Master's Thesis

Advancing Semantic Segmentation for Building Detection in Very High-Resolution Data

submitted by

Lennart Giessing

Social and Economic Data Science

01/921963

at

Deutsches Zentrum für Luft- und Raumfahrt
German Aerospace Center

Universität Konstanz

Earth Observation Center
Geo-Risks and Civil Security
Team City and Society

Moritz Hertrich
Dr.Dorothee Stiller

German Aerospace Center
Oberpfaffenhofen, Germany

Department of Politics and Public Administration

1st Supervisor: Prof. Dr. Bastian Goldlücke
2nd Supervisor: Prof. Dr. Nils Weidmann

University of Konstanz
Konstanz, Germany

September 19, 2025

Abstract

Accurate building footprints are essential for urban planning, crisis management, and social science research, yet Germany lacks a comprehensive and up to date nationwide register. Existing sources such as cadastral data, and OpenStreetMap remain incomplete or inconsistent. At the same time current deep learning models for automatic footprint extraction still suffer from systematic errors.

This thesis investigates whether optimizing preprocessing, postprocessing, and training data selection can improve the CNN-based extraction model proposed by Stiller et al. [2023]. Experiments with normalization strategies, LiDAR-based height layers, tile overlap, and threshold settings show that targeted adjustments enhance performance. The final pipeline, using DSM data and refined data normalization, improved Overall Accuracy by 7.0 percentage points, IoU by 5.2 percentage points, and F1 score by 3.3 percentage points compared to the baseline.

A case study on Berlin illustrates the practical value of the generated data for the social sciences by linking building geometries with demographic and building use data.

The findings highlight both the technical and applied relevance of the improved workflow: advancing footprint extraction toward official usability while enabling new insights in the social sciences.

Contents

List of Abbreviations

ALS	Airborne Laser Scanning
CNN	Convolutional Neural Network
DFK	Digitale Flurkarte (Digital Cadastral Map)
DLR	Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center)
DSM	Digital Surface Model
DTM	Digital Terrain Model
DOP	Digital Orthophoto
GWR	Gebäude- und Wohnungsregister (Building and Housing Register)
GWZ	Gebäude- und Wohnungszählung (Building and Housing Census)
ICC	Intraclass Correlation Coefficient
iDSM	image based digital surface model IoU
Intersection over Union	
LiDAR	Light Detection and Ranging
LoD2	Level of Detail 2 (3D building model standard)
ML	Maximum Likelihood (estimation)
nDSM	Normalized Digital Surface Model
NNs	Neural networks
NRW	North Rhine–Westphalia (Nordrhein-Westfalen)
NIR	Near Infrared
OSM	OpenStreetMap
pp	percentage point
RGB	Red, Green, Blue (color layers)
RQ	Research Question
RH	Research Hypothesis
SD	Standard Deviation
TIN	Triangulated Irregular Network
WMS	Web Map Service

List of Tables

List of Figures

1 Introduction

According to Li et al. [2024b, p. 2], building footprints are the "two-dimensional (2D) visual representation of a building, describing its exact location, size, and shape on the ground". Figure 1 illustrates this concept, showing footprints extracted from aerial imagery.

Figure 1: Example of building footprints derived from aerial imagery (RGB from Web map Service (WMS) Layer Berlin, DLR internal).

In the social sciences, accurate building footprint data are essential for tasks such as estimating population distributions in regions where census information is outdated or incomplete [Hertrich et al., 2025]. As shown by Wardrop et al. [2018], building outlines derived from remote sensing imagery serve as key covariates in models that predict population density, enabling more informed research and policy making [Stiller et al., 2021].

Yet, the relevance of building footprints extends far beyond data-poor contexts. In

highly developed countries such as Germany, they represent a critical data layer for multiple domains. Accurate and recent building footprint information underpins urban planning [Schiller et al., 2021], infrastructure monitoring, and environmental reporting [Milojevic-Dupont et al., 2023]. It also supports crisis management, for instance in flood risk assessment or disaster response [Bhuyan et al., 2023] and contributes to socio-economic analysis by serving as proxies for population distribution and housing supply [Bakillah et al., 2014]. In short, building footprints are indispensable geospatial information for both scientific research and public administration [Krause et al., 2022].

Despite their importance, no recent and comprehensive nationwide register of building footprints exists in Germany. Instead, information is fragmented across multiple administrative systems, and no single source provides complete or uniform coverage [Krause et al., 2022].

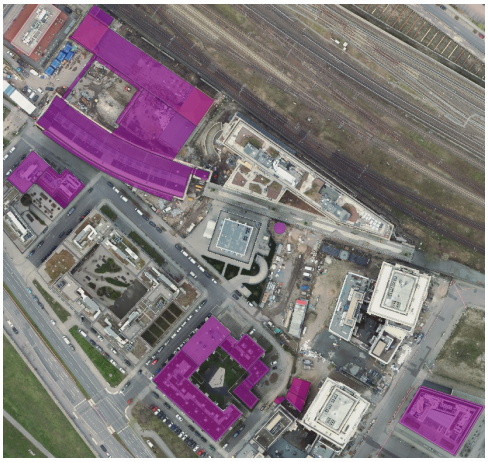
Cadastral offices, for example, record geolocations and selected structural attributes of buildings. Boundaries are measured through terrestrial surveys and documented as two dimensional ground plans in the official digital cadastral map (Digitale Flurkarte, DFK). This process is slow and labor intensive, often resulting in delays of several months between the completion of a building and its appearance in the records (e.g. Figure 2). Moreover, the system relies on building owners reporting construction or demolition activities, if these are not communicated, undocumented buildings remain absent from the DFK altogether [Li et al., 2020].

Building authorities (Bauaufsichtsbehörden) represent another potential source of footprint information, as they collect data during permit procedures or when projects are subject to notification. Depending on state regulations, this may include site plans, construction drawings, and other attributes [Maenning and Just, 2012]. However, their data is also incomplete. Many projects are now permit- or notification-exempt, so new builds or demolitions are not consistently recorded. In addition, discrepancies between approved plans and completed buildings are not always documented, further limiting their value for nationwide coverage [Li et al., 2020].

Finally, the census survey (Gebäude- und Wohnungszählung, GWZ) provides detailed information on building footprints and housing spaces, most recently in 2011 and 2022. Yet these data cannot be repurposed to build an administrative register, as they are protected by the principle of statistical secrecy (Statistikgeheimnis) [Krause et al., 2022].

To address the fragmented data landscape, the Federal Statistical Office (Statistisches Bundesamt) has proposed the creation of a nationwide *Gebäude- und Wohnungsregister* (GWR) [Krause et al., 2022]. This register shall be based on a survey of property owners, similar to the building and housing census, and shall be maintained by integrating updates from existing administrative sources such as cadastral data, construction permits, and tax records. To ensure consistency across these sources, unique identifiers for every building shall be introduced. However, this register is still far from being finished [Graaf

(a) Industrial area in Berlin



(b) Newly built area in Berlin

Figure 2: Example for buildings that are not included into the DFK. The purple areas represent the available cadastral data. Buildings that are visible but not covered by purple indicate structures that are missing from the official DFK dataset.

and Steuwer, 2025], and therefore not a feasible solution of the current lack of building footprint information.

Another available source for building footprints is Open Street Map (OSM) data, a global, collaborative mapping project that provides freely accessible geospatial data contributed by volunteers. The platform allows users to digitize and edit building footprints [Weber and Haklay, 2008]. But in the OSM data, many small buildings are missing, and those building footprints that exist, often show a strong offset to real world reference [Fan et al., 2014], making OSM data unsuitable for official tasks which require precise information of building footprints [Herfort et al., 2023].

In light of the shortcomings of administrative and volunteered data sources, global initiatives that leverage artificial intelligence and satellite imagery have emerged as promising alternatives [Touzani and Granderson, 2021]. One project is Microsoft's Building Footprints initiative, which aims to create detailed, open access maps of building outlines worldwide [Bing Maps, 2023]. Using deep learning algorithms trained on high resolution satellite imagery, the project provides an automated approach to map building footprints. For Germany, this means access to a regularly updated, unified layer of building footprint data.

However, limitations remain: building detection quality depends heavily on the resolution and recency of satellite imagery, and errors such as misclassifications, missing small structures, or footprint simplification can still occur especially in urban areas [Touzani and Granderson, 2021].

In response to these limitations of administrative and global datasets, the German Statistical Office in collaboration with the Federal Agency for Cartography and Geodesy (Bundesamt für Kartographie und Geodäsie) has initiated the research project *Sat4GWR*,

conducted by the German Aerospace Center (DLR) [Hennig et al., 2025]. The project applies CNN-based deep learning models to high resolution aerial imagery and LiDAR derived height profiles to automatically detect building footprints. Compared to global initiatives such as Microsoft's building footprints, Sat4GWR leverages the finer resolution of aerial data, allowing for more precise extraction of building footprints.

One contribution from the Sat4GWR project is the study by Stiller et al. [2023], which systematically examined the challenges of building footprint segmentation, by using a Deep Learning Model.

But their workflow and model also revealed limitations: frequent false positives in rural settings (e.g., water bodies, forests) and difficulties in complex urban areas such as bridges or railway infrastructure (Figure 3). These shortcomings highlight the need for further methodological refinements to ensure reliable nationwide application.

The main goal of this thesis is to reduce misclassifications while ensuring highly precise building detection. Rather than modifying the model architecture, which was already thoroughly optimized by [Stiller et al., 2023], this work will focus on improving the surrounding components of the data pipeline. Specifically, the investigation will target enhancements in pre- and post-processing procedures, as well as in the selection and composition of the training data. The main research question (RQ) is therefore:

- **RQ 1: Can optimization of pre-processing, post-processing and training data selection improve building footprint detection of the CNN model proposed by Stiller et al. [2023]?**

This main research question shall be targeted by several sub questions (presented in Table 1) which aim to identify which settings regarding pre- and post-processing, as well as training data selection, will lead to the best results.

Table 1: Summary of the main research question and its sub-questions.

RQ	**Research Question**
RQ1	Can optimization of pre-processing, post-processing and training data selection improve building footprint detection of the CNN model proposed by Stiller et al. [2023]?
RQ1.1	Does normalization of input data improve the accuracy of building footprint detection compared to unnormalized inputs?
RQ1.2	Which normalization strategy (Z-score or Min–Max) provides better results in heterogeneous landscapes?
RQ1.3	Do models that combine RGB and height information (DSM or nDSM) outperform models using only RGB data?
RQ1.4	Does the use of a DSM provide better performance than an nDSM by offering richer contextual information?
RQ1.5	Does introducing overlap between adjacent tiles improve building footprint detection compared to non-overlapping tiling?
RQ1.6	Can adjusting overlap differently in the training and inference phases lead to better performance?
RQ1.7	Does optimizing the decision threshold for class assignment in overlapping tile regions improve building footprint detection compared to default majority voting?
RQ1.8	Does increasing the diversity of training data by adding targeted regions that cover known failure cases improve overall detection accuracy?

Beyond these technical questions, this thesis also includes an applied case study (Section 8) to highlight the broader value of improved building footprint data for the social sciences. The case study focuses on Berlin and demonstrates how Convolutional Neural Network (CNN) derived footprints, when linked with demographic and land use information, can be used to investigate substantive urban questions. In particular, it examines the relationship between migration and the functional composition of the building stock. Migration has been one of the main drivers of demographic change in Berlin, raising questions about how migrant presence interacts with the city's housing supply and commercial infrastructure. The case study therefore addresses the following social science research question:

- **RQ2: Is a higher migration share associated with the balance between housing and commercial functions in Berlin's building stock?**

2 Literature

The first attempts to automate building footprint extraction from aerial or satellite imagery relied on rule based algorithms [Bouziani et al., 2010] and classical machine learning [Van Nguyen et al., 2015]. The first approach struggles with complex urban buildings and the second needs a wast amount of labeled data, when used in new environments [Dabove et al., 2024]. As a result, their generalizability to new environments was limited. Around the mid 2010s, deep CNNs began to transform aerial image analysis. In 2015–2016, pioneering studies [Saito et al., 2016, Alshehhi et al., 2017] showed that CNNs could significantly outperform earlier methods in detecting buildings by automatically learning features from large datasets, which made Deep Learning quickly the state of the art approach for this task [Luo et al., 2021]. By 2017, encoder-decoder CNNs with skip connections like U-Net [Maggiori et al., 2017] and SegNet [Bischke et al., 2019] emerged, yielding much higher accuracy and more coherent building shapes than earlier methods [Luo et al., 2021].

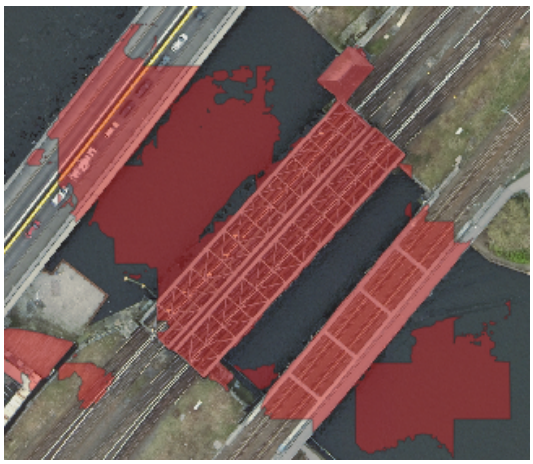
Between 2018 and 2020, research on building footprint segmentation focused on making models more accurate and robust by improving both architecture and training strategies [Luo et al., 2021]. A key direction was the use of multi scale models. Since building morphology range from tiny sheds to large complexes and are embedded in diverse surroundings, networks such as DeepLab introduced multi-scale feature fusion (e.g. Atrous spatial pyramid pooling), improving building detection at different scales [Ryuhei Hamaguchi et al., 2018]. At the same time, transfer learning with pretrained CNNs, based on large datasets (e.g. ImageNet) was used [Iglovikov and Shvets, 2018]. Additionally, researchers increasingly turned to data fusion: supplementing RGB imagery with additional sources such as infrared layerss or height information from LiDAR and stereo derived DSMs. These multi modal approaches helped separate buildings from visually similar objects and provided richer input features [Liu et al., 2019, Dabove et al., 2024].

Transformer based models for building footprint segmentation in aerial and satellite imagery started to appear around 2021. Their main advantage lies in the ability to capture long range contextual information, which led to more accurate footprint detection [Chen et al., 2021]. By 2024 transformer based segmentation of building footprints matured reported not only higher accuracy, but also a better generalization across different urban areas, compared to purely CNN based models [Gibril et al., 2024].

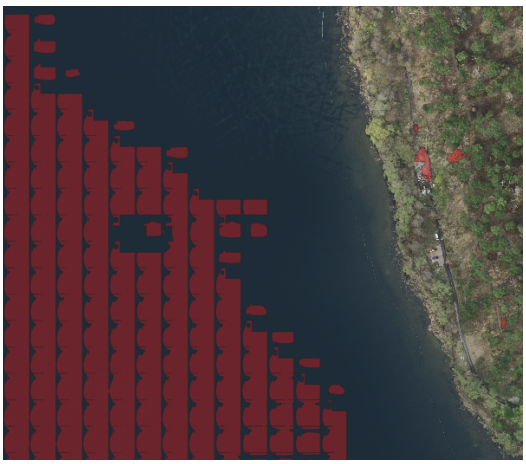
Nevertheless, CNNs remain highly competitive for building extraction from aerial imagery. As Angelis et al. [2023] point out, transformers are computationally more expensive and memory intensive than CNNs, which directly impacts inference speed. This is a critical consideration for applications involving high-resolution imagery at large scale. For projects aiming to map extensive, heterogeneous regions such as entire countries, CNN architectures thus continue to offer a favorable balance between accuracy and efficiency.

Additionally CNNs are more data efficient and better suited to limited label regimes [Rosy et al., 2025]. One representative example of a CNN-based study that was trying to approach the problem of building detection in aerial imagery was that of Stiller et al. [2023]. The authors systematically addressed the challenges of building segmentation from high resolution remote sensing images, focusing on urban environments with complex spatial and morphological features. The authors generated a comprehensive building dataset covering approximately 34.5 km^2 across various cities in North Rhine-Westphalia, integrating both orthophotos (including RGB and near infrared layers) and normalized digital elevation models (nDSM). This diverse dataset allowed the authors to design nearly 500 experimental configurations, where the authors analyzed systematically the influence of multiple factors such as model architecture, spatial resolution, input image size, training region diversity, and label accuracy. The authors systematically evaluated model accuracy. The findings revealed that the FPN architecture with the resnet50 encoder backbone achieved the best overall results, when combining orthophotos with nDSM data, substantially improving building detection and separation from background structures. Although the CNN models developed by Stiller et al. [2023] demonstrated strong performance on the North Rhine-Westphalia dataset, they show limitations in other settings. This is evident in rural areas characterized by natural features such as water bodies (Figure 3b), forests, and narrow roads. Moreover, the model also struggles in complex urban environments, for instance, around bridges or railway infrastructure (Figure 3a) where structural ambiguity increases the risk of misclassifications.

The primary problem observed is a high rate of false positives (FP): regions that do not contain buildings are frequently misidentified as such. These errors are especially prevalent in settings where non building objects share visual or textural similarities with built structures. Consequently, the model's reliability diminishes outside the specific urban contexts in which it was originally trained.



(a) Example for FP in an urban setting



(b) Example for FP in a rural setting

Figure 3: The red areas were detected as buildings by the model.

While the influence of different model architectures has already been thoroughly investigated by Stiller et al. [2023], it is crucial to recognize that network design is not the only factor influencing the performance of deep learning workflows. For CNN-based models, which input data is selected, how input data are prepared and how raw outputs are refined can be equally decisive. In image based deep learning tasks, preprocessing constitutes the initial stage in which raw data are transformed into a form more suitable for CNN training. The underlying mechanisms vary, but the overall goals are consistent: to improve data quality, ensure consistency across inputs, stabilize and accelerate the training process, enhance generalization to unseen data, and ensure compatibility with CNN architectures [Li et al., 2019]. Raw images frequently contain noise or low quality visual information caused by factors such as radiation effects, shadows, or varying illumination. Preprocessing mitigates these issues through techniques such as e.g. color, contrast, and brightness adjustments, augmentation (e.g., flipping, resizing), noise filtering, and resolution modifications. These operations enhance feature clarity, making patterns more recognizable for CNNs and often improving model performance [Koresh, 2024]. Postprocessing steps, on the other hand, refine the model's raw predictions. They address common CNN limitations such as e.g. blurred boundaries or false positives and produce vector ready building footprints suitable for GIS applications. Examples include morphological filtering, boundary regularization, and polygon simplification [Wei et al., 2020]. Comparative studies demonstrate that models trained and applied without robust pre- and postprocessing pipelines are prone to higher error rates and reduced transferability to new environments [Sakeena et al., 2023, Liu et al., 2018]. Equally important is the selection of training data. CNN based segmentation models depend on the representativeness of their training sets to learn discriminative patterns, and models trained on thematically limited data often struggle to generalize to new environments [Maggiori et al., 2017, Audebert et al., 2018]. Recent studies confirm that these supporting procedures strongly affect performance but are often overlooked in comparison to architecture [Chawda et al., 2018].

This thesis addresses this gap by systematically evaluating key pre-processing, postprocessing strategies, and training data selection for CNN-based building footprint extraction. Rather than proposing a new architecture, the focus lies on improving the efficiency and reliability of existing models by conducting a series of experiments designed to enhance the performance of the overall deep learning workflow. By analyzing aspects of these three components systematically, this thesis provides a structured framework for understanding how supporting procedures, in addition to architecture, govern the performance and robustness of building footprint extraction models [Clark et al., 2023].

H1. Main Research Hypothesis. Improving preprocessing routines, postprocessing

strategies, and training data selection enhances the accuracy of CNN based building footprint extraction.

The specific hypotheses presented in the following sections are formulated as sub-hypotheses that address this overarching research hypothesis. Each of them isolates and tests a distinct factor that may influence the performance of the model. Together, these sub-hypotheses provide a systematic framework to evaluate how supporting steps around the core CNN architecture contribute to the task of detecting building footprints in high resolution aerial imagery.

2.1 Normalization

Normalization is one of the most fundamental preprocessing steps in deep learning pipelines for semantic segmentation. It rescales or standardizes input values so that features are expressed on a comparable scale, preventing single layers or pixel intensities from disproportionately influencing model training. In computer vision, normalization has long been shown to stabilize optimization, accelerate convergence, and improve generalization across diverse test sets [Lecun et al., 1998]. In remote sensing applications, common approaches include Min-Max normalization, which rescales input data to a fixed interval (e.g., [0,1]), and Z-score normalization, which centers values around zero with unit variance [Jeon et al., 2021].

For building footprint detection, normalization has direct implications for segmentation quality. Buildings can occur in highly heterogeneous settings: dense urban cores with sharp contrasts, but also homogeneous rural landscapes where pixel values vary only slightly. Tile wise Min-Max normalization, for example, can distort these subtle differences by stretching narrow value ranges, effectively erasing discriminative information in uniform areas such as water bodies or large roofs. Z-score normalization, by contrast, tends to preserve variance across tiles, leading to more stable predictions [Jeon et al., 2021].

Theoretical Argument. If normalization ensures consistent scaling of training and testing data, models are better able to generalize across heterogeneous imagery, reducing false positives in spectrally similar non building areas. Conversely, inconsistent or poorly chosen normalization settings distort the input, degrade feature learning, and inflate commission errors. Thus, careful evaluation of normalization strategies is theoretically essential to advancing the quality of building footprint extraction.

Research Hypotheses.

H1.1. Normalization of input data improves the accuracy of building footprint detection compared to unnormalized inputs.

H1.2. Z-score normalization outperforms Min-Max normalization in heterogeneous landscapes, due to its robustness to varying pixel distributions and outliers.

2.2 Height Input Layer

Beyond spectral information, height cues play a central role in improving the separability of buildings from spectrally similar surfaces. Optical imagery alone often fails to distinguish light colored roofs from roads, industrial hardscapes, or even reflective water surfaces. To address this limitation, many studies have combined RGB imagery with height information derived from Light Detection and Ranging (LiDAR) data. Two versions of this derived data are Digital Surface Models (DSM) (Section 4.2.1) [Dabove et al., 2024] and normalized DSMs (nDSM) Section 4.2.3) [Stiller et al., 2023], which represent the elevation of objects above ground level. The inclusion of vertical structure provides a powerful discriminative feature, as buildings are typically elevated above their surroundings, while most non building classes lie close to ground level [Liu et al., 2019].

In remote sensing research, height information has repeatedly been shown to improve building extraction accuracy. Liu et al. [2019] found that CNNs trained with combined RGB and DSM inputs achieved significantly higher IoU and F1 scores than those trained with RGB data alone. However, despite the benefits, the literature still debates the most effective form of height representation. While DSMs encode absolute elevation, including terrain undulations, nDSMs isolate relative heights and thus emphasize vertical structures. Each representation carries advantages and drawbacks: DSMs can provide richer topographic context, but may introduce confusion in hilly landscapes; nDSMs simplify the height profile but may suppress useful contextual variation [Audebert et al., 2018].

Theoretical Argument. Theoretically, adding height information should enhance the model's ability to discriminate between buildings and spectrally similar non building surfaces. nDSMs are expected to improve completeness by ensuring that even spectrally ambiguous buildings are detected, while DSMs may reduce false positives by incorporating full terrain context. Conversely, removing height layers (removing nDSM or DSM data) should lead to worse results through all metrics, as the model is forced to rely solely on spectral cues. Thus, evaluating the contribution of different height representations is essential to determine which configuration maximizes the model results.

Research Hypotheses.

H1.3. Models using combined RGB and height information (DSM or nDSM) outperform models using only RGB data.

H1.4. The use of DSM data as the fourth layer improves the model results compared to nDSM, as absolute elevation provides richer contextual information.

2.3 Tile Overlap

In deep learning workflows for semantic segmentation of high-resolution aerial imagery, large images are typically partitioned into smaller tiles to fit GPU memory constraints. While this tiling step is unavoidable, it introduces boundary effects: objects cut at tile edges are often segmented incompletely, leading to discontinuities or missing detections. To mitigate these artifacts, many studies employ overlapping tiles, ensuring that boundary pixels are observed in multiple contexts during training and prediction. Overlap thus provides contextual redundancy that improves segmentation accuracy, particularly at object boundaries [Reina et al., 2020, Volpi and Tuia, 2016].

Increasing the overlap ratio has been shown to enhance boundary delineation and reduce edge artifacts [Sherrah, 2016]. However, higher overlaps also inflate the number of training and testing patches, which significantly increases computational cost. Consequently, overlap design involves balancing segmentation accuracy against runtime efficiency [Reina et al., 2020]. The work by Audebert et al. [2018] further suggests that the overlap ratio does not necessarily need to be identical during training and testing: a carefully chosen combination may provide accuracy gains while reducing computational load.

Theoretical Argument. From a theoretical perspective, overlap improves the model's ability to capture contextual cues across object boundaries, thereby reducing false negatives (missed buildings) and false positives (fragmented or spurious detections) near tile edges. Moreover, different overlap strategies between training and inference may shift the balance between redundancy and efficiency.

Research Hypotheses.

H1.5. Introducing overlap between adjacent tiles improves building footprint detection compared to non overlapping tiling.

H1.6. Changing the overlap between the training and inference phase can have a positive impact on the model results.

2.4 Decision Thresholds

In the deep learning workflow of Stiller et al. [2021], the model outputs hard class labels for each pixel. It produces predictions indicating whether a pixel belongs to the background or to the building class. Due to tile overlap, some pixels receive multiple predictions from different tiles. To transform these probabilities into a binary mask, a *decision threshold* must be applied. This threshold defines the cutoff above which a pixel is classified as building and below which it is classified as background [Reina et al., 2020]. Although several studies adopt the default value of 0.5 [Roth et al., 2018, Reina et al., 2020, Cira et al., 2024], and therefore a majority vote, research shows that varying the threshold parameter directly impacts model performance [Wu et al., 2019]. Lower thresholds increase sensitivity by labeling more pixels as positive, which improves recall but also increases false positives (commission errors). Higher thresholds, in contrast, reduce false positives but risk omitting true positives (omission errors) [Sherrah, 2016]. The work by Bohao Huang et al. [2018] has highlighted that threshold optimization can lead to measurable improvements in segmentation tasks, particularly in heterogeneous landscapes where class boundaries are ambiguous. Moreover, threshold tuning is computationally inexpensive compared to architectural modifications, making it attractive for large scale applications such as nationwide building mapping. The novelty of the approach in this study is that instead of relying solely on majority voting, all thresholds between 0 and 1 are systematically evaluated, and the setting yielding the best results is selected.

