Deep Learning-Based Building Detection Using High-Resolution RGBI Orthophotos and DSMs

Mohamed Fawzy[1,2], Attila Juhasz[1], Arpad Barsi[1]

[1] Department of Photogrammetry and Geoinformatics, Faculty of Civil Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary,
{mohamed.fawzy, juhasz.attila, barsi.arpad}@emk.bme.hu

[2] Civil Engineering Department, Faculty of Engineering, Qena University, 83523 Qena, Egypt,
mohamedfawzy@eng.svu.edu.eg

Keywords: U-Net, Deep Learning, Building Detection, Digital Surface Models, High-Resolution RGBI Orthophotos.

Abstract

Deep learning techniques have demonstrated a promising efficacy for building feature extraction, presenting practical strategies to lessen the labour-intensive work of map updating, change detection, and urban growth monitoring. To address the labour-consuming challenges, a U-Net-based framework is developed to generate building maps automatically using high-resolution RGBI orthophoto and DSM data. The approach shows the effectiveness of the U-Net-based semantic segmentation for urban scene analysis. The presented procedures collect, preprocess, and combine orthophoto with DSM to train, apply, and assess the U-Net model for building extraction in urban environments using two input scenarios: (1) solely RGBI orthophoto and (2) RGBI orthophoto integrated with DSM. Four standard metrics: completeness, correctness, quality, and overall accuracy are applied to evaluate the model outputs, comparing the single orthophoto input to the combined orthophoto with DSM for building detection. The significant impact of the DSM and RGBI pairing is demonstrated by the heightened reliability of the data integration strategy when estimating buildings within nearby similar objects like roads and impervious surfaces. However, a few challenges related to the model's generalisation are noticed across complex urban contexts, including tree occlusions, unreferenced building extensions, and height irregularities surrounding structures. The findings highlight the potential of multimodal data fusion in urban investigations and reveal how it can improve the built-up asset mapping. Final results argue that DSM incorporation significantly enhances building classification performance using deep learning frameworks for geospatial applications, particularly in complex urban environments where single data and traditional image-based segmentation methods are inadequate.

1. Introduction

Due to the recent advancements in data resolution, satellite data are widely employed for analysing urban features. Remote sensing data resolution has greatly improved; submeter-level spatial resolution, spectral resolution covering tens or hundreds of bands, radiometric resolution of 16-bit or higher, and daily or sub-daily temporal resolution are now feasible (Fawzy, Szabó et al., 2023). Image resolution has the greatest influence on the efficiency of metropolitan remote sensing-based processes such as monitoring urban activities, infrastructure development, and environmental 2D/3D mapping. Accurate feature extraction plays a pivotal role in various applications, including change detection, city planning, land use analysis, development strategies, and sustainable urban management. In addition, accurate building detection can serve as a basis for natural disaster management, urban environment modelling, or simulation in smart cities or digital twins (Deren, Wenbo et al., 2021). With the rapid advances of high-quality sensors, integrating both 2D and 3D datasets has become increasingly attainable for automated urban investigations. Previous studies have demonstrated that Very High Resolution (VHR) optical imagery provides rich spectral details, while Digital Surface Models (DSMs) offer essential elevation information to enhance object differentiation (Jin, Li et al., 2025). However, deep learning methods are often highly effective, relying only on RGB imagery restricts their ability to capture important height-related details in urban environments. To overcome this limitation, some approaches integrate additional data sources, such as DSM (Fawzy, Juhasz et al., 2025-a), LiDAR (Huang, Zhang et al., 2019, Fawzy, Dowajy et al., 2024), and synthetic aperture radar (Sun, Hua et al., 2021). On the other hand, using multi-source data for building footprint segmentation introduces several challenges that must be taken into account, including higher data acquisition costs, complex data alignment requirements, and the need for extensive algorithm calibration (Dabove, Daud et al., 2024). By automating feature extraction techniques, Deep Learning (DL) has revolutionised remote sensing applications in urban regions. