Evaluation of OpenStreetMap Data of the Built Environment with the Help of Spatio-Temporal Digital Elevation Models

Ruiqi Liu[1*], Paul Kuper[1], Mulhim Al-Doori[1], Martin Breunig[1]

[1] Geodetic Institute, Karlsruhe Institute of Technology, Germany
– (ruiqi.liu, paul.kuper, martin.breunig)@kit.edu, maldoori@gmail.com

Keywords: Spatio-temporal DEM, OSM, Geospatial Data Integration, Geospatial Data Validation, Self-Training Model.

Abstract

Recent advances in remote sensing have shifted the focus from the analysis of individual image scenes to the understanding of complex earth systems. That means the analysis of dynamic evolutions replaces previous static examinations for fixed time points. Furthermore, interdisciplinary research and the integration of heterogeneous data sources are characterizing this transformation process. Digital Elevation Models (DEMs) are predestined for supporting this process by supplementing orthophotos and map data. Promising applications include city planning, landslide analysis, and flood risk assessment where spatio-temporal change detection is a central concept to be applied. Concerning map data, the OpenStreetMap (OSM) project, based on the idea of Volunteered Geographic Information, has revolutionized the effective production and update of digital maps. However, OSM data do not include elevation information and often contains incorrect geometric information in the built environment. In this paper, we introduce a self-training framework for evaluating OSM building footprints with the aid of high-resolution DEMs. The framework supports building segmentation with a weakly supervised approach to improve the representation of OSM building footprints. The availability of DEMs is used to check the quality of OSM data. The applicability of the proposed approach is demonstrated through a case study in Karlsruhe, Germany, showing promising performance in evaluating OSM building footprints. With our approach, change detection of OSM data can also be carried out using different temporal versions of DEM and OSM data. Finally, conclusions are drawn from the presented approach and an outlook is presented on future research on spatio-temporal DEMs.

1. Introduction

Recent advances in remote sensing have shifted the focus from analysing individual image scenes to understanding complex earth systems by promoting interdisciplinary research work and integrated approaches (Reichstein et al., 2019). Geospatial data and model integration as well as advanced geospatial data analysis are contributing to the improvement of geospatial data quality in earth observation. Explicitly considering the dimensions of space and time strengthens this process. In this paper we are focusing on Digital Elevation Models (DEMs) representing the 2.5D terrain with $e = f(x,y)$, let e be the elevation of a point $P_{x,y}$ on the DEM. DEMs are indispensable to support geospatial analysis in the natural and built environment. As is common in geography literature (e.g. Weibel and Heller, 1991; Zhou, 2017), we are using the term “DEM” as the general term for the Digital Terrain Model (DTM, i.e. bare land surface model) and the Digital Surface Model (DSM, i.e. representing the earth´s surface including the elevations of objects from the natural (trees, bushes etc,) and the built environment (buildings, bridges, other infrastructures etc.). In this study, we are using both, DSM and DTM with the focus on the built environment. Liu et al. (2025) used spatio-temporal DEMs to monitor the built environment changes over time. They consist of a series of DEM datasets acquired at different timestamps. In recent years, many regions have publicly released high-quality spatio-temporal DEMs to support public services, urban planning, and environmental management. For example, in Germany, several regional governmental agencies have produced and made available high-resolution (e.g. 1 m) LiDAR-derived DEM datasets. OpenStreetMap (OSM, 2025), though similarly open and widely accessible, has lower spatial accuracy and consistency compared with these DEMs for the same regions. Therefore, we propose an integrated approach using high-resolution DEMs to identify errors in OSM data.

In the first step, we subtract the DTM from the DSM, receiving the so-called normalized DSM (nDSM), which purely represents features of the built environment being placed on top of the earth´s surface such as vegetation, buildings, and bridges. In the second step, we apply region-based segmentation methods to generate reliable labels based on the geometric characteristics of nDSM and OSM data. In the third step, we apply a weakly supervised learning approach to identify errors in OSM data. A self-training Convolutional Neural Network (CNN) is applied to spatio-temporal DEM data to generate semantic segmentation results, which can subsequently be used as a reference for evaluating corresponding OSM data.

The remainder of this paper is structured as follows: In section 2 we summarize prior research work on DEM and OSM data. In section 3 we introduce our integrated approach using DEM data to evaluate the quality of OSM data. Section 4 is dedicated to a use case demonstrating the automatic segmentation and validation of building footprints at Karlsruhe, Germany. In section 5 additionally the temporal evolution of OSM footprints is considered to evaluate the development of a city. Finally, in section 6 conclusions are drawn from the presented study and an outlook is given on our future research dealing with AI-based and database-supported spatio-temporal analysis of DEMs.

2. Related Work

Zhou (2017) described the differences between DEM, DTM, and DSM, the corresponding data acquisition techniques, and the three digital terrain representations “regular grid structures”, “triangulated irregular networks” and “contour structures”. Early research work on DTMs has been published by several authors (Doyle, 1978; Weibel and Heller, 1991; Zhilin et al., 2004). Weibel and Heller (1991) focused on their role as part of National Spatial Data Infrastructures (NSDI) and in hydrologic modelling, sediment transport, soil erosion estimation etc. Furthermore, the generation of DTMs as raster (field-based approach) or