Theoretical Argument. From a theoretical perspective, the decision threshold directly governs the trade-off between omission and commission errors in building footprint detection. A lower threshold prioritizes completeness, increasing the likelihood that even tiles with limited discriminative features (e.g., without visible roof edges) are still classified as buildings. Conversely, a higher threshold prioritizes correctness, ensuring that predicted building labels are more reliable, but at the expense of overlooking marginal or ambiguous cases. Identifying an optimal threshold is therefore crucial to balance these competing objectives.

Research Hypotheses.

H1.7. Optimizing the threshold for class assignment in overlapping tile regions improves building footprint detection performance compared to the default majority voting approach.

2.5 Additional Training Data

The quality and diversity of training data are among the most decisive factors in deep learning for remote sensing [Maggiori et al., 2017]. Unlike classical methods, where per-

formance depends heavily on handcrafted features, CNN-based segmentation models rely on the representativeness of the training set to learn discriminative patterns [Wu et al., 2019]. Studies have shown that models trained on thematically limited data often struggle to generalize to new environments. Errors frequently arise in settings that are underrepresented in the training data [Audebert et al., 2018, Liu et al., 2019].

One approach to mitigate these weaknesses is to enrich the training data with additional samples specifically targeting those areas or objects that have high error rates. Targeted sampling strategies have been applied in semantic segmentation to improve performance on rare or difficult classes. This method has been called "active learning" by [Tuia et al., 2016]. In building extraction, expanding training datasets with complementary regions has been shown to improve robustness across urban morphologies [Li et al., 2021]. However, the effectiveness of this strategy is not guaranteed: adding data that differs too strongly from the target distribution can lead to marginal gains or even degraded performance due to domain shift [Persello and Bruzzone, 2014].

Theoretical Argument. From a theoretical standpoint, training data diversity is expected to improve generalization by exposing the model to a wider range of spectral, structural, and contextual patterns. Specifically, areas or objects that the model does not perform well shall be added to the training data. Nonetheless, if the new samples differ substantially from the target domain, they may introduce noise or bias rather than improve performance. Careful evaluation is therefore necessary to determine whether targeted additions meaningfully contribute to accuracy or whether they instead dilute the representativeness of the training set.

Research Hypotheses.

H1.8. Increasing the diversity of training data by adding targeted regions addressing known failure areas/objects improves overall building footprint detection accuracy.

Summary of Research Hypotheses

Table 2: Summary of the main research hypothesis and its sub-hypotheses

Hypothesis	**Statement**
H1	Improving preprocessing routines, postprocessing strategies, and training data selection enhances the accuracy of CNN based building footprint extraction.
H1.1	Normalization of input data improves the accuracy of building footprint detection compared to unnormalized inputs.
H1.2	Z-score normalization outperforms Min-Max normalization in heterogeneous landscapes, due to its robustness to varying pixel distributions and outliers.
H1.3	Models using combined RGB and height information (DSM or nDSM) outperform models using only RGB data.
H1.4	The use of DSM data as the fourth layer improves the model results compared to nDSM, as absolute elevation provides richer contextual information.
H1.5	Introducing overlap between adjacent tiles improves building footprint detection compared to non overlapping tiling.
H1.6	Changing the overlap between the training and inference phase can have a positive impact on the model results.
H1.7	Optimizing the threshold for class assignment in overlapping tile regions improves building footprint detection performance compared to the default majority voting approach.
H1.8	Increasing the diversity of training data by adding targeted regions addressing known failure areas/objects improves overall building footprint detection accuracy.

3 Model

With the problem context and hypotheses established, follows the introduction of the neural network (NNs) foundation of this approach. To provide a clear methodological foundation for the subsequent tests in the methodology part, the following section introduces the NNs model underlying this approach, beginning with the basics of neural networks, Deep Learning and CNNs.

3.1 Neural Networks

NNs are a class of machine learning models designed to approximate complex, often non-linear relationships in large datasets [Wu and Feng, 2018]. In computer vision tasks like semantic segmentation, NNs can learn to map pixel values directly to semantic categories, thereby enabling automated extraction of structures such as buildings from aerial imagery [Audebert et al., 2018].

At the core of NNs lies the perceptron [Rosenblatt, 1958], a simple computational unit. Figure 4 illustrates its basic structure. Each perceptron receives a set of inputs $a_1, a_2, \ldots, a_N$, which may represent features such as pixel intensities in an image. Every input is multiplied by a corresponding weight $w_1, w_2, \ldots, w_N$, reflecting the importance of that feature for the current task. A bias term b is added, which shifts the decision boundary and allows more flexibility in learning. The weighted inputs are summed, and the result is passed through a non-linear activation function σ, producing the output a_{out} [Singh and Banerjee, 2019]. Depending on the task, this output can take different forms: in a binary classification setting it might be a single probability between 0 and 1 (e.g. the likelihood that a pixel belongs to a building), in multi-class classification it can be a probability distribution across several classes, and in regression tasks it represents a continuous value [Lecun et al., 2015].

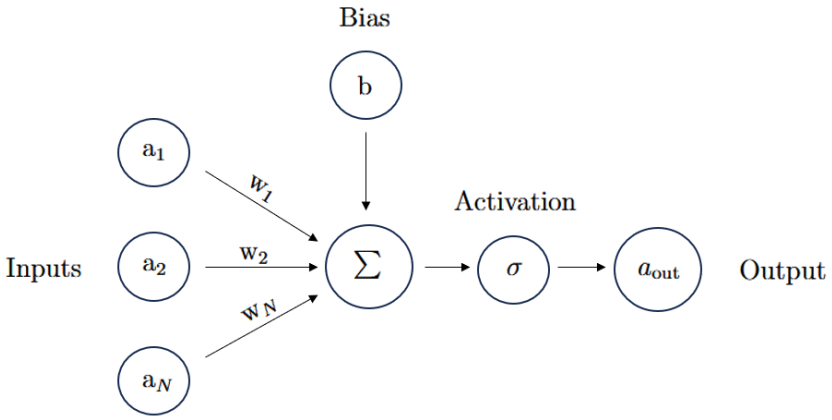


Figure 4: Perceptron visualization, adapted from Singh and Banerjee [2019].

During training, the weights and bias are initially randomized. Predictions are compared to the true labels using a *loss function* (L), which quantifies how far the model's predictions deviate from the correct outputs. In classification tasks, the most common choice is the *cross-entropy loss* [Goodfellow et al., 2016], defined as

$$L = -\sum_{c=1}^{M} y_c \log(\hat{y}_c),$$

where M is the number of classes, y_c is the true label (1 if the sample belongs to

class c, 0 otherwise), and $\hat{y}_c$ is the predicted probability for class c. This loss penalizes confident but incorrect predictions more strongly, encouraging the model to assign high probability to the correct class.

To minimize the loss, the network parameters $\theta = \{w_i, b\}$ are updated using *gradient descent*. In this procedure, the gradient $\nabla_\theta L(\theta)$ is computed [David E. Rumelhart et al., 1986], indicating how much the loss changes with respect to each parameter. The parameters are then adjusted in the opposite direction of the gradient:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta L(\theta),$$

where η is the learning rate, controlling the step size. Here, t denotes the iteration step in training: $\theta^{(t)}$ are the parameters before the update, and $\theta^{(t+1)}$ the parameters after applying one gradient descent step. This iterative process aims to converge toward a (local) minimum of the loss function. Training proceeds over multiple *epochs*, where one epoch corresponds to a complete pass through the training dataset [Goodfellow et al., 2016].

3.2 Deep Learning

A single perceptron can model only simple linear relationships. By combining many perceptrons into layers, and stacking these layers together, one obtains a *Multi-Layer Perceptron* (MLP). In an MLP, outputs from one layer of perceptrons serve as inputs to the next, enabling the network to learn increasingly complex, non-linear representations [Prince, 2023]. The intermediate layers between input and output are called *hidden layers*, because they are not directly observed in the data or labels but instead learn internal feature representations that facilitate the final prediction [Bishop, 1995]. This step from single units to deep architectures marks the essence of deep learning [Goodfellow et al., 2016].

Figure 5 illustrates a generic deep learning model with multiple hidden layers, highlighting how information flows from input features through successive transformations to the output layer.

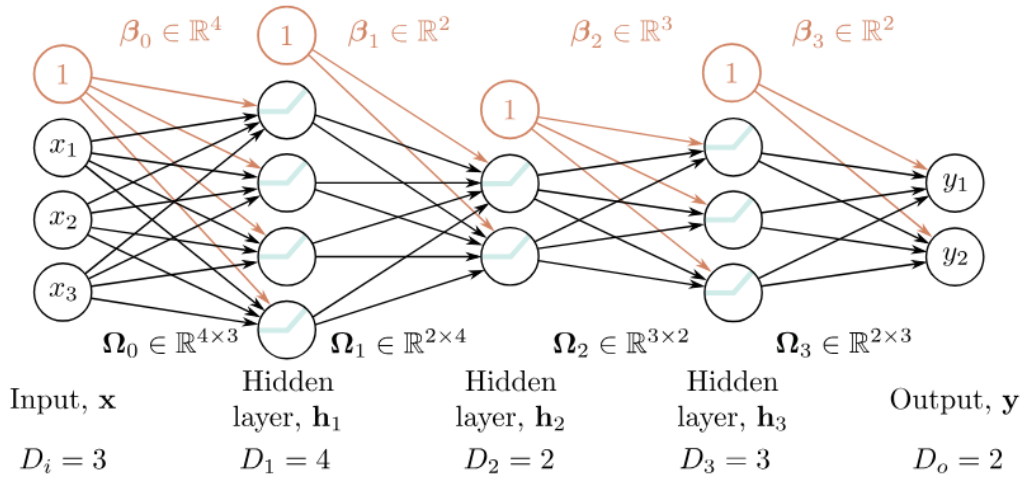


Figure 5: Example of a fully connected neural network with $D_i = 3$ input features $\mathbf{x}$, $D_o = 2$ output units $\mathbf{y}$, and $K = 3$ hidden layers h_1, h_2, h_3 of sizes $D_1 = 4$, $D_2 = 2$, and $D_3 = 3$ respectively. The weights are represented by matrices Ω_k that transform the activations from one layer into pre-activations for the next layer. For example, $\Omega_1 \in \boldsymbol{R}^{2\times4}$ maps the four activations in h_1 to the two units in h_2. Bias terms are stored in vectors β_k and have dimensions matching the layer they feed into; for instance, $\beta_2 \in \boldsymbol{R}^3$ corresponds to the three units in h_3. Information flows strictly from the input layer to the output layer, making this a simple example of a deep learning network structure. Image sourced from Prince [2023].

Deep learning excels at autonomously learning meaningful representations from raw input data, a capability that proves especially valuable in high-dimensional domains with abundant training samples [Lecun et al., 2015].

Modern deep learning extends beyond MLPs through specialized architectures designed for different data types.

3.3 Convolutional Neural Networks

CNNs are a class of neural networks designed to process grid structured data such as images. Unlike fully connected NNs, which connect every neuron in one layer to every neuron in the next (Figure 5), CNNs exploit local connectivity: each convolutional layer applies a set of learnable filters (kernels) to small regions of the input. This enables the network to detect simple patterns such as edges, corners, and textures in early layers, and more complex structures such as shapes or object parts in deeper layers [Lecun et al., 1998].

A key advantage of CNNs is parameter sharing. The same filter is applied across the entire input image, meaning that the number of parameters is independent of image size and that learned features recognized in one location can also be recognized elsewhere. Pooling layers are often added between convolutional layers to downsample the spatial resolution, thereby increasing the receptive field and reducing computational cost, while

retaining the most important features [Goodfellow et al., 2016].

For semantic segmentation, where each pixel is assigned a class label, CNNs act as the primary feature extractors. Combined with decoder mechanisms in encoder–decoder networks, they recover fine spatial detail after downsampling. This makes CNNs well suited to high-resolution remote sensing tasks such as building footprint detection [Prince, 2023].

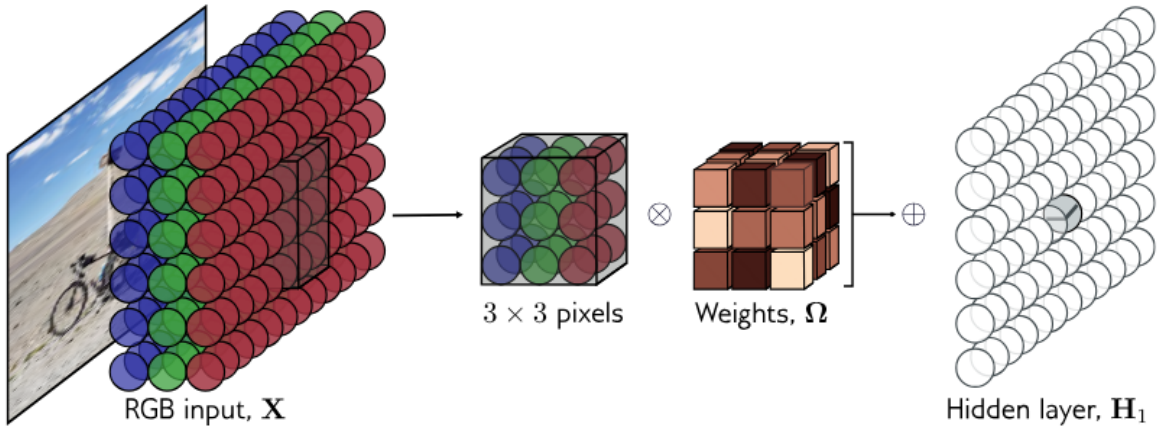


Figure 6: Example of a 2D convolution on an RGB image. A 3×3 kernel slides across the image to generate feature maps, enabling the network to detect local patterns efficiently. Image source: Prince [2023].

3.4 The Encoder–Decoder Principle

Many modern semantic segmentation architectures, including the model used in this work, follow an encoder-decoder principle (Figure 7). The encoder transforms the input image into a series of abstract feature representations by applying convolutional operations and downsampling. This process captures high level semantic information while reducing spatial resolution. The decoder then takes these abstract features and reconstructs a dense, pixel level prediction map, using upsampling operations.

Encoder–decoder networks benefit from combining features across resolutions: low-level features preserve spatial detail, while high-level features capture semantic context. Fusing these multi-scale representations, as in Feature Pyramid Networks (FPNs), yields segmentation maps that are both precise and semantically rich, enabling robust detection of objects of varying sizes and shapes [Lin et al., 2017].

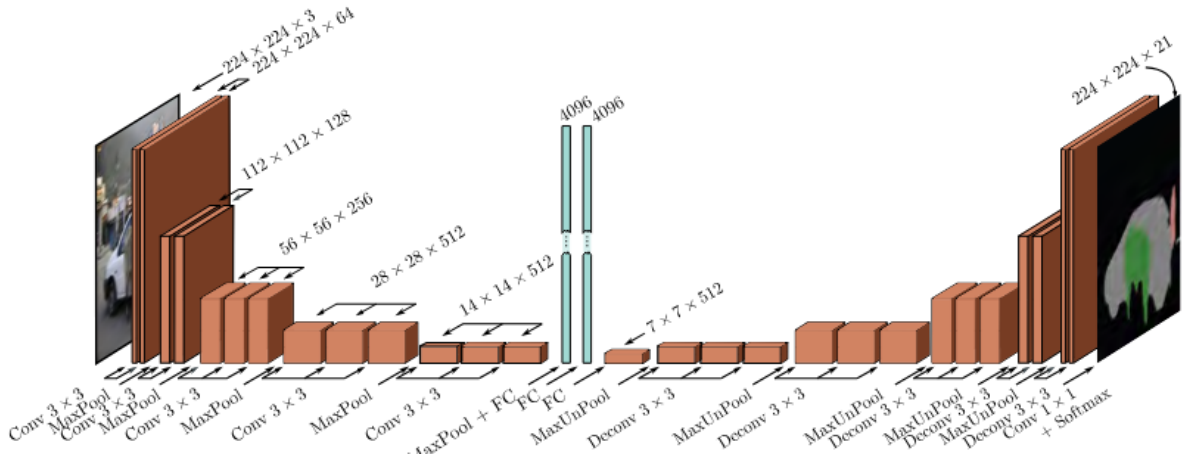


Figure 7: Example of an encoder–decoder network for semantic segmentation. The encoder (left part) reduces the input image to a compact feature representation using convolutions and pooling, while the decoder (right part) upsamples this representation through transposed convolutions to produce a pixel-wise classification map with class probabilities for each pixel. Figure and explanation Prince [2023]

3.5 Model Architecture and Usage

The model used in this work builds upon the framework established by [Stiller et al., 2023] (workflow in Figure 9), which identified an architecture for building footprint detection. It employs a ResNet-50 backbone as the encoder and a Feature Pyramid Network (FPN) as the decoder.

The ResNet-50 encoder is a deep convolutional network with residual connections that allow many layers to be trained without performance degradation [He et al., 2016]. These skip connections enable the network to learn residual mappings instead of direct transformations, resulting in more efficient optimization. As the input passes through the encoder, it is transformed into a hierarchy of feature maps with decreasing spatial resolution but increasing semantic abstraction. The FPN decoder reconstructs fine-grained prediction maps from this hierarchy using a top–down pathway with lateral connections (Figure 8) Lin et al. [2017].

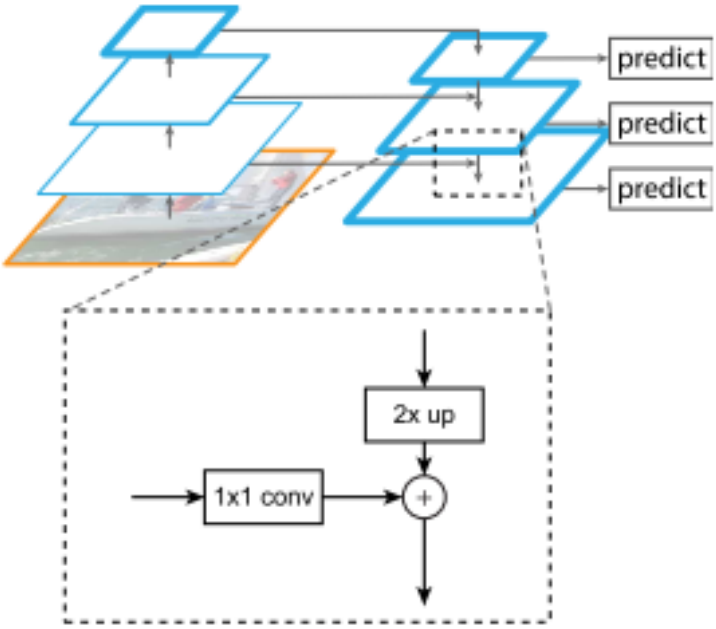


Figure 8: Schematic illustration of the Feature Pyramid Network (FPN) structure, showing the top–down pathway with upsampling, 1×1 convolutions for lateral connections, and element-wise addition for multi-scale feature fusion. In the encoder–decoder paradigm (Section 3.4), this represents the decoder stage, with the ResNet-50 backbone providing the encoder feature maps. Image and explanation are from Lin et al. [2017].

Coarse, high-level features are upsampled and merged with finer, lower-level features using 1×1 convolutions and element-wise addition. A 3×3 convolution is applied at each stage to make the features more consistent.

At the output, a two-class softmax activation produces per-pixel probability maps for binary segmentation

3.5.1 Training Phase

In the training phase (Figure 9), the model learns to separate building from non-building pixels by comparing predictions against ground truth labels. Input data consist of normalized RGB orthophotos combined with a height layer and the additional ground truth layer, cut into 320×320 pixel tiles with adjustable overlap. The available dataset is divided randomly into a training set (80%), used to update parameters, and a validation set (20%), used to monitor generalization performance. These tiles and their corresponding building masks are passed through the network, which produces probability maps for each pixel. The predictions are compared to the labels using a soft cross-entropy loss function, and the model parameters are updated through gradient based optimization.

To improve robustness, random flips, rotations, and color perturbations (data augmentation) are applied during training, while batch normalization layers stabilize learning by normalizing feature distributions. The encoder starts from ImageNet pre-trained weights

[Deng et al., 2009], transferring general visual features to the building extraction task. At the end of training, the model has learned a set of parameters that allow it to generalize from the training data to unseen images.

3.5.2 Testing Phase

In the testing phase (Figure 9), the trained model is applied to new imagery to evaluate its generalization ability. The inputs are, normalized Red Green, Blue layer (RGBs) and a height layer tiled into 320×320 size tiles. For each tile, the model outputs a probability map representing the likelihood that each pixel belongs to the building class. Where tiles overlap, predictions are aggregated, by averaging, to produce probabilities for each pixel.

To convert these probabilities into a binary building mask, a decision threshold is applied: pixels above the threshold are classified as buildings, while those below are labeled as background. The resulting masks are refined through post-processing steps, including Douglas–Peucker simplification with a tolerance of 0.2m, which reduces noise while preserving building geometry. To define how well the model works, the resulting masks are compared to the ground truth.

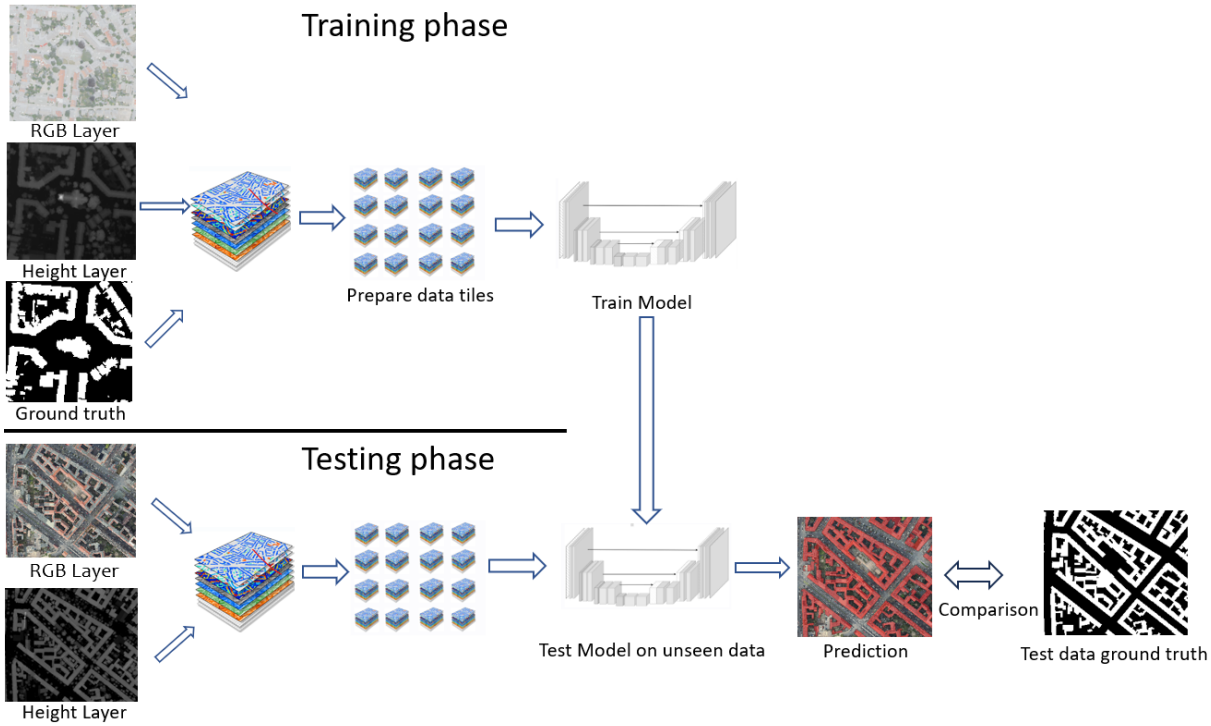


Figure 9: Training and testing workflow of the CNN model used in this work

4 Data

This chapter details the datasets used to train, validate, and test the building footprint model.

4.1 RGB

Digital orthophotos (DOP) are georeferenced, high-resolution aerial images produced in Germany according to state specific standards. For this work, TrueDOPs with a 10 cm ground resolution were used. Unlike traditional orthophotos, which are projected onto a terrain model, TrueDOPs employ a image based digital surface model (iDSM) that incorporates buildings, trees, and other elevated features. This corrects for radial displacement and the "leaning" of tall objects by filling hidden areas with information from overlapping images. The result is a geometrically accurate, complete representation of urban environments, well suited for mapping, planning, and GIS applications [NRW Geobasis, 2021]. A comparison to traditional DOPs is shown in Figure 10. All ortophotos used in this work are True DOPs. Orthophotos are provided in four layers: red, green, blue (RGB) and near-infrared (NIR). In this thesis, only the RGB layers of the TrueDOPs are used, as the NIR band does not improve model performance [Hertrich, 2024]. Throughout the text, the term RGBs refers to these three layers. Each image has a ground resolution of 0.1 m.

(a) Conventional DOP from 2016 (b) True DOP from 2019

Figure 10: Comparison between a traditional DOP and a True Orthophoto (True DOP) [NRW Geobasis, 2021].

4.2 LiDAR data

Light Detection and Ranging (LiDAR) estimates distances by emitting short laser pulses and measuring the time-of-flight of their returns. Airborne laser scanning (ALS) produces dense, georeferenced point clouds of (x, y, z) measurements that describe the three-dimensional structure of the surface. Individual pulses can generate multiple returns (first, intermediate, last), which facilitates the separation of canopy tops from the ground [Lillesand et al., 2015].

After acquisition, points are classified e.g., ground, vegetation, buildings, bridges, noise, using automated algorithms with subsequent manual quality control. These classes

are the base for the different height layers used in this work [Berlin Senatsverwaltung, 2021].

Because this work integrates datasets from multiple surveying authorities (NRW, Berlin, Brandenburg), the underlying classification schemes differ by region. For example, NRW provides a more granular set of classes than Brandenburg. To ensure comparability between training and testing, class selections were harmonized per region so that the derived DSM/DTM/nDSM layers have as similar semantics as possible. The resulting class mappings, and which classes where selected to create the different height layers, are documented in the Supplementary Materials (9).

4.2.1 DSM

A Digital Surface Model (DSM) represents elevations of the terrain including above-ground objects such as buildings and vegetation. In contrast to a DTM, a DSM captures the surface visible to the sensor and thus provides height cues that complement RGB appearance and aid the separation of spectrally similar classes (e.g., roads vs. rooftops) [Lillesand et al., 2015].

