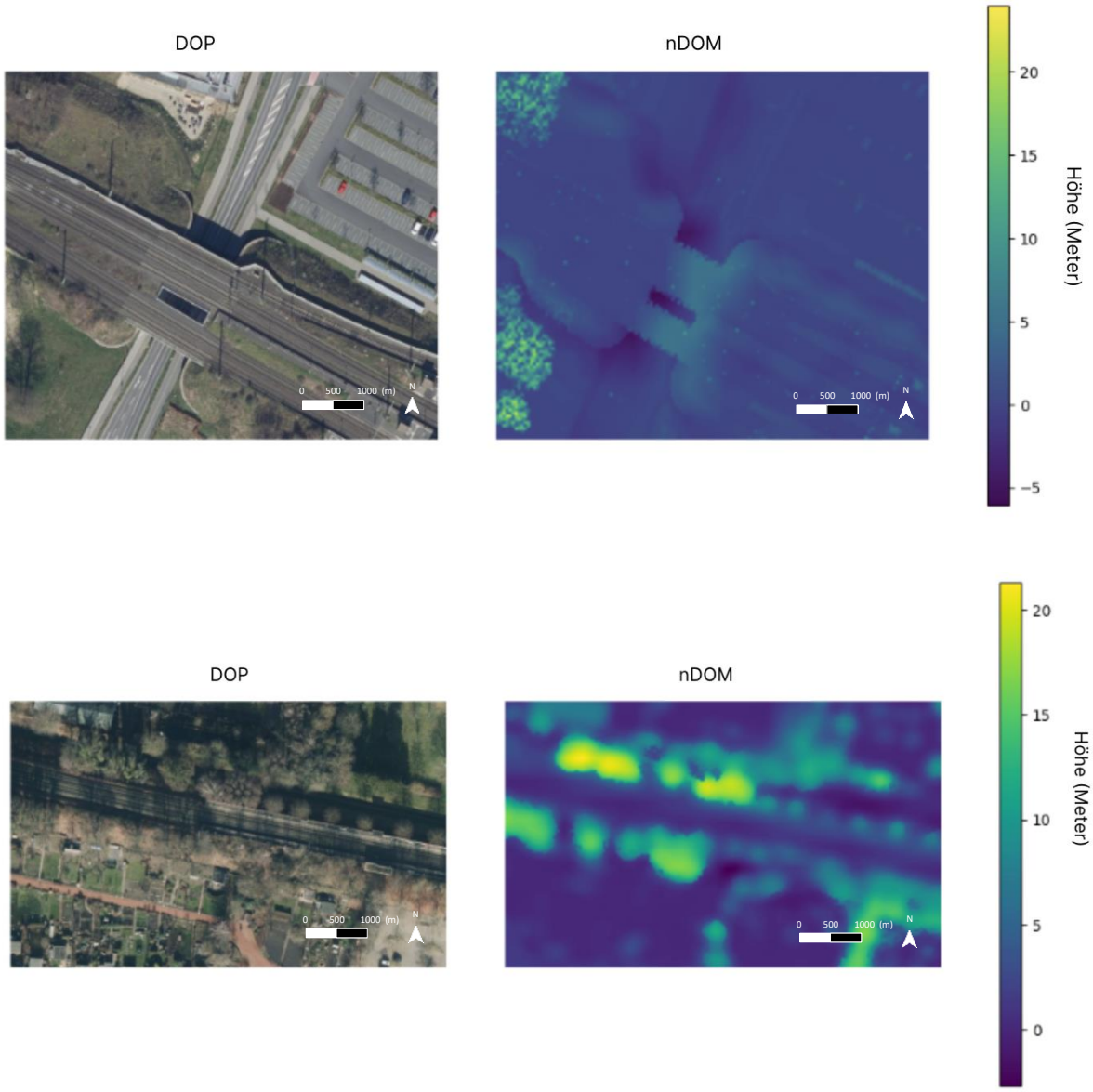


Abbildung 32: Beispielhafte Abbildung einer Bildkachel nach der Datenfusion von DOM, DGM und DOP

Backbone Vergleiche

Im Folgenden sind weitere Beispiele zu finden, die bei der Wahl des Backbone-Modells zur Entscheidung beigetragen haben.

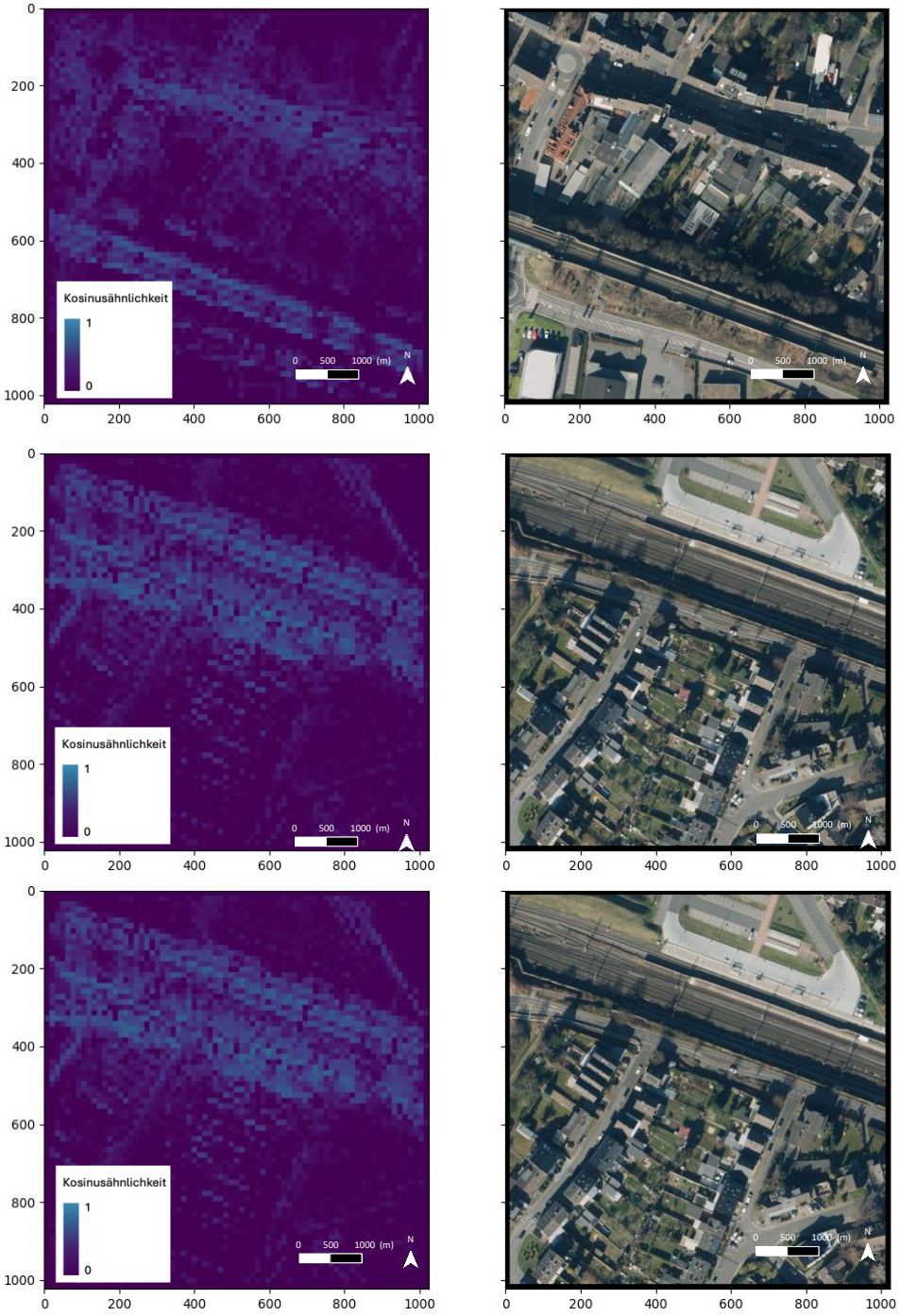


Abbildung 33: DINOv2 Klein (DOP: © GeoBasis-DE/BKG (2023)).

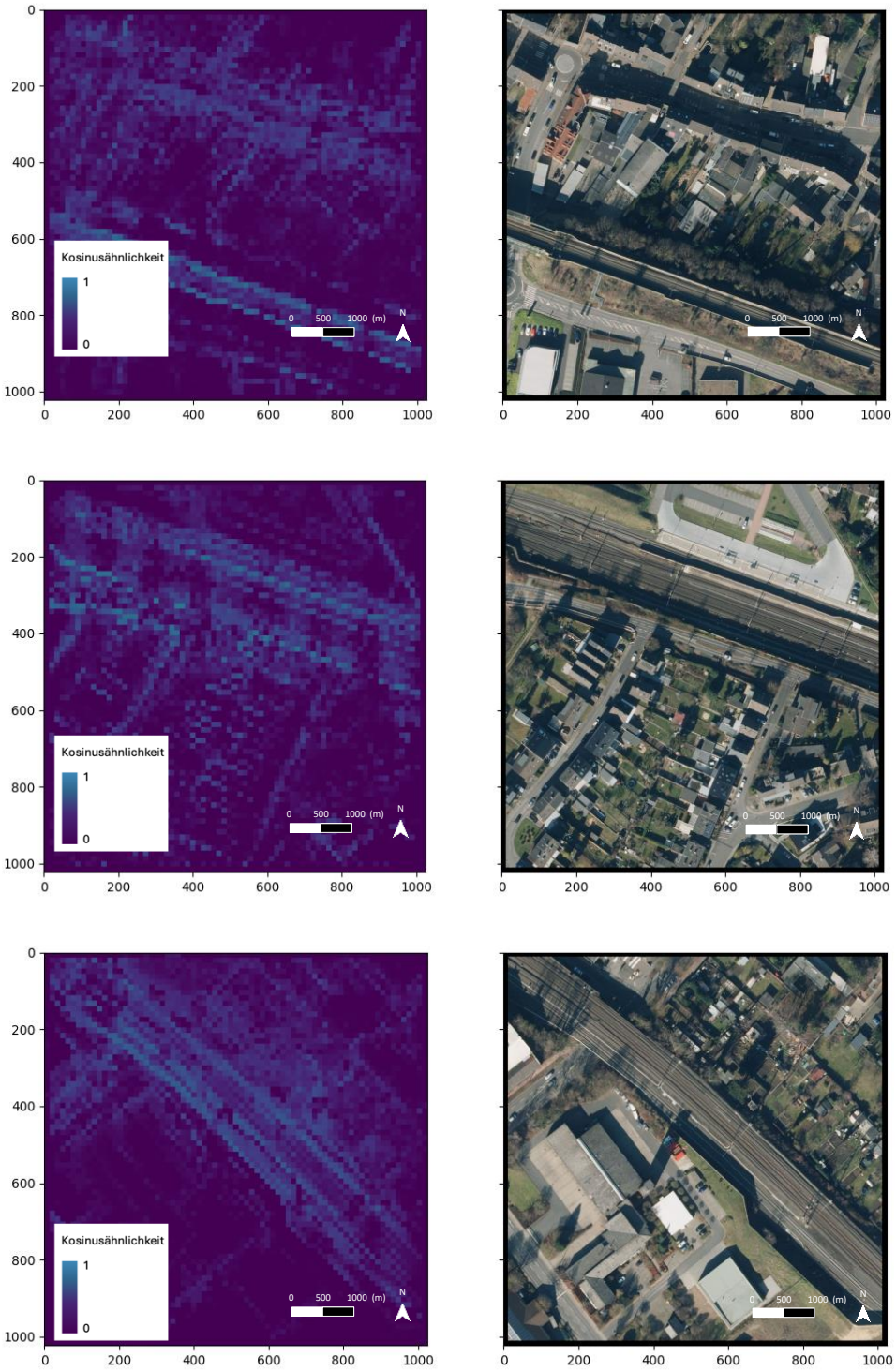


Abbildung 34: DINOv2 Basis (DOP: © GeoBasis-DE/BKG (2023)).