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-111-2026

Triangulated Irregular Networks (TINs - object-based approach) from point clouds derived from satellite image and laser scanning data has been discussed. Okolie et al. (2022) examined the role of DEMs for data fusion in remote sensing applications. However, due to the world-wide emerging demand of environmental monitoring in case of natural events such as landslides or earthquakes as well as of monitoring city development, the data structure of DEMs (i.e. DTMs and DSMs) has to be extended to model their temporal development and to allow change detection between different time steps of the DEM. Therefore, the notion of “Multi-Temporal DTM” has been created to examine uncertainties of DTMs for geomorphic studies and to estimate moving volumes during landslides (Tseng et al., 2013; Kamp et al., 2023). To manage multi-temporal DEMs efficiently in a geospatial database management system, corresponding spatio-temporal database operations for DEMs have to be provided. Liu et al. (2025) proposed an integrated approach for the management and analysis of spatio-temporal DEMs. Further research has applied DEMs to semantic segmentation tasks, such as building segmentation. Traditional methods often use nDSM as an additional input to improve orthophoto-based results (Buyukdemircioglu et al., 2022; Qiu et al., 2025). Vats et al. (2024) focused on LiDAR data and introduce a terrain-aware self-supervised learning approach, allowing the model to extract building footprints directly from LiDAR-derived DEMs and perform well with minimal labels. Since the generation of labels is usually time-consuming, one efficient approach is to directly use Volunteered Geographic Information (VGI) data (Goodchild, 2007). Undoubtedly, the OpenStreetMap (OSM, 2025) project, taking up the idea of VGI, has revolutionized the creation, update, and use of maps. However, the open approach makes OSM vulnerable to errors and gaps as noticed by many researchers such as (Helbich et al., 2012; Kaur et al., 2017). A database-supported analysis of changes for OSM data has been conducted by Martini et al. (2019). Uncertainty measures for geospatial data quality can be found in fundamental geospatial literature such as (Guptill and Morrison, 1995; Devillers and Jeansoulin, 2006). Finally, the quality of OSM geometries and the fitness of OSM data as labels for deep learning classification tasks have been examined in several studies (e.g., Hellbich et al., 2012; Schott et al., 2024). However, it seems that hitherto no or only few studies have been carried out to exploit the strengths of DEM data to improve the quality, i.e. completeness and preciseness, of geometric OSM data. This is the starting point of our study.

3. Integrated Approach: Using DEM Data to Evaluate the Quality of OSM Data

3.1 Description of Data Sources

3.1.1 Digital Elevation Models

The starting point for the development of DEMs typically is a LiDAR scan in the form of a point cloud. Such a scan is usually further processed in two main products: the DTM and the DSM. During this process, the raw point cloud is converted into a raster dataset with a fixed spatial resolution, e.g. 0.25m, providing one elevation value for each 2D-coordinate. A DTM represents the elevation of the bare ground surface, excluding vegetation and buildings, while the DSM includes the elevation of all surface features such as buildings and vegetation. The typical file format of DTM and DSM datasets is the raster format GeoTIFF. Another useful product is the nDSM, as mentioned above, it is conducted by subtracting the DTM from the DSM:

$$DSM - DTM = nDSM \quad (1)$$

The nDSM represents only surface features including vegetation and features of the built environment such as buildings and bridges.

3.1.2 OpenStreetMap Data

The data model of OSM is an especially open and flexible structure based on a key-value principle. This open design allows OSM contributors to continuously integrate new data categories into the overall dataset. Only general conventions and guidelines define how key attributes (e.g., building = yes) should be assigned to features. The absence of a predefined data schema, such as those typically found in traditional GIS datasets, offers significant advantages and enables numerous applications that were not originally anticipated. However, this flexibility can also lead to certain drawbacks, such as the need to extract subsets for specific applications. Smaller datasets are typically distributed in human-readable formats such as GeoJSON, while larger datasets are provided in compressed binary formats such as OSM PBF to reduce the storage requirements.

3.2 Integration of DEM and OSM Data

Provided that DEM and OSM datasets cover the same region, they can be integrated through spatial overlay. In this integration, OSM provides semantic information, while a high-resolution nDSM contributes accurate geometric characteristics of the real built environment.

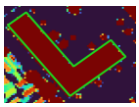
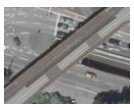
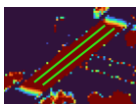
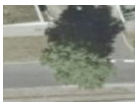
	Reference	OpenStreet Map	nDSM
Building			
Bridge			
Tree			

Table 1. Overlay of built environment features from nDSM and OSM data. For visualization clarity of buildings, the nDSM values are displayed within a height range of 2–5 m.

As shown in Table 1, after the overlay, the vector features in OSM correspond geometrically well to the elevation values provided by the nDSM. For example, buildings in OSM are typically represented as polygons, and their footprints can be directly compared with the elevation values of the corresponding building areas in the nDSM. Although bridges and trees, respectively, are usually represented as polylines and points in OSM, their corresponding surface features can be reconstructed by incorporating the relative elevation information from the surrounding areas in the nDSM.

3.3 Identification of Error Types in OSM Data

It is well known that the quality of OSM data differs significantly in different regions. When comparing nDSM with OSM data, we were able to identify six typical OSM error types (corresponding

to the built environment) as follows. (A) to (D) are defined at the object level, describing inconsistencies that affect the overall footprint of each feature, while (E) and (F) are defined at the part level, referring to local inconsistencies within individual features:
(A) Missing Feature refers to the case where the existing feature is completely absent from OSM.
(B) Extra Feature represents the case where the feature is mapped in OSM but does not exist in reality.
(C) Mismatch describes geometric discrepancies between OSM and the reference, where the general location is correct, but the shape differs significantly.
(D) Positional Misalignment refers to spatial shifts between OSM and the reference, while the overall shape remains accurate.
(E) Incomplete Shape indicates the feature that is only partially mapped in OSM, with some sections of the structure missing.
(F) Redundant Shape refers to the feature that contains extra parts within its geometry.