In this thesis, DSM rasters are derived from airborne LiDAR (ALS) point clouds. The procedure was adapted to the procedure that was used on the test data in Berlin. This procedure was explained by Matthias Weller from the Berlin surveying authority, who is the responsible for the LiDAR based height layers in Berlin. Using the same procedure, when working with the point clouds from different German federal states, ensures consistency between the different datasets. The procedure begins by reading LiDAR tiles (`.laz`/`.las`) and filtering the point clouds to retain surface-relevant classes (see the regional class tables; e.g., Tables 21, 22, and 23). The filtered LiDAR points (visualized in 11a) are then rasterized to a regular 0.5 m grid using a 2D binning method, where the maximum elevation value "z" within each cell is used to represent the surface height. The result of this binning method can be seen in Figure 11b. The problem with the result of this representation of the surface height is, that it contains NAs in those grid cells that did not contained points from the LiDAR data.

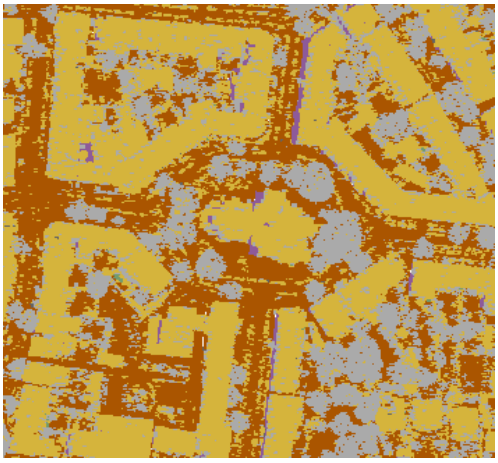
In order to produce a DSM at 1 m resolution, without NAs, the data is further processed using interpolation. All valid (non empty) cells from the 0.5 m DSM are extracted and converted into a set of (x, y, z) coordinates. A Delaunay triangulation is applied to these points, generating a Triangulated Irregular Network (TIN) [Evans et al., 2001]. Within each triangle, linear interpolation is performed using barycentric coordinates. A data point $\mathbf{p} = (x, y)$ lies inside a triangle defined by vertices $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, each with elevation values z_1, z_2, and z_3. The barycentric coordinates $(\lambda_1, \lambda_2, \lambda_3)$ are computed such that:

$$\mathbf{p} = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \lambda_3 \mathbf{v}_3, \quad \text{with} \quad \lambda_1 + \lambda_2 + \lambda_3 = 1$$

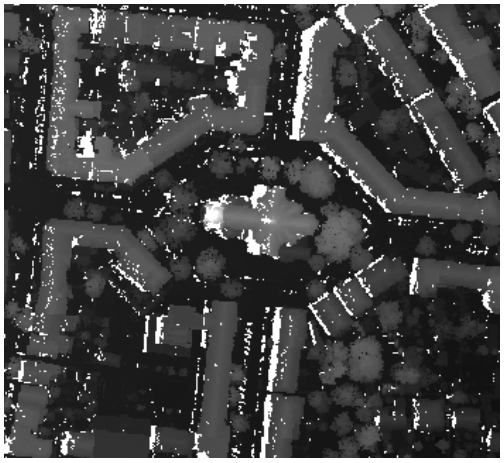
Then, the interpolated elevation value at $\mathbf{p}$ is given by:

$$z(\mathbf{p}) = \lambda_1 z_1 + \lambda_2 z_2 + \lambda_3 z_3$$

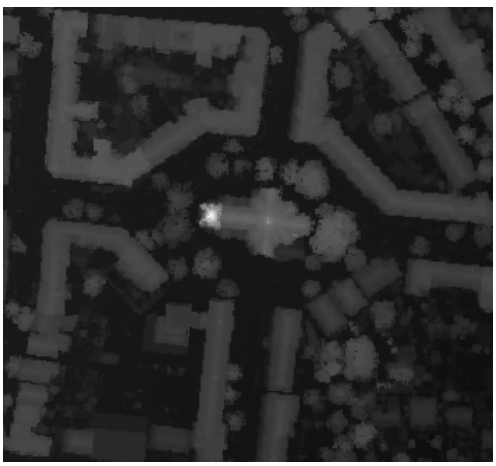
Once all raster cells in the 1m grid have been calculated through interpolation, the resulting DSM, with a resolution of 1m per pixel, is saved as a GeoTIFF (Figure: 11c). In a final step, a version of the dsm is created where the interpolated DSM is resampled to a resolution of 0.1 m. With a 0.1m resolution, the data fits to the RGB resolution (Figure: 11d).



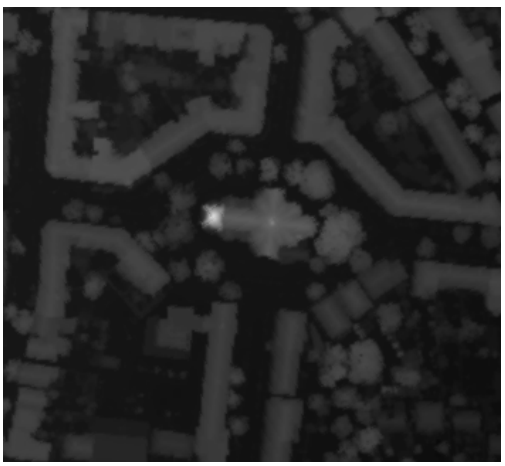
(a) Raw classified LiDAR points



(b) 0.5 m rasterized DSM (with nodata)



(c) 1 m interpolated DSM (TIN)



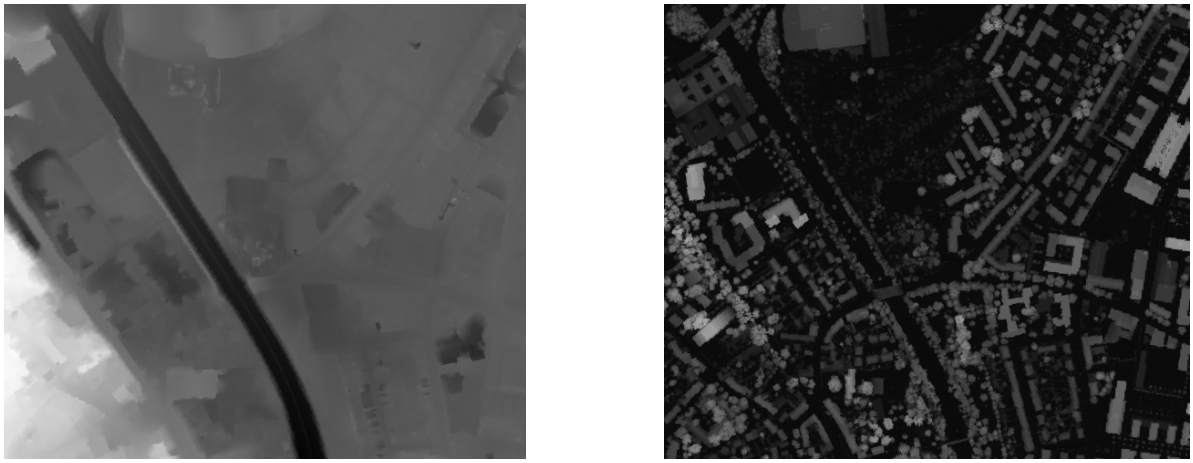
(d) Final DSM resampled to 0.1 m

Figure 11: DSM processing pipeline from ALS points to final 0.1m raster.

4.2.2 DTM

A Digital Terrain Model (DTM) is a representation of the Earth's bare surface, excluding buildings, vegetation, and other elevated objects. When derived from LiDAR point cloud data, DTMs are generated by selectively filtering points classified as ground, typically class 2 in standard classification schemes (see Table 24). These filtered points are then interpolated into a continuous raster surface that reflects the natural topography of the terrain.

The primary distinction between a DTM and a DSM lies in the type of features represented: while DSMs include the heights of all surface objects, DTMs are restricted to the underlying terrain Lillesand et al. [2015].



(a) Example for DTM data (b) Example for DSM data

Figure 12: Comparison of DTM and DSM data at the same extent in an area in Cologne.

The second difference of the DTM data compared to the DSM data is the resolution of the two-dimensional binning method. For the DSM data, a resolution of 0.5 grids has been used to determine the height per grid. On the DTM data a grid size of 1m was used. After that step the TIN procedure was used in order to fill the missing gaps. The data was resampled to a resolution of 0.1m

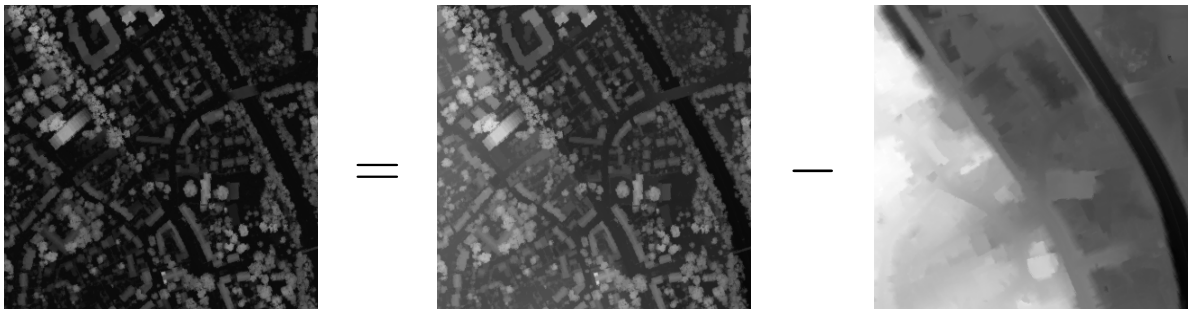
4.2.3 nDSM

A Normalized Digital Surface Model (nDSM) represents the height of objects above ground. It is obtained by subtracting the DTM from the DSM:

$$\text{nDSM} = \text{DSM} - \text{DTM}. \tag{1}$$

While the DSM contains elevations of all visible features (terrain, buildings, vegetation), the DTM contains only the underlying terrain. Their difference removes terrain influence and yields aboveground height, which is informative for distinguishing structures such as

buildings, tree canopies, and infrastructure [Lillesand et al., 2015]. Figure 13 visualizes this computation.



nDSM **DSM** **DTM**

Figure 13: Visual representation of nDSM calculation as DSM minus DTM.

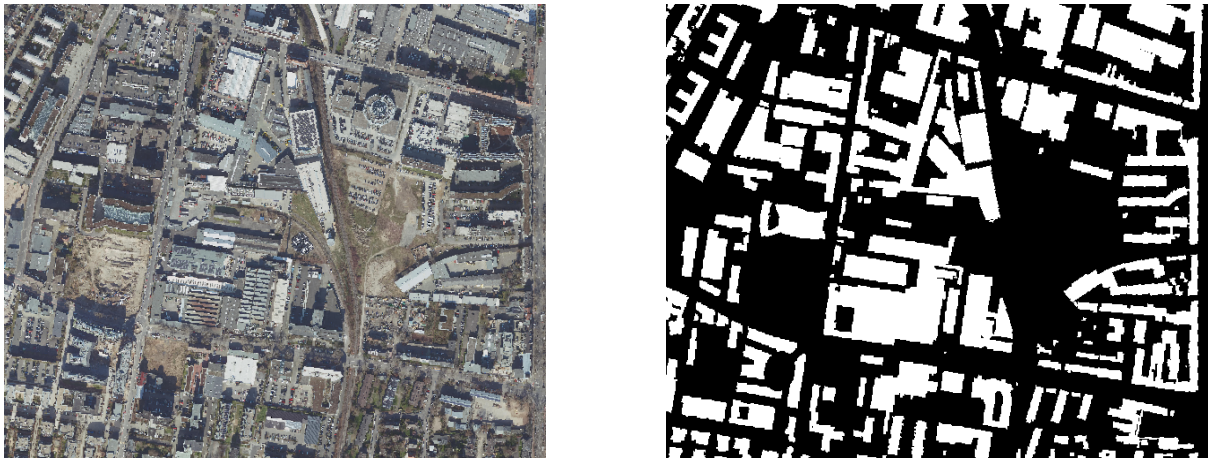
The DSM and DTM raster data is used at a resolution of 1m. The subtraction yields an nDSM of the same resolution. This raster is then resampled to 0.1m to match the RGB data.

4.3 Ground truth

To distinguish buildings from the background, the LoD2 definition was adopted, in which buildings are represented as 3D objects corresponding to man-made structures with roofs, exterior walls, and semantically structured boundary surfaces [Löwner et al., 2012]. The 2D building ground truth was derived from the RGB datasets in combination with official Level of Detail 2 (LoD2) building data [Open Geodata NRW, 2025, Open Geo Data Berlin, 2025].

Because the available LoD2 building datasets contains systematic issues (as explained in Section 1), missing or spurious buildings, manual correction was performed with the 0.1m RGB orthophotos as the primary reference. The nDSM (Section 4.2.3) was consulted in ambiguous cases; however, where conflicts arose, due to acquisition-date mismatches, the RGB imagery was preferred. This follows Stiller et al. [2023], who highlight RGB as the most informative predictor for the applied model.

Ground truth for the training regions was taken from previous DLR studies, while labeling of the test data was performed as part of this thesis. Corrected building vectors were rasterized to 0.1 m binary masks (building = 1, background = 0) and geo-aligned with the RGB grid.



(a) RGB patch from the training data (b) Building footprint ground truth

Figure 14: Example of derived ground truth from 0.1m RGB imagery [NRW Geobasis, 2021].

4.4 Training data

The training data for the model consists of five different layers: the three RGB layers (Red, Green and Blue), the height layer (LiDAR derived height profile) and the ground truth layer (Figure 15). The previous section gave an overview about the types of data used. This section shall explain which data combination was used for the different experiments and what sources they come from.

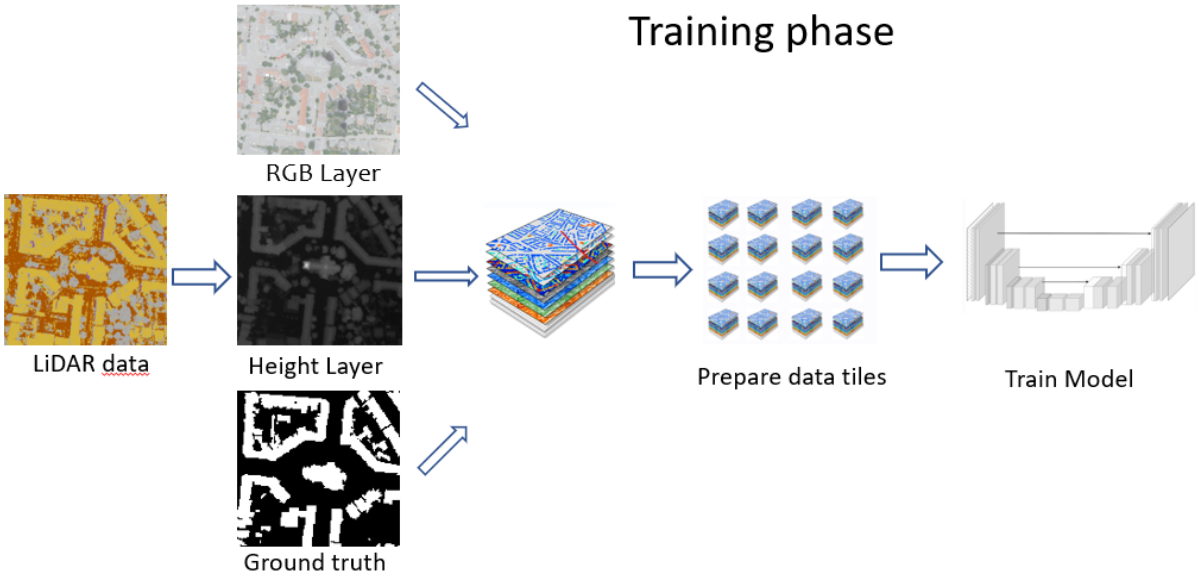


Figure 15: Overview of the training phase workflow

The main training data source, are the five regions in Figure: 16. Each regions has an area of 1.9 sqkm, making it 9.5 sqkm, of training data, in total.

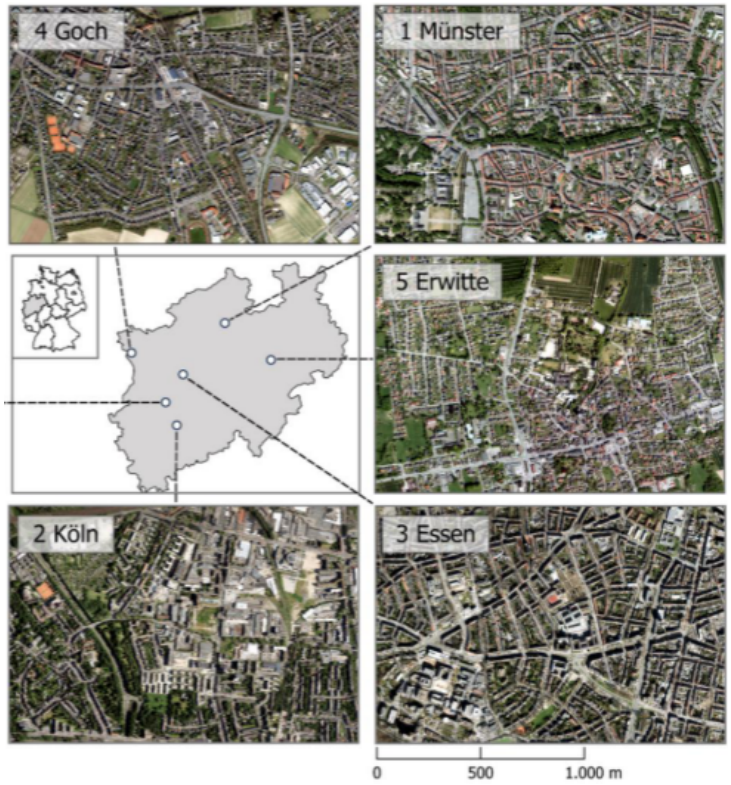


Figure 16: Overview of the RGB layers of the training regions adopted from [Hertrich, 2024].

The Baseline training data, consists out of the following layers and sources:

Table 3: Baseline training data summary

Datatype	**Year**	**Source**	**Comment**
RGB	2019	[NRW Geobasis, 2019]	0.1 m resolution, first statewide TrueDOP generation, worse data quality than later years.
nDSM	2017–2021	Statewide nDSM [NRW Geobasis, 2025c] 0.5 m	Downloaded as nDSM data, was all ready processed based on LiDAR data. Resampled to 0.1 m
Ground truth	2021	[Open Geodata NRW, 2025]	Footprints manually corrected against 0.1 m RGB, rasterized to 0.1 m.

To answer research question RQ1.4 and its following research hypotheses H1.4, a new training dataset, with DSM data instead of nDSM data had to be created (Table 4. Because the DSM was not available for the download from the NRW Geobasis [2025b], they had to be created based on the the LiDAR data from the same source (explained in Section 4.2.1). This new data comes from aerial imagery that was taken from 2021-2024, and therefore contains DSM data that is newer than the RGB tiles from 2019, that are displayed in Table 3. Out of this discrepancy in time a discrepancy in buildings emerged. Therefore, new RGB data (2023) had to be collected and the ground truth had to be adapted.

Table 4: DSM instead of nDSM training data summary

Datatype	Year	Source	Comment
RGB	2023	via WMS [NRW Geobasis, 2025a]	0.1 m True DOPs, manually corrected, higher overfly overlap as the RGB data from [NRW Geobasis, 2019]
DSM	2021–2024	Based on LiDAR point clouds [NRW Geobasis, 2025b]	Created as explained in Section 4.2.1. Point cloud class labels see Table 21.
Ground truth	2023	[Open Geodata NRW, 2025]	Footprints manually corrected against 0.1 m RGB, rasterized to 0.1 m.

4.4.1 Additional Training data

To answer research question RQ1.9 and its following H1.9 two additional training regions were added. One from Brandenburg and one from Duisburg in NRW.

Table 5: Additional training data — Duisburg (NRW)

Datatype	Year	Source	Comment
RGB	2023	via WMS [NRW Geobasis, 2025a]	0.1 m True DOPs, manually corrected, higher overfly overlap as the RGB data from [NRW Geobasis, 2019]
DSM	2017	Based on LiDAR point clouds NRW Geobasis [2025b]	creation as explained in Section 4.2.1. Point cloud class labels see Table 21.
Ground truth	2023	[Open Geodata NRW, 2025]	Footprints manually corrected against 0.1 m RGB, rasterized to 0.1 m.

Table 6: Additional training data — Brandenburg (BB)

Datatype	Year	Source	Comment
RGB	2018	[Brandenburg Geobasis, 2025]	Manually downloaded, data quality is much more blurry, especially edges of roofs
DSM	2016	Based on LiDAR point clouds [Brandenburg Geobroker, 2025]	creation as explained in Section 4.2.1. Point cloud class labels see Table 22
Ground truth	2018	Manually added footprints	

4.5 Test data

The main aim of this work is to improve the building footprint detection model and its surrounding components. To this end, regions had to be identified where the model struggled to correctly detect building footprints, particularly in cases of false positives. The selection procedure was as follows:

First, the baseline model [Stiller et al., 2023] was applied to the entire area of Berlin to obtain predicted building footprints. These predictions were then intersected with the 2018 LBM-DE land cover/use dataset [Bundesamt für Kartographie und Geodäsie, 2025], which provides 39 land-cover classes for Berlin. For each class, intersection areas with predicted building footprints were computed (see Supplementary Materials 9).

Tiles with disproportionately high overlaps in classes unlikely to contain buildings (e.g., streets and railways, water, trees) were selected. In addition, representative examples of urban morphology (housing, industrial, inner city) and known challenging cases (bridges and boats, forest roads, map edges, riverbanks) were included. The selection was carried out on a regular $1\,\mathrm{km} \times 1\,\mathrm{km}$ grid, with each chosen tile covering exactly $1\,\mathrm{km}^2$. All test areas are summmarized in Figure 17

(a) Forest (b) Streets and railway (c) Water

(d) Housing area (e) Industrial area (f) Inner city

(g) Bridges and boats (h) Street through forest (i) Water on edge

(j) Water riverbank

Figure 17: Examples of selected high false positive test areas in Berlin, ordered by category: (a–c) high FP land cover / land use classes; (d–f) characteristic Berlin urban scenes; (g–j) manually added complex falsepositive cases. Each tile covers 1 km^2.

The same information is needed in the test data, as compared to the training data. Below is an overview of the different layers used in the test data:

Table 7: Test data from Berlin

Datatype	**Year**	**Source**	**Comment**
RGB	2024	DLR internal WMS	TrueDOPs at 0.1 m resolution; most recent imagery available for Berlin.
nDSM/DSM	2021	Downloaded from Berlin Geoportal [2025]	Both the DSM and the nDSM layer were used in different experiments. Resampled to 0.1 m for compatibility with RGB imagery. Informations about the point cloud class labels in Table 23.
Ground truth	2024	LoD2 building footprints [Open Geo Data Berlin, 2025]	Footprints corrected manually against 0.1 m RGB

5 Research design

The testing of the different possible settings in pre-processing, post-processing and changes in training data, in this thesis follows a sequential evaluation strategy. Each component of the data pipeline is evaluated step by step: the configuration that achieves the best performance at one stage is retained and used as the baseline for the subsequent stage. This allows for assessing the effect of incremental improvements.

To better isolate the effect of individual components, the settings in question are, at the first experiment, initially deactivated, i.e., set to a neutral baseline such as no additional height layer or no normalization. Parameters are then introduced and varied one by one. This way, performance differences can be attributed solely to the parameter under investigation.

This procedure is known in machine learning research as an *ablation study*. Specific components of a trained neural network or data pipeline are systematically removed to observe how such changes affect the performance [Meyes et al., 2019].

For each experiment, results will be ranked according to the evaluation metrics, compared against both the baseline from the previous stage and the corresponding ablation. This stepwise ranking makes it possible to systematically assess the relative importance and effectiveness of each component in the overall pipeline.

Experiments were run on an NVIDIA RTX A4000 (16 GB VRAM; CUDA 12.8) using PyCharm 2024.1, with 64 GB system RAM. Each model setting will be trained with 50 epochs.

6 Methodology and Results

Chapter structure. In this chapter, methodology and results are presented together for each of the five experiments (Normalization, Height Layer, Tile Overlap, Decision Threshold, and Selection of Training Data). Although this is uncommon, integrating them here avoids redundancy and preserves the logical flow from research question to experimental setup and immediate outcome. Implications are synthesized subsequently in the Discussion (Section 7).

Evaluation Metrics The performance of the model and its different parameter settings is evaluated using common metrics in semantic segmentation: Intersection over Union (IoU), Precision, Recall, F1 Score, and Overall Accuracy. In this study, Overall Accuracy is considered the most important evaluation metric. Because some of the selected test areas do not contain any buildings, which are in this case the, true positives (TP). For such cases, the standard segmentation metrics, Intersection over Union (IoU), Precision, Recall and F1 Score, cannot be computed, as their formulas require at least one TP. Overall Accuracy, in contrast, can be calculated for *all* test images, regardless of whether true positives are present. Therefore, it serves as the most comprehensive indicator of model performance across the entire test set.

The IoU, Precision, Recall and F1 Score are still reported for those test areas that contain true positives, as they provide additional insight into the model's segmentation quality when buildings are present. In this dataset, images with TP correspond to Figure 17 (b, d, e, f, g), while images without TP are shown in Figure 17 (a, c, h, i, j).

Overall Accuracy. Overall Accuracy is the fraction of correctly classified pixels over all pixels:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

In the special case of test images without any true positives (TP = 0 and FN = 0), the formula reduces to:

$$\text{Accuracy} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Here, the metric reflects the proportion of background pixels correctly classified as background, providing a valid measure of performance even when no buildings are present in the ground truth.

Intersection over Union (IoU). IoU, also known as the Jaccard Index, measures the overlap between the predicted and ground truth building masks:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

It ranges from 0 (no overlap) to 1 (perfect match), with higher values indicating better segmentation accuracy.

Precision. Precision quantifies the proportion of predicted building pixels that are correct:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

A high precision means few false positives.

Recall. Recall measures the proportion of actual building pixels that are correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

A high recall means few false negatives.

F1 Score. The F1 Score is the harmonic mean of Precision and Recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

It balances false positives and false negatives.

After Overall Accuracy, the two most important secondary indicators are IoU and F1 Score. Both are widely used in semantic segmentation to assess the quality of predicted object shapes and boundaries, as for example in: Dabove et al. [2024], Chawda et al. [2018], Li et al. [2019]. IoU measures the proportion of overlap between predicted and ground truth masks relative to their union, making it a direct measure of spatial agreement. F1 Score, as the harmonic mean of Precision and Recall, balances the trade off between missing buildings (false negatives) and including non existent buildings (false positives), providing a more holistic measure of segmentation performance.

While Precision and Recall still provide valuable insight into model behavior, especially for understanding bias towards over or under segmentation, they are not used as primary ranking criteria for parameter selection in this work. Instead, they serve as supporting metrics to interpret why certain parameter configurations may perform better or worse.