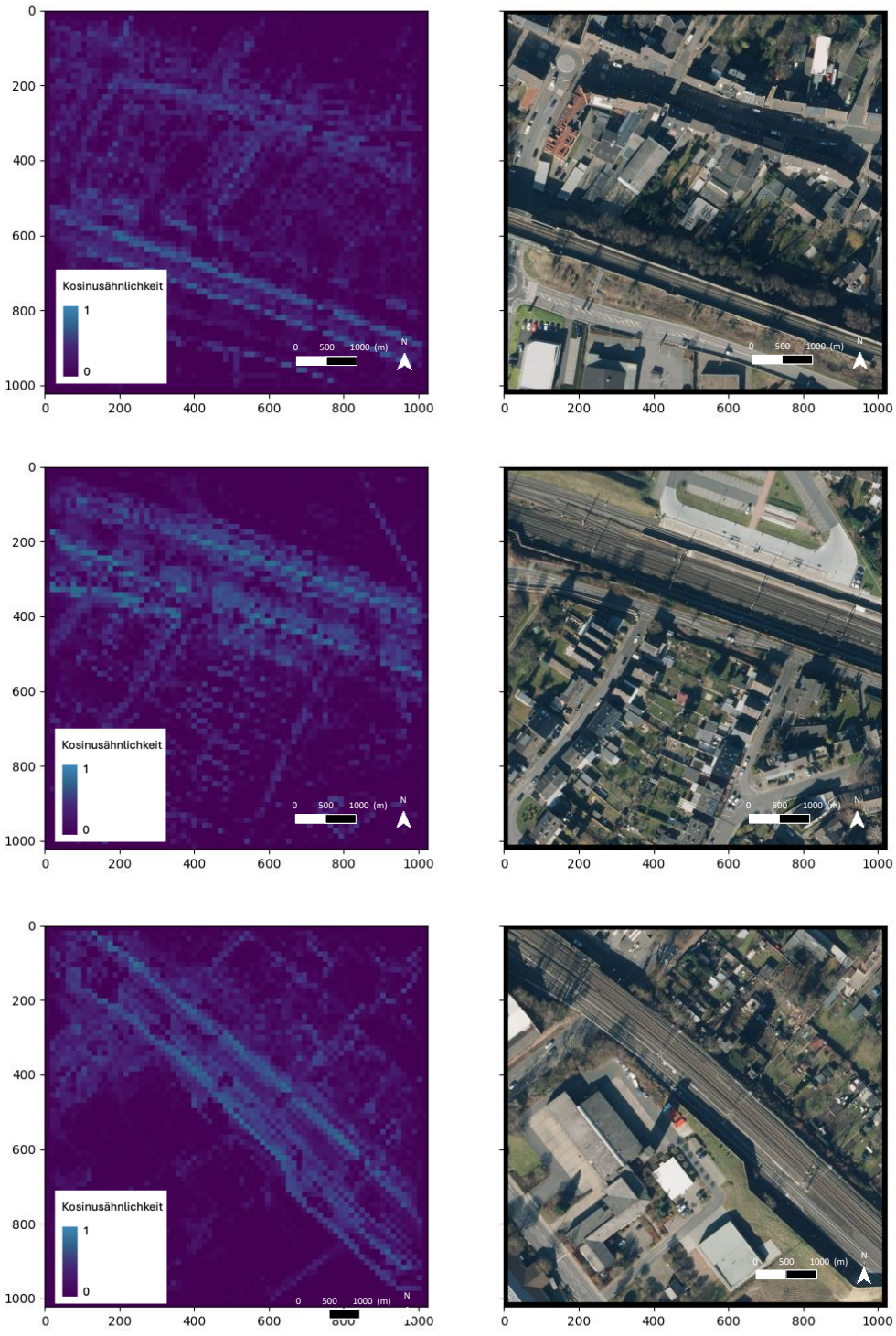


Abbildung 35: DINOv2 Groß (DOP: © GeoBasis-DE/BKG (2023)).