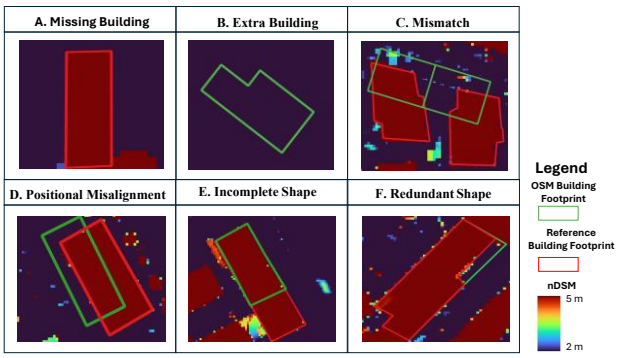


Figure 1. OpenStreetMap Error Types for Building Footprint.

To better describe the defined error types, we take building footprints as an example to demonstrate the typical types of OSM data errors, as shown in Figure 1, where red outlines represent ground truth shapes, and green outlines indicate OSM building footprints.

3.4 Automatic Segmentation of Built Environment

In principle, the quality of OSM data can be verified by visually interpreting and comparing it with the DEM to assess geometric accuracy and data completeness. However, such manual inspection is highly time-consuming for large-scale datasets. To address this limitation, we developed an approach based on CNNs for an automated segmentation of features within the nDSM. By comparing the segmentation results with OSM features, OSM error types can be detected. Since DEMs are usually provided in a raster format (e.g. GeoTIFF), they are particularly suitable to be processed with CNN-based architectures.
The training of CNNs typically relies on high-quality ground truth data. However, due to various errors, OSM data cannot be directly used as reliable labels. They can still serve as weak labels. Self-training is one of the most common semi- or weakly supervised approaches and has been widely used in remote sensing image segmentation. In this method, the model generates high-confidence predictions on unlabelled data and reuses them as pseudo labels to iteratively improve its performance, allowing the network to learn effectively even without complete ground truth (Amini et al., 2025; Zhou, 2018).

4. Use Case: Evaluation of OSM Building Footprints

4.1 Study Area

Our experiments were conducted in Karlsruhe, a city located in southwestern Germany, characterized by flat terrain and a regular urban layout dominated by low- and mid-rise buildings. As shown in Figure 2, OpenStreetMap records 9,437 buildings within the study area. Most building footprints correspond well with the high-resolution DEM (0.25 m DTM and 1 m DSM) acquired in 2016 from airborne laser scanning (ALS) data provided by the State Office for Geoinformation and Land Development Baden-Württemberg (LGL). This strong spatial consistency indicates a high level of completeness and accuracy in the OSM data, providing a reliable foundation for the subsequent self-training process.

Figure 2. Study area in Karlsruhe.

To ensure validation precision, a relatively small yet representative study area of 12 km² was selected. The area comprises diverse land-cover types, including urban centres, towns, and forested regions, allowing for a comprehensive evaluation of model performance across different environments. These datasets were divided into 60% for training, 20 % for validation, and 20% for testing, and data augmentation (random rotations as well as horizontal and vertical flips) was applied to the training set to enhance model generalization.

4.2 Self-Training Model for Automatic Segmentation of the Built Environment

As mentioned in Sections 3.2 and 4.1, some OSM building footprints contain geometric inaccuracies and therefore cannot be directly used as ground truth for CNN training. To address this issue, the geometric features of the building footprints were aligned with those of the DEM data to generate soft labels, where each pixel represents the likelihood of building presence, providing a more reliable supervision signal for model training. This section consists of two main parts. The first part introduces the process of generating soft labels. In the second part, the generated soft labels are used as pseudo ground truth and combined with DEM channels in a self-training framework based on the U-Net model, where the soft labels are iteratively updated until the model converges.

4.2.1 Generation of Soft Labels

Figure 3 describes the process of generating soft labels. First, the spatial and temporal ranges of the study area were derived from the DEM to enable data retrieval in *Overpass Turbo*, which is an API-based query interface of OpenStreetMap. We collected elevated objects in the built environment including building footprints, bridges, and forests. With this OSM information, as Figure 4 shows, soft labels are divided into three categories: *non-*

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-111-2026 | © Author(s) 2026. CC BY 4.0 License.

building, *likely building*, and *unknown area*. *Non-building* represents pixels with a probability of 0, such as the height lower than 1.5 meter, as well as bridges and forested areas. The *likely building* corresponds to areas where OSM building footprints exist, and the remaining regions are labelled as *unknown area*.

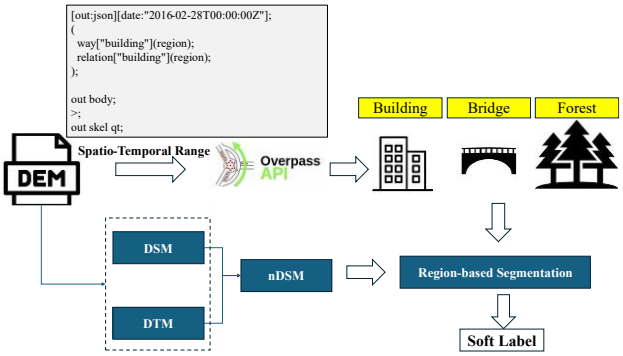


Figure 3. The integration of DEM data and OpenStreetMap data for the soft label generation.