6.1 Normalization

Normalization, of the input pixel ranges, is a key preprocessing step in semantic segmentation, ensuring input features share a common scale to stabilize training and improve generalization, as described in 2.1. Two widely used methods are Min–Max normalization and Z-score normalization. Min–Max rescales values to a fixed range, typically

$[0, 1]$:

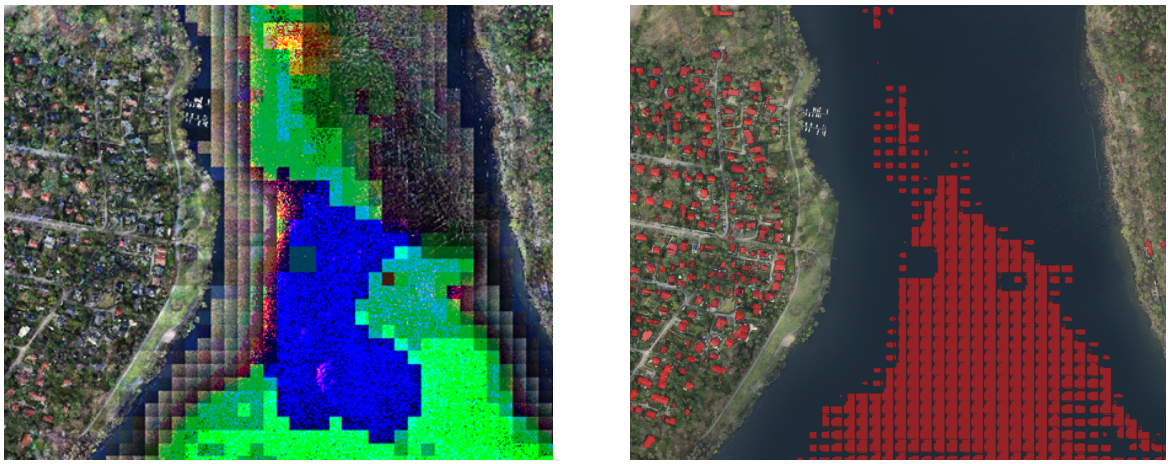
$$X_{\text{min-max}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{2}$$

where x is the input value and $x_{\min}$, $x_{\max}$ are the dataset-specific minima and maxima. Z-score normalization, by contrast, standardizes inputs to zero mean and unit variance:

$$X_{\text{z-score}} = \frac{x - \mu}{\sigma}, \tag{3}$$

where μ and σ are the mean and standard deviation of the training data.

In the baseline setup 3.5 adapted from Stiller et al. [2023], Min–Max normalization was applied per layer to large training images (16,755×11,399 px). As these images already spanned nearly the full RGB range, normalization had little effect. However, at test time the small tiles (320×320 px) were normalized independently, as they were drawn from the WMS layer. In homogeneous tiles (e.g., water), the very small pixel range made Min–Max collapse almost all pixels in a layer to zero, deleting that layer (e.g., an R-layer with 99.9% value 29 and 0.01% value 30 maps 29 to 0). This confused the model and produces systematic false positives, as visible in Figure 18.



(a) Min-Max normalized image tiles stitched together

(b) Corresponding building footprint prediction

Figure 18: Error from Min–Max normalization of small tiles.

A further issue was the sequential use of Min–Max and Z-score normalization, a non-standard practice. Since Min–Max rescales and Z-score re-centers and re-scales, applying both cancels out the benefits of each and may confuse the network, reducing interpretability and consistency. Best practice is to evaluate these approaches separately and select the most effective one empirically for the task [Jeon et al., 2021].

The baseline configuration nonetheless achieved solid results (Table 8), with overall accuracy of 0.912, IoU of 0.811, and F1 of 0.891. However, the imbalance between high

recall (0.953) and lower precision (0.843) indicates systematic overprediction, consistent with the normalization artifacts described above.

Table 8: Baseline normalization performance.

Trial	Parameter	Overall Accuracy	IoU	F1 Score	Precision	Recall	Ranking
Normalization	Baseline set-up	**0.912**	**0.810**	**0.890**	**0.842**	**0.952**	1

6.1.1 Ablation

In order to assess the importance of normalization in the preprocessing pipeline, an ablation experiment was conducted in which neither Min–Max normalization nor Z-score standardization was applied to the input data. In this configuration, the raw pixel values of the RGB layers and the nDSM data were fed directly into the network without any rescaling or centering.

Table 9: Comparison between Baseline and ablation experiment

Trial	Parameter	Overall Accuracy	IoU	F1 Score	Precision	Recall	Ranking
Normalization	**Baseline**	0.912	**0.810**	**0.895**	**0.842**	**0.952**	1
	Ablation	**0.958**	0.587	0.848	0.772	0.953	2

Results: The ablation experiment (Table 8) shows that, despite achieving the highest Overall Accuracy (0.958), the model's Intersection over Union (IoU) drops sharply to 0.587 compared to the baseline's 0.8109. The F1 Score also decreases to 0.848, with Precision falling to 0.772 while Recall remains high at 0.953. This indicates that the model continues to identify most building pixels (high recall), but does so with a much larger proportion of false positives, and with poorer spatial alignment between predicted and true building shapes.

Although Overall Accuracy is the primary metric for ranking in this study (Section 6), it becomes less meaningful in extreme cases such as this one. The high Overall Accuracy here is largely driven by correctly classifying background pixels in areas without buildings, while the actual segmentation quality for building footprints, as reflected by the low IoU, is poor. This highlights that without proper normalization, the model struggles to generalize well to the building class, leading to degraded object level performance despite seemingly strong pixel level accuracy.

These findings confirm that normalization plays a critical role in ensuring balanced model performance across both classes, and that Overall Accuracy alone is insufficient to capture quality in cases where object segmentation degrades severely.

6.1.2 Min–Max Normalization

In this experiment, only Min–Max normalization was applied to all four input layers, without any Z-score normalization. The normalization was performed independently for

each image or tile, using the tile specific minimum and maximum values. The minimum and maximum values from the training data were not transferred to the test data. This decision was based on two considerations: first, the RGB layers already spanned nearly the full 0–255 range, making global scaling unnecessary; and second, restricting nDSM values to a fixed range would risk clipping important elevation differences.

Table 10: Min-Max Normalization experiment

Trial	Parameter	Overall Accuracy	IoU	F1 Score	Precision	Recall	Ranking
Normalization	**Baseline**	0.912	**0.810**	**0.890**	**0.842**	**0.952**	1
	Ablation	**0.958**	0.587	0.848	0.772	0.953	2
	Min-Max Normalization	0.896	0.741	0.844	0.771	0.945	3

The results (Table 8) show that this approach yields an Overall Accuracy of 0.896, the lowest among the tested configurations. The IoU drops to 0.741 compared to 0.810 in the baseline, and the F1 Score decreases to 0.844. Precision (0.771) is substantially lower than in the baseline, while Recall (0.945) remains high.

6.1.3 Z-score Normalization

In this experiment, only Z-score normalization was applied to the input data. For each of the four layers (RGB + nDSM), the mean and standard deviation were computed from the training images, and these same statistics were used to normalize both the training and test data. This ensured that all input values were centered around zero with unit variance, and that the scaling applied to the test set was consistent with that of the training set.

Table 11: Comparison between Experiments

Trial	Parameter	Overall Accuracy	IoU	F1 Score	Precision	Recall	Ranking
Normalization	Baseline	0.912	0.810	0.890	0.842	0.952	2
	Ablation	0.958	0.587	0.848	0.772	**0.953**	3
	Min-Max Normalization	0.896	0.741	0.844	0.771	0.945	4
	Z-score Normalization	**0.978**	**0.840**	**0.911**	**0.875**	0.952	1

The results (Table 8) show that this configuration outperforms all other tested normalization approaches. It achieves the highest Overall Accuracy (0.978), IoU (0.840), F1 Score (0.911) and Precision (0.875) while maintaining a high Recall (0.952). Compared to the baseline, Precision improves markedly, indicating fewer false positives and better discrimination between building and background pixels. The simultaneous improvement in IoU and F1 Score demonstrates that the model is achieving both high spatial overlap with the ground truth and a balanced trade off between Precision and Recall.

A clear distinction emerges between test areas that contain true positives (TP) and those that do not. In TP containing areas (see Fig. 17 b, d, e, f, g), improvements in IoU and F1 Score indicate overall a slightly more accurate building delineations and a better balance between detecting buildings and avoiding false alarms. In contrast, for non TP

areas (Fig. 17 a, c, h, i, j), metrics such as IoU, F1 Score, Precision and Recall, remain at zero by definition, since no buildings are present in the ground truth. For these cases, Overall Accuracy becomes the only meaningful indicator, showing substantial reductions in false detections with the Z-score normalization.

6.2 Height Layer

The fourth input layer in the model contains height information in the form of a normalized Digital Surface Model (nDSM), as described in Section 4.2.1. In the baseline configuration, this layer provides per pixel elevation relative to the surrounding terrain, enabling the model to leverage vertical structure as an additional discriminative feature.

The experiments in this section builds upon the best performing configuration from the normalization experiments (Section 6.1), which now serves as the new baseline. The performance of this baseline, which uses Z-score normalization, is summarized in Table 12. These values form the reference point against which all subsequent height layer related modifications will be evaluated.

Table 12: Performance of the new baseline configuration, derived from the optimal normalization settings in Section 6.1.

Trial	**Parameter**	**Overall Accuracy**	**IoU**	**F1 Score**	**Precision**	**Recall**	**Ranking**
Height Layer	Baseline	0.978	0.840	0.911	0.875	0.952	1

6.2.1 Ablation

The ablation experiment removes the height layer entirely, leaving the model to rely solely on spectral information from the RGB layers. As shown in Table 13, the removal of the height information results in a noticeable decline in performance across all key metrics. Overall Accuracy drops from 0.978 to 0.957, but the more telling changes are observed in the spatially sensitive metrics: IoU decreases by 0.1, and F1 Score declines by 0.025, indicating a reduced ability to produce precise and complete building footprint delineations.

Precision decreases slightly, from 0.875 to 0.843, showing an increased rate of false positives when height cues are absent. Recall also drops from 0.952 to 0.933, reflecting a higher number of missed building pixels.

Table 13: Ablation of the height layer compared to the baseline model.

Trial	**Parameter**	**Overall Accuracy**	**IoU**	**F1 Score**	**Precision**	**Recall**	**Ranking**
Height Layer	Baseline	**0.978**	**0.840**	**0.911**	**0.875**	**0.952**	1
	Ablation height layer	0.957	0.740	0.886	0.843	0.933	2

6.2.2 DSM Data

In this experiment, the height layer that consisted of nDSM data was replaced with a Digital Surface Model (DSM), in the training and testing data. The training DSM data was created through the same procedure as the test data for Berlin (see Section 4.2.1. Through this procedure the training data closely matches the Berlin test data in resolution, preprocessing, and interpolation methodology.

Because the newly generated DSM data are based on LiDAR acquisitions from 2021–2023, new RGB orthophotos (2022–2024) were used for the corresponding training regions to ensure temporal consistency between height and optical inputs. Ground truth building masks were also adapted to reflect structural changes and building stock updates (see Table 4).

As shown in Table 14, replacing the nDSM with the DSM data leads to slightly higher Overall Accuracy (0.979 vs. 0.978), IoU (0.842 vs. 0.840), with a substantial improvement in Precision (0.913 vs. 0.875). This indicates a reduction in false positives. However, Recall decreases (0.912 vs. 0.952), suggesting a more conservative detection behavior that occasionally misses buildings.

Table 14: Comparison of the baseline against the ablation and DSM data.

Trial	Parameter	Overall Accuracy	IoU	F1 Score	Precision	Recall	Ranking
nDSM	Baseline	0.978	0.840	0.911	0.875	**0.952**	2
	Ablation	0.957	0.740	0.886	0.843	0.933	3
	DSM Data	**0.979**	**0.842**	**0.912**	**0.913**	0.912	1

6.3 Tile Overlap

The *overlap* parameter controls the fraction of shared pixels between adjacent tiles when partitioning large orthophotos into smaller patches for training and testing. Choosing an overlap affects both the quantity of unique training samples and the consistency of model predictions along tile boundaries. Larger overlap ratios increase sample redundancy and ensure that boundary pixels are seen in multiple contexts, often improving edge segmentation performance and reducing boundary artifacts. However, higher overlap also inflates the dataset size and increases training and inference time. Conversely, minimal or zero overlap yields fewer patches and faster processing but may degrade segmentation quality at tile edges, manifesting as discontinuities or missing detections. An example of how the overlap works, can be seen in Figure 19.

To quantify these trade offs, experiments were conducted with different overlap sizes. In the training phase, only overlaps of 0%, 10%, 20%, 30%, and 50% were tested, as larger values would be unreasonable from a time efficiency perspective. In the testing phase, overlaps from 0% up to 80% were evaluated to better understand their impact on

prediction quality.

Figure 19: Example of tiles with 20% overlap marked by the yellow bands. Image is from the test data in Berlin.

As shown in Figure 20, the result for a 0% overlap in the testing phase is substantially below all other settings; the panel displays IoU, but the same pattern holds for Overall Accuracy, F1, Precision, and Recall. To make the smaller differences among the non-zero settings visible, the testing overlap of 0 is excluded in the visualization. Figure 21 plots all metrics with the test overlap on the x-axis, while line colours encode the training overlap (0–50%). This design separates the effect of test overlap, described through the x-axis from training overlap (colour), clarifying how performance (y-axis) varies within the non-zero range.

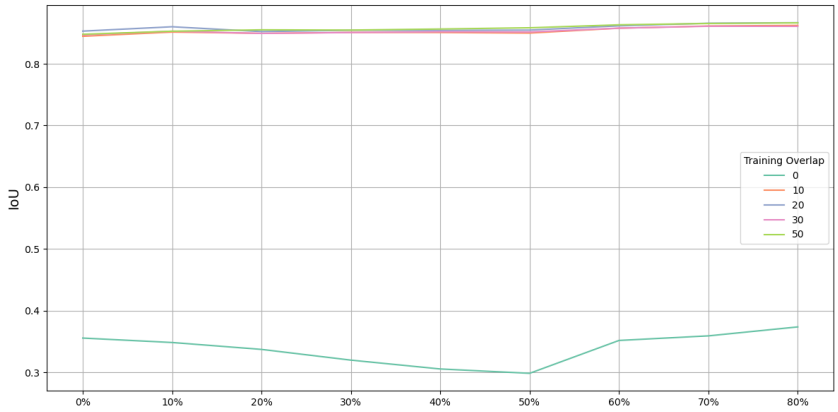
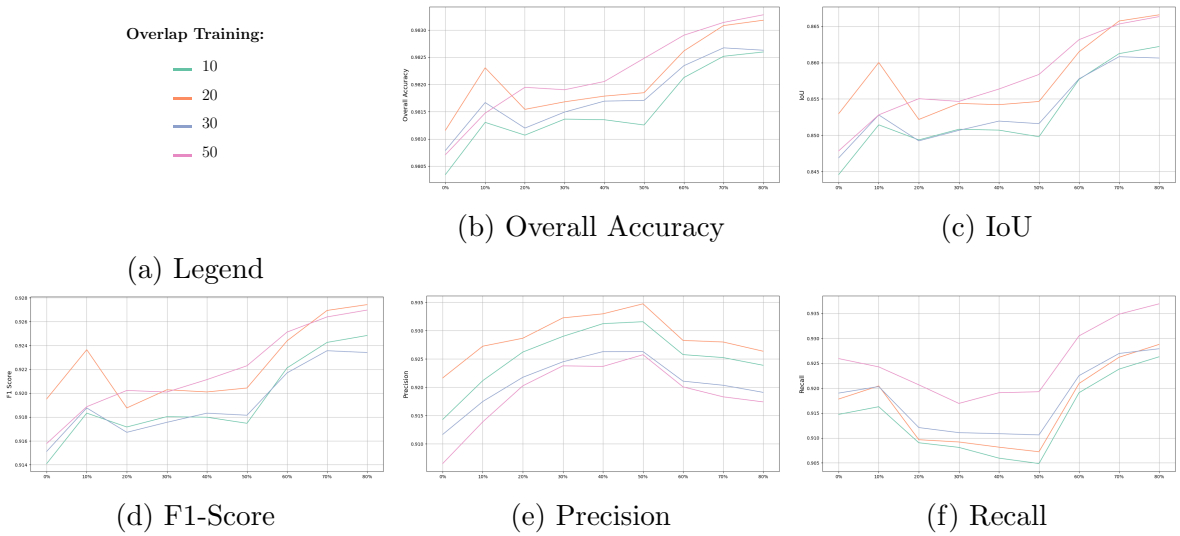


Figure 20: Comparison of model performance across different overlap settings. The 0% overlap case is significantly below the others, making differences between higher overlaps harder to distinguish.



(a) Legend (b) Overall Accuracy (c) IoU

(d) F1-Score (e) Precision (f) Recall

Figure 21: Performance metrics across overlap settings.

Two optima emerge, both visible in Figure 21 and quantified in Table 15. Overall Accuracy, IoU, and F1-Score exhibit an early local peak when the training overlap is set to 20% and the testing overlap to 10%. Performance then declines slightly for intermediate settings and reaches a global optimum at a training overlap of 50% combined with a testing overlap of 80%.

However, the computational cost of the global optimum configuration, with 50% training overlap and an 80% test overlap, is computationally difficult. Extrapolation from this test run indicated that the 50% training overlap requires 105 hours of training, and using this model for bigger areas like Berlin at 80% test overlap would take about 11 days. Given that the gain in Overall Accuracy over the early peak (train 20%, test 10%) is only ≈ 0.001, this additional cost is not reasonable. Therefore the first peak with 20% overlap in training and 10% in testing is used as the improved setting for subsequent experiments (see also Section 9).

Table 15: Comparison of optimal Overlaps with the Baseline from last experiment

Trial	Parameter	Overall Accuracy	IoU	F1 Score	Precision	Recall	Ranking
Overlap	Baseline	0.979	0.842	0.912	0.913	0.912	3
	Ablation	0.365	0.124	0.216	0.187	0.271	4
	Training 20% Testing 10%	0.982	0.860	0.923	**0.927**	0.920	2
	Training 50% Testing 80%	**0.983**	**0.866**	**0.926**	0.917	**0.937**	1

6.4 Decision Threshold

In the baseline, predictions from overlapping tiles were merged by majority vote. For each pixel, it was counted how many of the tiles covering it predicted "building" versus "background" and assigned the class with more votes. This is equivalent to using a fixed decision threshold of $\tau = 0.5$ on the aggregated score: a pixel is labeled "building" if at

least half of the votes (or the mean probability) favor that class. Because τ is fixed at the conventional default and not tuned, this baseline also serves as an ablation.

To improve on this, a configurable decision threshold is introduced. In this setting, each pixel may have multiple prediction probabilities due to overlapping tiles. These probabilities are aggregated (e.g., by averaging or summing), and the pixel is classified as a building only if the aggregated score exceeds a defined threshold (e.g., 0.7). Varying the decision threshold allows explicit control over the balance between precision and recall: lower thresholds increase sensitivity but raise false positives, while higher thresholds reduce false positives but risk omitting true positives. In contrast to majority voting, this threshold based approach provides finer control of this trade off.

A visualization with the $\tau = 0$ value is shown in Figure 22. Another visualization without the $\tau = 0$ can be seen in Figure 23 for better distinguishability between the values.

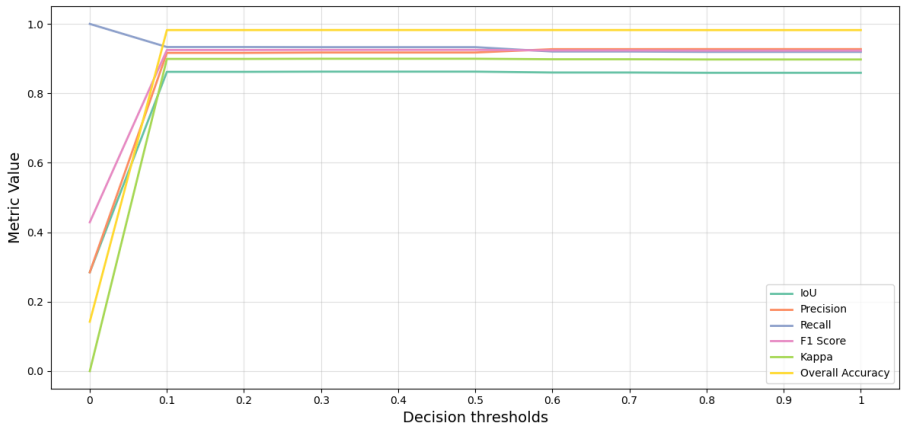


Figure 22: Impact of varying the decision threshold (0.0 to 1.0) on different evaluation metrics. Lower thresholds increase recall but reduce precision, while higher thresholds improve precision at the expense of recall.

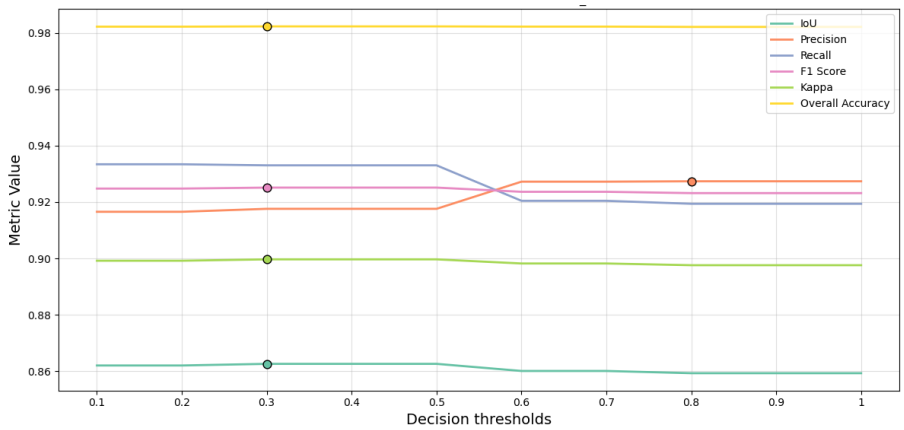


Figure 23: Impact of varying the decision threshold (0.1 to 1.0) on evaluation metrics. The 0.0 case is omitted to make the differences between thresholds more distinguishable.

The results show that Overall Accuracy, F1-Score, and IoU all have the highest value

at a decision threshold of about 0.3, suggesting this value provides the best overall trade off. Precision reaches its maximum at a much higher threshold of 0.8, meaning stricter requirements for building classification reduce false positives but at the cost of recall. Recall itself peaks at the lowest threshold (0.0), since every potential building pixel is labeled positive, but this naturally introduces large numbers of false detections. For practical purposes, the 0.0 threshold is not considered, as it visually masks the smaller but more relevant differences across the other thresholds. Hence, a decision threshold of 0.3 emerges as the most balanced choice for this task.

Table 16: Comparison of baseline (ablation at fixed 0.5 threshold) with the optimized threshold setting.

Parameter	**Overall Accuracy**	**IoU**	**F1 Score**	**Precision**	**Recall**	**Ranking**
Baseline (0.5 threshold)	0.982	0.860	0.923	**0.927**	0.920	2
Decision Threshold (0.3)	**0.982**	**0.862**	**0.923**	0.917	**0.933**	1

6.5 Selection of training data

The baseline setup is trained on imagery from five predominantly urban regions in North Rhine-Westphalia (NRW). In the work by Stiller et al. [2023], it was shown, that all 5 selected training regions improve the model performance. However, it still often misclassifies specific types of regions. To address this, two additional 1 km^2 training regions were curated to specifically target the remaining failure modes. First, very large rooftops, where tiles that do not touch roof edges tend to be missed (false negatives), and second large vessels that are still sometimes predicted as buildings (false positives). One area was selected in Brandenburg to cover extensive roof structures, and one in Duisburg to capture harbour scenes with large boats.

Figure 24: Example of the additional Brandenburg training region targeting very large and green roofs.

Figure 25: Example of the additional Duisburg training region targeting harbour scenes with large vessels.

From each of the two regions, tiles were added to the training set in 20% increments (20, 40, 60, 80, 100% of the available tiles) to test whether gradually increasing the share of these targeted scenes improves performance. Across all increments and for both regions, no configuration increased the overall accuracy on the Berlin test set compared to the baseline. The detailed metric deltas are summarized in Figures 26 and 27, while Figures 24 and 25 show the two new training regions.

The creation of the additional training data are explained in the data Section in Table 3 for the data from Duisburg and Table 6 for the data from Brandenburg.

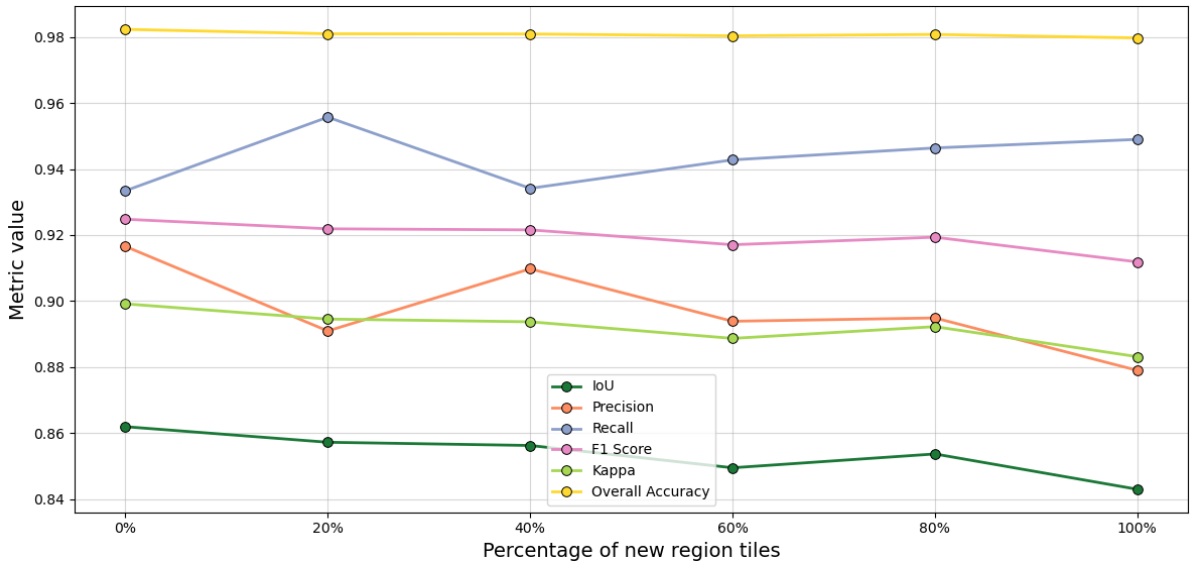


Figure 26: Change in evaluation metrics when adding tiles from the Brandenburg region (large roofs) in 20% increments.