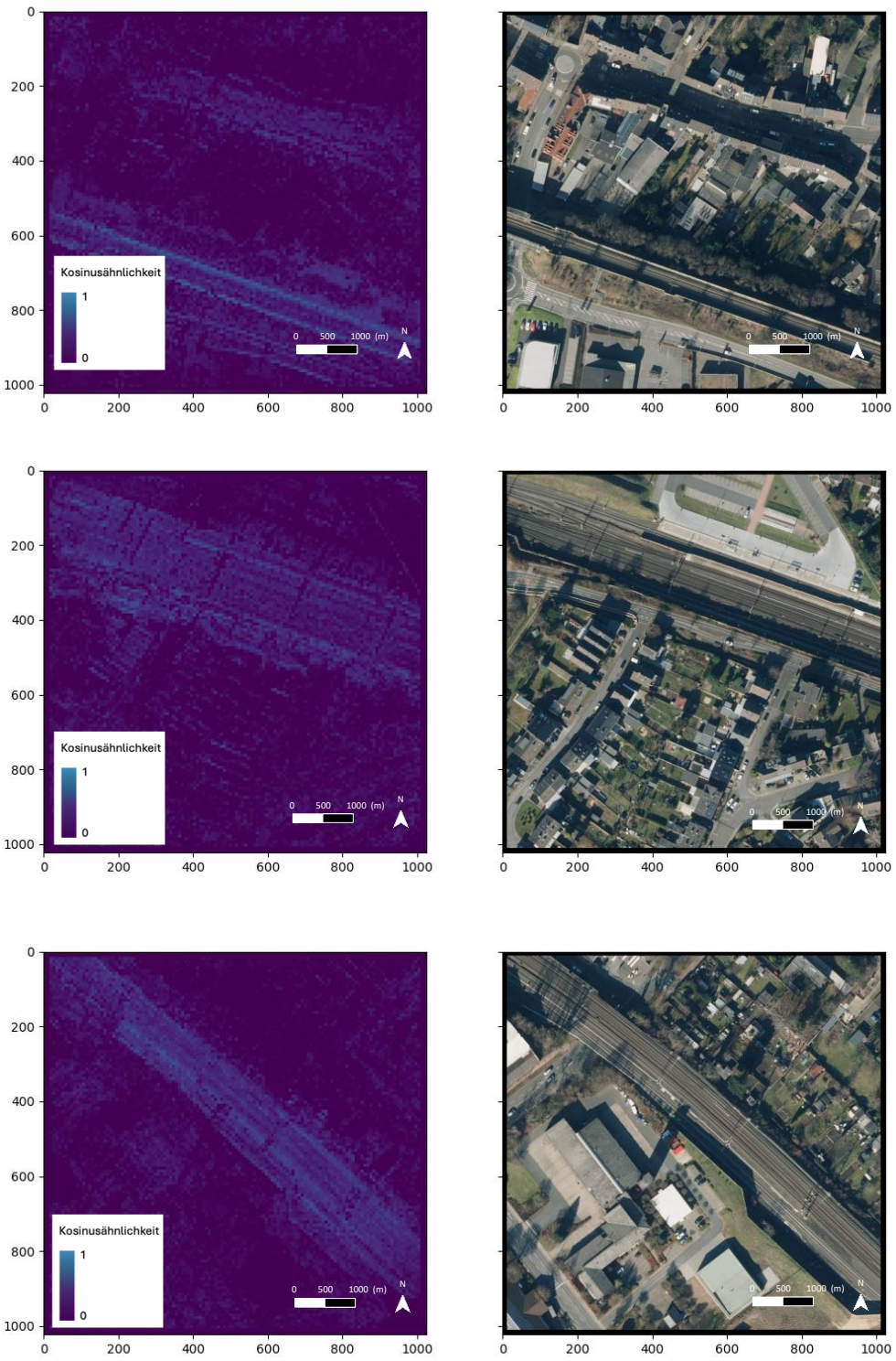


Abbildung 36: DINOv1 Klein (DOP: © GeoBasis-DE/BKG (2023)).

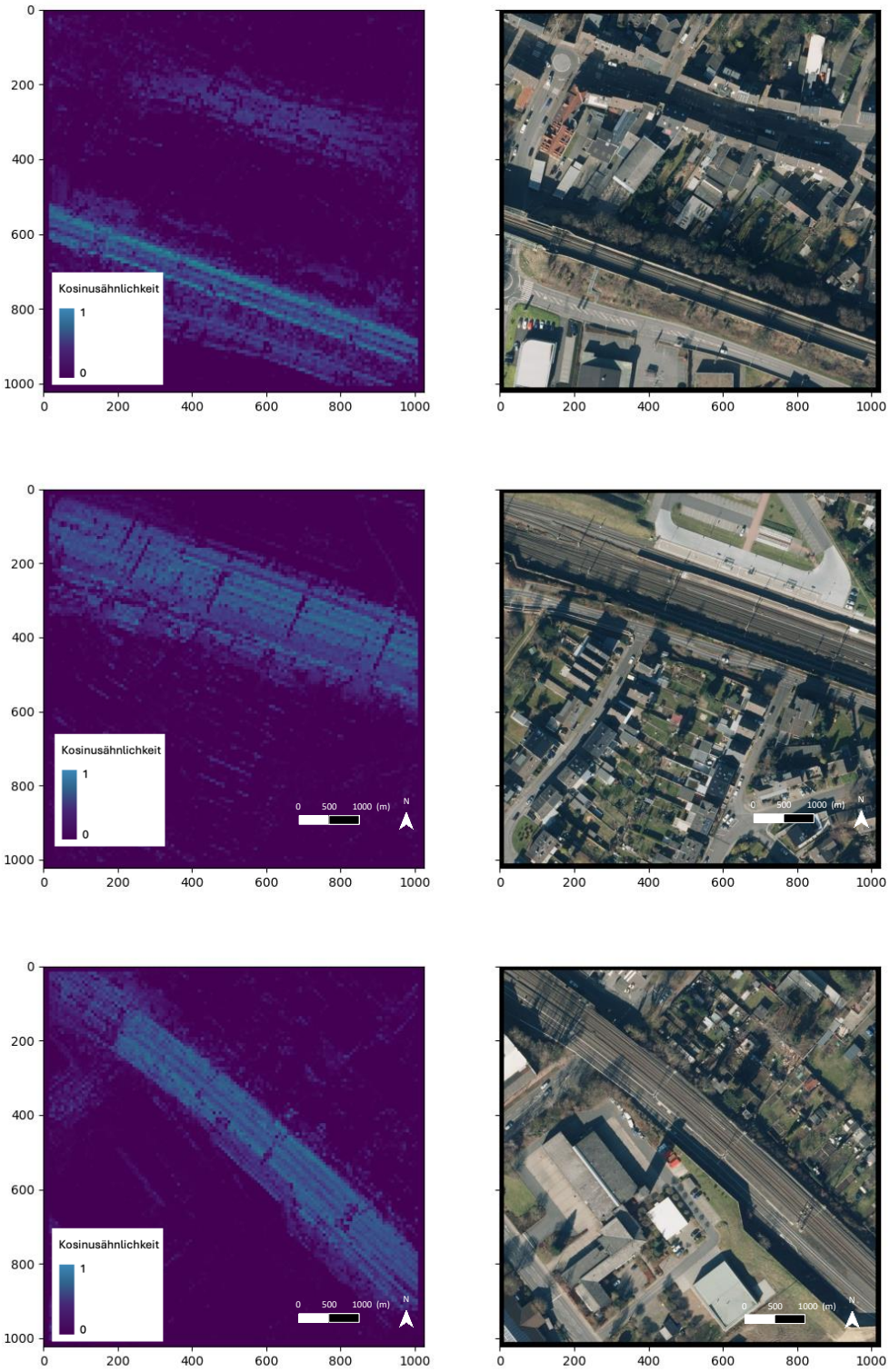


Abbildung 37: DINOv1 Basis (DOP: © GeoBasis-DE/BKG (2023)).