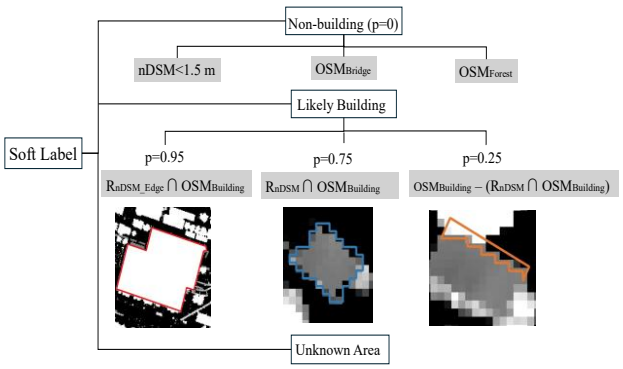


Figure 4. The composition of soft labels. R_{nDSM_Edge} denotes regions derived from watershed segmentation on the nDSM edge map, and R_{nDSM} denotes regions obtained by region growing on the nDSM using OSM building centroids as seeds.

The probability of each pixel in the *likely building* area differs from the reliability level of OSM building labels. It is estimated based on the spatial overlap between the OSM building footprints and the regions extracted from the nDSM using region-based segmentation methods including watershed segmentation and region growing (Khalid, 2022). Pixels with high confidence (p = 0.95) were first identified by the watershed segmentation based on nDSM edges representing regions with clear building boundaries. A significant spatial overlap between these regions and the OSM building footprints indicates a high probability of building presence. Then, the centroids of the OSM building footprints were used as seeds for region growing on the nDSM to identify remaining potential building regions, where height similarity guided the expansion of connected regions. As this approach relies mainly on height information, the extracted regions may be affected by other elevated objects such as trees, leading to relatively lower reliability. Their spatial overlap with the OSM building footprints was therefore assigned a building existence probability of p = 0.75. For p = 0.25, pixels are regarded as less reliable, corresponding to the areas of OSM buildings outside the intersection with the region growing results.

4.2.2 Self-Training Workflow

Based on OSM soft labels and DEM data, the self-training workflow is described in Figure 5. First, the DEM characteristics were inputted into multiple channels, including nDSM, DTM slope, DTM aspect. If orthophotos are available, they can also be used as additional input channels. The soft labels were used as pseudo ground truth for training. Since the DTM (0.25 m) and DSM (1 m) have different spatial resolutions, the DTM was adopted to match the DSM resolution. Then the data were divided into patches, which were subsequently split into training, validation, and testing datasets. During the self-training process, a U-Net model was employed as the backbone network, which follows an encoder–decoder architecture with skip connections to capture multi-scale contextual information (Ronneberger et al., 2015). In the initial stage, the soft labels were used as inputs to compute the loss and obtain the model predictions. The predicted probabilities were then updated and converted into new pseudo labels, which were added to the existing training labels for the next training iteration. The process is shown in Algorithm 1.

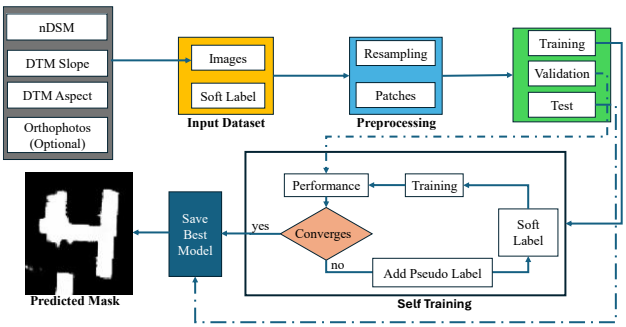


Figure 5. Self-training workflow for OSM building footprints.

Algorithm 1. Pseudo code of self-training algorithm.

Algorithm 1 Self-training algorithm

Require: Labeled data L; Unlabeled data U; prediction probability p; label correction threshold $T_l = 0.99$; pseudo-label threshold $T_p = 0.995$
Ensure: Final model M

Initialize best_F1 $\leftarrow -\infty$, no_improve_stages $\leftarrow 0$
repeat
  Train model M on L until no F1 improvement for 10 consecutive epochs
  Evaluate on validation set and record best F1 of this stage
  for L **do**
    if $L \geq T_l$ and $p \leq 0.1$ **then**
      $L \leftarrow 0.1$
    else if $L < 0$ and $p \geq T_l$ **then**
      $L \leftarrow 0.9$
    end if
  end for
  for U **do**
    if $p \geq T_p$ **then**
      move the sample to L
    end if
  end for
  Train model M on the updated labeled set L
  if best F1 improves by ≥ 0.002 **then**
    best_F1 $\leftarrow$ new value
    no_improve_stages $\leftarrow 0$
  else
    no_improve_stages $\leftarrow$ no_improve_stages $+ 1$
  end if
until no_improve_stages $= 2$
return M with best F1

Within each stage, training stopped if the F1 score did not improve for 10 consecutive epochs. The self-training process stopped when the validation F1 improvement was smaller than 0.002 for two consecutive stages. Finally, the model with the best F1 score was saved.

4.3 Results of the Self-Training Model

We implemented automatic segmentation for buildings of the year 2016 based on the workflow illustrated in Figure 5. The ground truth data for the validation and test datasets were manually interpreted from orthophotos and nDSMs. The segmentation results are presented as follows.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-111-2026 | © Author(s) 2026. CC BY 4.0 License.