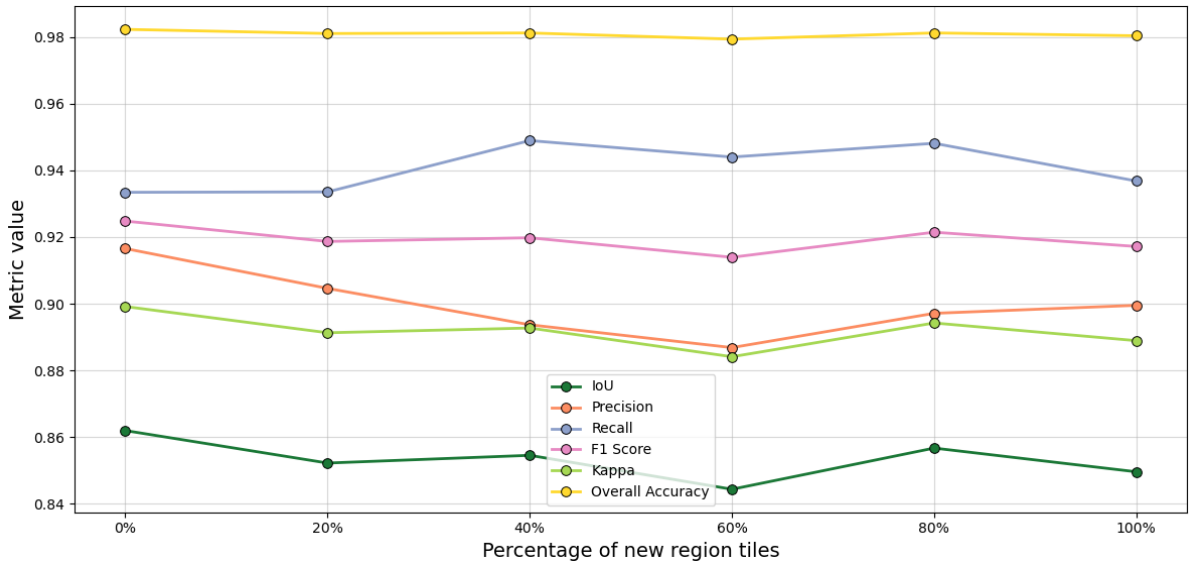


Figure 27: Change in evaluation metrics when adding tiles from the Duisburg region (harbour with large vessels) in 20% increments.

In summary, while the additional datasets were specifically chosen to address persistent failure modes, none of the increments of added training tiles whether from Brandenburg or Duisburg led to an improvement in the overall accuracy of the model. This is as well testified by table 17. The ablation which is at the same time the baseline, is compared against the best resulting values of the additional training data.

Table 17: Comparison of baseline (ablation) with best performance of additional training data.

Parameter	**Overall Accuracy**	**IoU**	**F1 Score**	**Precision**	**Recall**	**Ranking**
Ablation	**0.982**	**0.862**	**0.923**		**0.917**	**0.933**
1						
80% tiles from Duisburg	0.981	0.856	0.921	0.897	0.948	2

6.6 Summary of Results

Table 18 summarizes all Results that have been gathered. It makes clear, that the biggest gain in performance could be achieved through the correct Normalization and the use of DSM data instead of nDSM data as the height Layer.

Table 18: Summary table of the different tests.

Trial	**Parameter**	**Overall Accuracy**	**IoU**	**F1 Score**	**Precision**	**Recall**	**Ranking**
Normalization	Baseline	0.912	0.810	0.890	0.842	0.952	2
	Ablation	0.958	0.587	0.848	0.772	**0.953**	3
	Min-Max Normalization	0.896	0.741	0.844	0.771	0.945	4
	Z-score Normalization	**0.978**	**0.840**	**0.911**	**0.875**	0.952	1
Height Information	Ablation	0.957	0.740	0.886	0.843	0.933	2
	DSM data	**0.979**	**0.842**	**0.912**	**0.913**	0.912	1
Overlap	Ablation	0.365	0.124	0.216	0.187	0.271	3
	Training 20% / Testing 10%	0.982	0.860	0.923	**0.927**	0.920	2
	Training 50% / Testing 80%	**0.983**	**0.866**	0.926	0.917	**0.937**	1
Decision Threshold	Ablation	0.982	0.860	0.923	**0.927**	0.920	2
	Decision Threshold (0.3)	**0.982**	**0.862**	**0.923**	0.917	**0.933**	1
Additional Training Data	**Ablation**	**0.982**	**0.862**	**0.923**	0.917	**0.933**	1
	80% tiles from Duisburg	0.981	0.856	0.921	0.897	0.948	2

The comparison of baseline and final results (Table 19) highlights the magnitude of the improvements. Overall Accuracy increased by 7.0 percentage points (pp) from 0.912 to 0.982, IoU by 5.2 pp from 0.810 to 0.862, F1 Score by 3.3 pp from 0.890 to 0.923. Precision showed the strongest gain with 7.5 pp from 0.842 to 0.917, while Recall declined by 1.9 pp from 0.952 to 0.933.

Table 19: Comparison of first baseline and final improved version.

Trial	**Parameter**	**Overall Accuracy**	**IoU**	**F1 Score**	**Precision**	**Recall**	**Ranking**
Final comparison	Baseline	0.912	0.810	0.890	0.842	**0.952**	2
	Final results	**0.982**	**0.862**	**0.923**	**0.917**	0.933	1

7 Discussion

The sequential experiments conducted in this thesis directly addressed the main research question (RQ1) and its sub-questions by systematically refining preprocessing, postprocessing, and training data selection. The results provide partially evidence on the validity of the proposed hypotheses (H1–H1.9), which will be explained in the following paragraphs.

Normalization The first set of experiments investigated the role of input data normalization. The results clearly support the hypotheses that normalization improves model accuracy (H1.1) and that Z-score normalization outperforms Min-Max normalization in heterogeneous landscapes (H1.2). Among the tested strategies, Z-score normalization achieved the highest overall accuracy and the best balance between precision and recall, while avoiding the systematic false positives introduced by per tile Min-Max scaling.

This pattern can be explained by a normalization mismatch between training and test data. When Min-Max normalization is computed on a per tile basis, particularly in small or homogeneous tiles such as open water areas, the dynamic range of certain layers is compressed toward zero. This erases informative variation and encourages the model to misclassify large regions as buildings. Although recall remains high, since the model is inclined to label many pixels as buildings, the decline in IoU and precision indicates reduced segmentation quality and an increase in false positives.

A regional comparison confirms this mechanism. The weakest performance of per tile Min-Max normalization occurs in very homogeneous environments, such as water bodies, where local variation is minimal. By contrast, results in heterogeneous urban areas, such as inner city tiles, are relatively better, since the stronger local contrast partly is not leading to the scaling problem. In other words, Min-Max normalization may be acceptable on heterogeneous tile, but it performs poorly in homogeneous landscapes.

Z-score normalization, in contrast, avoids these pitfalls by relying on training set statistics, ensuring consistent scaling between training and test data. This stabilizes the input distribution and allows the model to generalize more effectively. As a result, building delineations are more accurate and overprediction errors are reduced. While Z-score and Min-Max normalization yield similar results in heterogeneous tiles, Z-score provides a clear advantage in homogeneous tiles. For this reason, Z-score normalization emerges as the most effective strategy for this dataset and model configuration.

Height Layer The inclusion of height information alongside RGB data was examined. As hypothesized (H1.3), the addition of vertical structure significantly improved segmentation accuracy compared to the ablation. Between the two tested representations, DSM data outperformed nDSM (supporting H1.4), particularly by reducing false positives in

spectrally ambiguous areas such as water bodies and forests. While recall decreased slightly for test areas with huge homogeneous rooftops, the precision gains more than compensated for this drawback, leading to more reliable predictions overall.

The benefits of DSM input are especially visible water and forest areas, false positives were almost completely eliminated, with a false positive rate of 0% in pure water bodies and dense forest regions, of the test areas. In structurally complex settings such as bridges, industrial facilities, and mixed street–railway areas, DSM data improved the model's geometric understanding and reduced misclassifications of infrastructure elements as buildings. For instance, in the "Streets & railway" category, IoU increased by more than 0.04, reflecting a clearer separation between buildings and their surroundings.

The main limitation of the DSM input was observed in homogeneous rooftop tiles without visible edges, where some marginal building pixels were missed, creating gaps in otherwise continuous building outlines. This slight loss in recall, however, is outweighed by the consistent reduction in false positives across other classes. Overall, these findings demonstrate that incorporating DSM data provides substantial added value for building footprint segmentation.

To verify that these improvements were due to the use of DSM data itself rather than differences in acquisition times between the DSM (2023) and nDSM (2021) (see for more information about the two different data layers in Section 4) datasets, an additional comparison experiment was conducted. As documented in the supplementary materials (Section 9), this analysis confirms that the observed performance differences stem from the DSM input rather than from variations in the RGB imagery or building footprint data.

Overlap The experiments confirmed that choosing the size of the overlap between adjacent tiles, has a notable impact on detection quality. As hypothesized (H1.5), introducing overlap between adjacent tiles substantially improved the segmentation of large roofs compared to non-overlapping tiling. Moreover, varying the amount of overlap during training and testing revealed that asymmetric settings can provide additional benefits (H1.6).

While the global optimum was reached with 50% training overlap and 80% testing overlap, this configuration proved computationally prohibitive. A more practical solution was found at 20% overlap in training and 10% in inference, which captured most of the performance gains of larger overlaps while keeping runtime manageable.

These improvements are consistent with the use of DSM input. Tiles centered on large, uniform roofs were sometimes missed, possibly because they contained too little height contrast. Increasing overlap raised the likelihood that a tile also included roof edges, where height transitions provided crucial cues for separating buildings from the background. Interesting for future research can be, how increased tile sizes could positively influence the model results. As bigger tiles have a higher chance of detecting roof

edges as well.

Decision Threshold A further refinement was introduced by optimizing the decision threshold for classifying building pixels. As hypothesized (H1.7), aggregating probabilities and applying a flexible threshold outperformed the default majority voting. A threshold of 0.3 improved IoU and F1 with a slightly lower recall. Although the improvements were modest, this adjustment provided finer control over the precision recall balance.

Regional comparisons illustrate how this mechanism plays out in practice. In heterogeneous urban areas such as applying a threshold of 0.3 improved IoU and F1 while recall was maintained or even slightly increased. In homogeneous both the baseline and optimized thresholds already achieved perfect accuracy, resulting in no measurable difference. In dense built-up areas, improvements were smaller but consistent, with recall benefitting slightly at the expense of a modest precision drop.

While the absolute gains over the majority vote baseline were small, they were systematic, computationally inexpensive, and particularly valuable in complex regions.

Future work could further explore the interaction between decision thresholds and tile overlap. It is assumed that threshold effects might become more pronounced with larger overlaps, although testing this systematically was beyond the computational resources available for this study. With greater computing power, such experiments could provide valuable insights and potentially yield further performance improvements.

Additional Training Data In contrast to the other experiments, the hypothesis on training data diversity (H1.8) was not supported. Adding targeted training regions did not improve generalization to the Berlin test set. This outcome suggests that the original NRW regions already provided sufficient diversity, and that the new regions either failed to contribute meaningful variability or did not possess the necessary quality to enhance model performance.

One likely explanation lies in the quality of the additional data. The Brandenburg imagery, for instance, often contains blurry roof edges, in contrast to the NRW dataset, which benefits from manual post-processing and a high overflight rate [NRW Geobasis, 2021]. As a result, the added tiles may have introduced more noise than useful signal. Furthermore, the new regions were relatively large, but only small parts of them actually contained structures of interest, such as vessels or extensive green roofs. Since the training tiles were sampled randomly in 20% increments, many of the added tiles may not have represented these challenging cases at all. As for example the additional training data from NRW, it contained a lot of industrial areas, which might have confused the model.

A more targeted strategy could therefore be more effective. Instead of sampling broadly from entire regions, future work should focus on selectively including tiles that

truly capture the specific structures or contexts where the model still struggles. Such a tailored approach may provide the intended additional aspects, that the model can learn, without diluting the training set with irrelevant examples.

Answer of the Main RQ and Hypothesis Taken together, the experiments provide partial confirmation of the main hypothesis (H1). Systematic optimization of preprocessing and post-processing substantially enhanced the accuracy of CNN-based building footprint detection, whereas the addition of new training data did not yield improvements. The final optimized configuration, combining Z-score normalization, DSM input, a moderate overlap of 20% during training and 10% during testing, and a decision threshold of 0.3, outperformed the baseline across nearly all evaluation metrics. This demonstrates that careful refinement of data preparation and model settings can significantly improve segmentation performance, even in heterogeneous environments, while simply increasing the size of the training dataset is not necessarily beneficial.

8 Case study

This case study illustrates how the improved CNN-Model for the detection of building footprints can be applied in the social sciences. The aim is not a full scale research project, but to demonstrate the potential of the Model and the building footprint data it can create. By linking detailed building geometries to demographic and building use data, it becomes possible to revisit research questions at an unprecedented spatial resolution.

Introduction and Literature

Over the past decade Germany has received the largest absolute number of migrants in Europe, making it a key case for studying how migration transforms societies [Federal Ministry of the Interior and Community, 2023]. Migrants predominantly settle in metropolitan areas, with Berlin standing out for the share of residents with a migration background. In 2022, 24.9% of Berlin's population were foreign citizens, around 972,000 people [Amt für Statistik Berlin-Brandenburg, 2022].

Research emphasizes that migration reshapes the urban landscape by altering land use, public space, economic activity, and social dynamics [Tawil et al., 2025]. For cities like Berlin, understanding these processes is crucial for evidence based urban planning and policymaking [Barbarino et al., 2021]. While housing pressures are a well-known consequence of immigration [International Organization for Migration, 2022], migration also reshapes local economies. Migrants frequently turn to self employment when facing exclusion from labor markets, while at the same time, ethnic communities create demand for goods and services not supplied by the local economy [Schmiz and Kitzmann,

2017, Portes and Manning, 2019]. These dynamics contribute to the formation of ethnic economies and concentrations of shops and workplaces in migrant dense neighborhoods.

Theoretical frameworks such as middleman minority theory highlight how immigrants historically occupied intermediary roles in trade and commerce, reinforcing the expectation that migrant presence correlates with commercial activity [Bonacich, 1973]. Yet despite extensive theorizing, empirical evidence at the urban scale remains scarce. For example, Olney [2013] shows for the U.S. that areas with rising immigration experienced growth in small shops. However, it remains unclear whether this also holds true for Berlin, as it has not been studied to date. The research question investigated in this case study is therefore:

- **RQ2: Is a higher migration share associated with the balance between housing and commercial functions in Berlin's building stock?**

The hypothesis is that, due to particular needs and exclusion from mainstream professions, migrant communities contribute to a relative increase in shops and workplaces, which is reflected in a lower housing share, compared to the amount of shops and working places, in their neighborhoods.

Data

Answering the research question of the case study requires data that link building footprints, building use, and demographic composition at a fine spatial scale.

The first dataset consists of CNN-derived building footprints generated with the final improved model described in Chapter 6.6. Using RGB orthophotos at 10 cm resolution and LiDAR-based DSM data (Table 7), the model was applied to the entire Berlin area through an inference pipeline, producing a shapefile of several hundred thousand building footprints. This dataset provides the base for the subsequent analysis, with each building footprint polygon, representing one unit of observation.

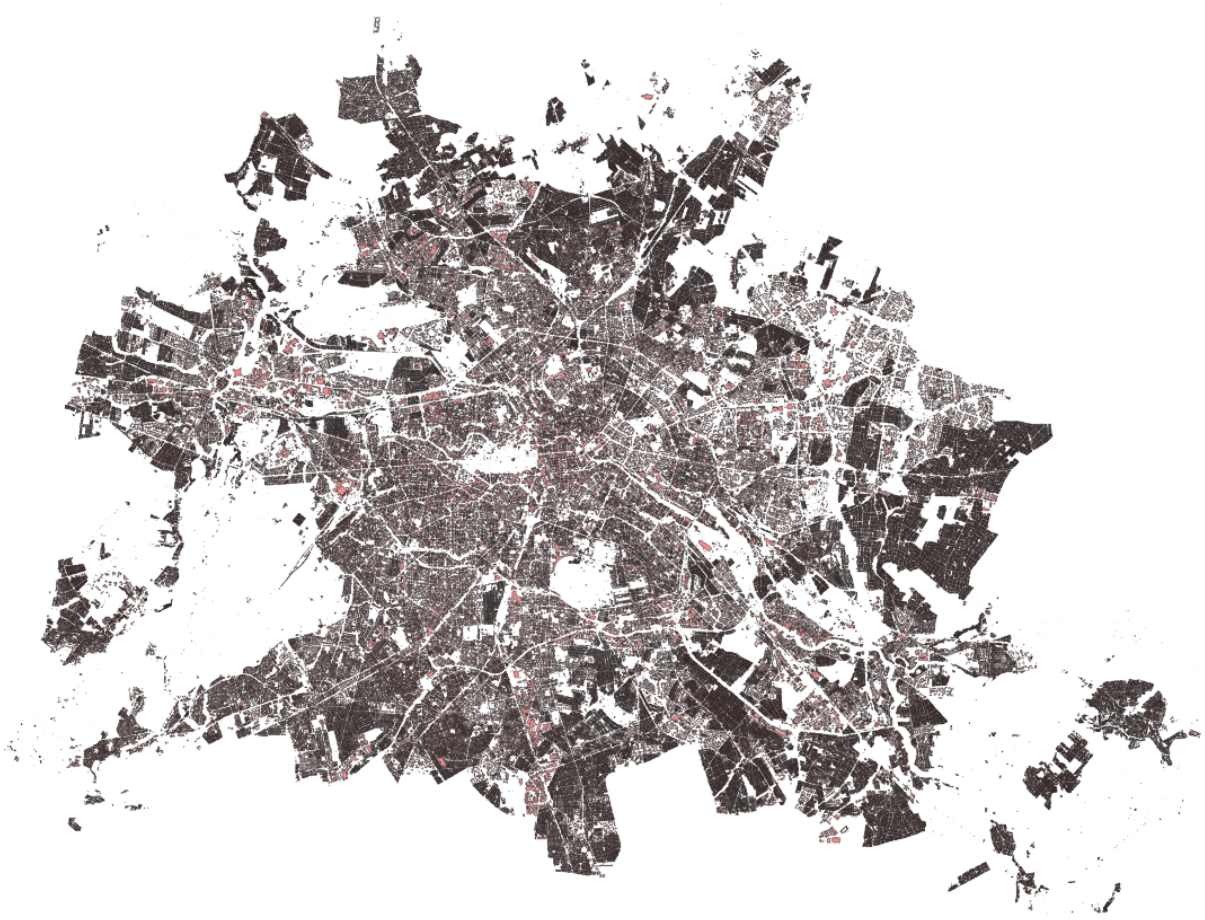


Figure 28: Building footprints of Berlin as extracted by the improved CNN-based model.

The second dataset is composed of LoD2 [Open Geo Data Berlin, 2025] and OSM [Boeing, 2025] building layers. Although less precise and partly incomplete (see Section 1), they provide functional information on building use. Both cadaster and OSM contain highly detailed building use classes (over 300 in cadaster and 215 in OSM), which were harmonized and mapped to a common set of aggregated categories. This enrichment procedure was developed by the DLR and made available for this work. CNN footprints were spatially joined with the harmonized dataset, and class percentages were assigned based on intersection areas. Footprints without a direct match were linked to OSM land use polygons, and in cases of multiple overlaps, percentages were normalized to sum to 1. Remaining unmatched footprints were assigned the class *Other*. For the analysis, only the classes *Housing* and *Shopping/Working* were retained, while footprints classified as completly part of the class "Other", were excluded. The two retained categories were renormalized to sum to 1. s By using percentage shares per building, the data capture that buildings are not necessarily used exclusively for housing or for shopping/working. This provides a more realistic representation of urban land use than a simple binary classification.

The third dataset provides demographic composition from the 2022 Zensus [Amt für

Statistik Berlin-Brandenburg, 2022], reporting the share of migrants in the total population. Zensus data is only available at the block level Amt für Statistik Berlin-Brandenburg [2016]. Figure 29 illustrates this integration, where census polygons (blue) provide demographic information and building footprints (red) supply geometry and functional attributes.

Figure 29: Integration of data sources: census polygons (blue) provide demographic information on migrant population, while building footprints (red) supply geometry and functional attributes from LoD2 and OSM. The merge enables building level analysis of how migrant share relates to building use.

To account for further context, three control variables were included. Each observation was assigned its district identifier to capture broad planning and regulatory differences across the city. Building polygon area size was added because housing complexes tend to consist of smaller buildings. Whereby smaller building complexes tend to only contain housing. Finally, the building to city center distance (defined as the Berliner Alexanderplatz), was included to reflect Berlin's monocentric structure, where central areas host a stronger mix of economic uses. .

Research Design and Methodology

The central hypothesis is that higher migrant shares will be linked to lower proportions of housing use within building footprints.

The dependent variable is therefore the proportion of each building footprint classified as Housing, measured as the share of footprint area. Because this outcome is a bounded continuous variable between 0 and 1, it is modeled using a *multilevel beta regression* rather than a linear model. This ensures that predictions remain between 0 and 1 and that the variance structure, which depends on the mean, is correctly represented [Heiss, 2021].

Buildings are nested within blocks and districts, and use patterns are known to cluster spatially. To capture this hierarchical structure and unobserved contextual influences, the model includes random intercepts at both block and district levels. This allows each district and block to have its own baseline housing share, absorbing unobserved heterogeneity and reducing the risk of omitted variable bias [Harrison et al., 2018].

(1) Distribution of housing share

$$y_{ijk} \sim \text{Beta}(\alpha_{ijk},\, \beta_{ijk}), \qquad \alpha_{ijk} = \mu_{ijk}\,\phi, \quad \beta_{ijk} = (1 - \mu_{ijk})\,\phi$$

(2) Linking predictors to the mean

$$\text{logit}(\mu_{ijk}) = \alpha + \beta_1\,\text{MigrantShare}_j + \beta_2\,\text{DistCenter}_{ijk} + \beta_3\,\text{Area}_{ijk} + u_k + v_j$$

(3) District and block variation

$$u_k \sim \mathcal{N}(0, \sigma^2_{\text{district}}), \qquad v_j \sim \mathcal{N}(0, \sigma^2_{\text{block}})$$

Here y_{ijk} is the proportion of housing use in building i in block j and district k, μ_{ijk} is its expected mean, and ϕ is the precision parameter, which governs how tightly the observed values are distributed around the mean. The mean is linked to the predictors through the logit function, ensuring that fitted values remain between 0 and 1. Random intercepts u_k and v_j are modeled as normally distributed with mean zero and variances $\sigma^2_{\text{district}}$ and σ^2_{block}, respectively. This reflects the standard hierarchical modeling approach, where most groups are expected to cluster around the overall mean with symmetric deviations, while allowing some districts and blocks to differ more strongly [Harrison et al., 2018].

All parameters, including the precision ϕ, were estimated in a Bayesian framework, with results summarized as posterior means and 95% credible intervals.

Results

Table 20 reports the fixed effects of the multilevel beta regression with random intercepts for districts and blocks. The dependent variable is the proportion of each building footprint classified as Housing.

Table 20: Multilevel beta regression for housing share.

Predictor	Mean	95% CrI
Intercept	0.731	[0.683, 0.775]
Migrant share (per +10 pp)	0.001	[-0.024, 0.026]
Distance to center (per km)	0.002	[-0.004, 0.004]
Building area (sqm)	0.015	[-0.030, 0.027]
District SD ($\sigma_{district}$)	0.023	[0.003, 0.059]
Block SD (σ_{block})	0.018	[0.001, 0.054]
Precision (ϕ)	0.215	[0.207, 0.223]

Notes: Coefficients are on the logit scale of the mean housing share.

The posterior means indicate that block-level migrant share, distance to the city center, and building area do not show credible associations with housing share, as all 95% credible intervals overlap zero.

The estimated standard deviations of the random intercepts are $\sigma_{\text{district}} = 0.023$ and $\sigma_{\text{block}} = 0.018$. These values indicate how much the baseline housing share (on the logit scale) typically varies across districts and blocks, after accounting for the fixed predictors. The variation is modest, meaning that most districts and blocks differ only slightly from the overall average housing share.

The estimated precision parameter is $\phi \approx 0.215$, which indicates a great dispersion around the predicted means. This is consistent with the observed distribution of housing shares, which includes many values close to the extremes of 0 or 1.

Overall, the analysis provides no evidence for the central hypothesis that higher migrant share is negatively associated with housing use. None of the predictors display a robust relationship with housing share, while contextual variation across districts and blocks remains present but modest in magnitude.

Summary and Discussion

This case study demonstrates how the CNN-Model and the building footprints it can create, can improve research in the social sciences.

The combination of building footprints, harmonized building use classes, and demographic information made it possible to test the hypothesis: that migrant presence is associated with a lower proportion of housing space relative to working places and shops.

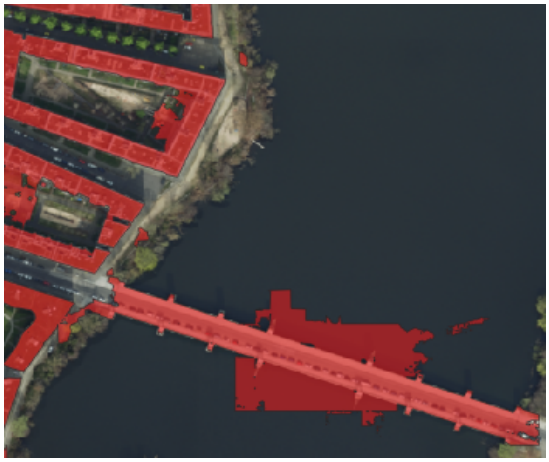
The empirical results, however, do not support this hypothesis. In the multilevel beta regression, block level migrant share was not credibly associated with housing share, and neither distance to the city center nor building size showed robust effects. Instead, most of the variation is explained by building level extremes in housing share (close to 0 or 1), with only modest baseline differences across districts and blocks. These findings highlight that, once hierarchical structure and data dispersion are properly accounted for, the expected negative relationship between migrant share and housing share does not emerge.