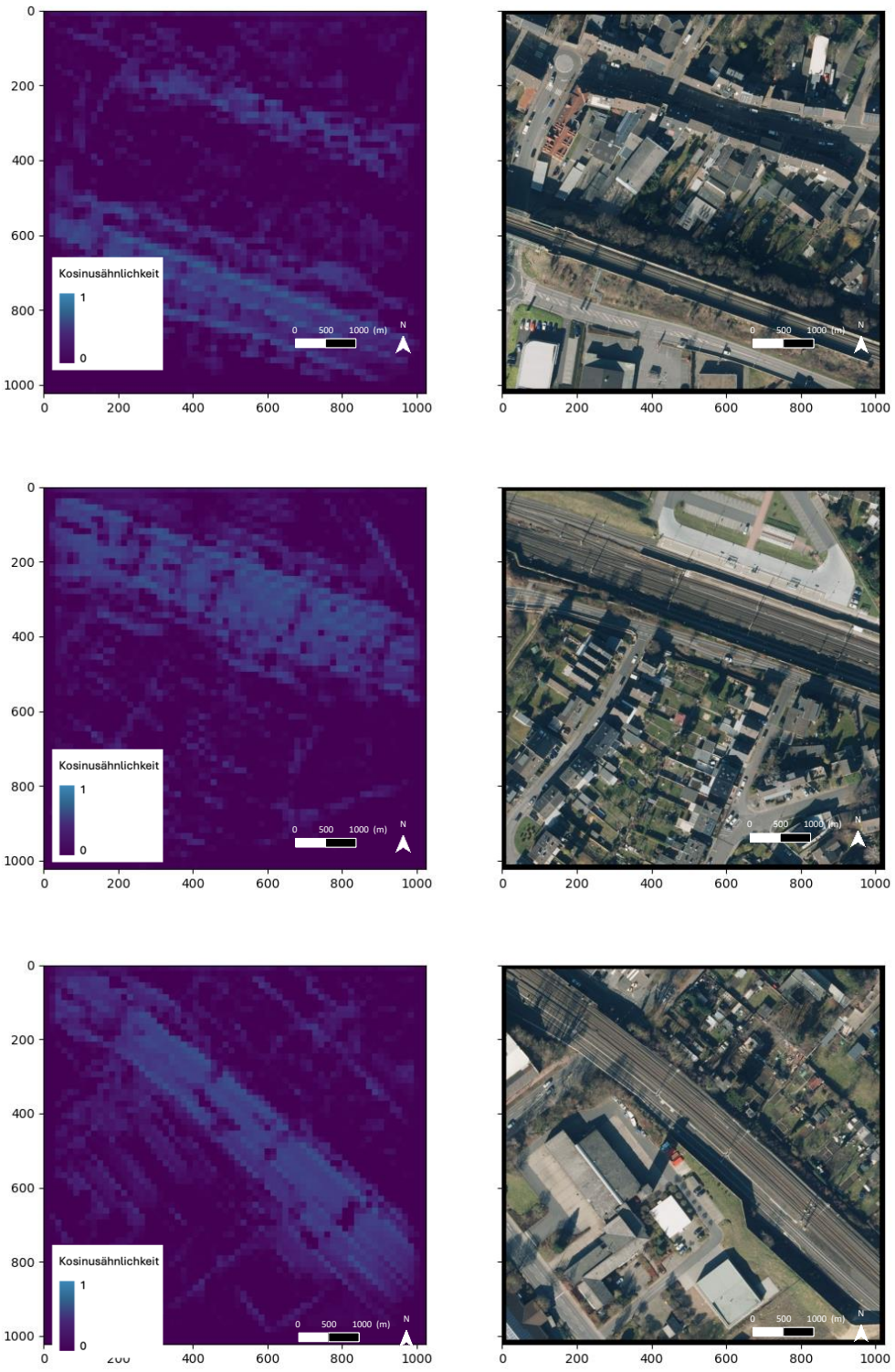


Abbildung 38: SAM Basis (DOP: © GeoBasis-DE/BKG (2023)).

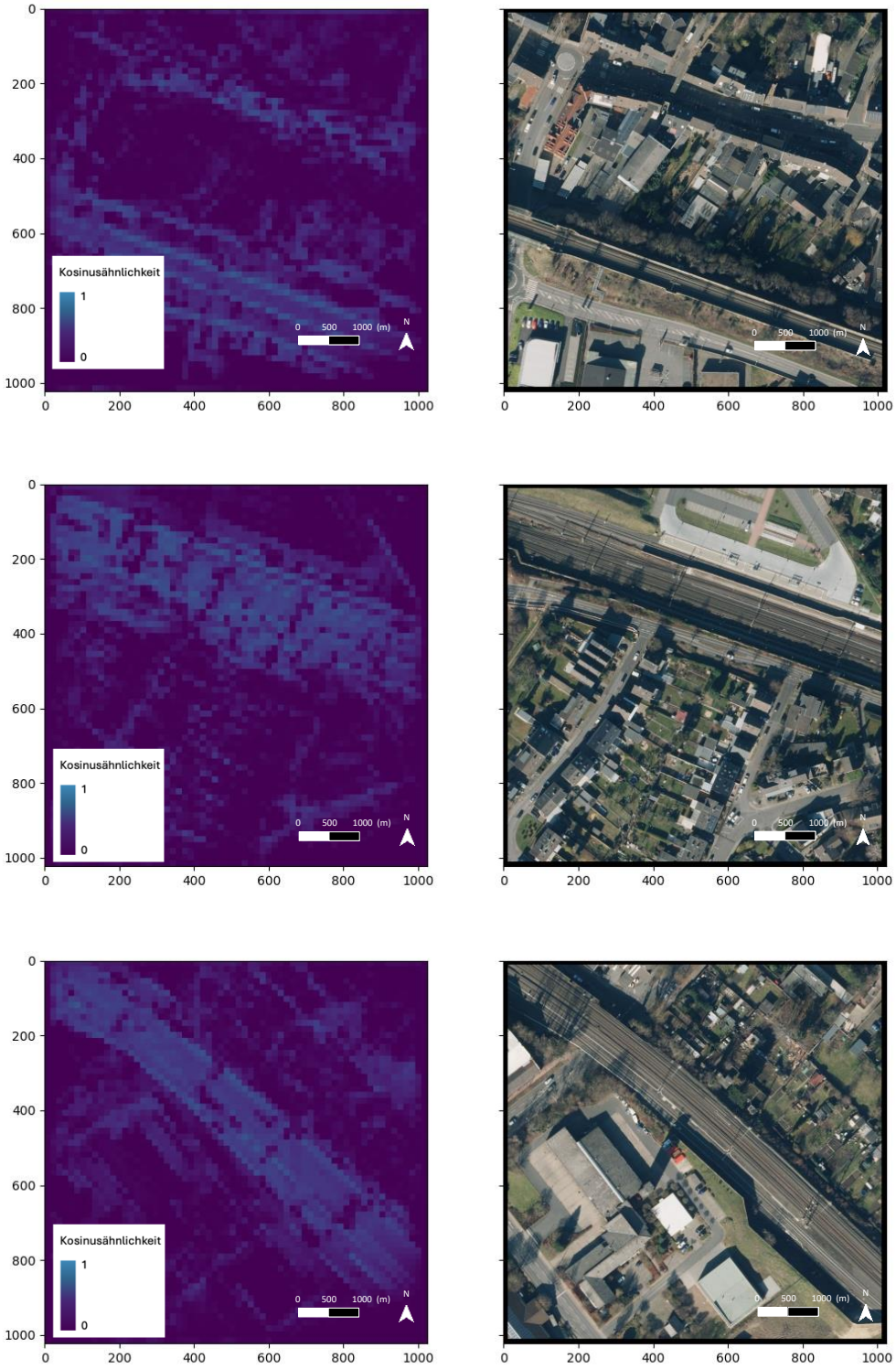
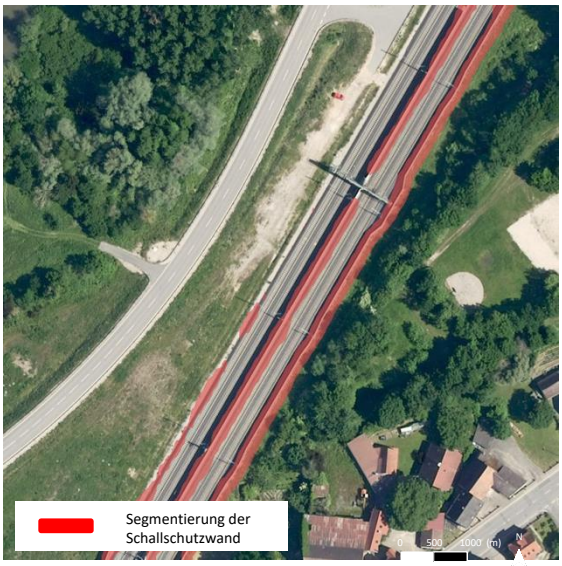
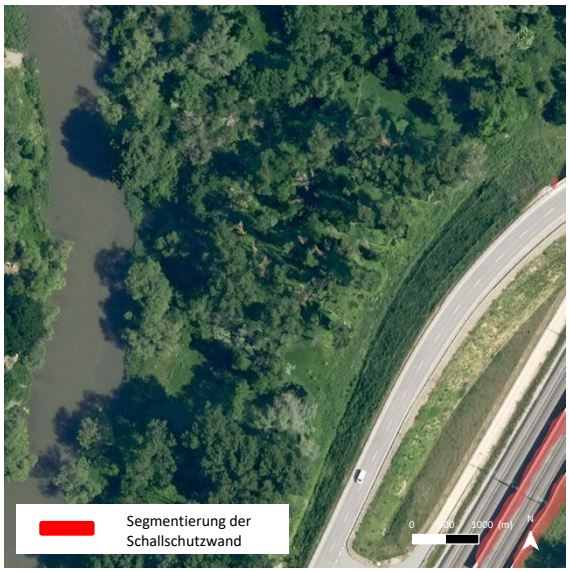