Model	U-Net
Learning Rate	0.0005
Optimizer	Adam
Metrics	F1-Score
Loss Function	BCE + Dice + Low height penalty
Batch Size	8
Input Size	128×128

(a). Parameters used in the self-training model.

Epoch	168	Validation Precision	0.8848
Best Stage	5	Validation IoU	0.8110
Training F1 Score	0.9484	Test F1 Score	0.8983
Training Precision	0.9282	Test Precision	0.8888
Training IoU	0.8931	Test IoU	0.8154
Validation F1 Score	0.8956	—	

(b). Results of the self-training model.

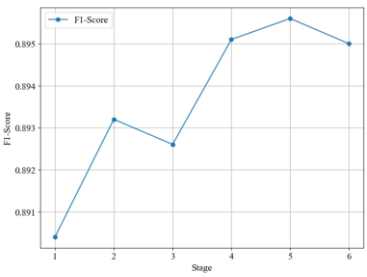
Table 2. Parameters and results of the self-training model.

Table 2 (a) presents the main parameter configurations of the model. We applied a hybrid loss function:

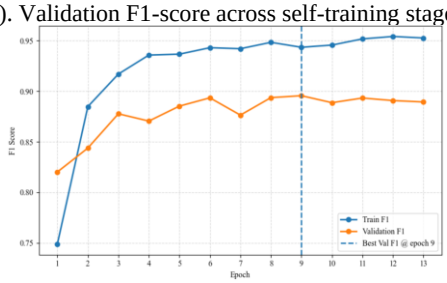
$$L = \alpha L_{BCE} + (1-\alpha) L_{Dice} + \lambda L_{lowH} \quad (2)$$

Where L_{BCE} denotes the Binary Cross-Entropy loss and L_{Dice} represents the Dice loss that evaluates the overlap between predicted and ground truth regions. The combination is widely used in U-Net-based segmentation models to balance pixel-wise classification and region overlap (Zhou et al., 2018). Besides, we add a low height penalty term L_{lowH} applied to regions where the nDSM height lower than 1.5 m. α and λ are hyperparameters empirically determined based on validation performance.

Table 2 (b) shows the training results, where the highest validation F1-score of 0.8956 was achieved at Stage 5, indicating a good balance between precision and recall. The corresponding test result is also consistent, with a test F1-score of 0.8983.



(a). Validation F1-score across self-training stages.



(b). Training and validation F1-score during stage 5.

Figure 6. Performance of the self-training model.

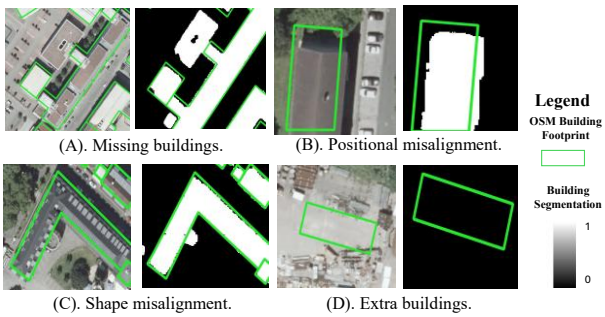
Figure 6 (a) illustrates the F1-score variation from Stage 1 to Stage 6, demonstrating that the model performance improved and gradually converged; Figure 6(b) illustrates the training and validation F1 curves during the best Stage 5. The validation performance increases steadily and reaches its peak at epoch 9, indicating stable training behaviour and a well-defined early stopping point. These results show the effectiveness and reliability of the proposed self-training strategy for automatic building segmentation.

5. Spatio-Temporal Evaluation

DEM datasets provided by LGL for the years 2001, 2016, and 2022 in the same study area were used to evaluate the spatio-temporal quality of OSM building footprints and automatic segmentation results, as well as the cross-temporal performance of the self-training model.

5.1 Spatio-Temporal Quality of OSM Building Footprints

According to the six error types illustrated in section 3.3, we further classified them into four principal categories: *missing buildings*, *positional misalignment*, *shape misalignment*, and *extra buildings*. These error types were identified through manual visual inspection of the automatic segmentation results, as illustrated in Figure 7.

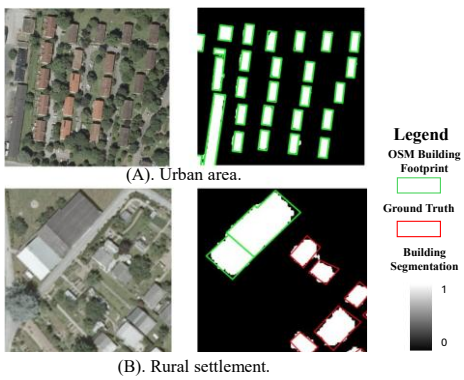


(A). Missing buildings. (B). Positional misalignment.

(C). Shape misalignment. (D). Extra buildings.

Figure 7. Examples of four error types (A–D) identified through manual visual inspection of the automatic segmentation results.

To further analyse this spatial variability, we divided study areas into urban areas and rural settlements with the help of OSM land-use tags to detect OSM and our model’s errors. Specifically, areas labeled as *residential*, *commercial* and *industrial* were defined as urban areas, while areas tagged as *allotments*, *farmland*, and other low-intensity land-use types were classified as rural settlements. As expected in our observations OSM data performs greater accuracy in urban areas, whereas their performance is poorer in rural settlements.



(A). Urban area.

(B). Rural settlement.

Figure 8. Comparison of OSM building footprints and building segmentation performance in urban areas and rural settlements.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-111-2026 | © Author(s) 2026. CC BY 4.0 License.