Besides that, several limitations should be acknowledged. First, the model tries to capture associations rather than causal effects, even if the results would have shown robust positive or negative results, they cannot determine whether migrant presence influences building use, if building use influences migrant presence, or whether both are shaped by unobserved third factors. Second, functional classifications from LoD2 and OSM, while harmonized, are not free of error and may bias category proportions. Third, the model uses a simplified specification, contrasting housing with all other use cases. Future work could explore multinomial regression with more detailed categories of building functions to better capture land-use mixes. Finally, the underlying data remain incomplete: OSM and LoD2 contain gaps and inconsistencies at the building level. Future releases of the Gebäudewirtschaftsregister (GWR) could offer a more powerful data source when combined with CNN-derived building footprints [Krause et al., 2022].

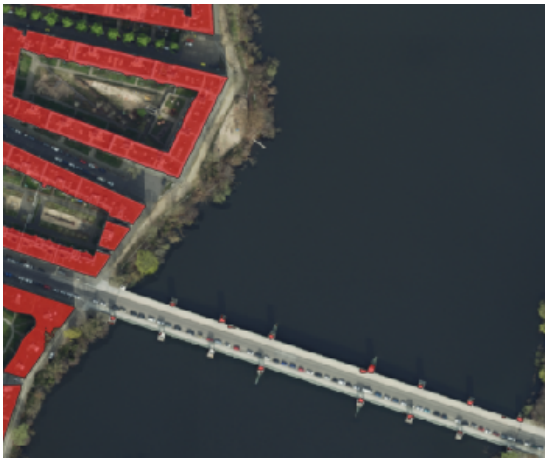
9 Conclusion

This thesis set out to evaluate whether improvements in pre-processing, post-processing, and training data selection can enhance the CNN-based building footprint detection model proposed by Stiller et al. [2023]. The results demonstrate that targeted workflow adjustments do indeed lead to improvements. In particular, the improvement of normalization strategies and the use of DSM instead of nDSM as the height layer data, had the strongest positive impact, underscoring that model performance depends not only on architecture but also on the design of the surrounding data pipeline.

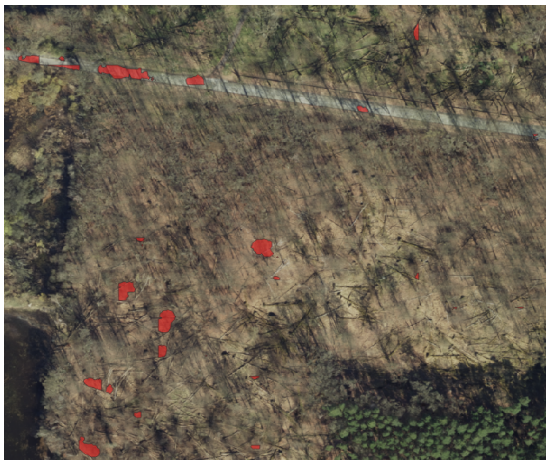
In addition to the quantitative improvements, visual inspection shows that the pipeline performs more robustly across heterogeneous landscapes. Figure 30 illustrates examples from bridges, forests and roads, and water bodies, where the improved workflow clearly reduced false positives and yields more accurate building outlines.



(a) Bridge (baseline)

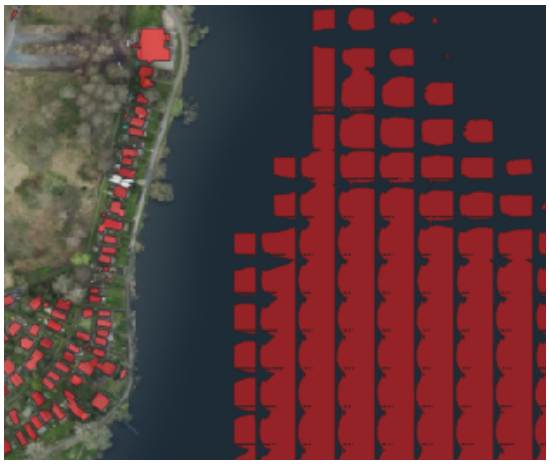


(b) Bridge (improved)



(c) Forests and roads (baseline)

(d) Forests and roads (improved)



(e) Water (baseline)

(f) Water (improved)

Figure 30: Examples of improvements across heterogeneous landscapes. In each case (bridges, forests/roads, and water), the improved pipeline reduces false positives and refines building footprint predictions.

For practical applications, these improvements address the central challenge outlined in the introduction: the lack of a reliable and up-to-date nationwide source of building

footprints in Germany. The improved CNN-based pipeline developed in this thesis provides a scalable alternative that produces building footprints directly from recent aerial imagery and LiDAR data.

By reducing false positives in heterogeneous landscapes and generating more consistent predictions, the approach brings automatic building detection closer to the level of completeness and precision required for official use. The model was able to deliver detailed, citywide estimates of building areas. While OSM and LoD2 data remain valuable for functional attributes and administrative purposes, the aerial, image based pipeline established here contributes the missing piece: an accurate, up to date footprint layer that can be integrated into future registers such as the planned GWR.

Beyond its technical contribution, this thesis also demonstrated the research value of the generated data through a case study in Berlin. By linking building footprints with demographic and building use information, the case study showed that the dataset can help to answer substantive social science questions. For example, how migrant presence interacts with the urban economic landscape. This underlines the twofold relevance of the improved model and the building footprints it can create: as a step toward closing critical data gaps in official statistics, and as a resource for applied research in urban and social sciences.

Nevertheless, several limitations remain. The most fundamental issue lies in the very definition of a building footprint. In the introduction, footprints were defined as the "two-dimensional (2D) visual representation of a building, describing its exact location, size, and shape on the ground" [Li et al., 2024b]. However, as this work has shown, aerial imagery cannot directly capture the building's base; it only records the extent of the roof as seen from above. In practice, the predictions produced by the model therefore correspond to "roofprints".

In regions such as Berlin or NRW, this distinction has little practical impact, as roof outlines usually align closely with the building's ground footprint. Yet in areas where roofs extend far beyond the walls, the roof-based representation may overestimate the true ground level footprint, highlighting a structural limitation of aerial image based extraction approaches. This could be solved through more sophisticated approaches, including a 3D Model based on LiDAR data as for example in Li et al. [2024a].

Another limitation lies in the simple implementation of the decision threshold. More advanced approaches, such as adaptive or context-aware thresholds [Wu et al., 2019], could further improve results. The interaction between thresholding, tile size, and overlap was also not systematically explored, although larger tiles or overlaps may capture more context at object boundaries [Bohao Huang et al., 2018]. Computational constraints prevented such tests, underscoring the need for stronger GPU resources or distributed computing.

The quality and selection of training data also posed challenges. Additional data

from Brandenburg was of lower quality than imagery from Berlin and NRW, which may explain why it failed to improve performance. Furthermore, tiles from the additional training data were choosen randomly, diluting the training signal with heterogeneous scenes. A more targeted sampling strategy focusing specifically on known failure modes such as large roofs or vessels would likely be more effective. This underlines that simply adding more data is not sufficient, the quality, representativeness, and sampling strategy are equally critical.

Finally, this thesis was limited to CNN-based architectures. Recent research suggests that transformer models can achieve higher accuracy and better generalization, particularly in capturing long range dependencies [Gibril et al., 2024]. The workflow developed here provides a solid foundation for systematic comparisons, and future work should directly contrast CNNs and transformers to more fully assess the state of the art in semantic building footprint extraction.

This thesis has shown that strengthening the robustness of CNN based extraction and demonstrating its value in applied research enhances the groundwork for future geospatial infrastructures, where up to date building footprints are no longer a bottleneck but a catalyst for innovation across science, policy, and society.

References

Rasha Alshehhi, Prashanth Reddy Marpu, Wei Lee Woon, and Mauro Dalla Mura. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:139–149, 8 2017. ISSN 09242716. doi: 10.1016/j.isprsjprs.2017.05.002.

Amt für Statistik Berlin-Brandenburg. Statistische Blöcke des RBS, 2016. URL `https://gdi.berlin.de/geonetwork/srv/eng/catalog.search#/metadata/ccfcd96d-31dd-3b23-b8a0-dc476a4c4f92`.

Amt für Statistik Berlin-Brandenburg. Zensus Daten Berlin, 2022. URL `https://www.statistik-berlin-brandenburg.de/zensus22/lokale-daten-berlin`.

Georgios Fotios Angelis, Armando Domi, Alexandros Zamichos, Maria Tsourma, Ioannis Manakos, Anastasios Drosou, and Dimitrios Tzovaras. A Comparative Study on Vision Transformers in Remote Sensing Building Extraction. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 3, pages 222–229. Science and Technology Publications, Lda, 2023. doi: 10.5220/0011787800003417.

Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of*

Photogrammetry and Remote Sensing, 140:20–32, 6 2018. ISSN 09242716. doi: 10.1016/j.isprsjprs.2017.11.011.

Mohamed Bakillah, Steve Liang, Amin Mobasheri, Jamal Jokar Arsanjani, and Alexander Zipf. Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, 28(9):1940–1963, 9 2014. ISSN 13623087. doi: 10.1080/13658816.2014.909045.

Robert Barbarino, Charlotte Räuchle, and Wolfgang Scholz. Migration-led institutional change in urban development and planning. *Urban Planning*, 6(2):1–6, 2021. ISSN 21837635. doi: 10.17645/up.v6i2.4356.

Berlin Geoportal. FIS Broker, 2025. URL `https://gdi.berlin.de/viewer/main/`.

Berlin Senatsverwaltung. Digitales Oberflächenmodell – DOM aus Airborne Laserscanning, 2021. URL `https://www.berlin.de/sen/sbw/stadtdaten/geoinformation/landesvermessung/geotopographie-atkis/dom-digitales-oberflaechenmodell/`.

Kushanav Bhuyan, Cees Van Westen, Jiong Wang, and Sansar Raj Meena. Mapping and characterising buildings for flood exposure analysis using open-source data and artificial intelligence. *Natural Hazards*, 119(2):805–835, 11 2023. ISSN 15730840. doi: 10.1007/s11069-022-05612-4.

Bing Maps. Global Building Footprints, 2023. URL `https://blogs.bing.com/maps/2023-06/Bing-Maps-Global-Building-Footprints-released`.

Benjamin Bischke, Patrick Helber, Joachim Folz, Andreas Dengel, and Damian Borth. Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks. Presented at the ICLR AI for social good workshop 2019, 2019. doi: 10.1109/ICIP.2019.8803050. URL `https://aiforsocialgood.github.io/iclr2019/accepted/track1/pdfs/48_aisg_iclr2019.pdf`.

Christopher M Bishop. *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

G Boeing. Modeling and Analyzing Urban Networks and Amenities with OSMnx. *Geographical Analysis*, 2025. doi: https://doi.org/10.1111/gean.70009.

Bohao Huang, Jordan M. Malof, Kyle Bradbury, Leslie M. Collins, and Daniel Reichman. Tiling and Stitching Segmentation Output for Remote Sensing: Basic Challenges and Recommendations. *arXiv*, 2018. URL `https://doi.org/10.48550/arXiv.1805.12219`.

Edna Bonacich. A Theory of Middleman Minorities. *American sociological review*, 38(5): 583–594, 1973.

Mourad Bouziani, Kalifa Goita, and Dong Chen He. Rule-based classification of a very high resolution image in an urban environment using multispectral segmentation guided by cartographic data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(8): 3198–3211, 8 2010. ISSN 01962892. doi: 10.1109/TGRS.2010.2044508.

Brandenburg Geobasis. 10cm DOP accessed 22.07.25, 2025. URL `https://geobroker.geobasis-bb.de/basiskarte.php?mode=startup&aProductId=b932fcd9-f1c7-4472-b473-767f0c440ef9`.

Brandenburg Geobroker. LiDAR data Brandenburg accessed 22.07.25, 2025. URL `https://geobroker.geobasis-bb.de/basiskarte.php?mode=startup&aProductId=d9895ec2-7039-4c0d-914c-a68f227a7069`.

Bundesamt für Kartographie und Geodäsie. Landcover data Berlin, 2025. URL `https://www.bkg.bund.de/SharedDocs/Produktinformationen/BKG/DE/P-2025/250528_LB-DE.html`.

Chandan Chawda, Aghav Jagannath, and Udar Swapnil. Extracting Building Footprints from Satellite Images using Convolutional Neural Networks. Bangalore, India, 2018. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). ISBN 9781538653142. doi: 10.1109/ICACCI.2018.8554893.

Keyan Chen, Zhengxia Zou, Zhenwei Shi, Citation : Chen, K ; Zou, and Z ; Shi. Building extraction from remote sensing images with sparse token transformers. *Remote Sensing*, 13(21):4441, 2021. doi: 10.3390/rs1010000. URL `https://doi.org/10.3390/rs1010000`.

Calimanut Ionut Cira, Miguel Ángel Manso-Callejo, Naoto Yokoya, Tudor Sălăgean, and Ana Cornelia Badea. Impact of Tile Size and Tile Overlap on the Prediction Performance of Convolutional Neural Networks Trained for Road Classification. *Remote Sensing*, 16(15), 8 2024. ISSN 20724292. doi: 10.3390/rs16152818.

Andrew Clark, Stuart Phinn, and Peter Scarth. Pre-Processing Training Data Improves Accuracy and Generalisability of Convolutional Neural Network Based Landscape Semantic Segmentation. *Land*, 12(7), 7 2023. ISSN 2073445X. doi: 10.3390/land12071268.

P. Dabove, M. Daud, and L. Olivotto. Revolutionizing urban mapping: deep learning and data fusion strategies for accurate building footprint segmentation. *Scientific Reports*, 14(1), 12 2024. ISSN 20452322. doi: 10.1038/s41598-024-64231-0.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. URL https://blog.waqasrana.me/assets/papers/rumelhart1986.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. Technical report, 2009. URL http://www.image-net.org.

W. Evans, D. Kirkpatrick, and G. Townsend. Right-triangulated irregular networks. *Algorithmica*, 30(2):264–286, 2001. ISSN 01784617. doi: 10.1007/s00453-001-0006-x.

Hongchao Fan, Alexander Zipf, Qing Fu, and Pascal Neis. Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4):700–719, 4 2014. ISSN 13658816. doi: 10.1080/13658816.2013.867495.

Federal Ministry of the Interior and Community. Migration Report of the Federal Government 2023. Technical report, 2023. URL https://www.bamf.de/SharedDocs/Anlagen/EN/Forschung/Migrationsberichte/migrationsbericht-2023-kurzfassung.pdf?__blob=publicationFile&v=9#:~:text=Germany%20remains%20the%20leading%20migration,with%20high%20im%02migration%20figures%20in.

Mohamed Barakat Gibril, Rami Al-Ruzouq, Abdallah Shanableh, Ratiranjan Jena, Jan Bolcek, Helmi Zulhaidi Mohd Shafri, and Omid Ghorbanzadeh. Transformer-based semantic segmentation for large-scale building footprint extraction from very-high resolution satellite images. *Advances in Space Research*, 73(10):4937–4954, 5 2024. ISSN 18791948. doi: 10.1016/j.asr.2024.03.002.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. Technical report, 2016.

Lisa Graaf and Sibyl Steuwer. Aufbau einer Datenbank über die Gesamtenergieeffizienz von Gebäuden in Deutschland. Technical report, 2025. URL https://www.bpie.eu/wp-content/uploads/2025/01/bpie-report-2501-250116-2_geschuetzt.pdf.

Xavier A. Harrison, Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N. Fisher, Cecily E.D. Goodwin, Beth S. Robinson, David J. Hodgson, and Richard Inger. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6(5):4794, 2018. ISSN 21678359. doi: 10.7717/peerj.4794.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. URL http://image-net.org/challenges/LSVRC/2015/.

Andrew Heiss. A guide to modeling proportions with Bayesian beta and zero-inflated beta regression models. *Political science*, 2021. URL https://www.andrewheiss.com/blog/2021/11/08/beta-regression-guide/.

Florian Hennig, Maren Köhlmann, Julius Weißmann, and Marianne Schepers. Erforschung von Satelliten-und weiteren Fernerkundungsdaten zur Ermittlung von Gebäudeangaben. *WISTA Wirtschaft und Statistik*, 77(4):70–82, 2025. URL https://openurl.ebsco.com/EPDB%3Agcd%3A11%3A36739707/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A187425079&crl=c&link_origin=scholar.google.de.

Benjamin Herfort, Sven Lautenbach, João Porto de Albuquerque, Jennings Anderson, and Alexander Zipf. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. *Nature Communications*, 14(1):3985, 12 2023. ISSN 20411723. doi: 10.1038/s41467-023-39698-6.

Moritz Hertrich. CNN-basierte semantische Segmentierung von Gebäudetypen mittels hochaufgelöster Luftbilder und normalisierten digitalen Oberflächenmodellen CNN-based semantic segmentation of building types using high-resolution aerial imagery and normalized digital surface models Masterarbeit. Technical report, Deutsches Zentrum für Luft- und Raumfahrt, 1 2024.

Moritz Hertrich, Dorothee Stiller, Marta Sapena, Thomas Stark, Patrick Rose, Michael Wurm, and Hannes Taubenbock. Instance Segmentation of Informal Buildings in Medellín for Assessing Population at Risk from Landslides. In *2025 Joint Urban Remote Sensing Event, JURSE 2025*, pages 1–4. Institute of Electrical and Electronics Engineers Inc., 2025. ISBN 9798350371833. doi: 10.1109/JURSE60372.2025.11076048.

Vladimir Iglovikov and Alexey Shvets. TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. *arXiv*, 1 2018. URL http://arxiv.org/abs/1801.05746.

International Organization for Migration. *Integrating Migration into Urban Development Interventions: A Toolkit for International Cooperation and Development Actors*. IOM, Brussels, 2022. ISBN 978-92-9268-244-6.

Eui ik Jeon, Sunghak Kim, Soyoung Park, Juwon Kwak, and Imho Choi. Semantic segmentation of seagrass habitat from drone imagery based on deep learning: A comparative study. *Ecological Informatics*, 66:101430, 12 2021. ISSN 15749541. doi: 10.1016/j.ecoinf.2021.101430.

H. James Deva Koresh. Impact of the Preprocessing Steps in Deep Learning-Based Image Classifications. *National Academy Science Letters*, 47(6):645–647, 12 2024. ISSN 22501754. doi: 10.1007/s40009-023-01372-2.

Anja Krause, Markus Zimmermann, and Ingmar Herda. Überlegungen zu einem Gebäude- und Wohnungsregister: Aufbau, Pflege und Nutzung. *WISTA-Wirtschaft und Statistik*, 74(4):25–38, 2022. URL `https://www.econstor.eu/bitstream/10419/263205/1/wista-2022-4-025-038.pdf`.

Yann Lecun, L Eon Bottou, Yoshua Bengio, and Patrick Haaner Abstract—. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 11 1998. doi: 10.1109/5.726791.

Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 5 2015. ISSN 14764687. doi: 10.1038/nature14539.

Kai Li, Yupeng Deng, Yunlong Kong, Diyou Liu, Jingbo Chen, Yu Meng, Junxian Ma, and Chenhao Wang. Prompt-Driven Building Footprint Extraction in Aerial Images with Offset-Building Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 10 2024a. doi: 10.1109/TGRS.2024.3487652. URL `http://arxiv.org/abs/2310.16717http://dx.doi.org/10.1109/TGRS.2024.3487652`.

Qingyu Li, Yilei Shi, Stefan Auer, Robert Roschlaub, Karin Möst, Michael Schmitt, Clemens Glock, and Xiao Xiang Zhu. Detection of undocumented building constructions from official geodata using a convolutional neural network. *Remote Sensing*, 12 (21):1–21, 11 2020. ISSN 20724292. doi: 10.3390/rs12213537.

Qingyu Li, Lichao Mou, Yao Sun, Yuansheng Hua, Yilei Shi, and Xiao Xiang Zhu. A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 3 2024b. ISSN 15580644. doi: 10.1109/TGRS.2024.3369723.

Weijia Li, Conghui He, Jiarui Fang, Juepeng Zheng, Haohuan Fu, and Le Yu. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing*, 11(4), 2 2019. ISSN 20724292. doi: 10.3390/rs11040403.

Ziming Li, Qinchuan Xin, Ying Sun, and Mengying Cao. A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery. *Remote Sensing*, 13(18):3630, 9 2021. ISSN 20724292. doi: 10.3390/rs13183630.

Thomas Lillesand, Ralph Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2015.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125. Cornell University and Cornell Tech, 7 2017.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path Aggregation Network for Instance Segmentation. *CVPR*, 2018.

Wei Liu, Meng Yuan Yang, Meng Xie, Zihui Guo, Er Zhu Li, Lianpeng Zhang, Tao Pei, and Dong Wang. Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network. *Remote Sensing*, 11(24), 12 2019. ISSN 20724292. doi: 10.3390/rs11242912.

Marc-O. Löwner, Joachim Benner, Gerhard Gröger, Ulrich Gruber, Karl-Heinz Häfele, and Sandra Schlüter. CityGML 2.0 Ein internationaler Standard für 3D-Stadtmodelle. *ZfV-Zeitschrift für Geodäsie, Geoinformation und Landmanagement*, 6 2012.

Lin Luo, Pengpeng Li, and Xuesong Yan. Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies*, 14(23):7982, 12 2021. ISSN 19961073. doi: 10.3390/en14237982.

Wolfgang Maenning and Tobias Just. *Understanding German Real Estate Market*. Springer, 2012.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. High-Resolution Aerial Image Labeling with Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7092–7103, 12 2017. ISSN 01962892. doi: 10.1109/TGRS.2017.2740362.

Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation Studies in Artificial Neural Networks. *arXiv*, 2 2019. URL `http://arxiv.org/abs/1901.08644`.

Nikola Milojevic-Dupont, Felix Wagner, Florian Nachtigall, Jiawei Hu, Geza Boi Brüser, Marius Zumwald, Filip Biljecki, Niko Heeren, Lynn H. Kaack, Peter Paul Pichler, and Felix Creutzig. EUBUCCO v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings. *Scientific Data*, 10(1):147, 12 2023. ISSN 20524463. doi: 10.1038/s41597-023-02040-2.

NRW Geobasis. Digital Orthophotos (DOP) 10cm resolution, 2019. URL `https://www.bezreg-koeln.nrw.de/geobasis-nrw/produkte-und-dienste/luftbild-und-satellitenbildinformationen/aktuelle-luftbild-und-0`.

NRW Geobasis. Nutzerinformation für Digitale Orthophotos in der Qualitätsstufe TrueDOP, 2021. URL https://www.bezreg-koeln.nrw.de/system/files/media/document/file/geobasis_nutzerinfo_true_dop.pdf.

NRW Geobasis. From 2023, 10cm DOP Data accessed 15.06.25, 2025a. URL https://www.bezreg-koeln.nrw.de/geobasis-nrw/produkte-und-dienste/luftbild-und-satellitenbildinformationen/aktuelle-luftbild-und-0.

NRW Geobasis. Laserscandaten Data accessed 15.07.25, 2025b. URL https://www.opengeodata.nrw.de/produkte/geobasis/hm/3dm_l_las/.

NRW Geobasis. nDSM Data NRW, 2025c. URL https://advmis.geodatenzentrum.de/trefferanzeige?docuuid=31a3b5a0-7264-4a85-a69e-580fe339f908.

William W. Olney. Immigration and firm expansion. *Journal of Regional Science*, 53(1): 142–157, 2 2013. ISSN 00224146. doi: 10.1111/jors.12004.

Open Geo Data Berlin. LOD2 Data Berlin, 2025. URL https://gdi.berlin.de/geonetwork/srv/eng/catalog.search#/metadata/8a7ea996-7955-4fbb-8980-7be09be6f193.

Open Geodata NRW. LOD2 Data NRW, 2025. URL https://www.opengeodata.nrw.de/produkte/.

Claudio Persello and Lorenzo Bruzzone. Active and semisupervised learning for the classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):6937–6956, 2014. ISSN 01962892. doi: 10.1109/TGRS.2014.2305805.

Alejandro Portes and Robert D Manning. The Immigrant Enclave: Theory and Empirical Examples. In *Social Stratification, Class, Race, and Gender in Sociological Perspective, Second Edition*, pages 568–579. 2019. URL https://pages.nyu.edu/jackson/analysis.of.inequality/Readings/Portes%20-%20Immigrant%20Enclave-Theory%20and%20Empirical%20Examples%20-%2086.pdf.

Simon J D Prince. *Understanding Deep Learning*. MIT press, 2025 edition, 2023. URL http://udlbook.com.

G. Anthony Reina, Ravi Panchumarthy, Siddhesh Pravin Thakur, Alexei Bastidas, and Spyridon Bakas. Systematic Evaluation of Image Tiling Adverse Effects on Deep Learning Semantic Segmentation. *Frontiers in Neuroscience*, 14, 2 2020. ISSN 1662453X. doi: 10.3389/fnins.2020.00065.

F Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

URL https://me8xh5ls2s.scholar.serialssolutions.com/?sid=google&auinit=F&aulast=Rosenblatt&atitle=The+perceptron:+a+probabilistic+model+for+information+storage+and+organization+in+the+brain.&id=doi:10.1037/h0042519&title=Psychological+review&volume=65&issue=6&date=1958&spage=386&issn=0033-295X.

N. Arockia Rosy, K. Balasubadra, and K. Deepa. Are vision transformers replacing convolutional neural networks in scene interpretation?: A review. *Discover Applied Sciences*, 7(9):1–21, 9 2025. ISSN 30049261. doi: 10.1007/s42452-025-07574-1.

Holger R. Roth, Le Lu, Nathan Lay, Adam P. Harrison, Amal Farag, Andrew Sohn, and Ronald M. Summers. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Medical Image Analysis*, 45:94–107, 4 2018. ISSN 13618423. doi: 10.1016/j.media.2018.01.006.

Ryuhei Hamaguchi, Fujita Keisuke, Nemoto Tomoyuki, and Imaizumi Shuhei Hikosaka. Effective Use of Dilated Convolutions for Segmenting Small Object Instances in }Remote Sensing Imagery. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1442–1450, 3 2018. doi: 10.1109/WACV.2018.00162. URL https://arxiv.org/pdf/1709.00179.

Shunta Saito, Takayoshi Yamashita, and Yoshimitsu Aoki. Multiple object extraction from aerial imagery with convolutional neural networks. *Journal of Imaging Science and Technology*, 60(1):1–9, 1 2016. ISSN 19433522. doi: 10.2352/J.ImagingSci.Technol.2016.60.1.010402.

Muntaha Sakeena, Eric Stumpe, Miroslav Despotovic, David Koch, and Matthias Zeppelzauer. On the Robustness and Generalization Ability of Building Footprint Extraction on the Example of SegNet and Mask R-CNN. *Remote Sensing*, 15(8):2135, 4 2023. ISSN 20724292. doi: 10.3390/rs15082135.