Abbildung 39: SAM Groß (DOP: © GeoBasis-DE/BKG (2023)).

Modell Inferenzen

Zur Veranschaulichung der Modell-Ergebnisse sind im Folgenden einige Beispiele der Segmentierungen auf Basis des Validierungs-Datensatzes beigefügt.





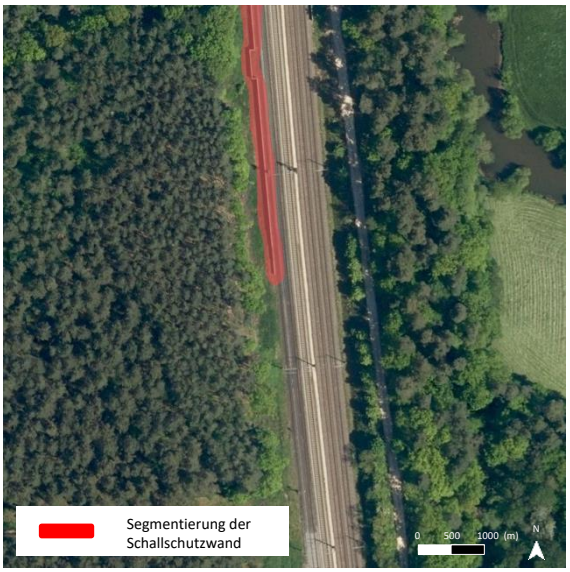


Abbildung 40: Modell Inferenzen (DOP: © GeoBasis-DE/BKG (2023)).