After converting the segmentation results into building objects through morphological post-processing, we counted buildings with the above error types at the object level using QGIS. As listed in Table 3, both of OSM building footprints and automatic segmentation results show less errors in urban areas, as Figure 8 (A) shows. However, in rural settlements as shown in Figure 8 (B), OSM data contain a large number of missing buildings. Although the automatic segmentation is able to capture the general shapes of buildings, the presence of dense trees surrounding many rural houses interferes with boundary recognition. As a result, some building edges extend into vegetated areas, leading to higher rates of shape misalignment and extra building errors.

Region	Data Source	Missing	Positional	Shape	Extra
Urban Area	**OSM Building Footprint**	119/1604	4/1604	43/1604	16
	Automatic Segmentation	2/1604	0/1604	26/1604	24
Rural Settlement	**OSM Building Footprint**	432/709	2/709	21/709	0
	Automatic Segmentation	8/709	4/709	52/709	13
Overall	**OSM Building Footprint**	551/2313	6/2313	50/2313	16
	Automatic Segmentation	10/2313	4/2313	78/2313	37

* Missing, Positional, Shape, and Extra represent missing buildings, positional misalignment, shape misalignment, and extra buildings. The denominators represent the total number of reference buildings in each region.

Table 3. Spatial evaluation of OSM building footprints and automatic segmentation results.

We also evaluated the temporal evolution of error types for OSM building footprints, as Figure 9 shows. Because the OpenStreetMap project was founded in 2004, OSM data were completely absent in 2001. From 2016 to 2022, the number of missing buildings decreased, while no significant improvement was observed in the other error types.

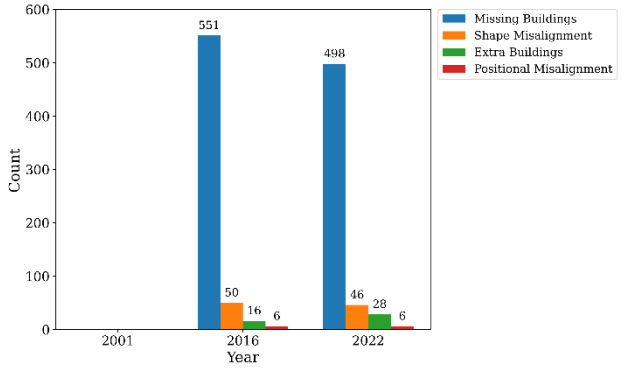


Figure 9. Temporal evolution of error types for OSM building footprints.

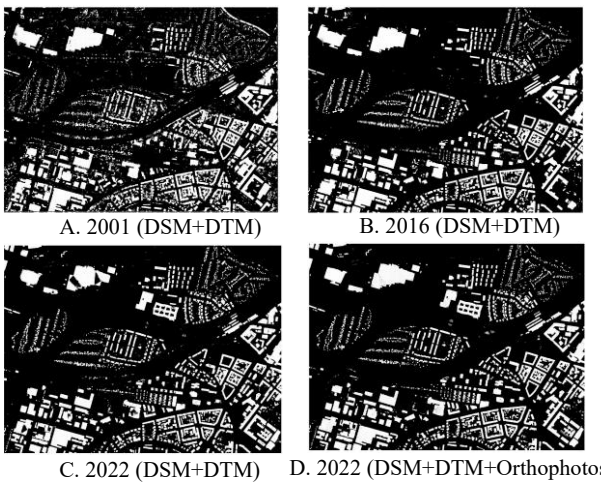
Overall, our model significantly improves the detection of buildings in rural settlements and enhances the accuracy of building shapes in urban areas. Nevertheless, boundary recognition in rural regions still requires further improvement.

5.2 Cross-Temporal Evaluation of the Self-Training Model

Table 4 presents the cross-temporal evaluation of the self-training model. We applied the model to the same test area for the years 2001 and 2022. As orthophotos were available for 2022, an additional experiment was conducted by incorporating them as another input channel for model training. The reference data were generated through manual visual interpretation using nDSM data, orthophotos and Google Earth® imagery.

Year	Precision	Recall	IoU	F1 score
2001 (DSM+DTM)	0.7835	0.8502	0.6887	0.8155
2016 (DSM+DTM)	0.8888	0.9080	0.8154	0.8983
2022 (DSM+DTM)	0.8850	0.9060	0.8105	0.8953
2022 (DSM+DTM+ Orthophotos)	0.9046	0.9197	0.8384	0.9121

Table 4. Cross-temporal evaluation of the self-training model.



A. 2001 (DSM+DTM) B. 2016 (DSM+DTM)

C. 2022 (DSM+DTM) D. 2022 (DSM+DTM+Orthophotos)

Figure 10. Visualization of building segmentation from different years in the test area.

The results in Table 4 and the corresponding visualization in Figure 10 indicate that the model performed less effectively on the 2001 data, mainly due to the lower DTM resolution (1 m) and the lack of reliable orthophotos for visual validation. Consequently, a larger number of tree-covered areas were misclassified as buildings. In contrast, the results for 2022 were much better, with building distributions more accurately identified. The inclusion of orthophotos led to a slight improvement in overall performance, particularly in distinguishing buildings from trees.