Georg Schiller, Andreas Blum, Robert Hecht, Holger Oertel, Uwe Ferber, and Gotthard Meinel. Urban infill development potential in Germany: comparing survey and GIS data. *Buildings and Cities*, 2(1):36–54, 2021. ISSN 26326655. doi: 10.5334/bc.69.

Antonie Schmiz and Robert Kitzmann. Negotiating an Asiatown in Berlin: Ethnic diversity in urban planning. *Cities*, 70:1–10, 10 2017. ISSN 02642751. doi: 10.1016/j.cities.2017.06.001.

Jamie Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv*, 6 2016. URL http://arxiv.org/abs/1606.02585.

Jaswinder Singh and Rajdeep Banerjee. A Study on Single and Multi-layer Perceptron Neural Network. In *3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 35–40, 2019. ISBN 9781538678084.

Dorothee Stiller, Pablo D'angelo, Hannes Taubenböck, Michael Wurm, Karsten Stebner, Thomas Stark, and Stefan Dech. Spatial parameters for transportation. *Journal of Transport and Land Use*, 14(1):777–803, 2021. doi: 10.2307/48646209. URL `https://www.jstor.org/stable/10.2307/48646209`.

Dorothee Stiller, Thomas Stark, Verena Strobl, Maike Leupold, Michael Wurm, and Hannes Taubenböck. Efficiency of CNNs for Building Extraction: Comparative Analysis of Performance and Time. *Joint Urban Remote Sensing Event (JURSE)*, 2022:1–4, 6 2023. doi: 10.1109/JURSE57346.2023.10144140. URL `https://ieeexplore.ieee.org/abstract/document/10144140`.

Maram Tawil, Christa Reicher, Eva Krings, Fabio Bayro Kaiser, Motez Amayreh, and Qais Ismail. Urban Contestation in Migrants' Settings: Towards More Resilience Through Fluid Planning in Aachen, Germany. *Urban Science*, 9(9):346, 8 2025. ISSN 2413-8851. doi: 10.3390/urbansci9090346. URL `https://www.mdpi.com/2413-8851/9/9/346`.

Samir Touzani and Jessica Granderson. Open data and deep semantic segmentation for automated extraction of building footprints. *Remote Sensing*, 13(13):2578, 7 2021. ISSN 20724292. doi: 10.3390/rs13132578.

Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, 6 2016. ISSN 21686831. doi: 10.1109/MGRS.2016.2548504.

Hoai Van Nguyen, Thanh Cong Luu, and Quang Dung Pham. Solving the TimeTabling problem at FPT university. In *ACM International Conference Proceeding Series*, volume 03-04-December-2015, pages 98–104. Association for Computing Machinery, 12 2015. ISBN 9781450338431. doi: 10.1145/2833258.2833311.

Michele Volpi and Devis Tuia. Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, pages 881–893, 10 2016. doi: 10.1109/TGRS.2016.2616585. URL `http://arxiv.org/abs/1608.00775http://dx.doi.org/10.1109/TGRS.2016.2616585`.

N. A. Wardrop, W. C. Jochem, T. J. Bird, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, and A. J. Tatem. Spatially disaggregated population estimates in the absence of national population and housing census data. *Pro-*

ceedings of the National Academy of Sciences of the United States of America, 115(14): 3529–3537, 4 2018. ISSN 10916490. doi: 10.1073/pnas.1715305115.

Patrick Weber and Mordechai Haklay. openstreetMap: User-Generated street Maps. *IEEE Pervasive computing*, 7(4):12–18, 2008. URL `www.openstreetmap.`

Shiqing Wei, Shunping Ji, and Meng Lu. Toward Automatic Building Footprint Delineation from Aerial Images Using CNN and Regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):2178–2189, 3 2020. ISSN 15580644. doi: 10.1109/TGRS.2019.2954461.

Yu chen Wu and Jun wen Feng. Development and Application of Artificial Neural Network. *Wireless Personal Communications*, 102(2):1645–1656, 9 2018. ISSN 1572834X. doi: 10.1007/s11277-017-5224-x.

Zhihuan Wu, Yongming Gao, Lei Li, Junshi Xue, and Yuntao Li. Semantic segmentation of high-resolution remote sensing images using fully convolutional network with adaptive threshold. *Connection Science*, 31(2):169–184, 4 2019. ISSN 13600494. doi: 10.1080/09540091.2018.1510902.

Supplementary Materials

LiDAR Classification tables

DSM LiDAR Classification tables

Table 21: LiDAR Classification Values Used in NRW and Their Use in DSM Creation

Class	Definition	Used in DSM
1	Unclassified points. These are not assigned to a specific category and may include reflections or vegetation.	No
2	Ground points. These describe the natural terrain, excluding buildings, vegetation, and water. Relevant for DGM.	Yes
9	Synthetic water points. Interpolated water surface points located under bridges.	No
17	Bridge points. Represent the surface of bridges such as roadways, but not piers or supports.	Yes
18	Noise. Points caused by measurement errors, such as birds, fog, or clouds.	No
20	Last return not ground. Points from the last laser return not classifiable into other specific categories, such as cars or rooftops.	Yes
21	Synthetic building points. Interpolated points under large buildings. Only used until 2019.	No
24	Basement points. Located in basement shafts or below natural terrain, such as light wells.	Yes
26	Synthetic ground points. Interpolated points under bridges or dense vegetation.	Yes

Table 22: LiDAR Point Classification Scheme and Their Use in DSM Creation (Brandenburg)

Class	**Definition**	**Used in DSM**
0	Created, never classified	Yes
1	Unclassified	Yes
2	Ground	Yes
20	Building	Yes

Table 23: LiDAR Classification Values Used in Berlin and Their Use in DSM Creation

Class	**Definition**	**Used in DSM**
0	Created, never classified	No
2	Ground	Yes
3	Low vegetation. Typically vegetation under 0.5 meters.	Yes
4	Medium vegetation. Vegetation between 0.5 and 2 meters in height.	Yes
5	High vegetation. Vegetation taller than 2 meters.	Yes
7	Low point. Often erroneous or noise points located below the surface.	No

DTM LiDAR Classification tables

Table 24: LiDAR Classification Values Used in NRW and Their Use in DSM Creation

Class	Definition	Used in DTM
1	Unclassified points. These are not assigned to a specific category and may include reflections or vegetation.	No
2	Ground points. These describe the natural terrain, excluding buildings, vegetation, and water. Relevant for DGM.	Yes
9	Synthetic water points. Interpolated water surface points located under bridges.	No
17	Bridge points. Represent the surface of bridges such as roadways, but not piers or supports.	No
18	Noise. Points caused by measurement errors, such as birds, fog, or clouds.	No
20	Last return not ground. Points from the last laser return not classifiable into other specific categories, such as cars or rooftops.	No
21	Synthetic building points. Interpolated points under large buildings.	No
24	Basement points. Located in basement shafts or below natural terrain, such as light wells.	No
26	Synthetic ground points. Interpolated points under bridges or dense vegetation.	No

New nDSM Data

To ensure that the performance gains observed with DSM input were not merely due to the use of newer RGB imagery or updated ground truth labels, a new nDSM dataset was created using exactly the same LiDAR source data as for the DSM experiment. The DSM and DTM inputs for this nDSM were produced with the same workflow described in Section 4.2.1 for DSM generation and Section 9 for DTM generation. The normalized Digital Surface Model was then computed by subtracting the DTM from the DSM following the method in Section 4.2.3 and, and finally resampled to 0.1,m resolution to match the RGB imagery. This ensures that the nDSM data for the NRW data was created in

the same way as the test data in Berlin.

This controlled setup ensured that the only variable that was different between the DSM and new nDSM experiments was the type of height representation, while the RGB inputs, acquisition years, and label updates remained identical.

The results in Table 25 demonstrate that the new nDSM configuration does not match the performance of the DSM based model. While recall is slightly higher (0.957, the best of all settings), other key metrics decline. IoU drops to 0.817, F1 score to 0.893, placing this setup below both the DSM and even the original baseline nDSM. These outcomes confirm that the improvements observed with the DSM input stem from the richer absolute elevation information it provides, rather than from newer RGB imagery or updated ground truth labels.

Trial	**Parameter**	**Overall Accuracy**	**IoU**	**F1 Score**	**Precision**	**Recall**	**Ranking**
	Baseline	0.978	0.840	0.911	0.875	**0**.952	3
	Ablation	0.957	0.740	0.886	0.843	0.933	4
nDSM	DSM Data	**0.979**	**0.842**	**0.912**	**0.913**	0.912	1
	new nDSM	0.973	0.817	0.893	0.845	**0.957**	2

Table 25: Comparison of the baseline and alternative height inputs, including the old and new DSM data instead of nDSM.

Intersection Test Areas

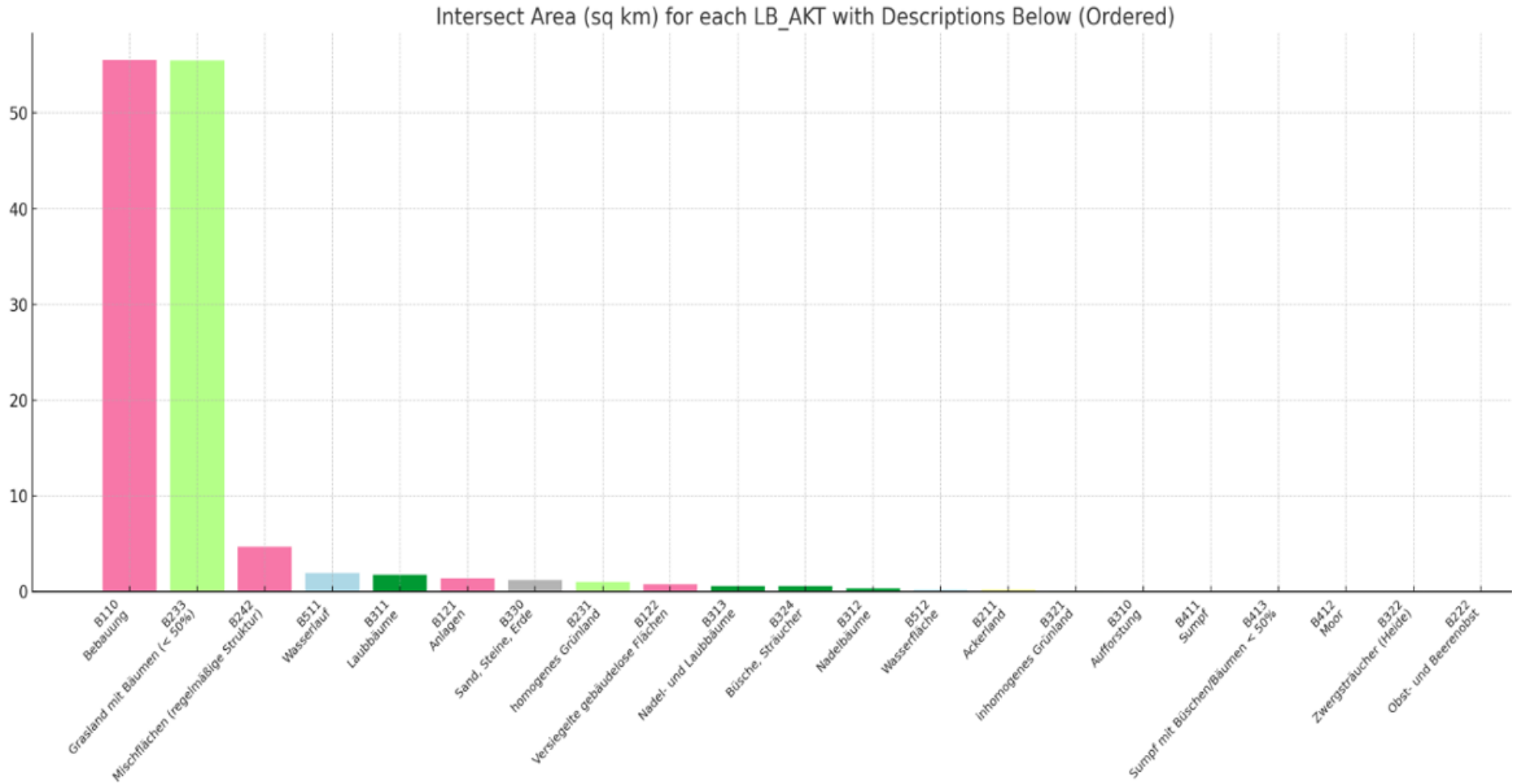


Figure 31: Intersection areas between predicted building footprints and land cover classes for Berlin. Y-Axis shows the amount of intersecting square kilometers.

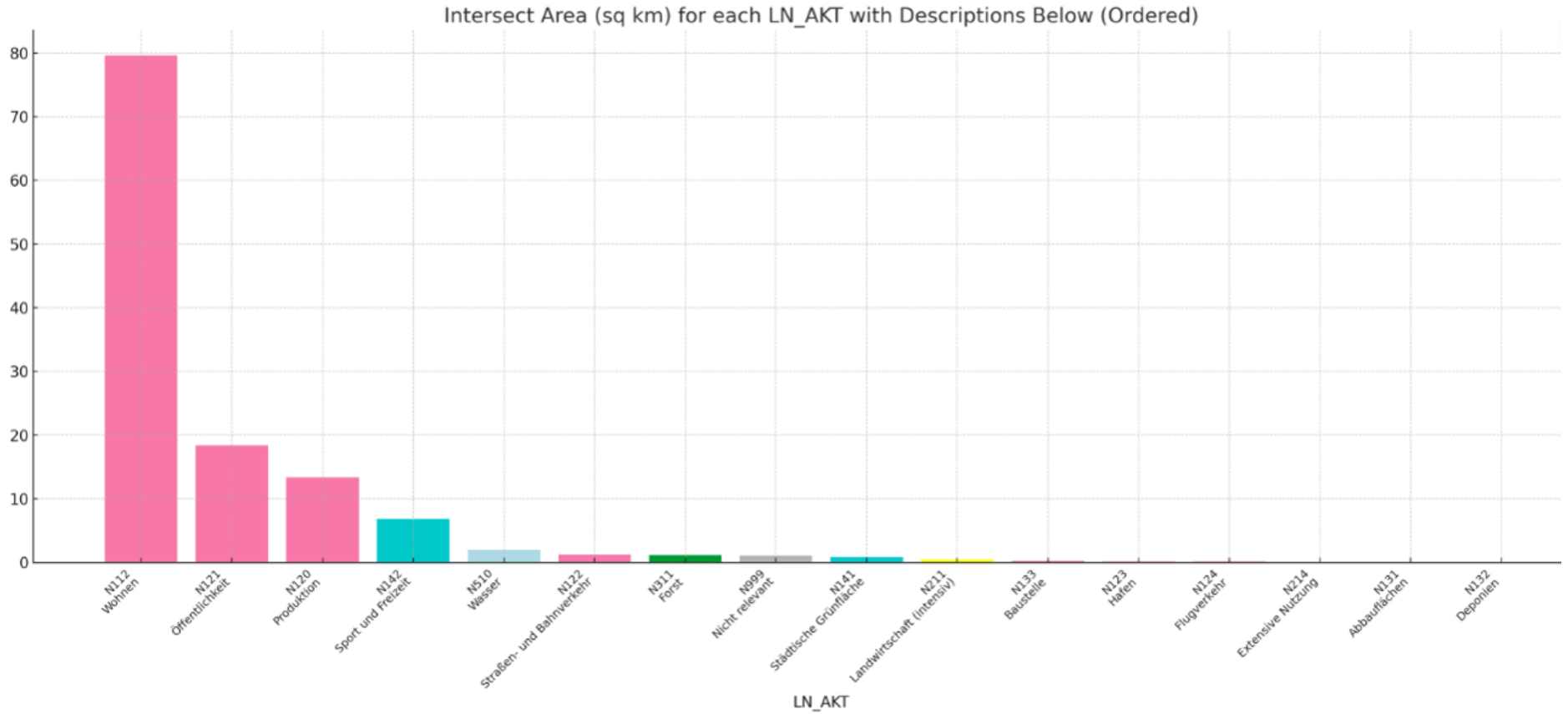


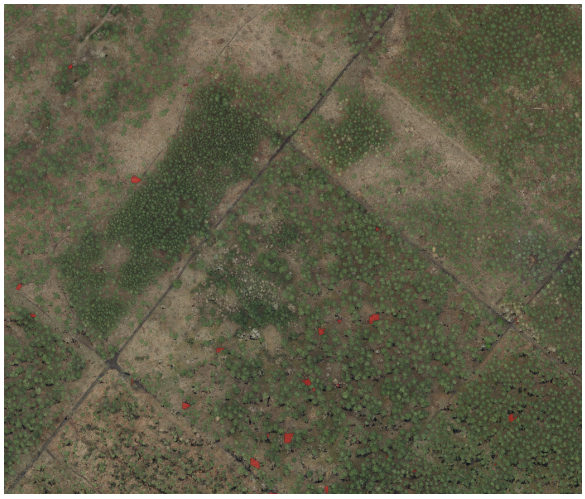
Figure 32: Intersection areas between predicted building footprints and land use classes for Berlin. Y-Axis shows the amount of intersecting square kilometers.

Test Images with Intersection

(a) Original – Bridges and boats

(b) Intersection – Bridges and boats

(a) Original – Forest (B311)



(b) Intersection – Forest (B311)

(a) Original – Housing area (N112)

(b) Intersection – Housing area (N112)

(a) Original – Industrial area (N120)

(b) Intersection – Industrial area (N120)

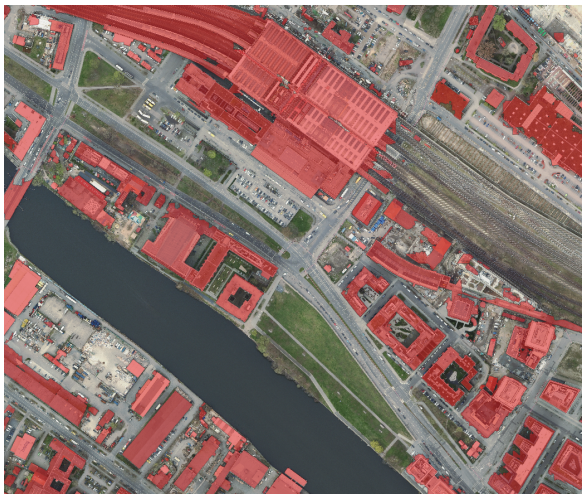
(a) Original – Inner city (B110)

(b) Intersection – Inner city (B110)

(a) Original – Street through forest

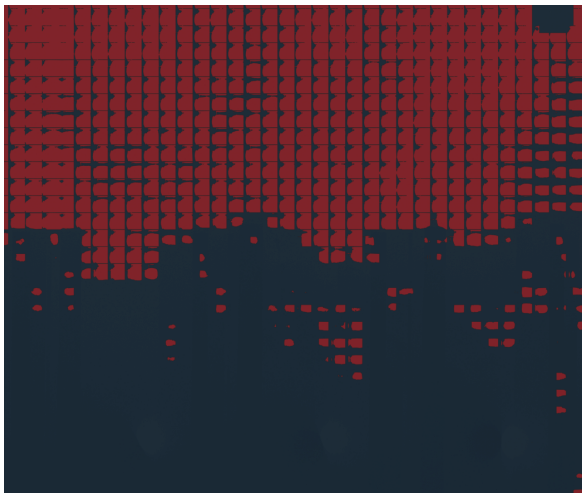
(b) Intersection – Street through forest

(a) Original – Streets and railway (N122)

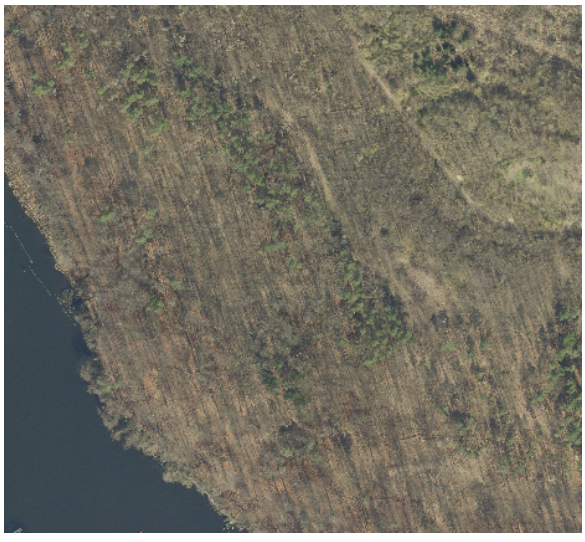


(b) Intersection – Streets and railway (N122)

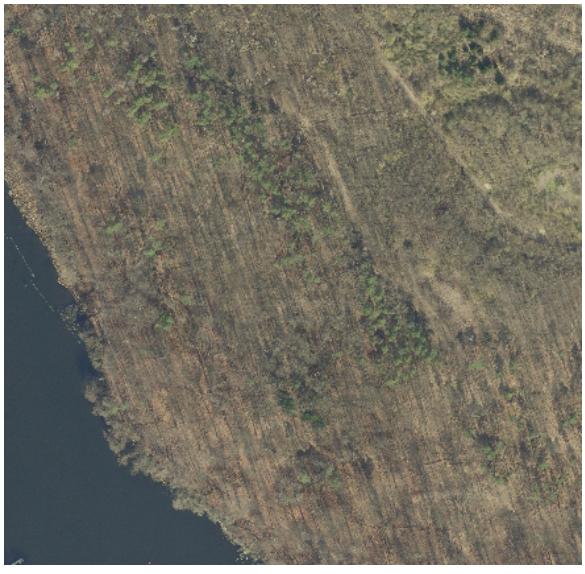
(a) Original – Water (B511)



(b) Intersection – Water (B511)

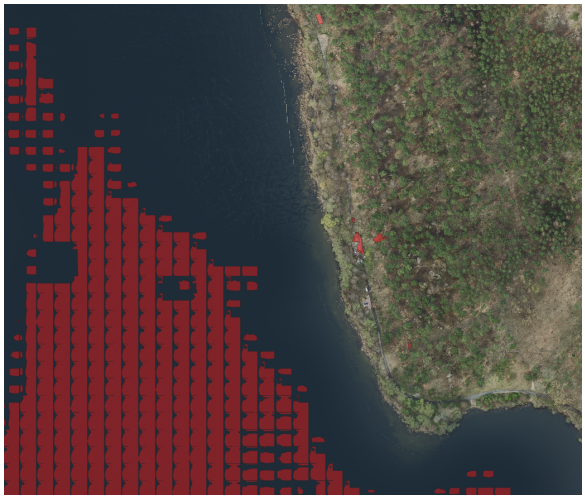


(a) Original – Water on edge



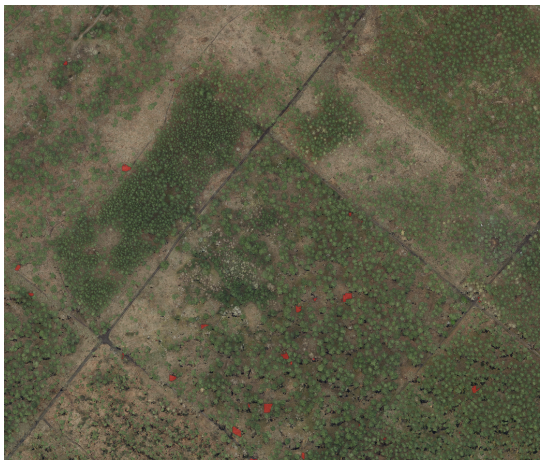
(b) Intersection – Water on edge

(a) Original – Water riverbank



(b) Intersection – Water riverbank

Normalization Experiments: Comparison Baseline and best results



(a) Baseline — Forest

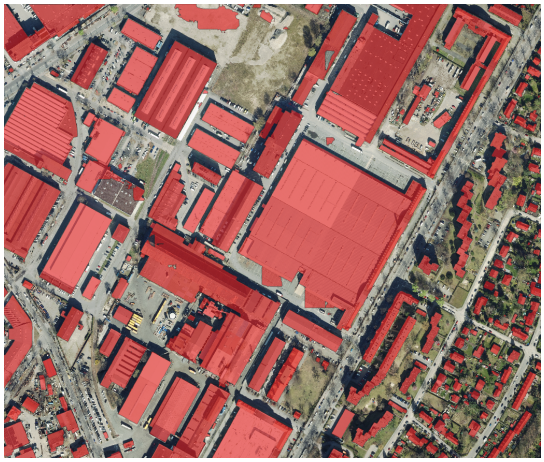
(b) Z-score — Forest

Figure 43: Forest: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.996	1.000	+0.004
False Positive Rate	0.004	0	-0.004

Table 26: Region: Forest

(a) Baseline — Industrial area



(b) Z-score — Industrial area

Figure 44: Industrial area: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.968	0.936	-0.032
IoU	0.927	0.852	-0.075
F1 Score	0.962	0.920	-0.042
Precision	0.948	0.960	+0.012
Recall	0.977	0.883	-0.094
False Positive Rate	0.022	0.015	-0.007

Table 27: Region: Industrial area

(a) Baseline — Inner city

(b) Z-score — Inner city

Figure 45: Inner city: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.964	0.952	-0.012
IoU	0.924	0.898	-0.025
F1 Score	0.960	0.946	-0.014
Precision	0.939	0.933	-0.006
Recall	0.983	0.961	-0.022
False Positive Rate	0.028	0.031	+0.002

Table 28: Region: Inner city

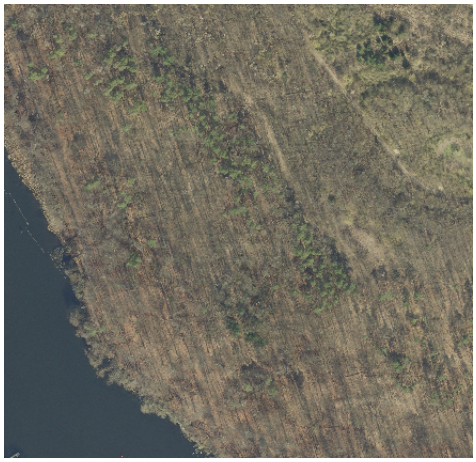
(a) Baseline — Small street through forest

(b) Z-score — Small street through forest

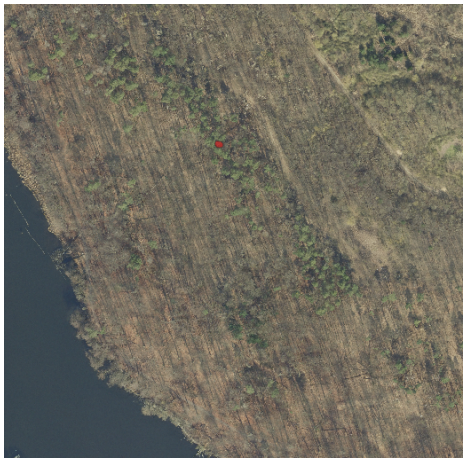
Figure 46: Small street through forest: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.994	0.999	+0.006
False Positive Rate	0.006	0.001	-0.006