6. Conclusions and Future Work

In this paper we proposed a self-training framework for validating OSM building footprints with the aid of high-resolution DEMs. The framework performs building

segmentation in a self-supervised manner by ensuring geometric consistency between OSM footprints and DEM data. It enables the automatic detection of both errors and omissions in OSM building datasets. The lessons learned from our examinations so far are: The proposed approach effectively identifies missing and erroneous buildings in OSM data, providing an automated approach for assessing and improving OSM building footprint quality. Furthermore, the model supports temporal generalization. For example, it can be applied to regions or years in which OSM data are lacking to predict building locations directly from DEM data. The approach confirms that OSM building footprints show high completeness and accuracy in well-structured urban centers, whereas significant omissions occur in suburban and rural areas with small detached houses during the last ten years. However, several limitations remain. First, the performance of the proposed framework strongly depends on the quality and resolution of the DEM data. Any noise or inaccuracy in the DEM can directly impact the geometric consistency and thus affect the validation results. Second, the reference data obtained through manual visual inspection inevitably contain uncertainties, introducing potential bias into the evaluation. Third, in rapidly changing urban areas with a high variability, this approach has its limit as involved datasets were acquired at different timestamps. Last but not least, the model may still misclassify non-building objects with similar elevation characteristics, such as shipping containers and construction materials.

Future research will focus on advancing the presented framework toward a comprehensive, data-driven methodology for the spatio-temporal analysis of DEMs. The integration of artificial intelligence and advanced statistical modelling techniques offers considerable potential to further enhance the precision, interpretability, and scalability of DEM-based geospatial validation and analysis. Furthermore, comparisons with other buildings footprint validation methods are planned for the future. Firstly, the current self-training approach should be extended to a multimodal learning framework that integrates multiple data sources, including DEMs, optical imagery, and synthetic aperture radar (SAR). Such integration will facilitate the exploitation of complementary spatial and spectral information for improved segmentation and change detection. Deep neural architectures - particularly attention-based encoder–decoder networks and graph-convolutional models - can be investigated to capture the complex spatial and temporal dependencies inherent in multi-epoch DEM time-series.

Secondly, a focus will be placed on explainable and uncertainty-aware artificial intelligence. Techniques from the emerging field of geospatial XAI (e.g., spatially explicit GeoAI with SHAP visualisations) provide means to interpret model outputs in geographical contexts. For example, Roussel (2024) presents visualisation of Shapley values over geographic maps to locate spatial clusters of feature importance in GeoAI workflows. In parallel, probabilistic and Bayesian inference methods should be explored to quantify elevation and segmentation uncertainties and to propagate these through subsequent spatial and temporal analyses. Deep Gaussian process networks and their uncertainty-propagation capabilities have recently been used to model spatial interpolation with full distributional output instead of point estimates (Gonçalves and Wellmann, 2025; Van der Lende et al., 2025). This dual focus on interpretability and uncertainty will enable the identification of systematic biases and temporal inconsistencies within both DEM and OSM datasets.

Thirdly, advanced statistical and algorithmic solutions will be developed for spatio-temporal DEM analysis. Techniques such as Gaussian process regression, hierarchical Bayesian models, and spatial autoregressive formulations will be employed to model terrain evolution and structural changes in the built environment. Studies quantifying mixed uncertainty in DEM-derived products (e.g., slope calculations) underscore the need for rigorous uncertainty modelling in elevation analysis (Zhang, 2025). These methods will allow for more robust statistical inference on temporal dynamics and enhance the reliability of DEM-derived urban and environmental indicators. Classical methods such as Monte Carlo–based DEM noise modelling also remain valuable for evaluating elevation uncertainty under variable input accuracies (Gatziolis, 2004).

Overall, the integration of artificial intelligence, advanced statistical inference, and spatio-temporal database technologies will enable a new generation of analytical tools for DEM-based Earth observation.

Acknowledgement

We sincerely thank the LGL (Landesamt für Geoinformation und Landentwicklung Baden-Württemberg, Karlsruhe, Germany) for providing the DEM datasets and for accompanying the DEM project. Map data were obtained from OpenStreetMap contributors https://www.openstreetmap.org/copyright.

References

Amini, M.R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E. and Maximov, Y., 2025. Self-training: A survey. *Neurocomputing*, 616, 128904. doi.org/ 10.1016/j.neucom.2024.128904.

Buyukdemircioglu, M., Can, R., KoScaman, S. and Kada, M., 2022. Deep learning based building footprint extraction from very high resolution true orthophotos and nDSM. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, *2*, pp.211-218. doi.org/10.5194/isprs-annals-V-2-2022-211-2022.

Devillers, R. and Jeansoulin, R. eds., 2006. *Fundamentals of spatial data quality* (No. ISBN: 1905209568). London: ISTE. doi.org/10.1002/9780470612156.

Doyle, F.J., 1978. Digital Terrain Models: An Overview. *Photogrammetric Engineering and Remote Sensing*, Vol. 44, No. 12, December 1978, pp. 1481-1485.

Gatziolis, D., 2004. Adding Gaussian noise to inaccurate digital elevation models: A Monte Carlo evaluation, *Water Resources Research*, vol. 40, no. 12, W12202, 2004. doi.org/ 10.1029/2002WR001735.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography, *GeoJournal* 69, Springer Nature, pp. 211-221. doi.org/10.1007/s10708-007-9111-y.

Gonçalves, Í. and F. Wellmann, 2025. Uncertainty propagation in deep Gaussian process networks, *Mathematical Geosciences*, 2025. doi.org/10.1007/s11004-025-10187-4.

Guptill, S.C., Morrison, J., 1995. Elements of Spatial Data Quality, Elsevier Science, New York, 202p. doi.org/10.1016/C2009-0-14900-0.