Table 29: Region: Small street through forest



(a) Baseline — Water at mosaic edge

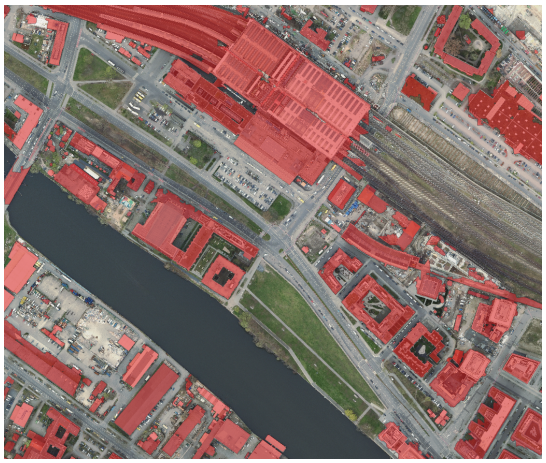


(b) Z-score — Water at mosaic edge

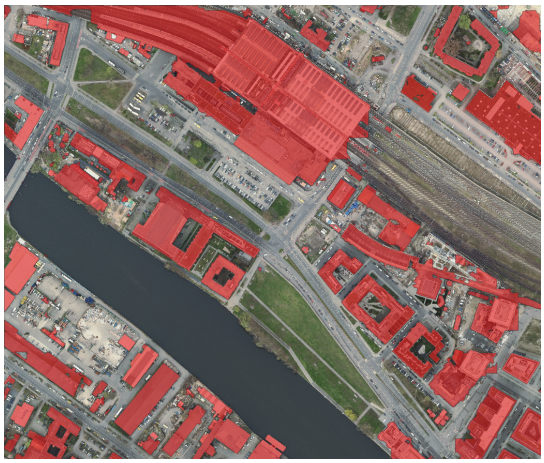
Figure 47: Water at mosaic edge: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	1.000	1.000	0
False Positive Rate	0	0	0

Table 30: Region: Water at mosaic edge



(a) Baseline — Streets & railway

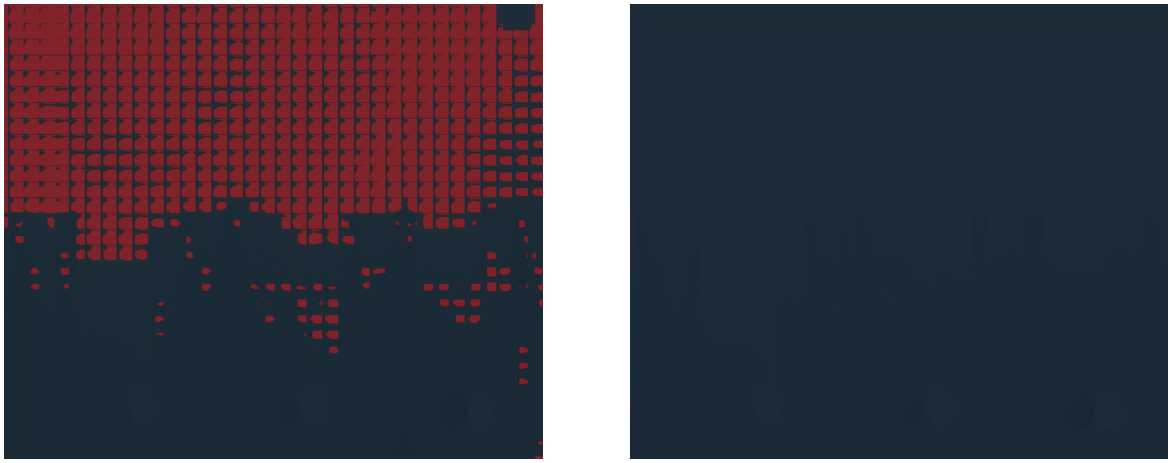


(b) Z-score — Streets & railway

Figure 48: Streets & railway: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.922	0.938	+0.016
IoU	0.753	0.794	+0.041
F1 Score	0.859	0.885	+0.026
Precision	0.786	0.827	+0.041
Recall	0.947	0.953	+0.006
False Positive Rate	0.065	0.050	-0.015

Table 31: Region: Streets & railway



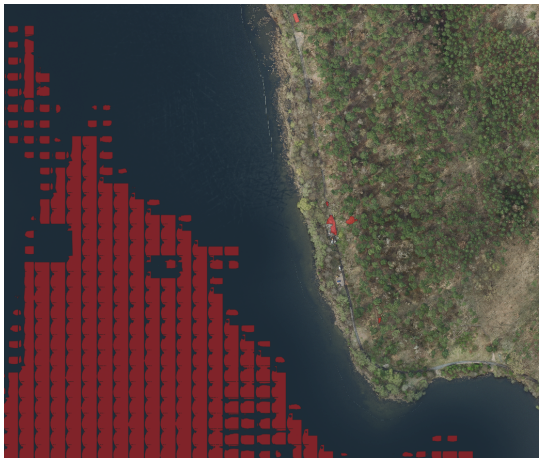
(a) Baseline — Water

(b) Z-score — Water

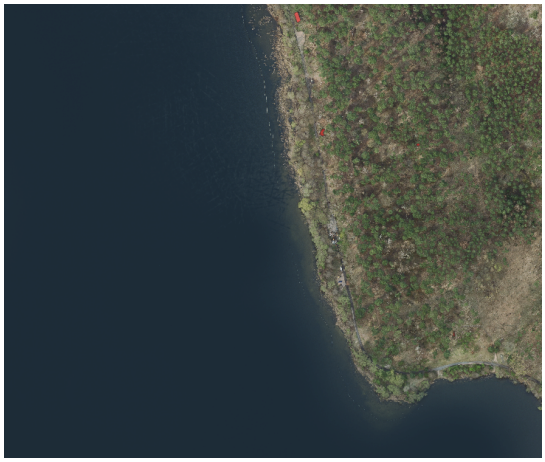
Figure 49: Water: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.628	1.000	+0.372
False Positive Rate	0.372	0	-0.372

Table 32: Region: Water



(a) Baseline — Water riverbank



(b) Z-score — Water riverbank

Figure 50: Water riverbank: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.771	1.000	+0.229
False Positive Rate	0.229	0	-0.229

Table 33: Region: Water riverbank

(a) Baseline — Housing area

(b) Z-score — Housing area

Figure 51: Housing area: baseline vs. Z-score normalized prediction.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.972	0.973	+0.001
IoU	0.853	0.858	+0.005
F1 Score	0.921	0.924	+0.003
Precision	0.913	0.916	+0.003
Recall	0.928	0.931	+0.003
False Positive Rate	0.016	0.015	-0.001

Table 34: Region: Housing area

DSM Experiments: Comparison Baseline and best results

(a) Baseline — Bridges & boats

(b) DSM — Bridges & boats

Figure 52: Bridges & boats: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	0.946	0.945	-0.001
IoU	0.700	0.671	-0.030
F1 Score	0.824	0.803	-0.021
Precision	0.748	0.787	+0.039
Recall	0.917	0.820	-0.098
False Positive Rate	0.042	0.030	-0.012

Table 35: Region: Bridges and boats

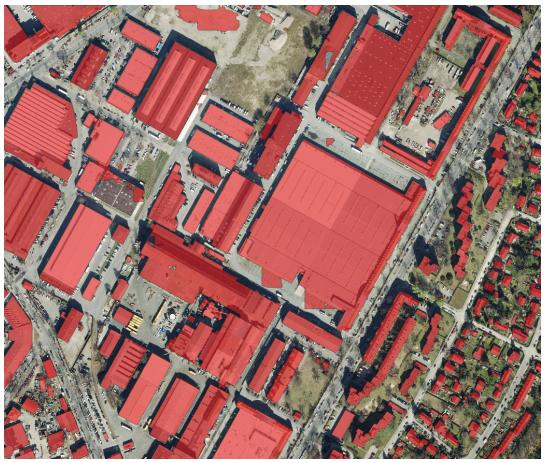
(a) Baseline — Forest

(b) DSM — Forest

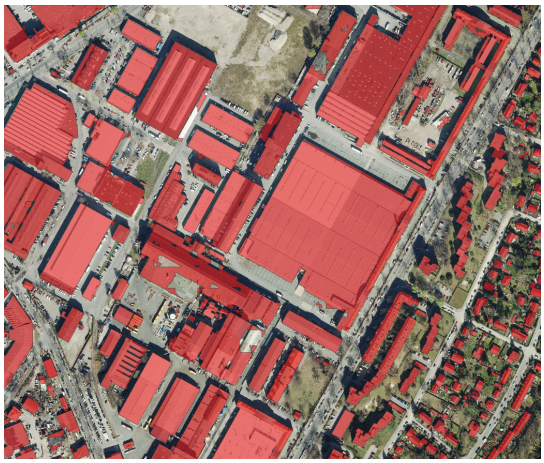
Figure 53: Forest: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	0.996	1.000	+0.004
False Positive Rate	0.004	0	-0.004

Table 36: Region: Forest



(a) Baseline — Industrial area



(b) DSM — Industrial area

Figure 54: Industrial area: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	0.968	0.936	-0.032
IoU	0.927	0.852	-0.075
F1 Score	0.962	0.920	-0.042
Precision	0.948	0.960	+0.012
Recall	0.977	0.883	-0.094
False Positive Rate	0.022	0.015	-0.007

Table 37: Region: Industrial area

(a) Baseline — Inner city

(b) DSM — Inner city

Figure 55: Inner city: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	0.964	0.952	-0.012
IoU	0.924	0.898	-0.025
F1 Score	0.960	0.946	-0.014
Precision	0.939	0.933	-0.006
Recall	0.983	0.961	-0.022
False Positive Rate	0.028	0.031	+0.002

Table 38: Region: Inner city

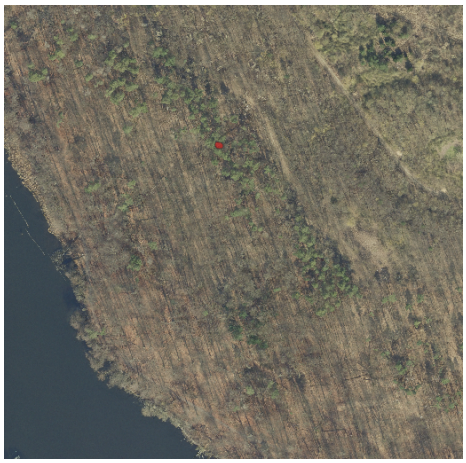
(a) Baseline — Small street through forest

(b) DSM — Small street through forest

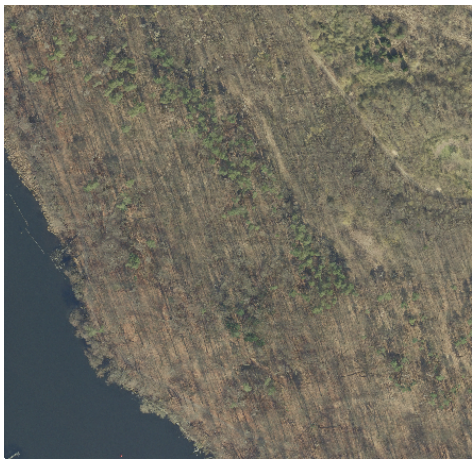
Figure 56: Small street through forest: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	0.994	0.999	+0.006
False Positive Rate	0.006	0.001	-0.006

Table 39: Region: Small street through forest



(a) Baseline — Water at mosaic edge

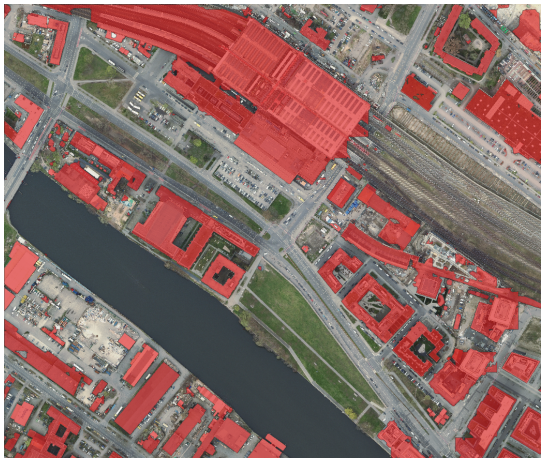


(b) DSM — Water at mosaic edge

Figure 57: Water at mosaic edge: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	1.000	1.000	0
False Positive Rate	0	0	0

Table 40: Region: Water at mosaic edge



(a) Baseline — Streets & railway

(b) DSM — Streets & railway

Figure 58: Streets & railway: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	0.922	0.938	+0.016
IoU	0.753	0.794	+0.041
F1 Score	0.859	0.885	+0.026
Precision	0.786	0.827	+0.041
Recall	0.947	0.953	+0.006
False Positive Rate	0.065	0.050	-0.015

Table 41: Region: Streets & railway

(a) Baseline — Water

(b) DSM — Water

Figure 59: Water: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	1.000	1.000	0
False Positive Rate	0	0	0

Table 42: Region: Water

(a) Baseline — Water riverbank

(b) DSM — Water riverbank

Figure 60: Water riverbank: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	1.000	1.000	0
False Positive Rate	0	0	0

Table 43: Region: Water riverbank

(a) Baseline — Housing area

(b) DSM — Housing area

Figure 61: Housing area: baseline (Z-score) vs. DSM input.

Metric	Baseline	DSM	Difference
Overall Accuracy	0.972	0.973	+0.001
IoU	0.853	0.858	+0.005
F1 Score	0.921	0.924	+0.003
Precision	0.913	0.916	+0.003
Recall	0.928	0.931	+0.003
False Positive Rate	0.016	0.015	-0.001

Table 44: Region: Housing area

Overlap Experiments: Comparison Baseline and best results

(a) Baseline — Bridges & boats

(b) Improved (train 20%, test 10%) — Bridges & boats

Figure 62: Bridges & boats: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.946	0.945	-0.001
IoU	0.700	0.671	-0.030
F1 Score	0.824	0.803	-0.021
Precision	0.748	0.787	+0.039
Recall	0.917	0.820	-0.098
False Positive Rate	0.042	0.030	-0.012

Table 45: Region: Bridges & boats

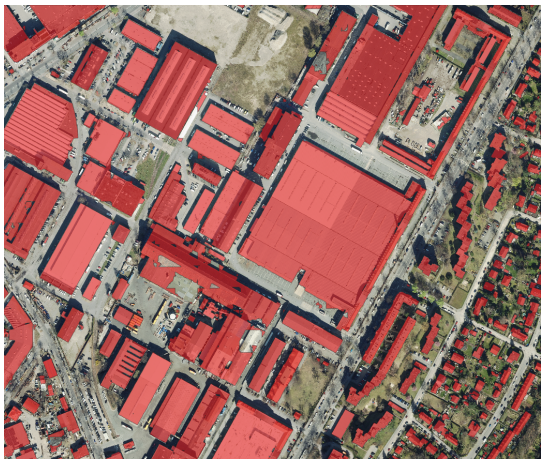
(a) Baseline — Forest

(b) Improved (train 20%, test 10%) — Forest

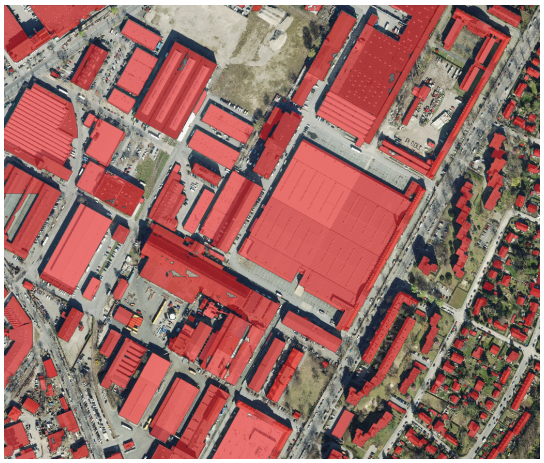
Figure 63: Forest: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.996	1.000	+0.004
False Positive Rate	0.004	0	-0.004

Table 46: Region: Forest



(a) Baseline — Industrial area



(b) Improved (train 20%, test 10%) — Industrial area

Figure 64: Industrial area: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.968	0.936	-0.032
IoU	0.927	0.852	-0.075
F1 Score	0.962	0.920	-0.042
Precision	0.948	0.960	+0.012
Recall	0.977	0.883	-0.094
False Positive Rate	0.022	0.015	-0.007

Table 47: Region: Industrial area

(a) Baseline — Inner city

(b) Improved (train 20%, test 10%) — Inner city

Figure 65: Inner city: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.964	0.952	-0.012
IoU	0.924	0.898	-0.025
F1 Score	0.960	0.946	-0.014
Precision	0.939	0.933	-0.006
Recall	0.983	0.961	-0.022
False Positive Rate	0.028	0.031	+0.002

Table 48: Region: Inner city

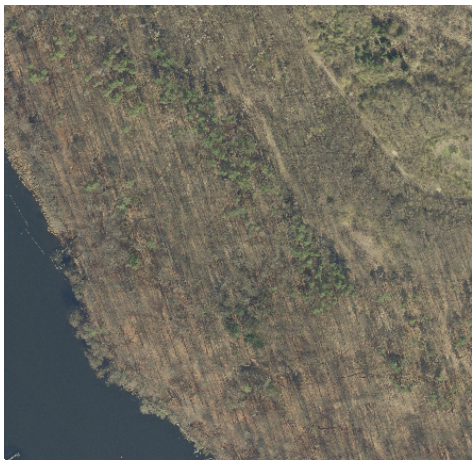
(a) Baseline — Small street through forest

(b) Improved (train 20%, test 10%) — Small street through forest

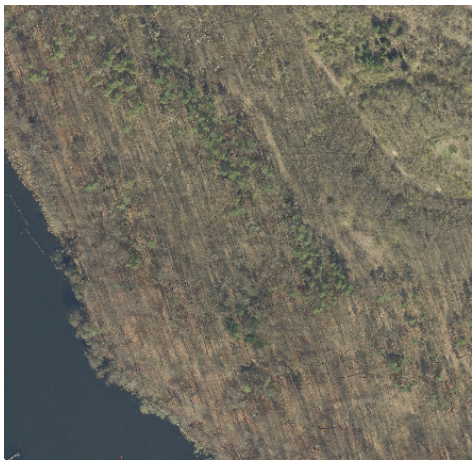
Figure 66: Small street through forest: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.994	0.999	+0.006
False Positive Rate	0.006	0.001	-0.006

Table 49: Region: Small street through forest



(a) Baseline — Water at mosaic edge



(b) Improved (train 20%, test 10%) — Water at mosaic edge

Figure 67: Water at mosaic edge: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	1.000	1.000	0
False Positive Rate	0	0	0

Table 50: Region: Water at mosaic edge

(a) Baseline — Streets & railway

(b) Improved (train 20%, test 10%) — Streets & railway

Figure 68: Streets & railway: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.922	0.938	+0.016
IoU	0.753	0.794	+0.041
F1 Score	0.859	0.885	+0.026
Precision	0.786	0.827	+0.041
Recall	0.947	0.953	+0.006
False Positive Rate	0.065	0.050	-0.015

Table 51: Region: Streets & railway

(a) Baseline — Water

(b) Improved (train 20%, test 10%) — Water

Figure 69: Water: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	1.000	1.000	0
False Positive Rate	0	0	0

Table 52: Region: Water

(a) Baseline — Water riverbank

(b) Improved (train 20%, test 10%) — Water riverbank

Figure 70: Water riverbank: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	1.000	1.000	0
False Positive Rate	0	0	0

Table 53: Region: Water riverbank

(a) Baseline — Housing area

(b) Improved (train 20%, test 10%) — Housing area

Figure 71: Housing area: baseline (DSM) vs. improved overlap configuration.

Metric	Baseline	Improved	Difference
Overall Accuracy	0.972	0.973	+0.001
IoU	0.853	0.858	+0.005
F1 Score	0.921	0.924	+0.003
Precision	0.913	0.916	+0.003
Recall	0.928	0.931	+0.003
False Positive Rate	0.016	0.015	-0.001

Table 54: Region: Housing area

(a) Baseline — Bridges & boats

(b) Decision threshold 0.3 — Bridges & boats

Figure 72: Bridges & boats: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Decision threshold: Comparison Baseline and best results

Decision threshold: Comparison Baseline and Threshold 0.3

(a) Baseline — Bridges & Boats

(b) Decision threshold 0.3 — Bridges & Boats

Figure 73: Bridges & Boats: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	0.948	0.965	+0.017
IoU	0.703	0.777	+0.073
F1 Score	0.826	0.874	+0.048
Precision	0.756	0.856	+0.100
False Positive Rate	0.040	0.028	-0.012

Table 55: Region: Bridges & Boats

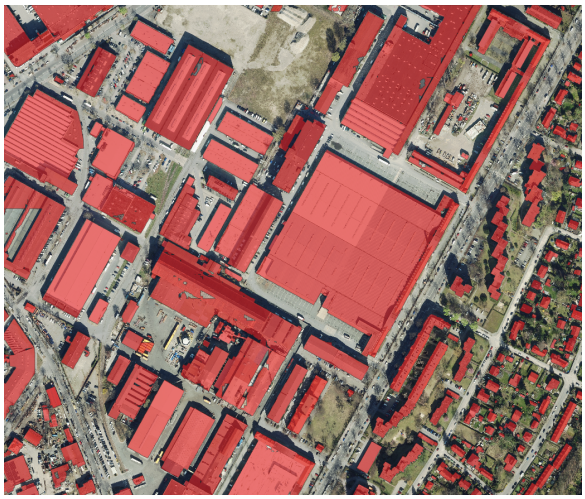
(a) Baseline — Forest

(b) Decision threshold 0.3 — Forest

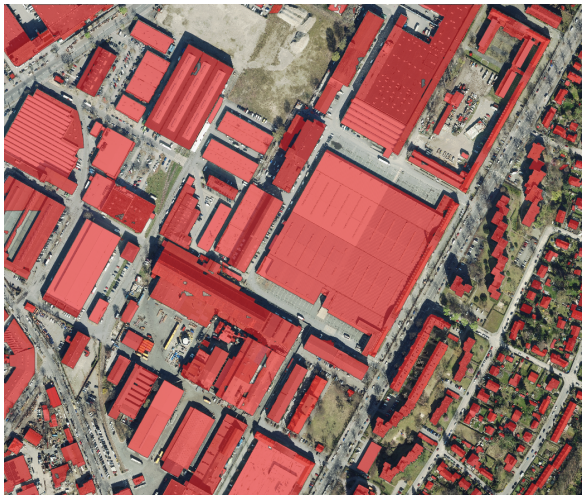
Figure 74: Forest: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	1.000	1.000	0.000
False Positive Rate	0.000	0.000	0.000

Table 56: Region: Forest



(a) Baseline — Industrial area



(b) Decision threshold 0.3 — Industrial area

Figure 75: Industrial area: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	0.974	0.975	0.000
IoU	0.940	0.941	0.001
F1 Score	0.969	0.970	0.001
Precision	0.967	0.962	-0.005
Recall	0.971	0.977	0.006
False Positive Rate	0.014	0.016	0.002

Table 57: Region: Industrial area

(a) Baseline — Inner city

(b) Decision threshold 0.3 — Inner city

Figure 76: Inner city: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	0.960	0.959	-0.001
IoU	0.915	0.914	-0.001
F1 Score	0.956	0.955	-0.001
Precision	0.941	0.935	-0.006
Recall	0.971	0.976	+0.005
False Positive Rate	0.027	0.030	+0.003

Table 58: Region: Inner city

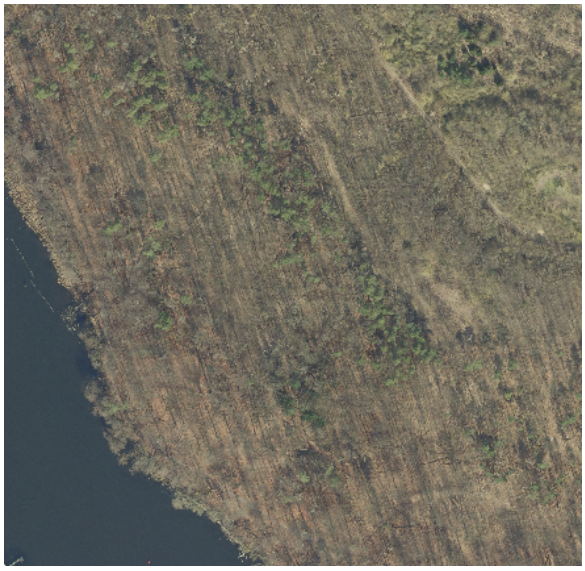
(a) Baseline — Small street through forest

(b) Decision threshold 0.3 — Small street through forest

Figure 77: Small street through forest: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	1	1	0
False Positive Rate	0	0	0

Table 59: Region: Small street through forest



(a) Baseline — Water at mosaic edge

(b) Decision threshold 0.3 — Water at mosaic edge

Figure 78: Water at mosaic edge: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	1.000	1.000	0.000
False Positive Rate	0.000	0.000	0.000

Table 60: Region: Water at mosaic edge

(a) Baseline — Streets & railway

(b) Decision threshold 0.3 — Streets & railway

Figure 79: Streets & railway: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	0.954	0.955	+0.001
IoU	0.829	0.835	+0.006
F1 Score	0.906	0.910	+0.004
Precision	0.925	0.914	-0.011
Recall	0.888	0.905	+0.017
False Positive Rate	0.018	0.021	+0.003

Table 61: Region: Streets & railway

(a) Baseline — Water

(b) Decision threshold 0.3 — Water

Figure 80: Water: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	1.000	1.000	0.000
False Positive Rate	0.000	0.000	0.000

Table 62: Region: Water

(a) Baseline — Water riverbank

(b) Decision threshold 0.3 — Water riverbank

Figure 81: Water riverbank: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	1.000	1.000	0.000
False Positive Rate	0.000	0.000	0.000

Table 63: Region: Water riverbank

(a) Baseline — Housing area

(b) Decision threshold 0.3 — Housing area

Figure 82: Housing area: baseline (train 20%, test 10% overlap) vs. decision threshold 0.3.

Metric	Baseline	Threshold 0.3	Difference
Overall Accuracy	0.970	0.971	+0.001
IoU	0.839	0.847	+0.008
F1 Score	0.912	0.917	+0.004
Precision	0.931	0.920	-0.011
Recall	0.895	0.914	+0.019
False Positive Rate	0.012	0.014	+0.002

Table 64: Region: Wohngebiet (Housing area) — Baseline vs. Decision Threshold 0.3