Helbich, M., Amelunxen, C., Neis, P., Zipf, A., 2012. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. In *Proceedings of GI_Forum 2012*, Salzburg, Austria, pp. 24-33.

Kamp, N., Krenn, P., Avian, M., Sass, O., 2023. Comparability of multi-temporal DTMs derived from different LiDAR platforms: Error sources and uncertainties in the application of geomorphic impact studies. *Earth Surface Processes and Landforms*. 2023, May, 48(6), pp. 1152-1175. doi.org/10.1002/esp.5540.

Kaur, J., Singh, J., Sehra, S.S., Rai, H.S., 2017. Systematic Literature Review of Data Quality within OpenStreetMap, Int. Conference on Next Generation Computing and Information Systems (ICNGCIS). doi.org/ 10.1109/icngcis.2017.35.

Khalid, N., 2022. Review on region-based segmentation using watershed and region growing techniques and their applications in different fields. *Journal La Multiapp*, *3*(5), pp.241-249. doi.org/10.37899/journallamultiapp.v3i5.714.

Li, Z., Zhu, C. and Gold, C., 2004. *Digital Terrain Modeling: Principles and Methodology* (1st ed.). CRC Press, Boca Raton, London, New York, Washington, D.C., 318p. doi.org/10.1201/9780203357132.

Liu, R., Kuper, P., Grombein, Th., Naab, C., Al-Doori, M., Breunig, M., 2025. Towards Integrated Data Management and Analysis for Spatio-Temporal Digital Terrain Models. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 543-550. doi.org/10.5194/isprs-annals-X-G-2025-543-2025.

Martini, A., Kuper, P.V., Breunig, M., 2019. Database-Supported Change Analysis and Quality Evaluation of OpenStreetMap Data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume IV-2/W5, 2019 ISPRS Geospatial Week 2019, 10–14 June 2019, Enschede, The Netherlands, Best Paper Award, pp. 535–541, doi.org/10.5194/isprs-annals-IV-2-W5-535-2019.

Okolie, C.J., Smit, J.L., 2022. A systematic review and meta-analysis of Digital Elevation Model (DEM) fusion: pre-processing, methods and applications. ISPRS Journal of Photorgrammetry and Remote Sensing, Vol. 188, June 2022, pp. 1-29. doi.org/10.1016/j.isprsjprs.2022.03.016.

OSM, 2025. OpenStreetMap Foundation – https://osmfoundation.org.

Qiu, K., Wagenbach, L., Bulatov, D., Iwaszczuk, D., 2025. Improving self-supervised segmentation of urban scenes using NDSM-enhanced CRF. In: Land Surface and Cryosphere Remote Sensing V, Vol. 13263, pp. 35-44. SPIE. doi.org/10.1117/12.3040373.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. and Prabhat, F., 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), pp.195-204. doi.org/10.1038/s41586-019-0912-1.

Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234-241. Cham: Springer international publishing. doi.org/10.1007/978-3-319-24574-4_28.

Roussel, C., 2024. Visualization of explainable artificial intelligence for GeoAI, *Frontiers in Computer Science*, vol. 6, art. 1414923, 2024. doi.org/10.3389/fcomp.2024.1414923.

Schott, M., Zell, A., Lautenbach, S., Sumbul, G., Schultz, M., Zipf, A., 2024. Analyzing and Improving the Quality and Fitness for Purpose of OpenStreetMap as Labels in Remote Sensing Applications, in: D. Burghardt, E. Demodova, D.A. Keim (eds.), Volunteered Geographic Information – Interpretation, Visualization and Social Context, Springer Nature, pp. 21-42. doi.org/10.1007/978-3-031-35374-1_2.

Tseng, C.-M., Lin, C.-W., Stark, C.P., Liu, J.-K., Fei, L.-Y., Hsieh, Y.-C., 2013. Application of a multi-temporal, LiDAR-derived digital terrain model in a landslide-volume estimation. Earth Surf. Process. Landforms 38, Wiley & Sons, Ltd., pp. 1587-1601. doi.org/10.1002/esp.3454.

Van der Lende, M., Ferrao, J.L. and N. Müller-Hof, 2025. Evaluating uncertainty in deep Gaussian processes, *arXiv preprint*, arXiv:2504.17719, 2025. https://arxiv.org/abs/2504.17719.

Vats, A., Völgyes, D., Vermeer, M., Pedersen, M., Raja, K., Fantin, D.S. and Hay, J.A., 2024. Terrain-informed self-supervised learning: Enhancing building footprint extraction from LiDAR data with limited annotations. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, pp.1-10. doi.org/10.1109/TGRS.2024.3391391.

Weibel, R., Heller, M., 1991. Digital terrain modelling. Geographic Information Systems, Principles and Applications, Longman, pp. 269-297.

Zhang, J., 2025. Quantifying the mixed uncertainty for calculating slope from DEMs, *Environmental Modelling & Software*, vol. 184, art. 106032, 2025. doi.org/ 10.1080/10106049.2025.2480307.

Zhou, Q., 2017. Digital Elevation Model and Digital Surface Model, *The International Encyclopedia of Geography*, John Wiley & Sons, Ltd. doi.org/10.1002/9781118786352.wbieg0768.

Zhou, Z., 2018. A brief introduction to weakly supervised learning, *National Science Review*, Volume 5, Issue 1, pp. 44–53. doi.org/10.1093/nsr/nwx106.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. and Liang, J., 2018, September. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *International workshop on deep learning in medical image analysis* (pp. 3-11). Cham: Springer International Publishing. doi.org/10.1007/978-3-030-00889-5_1.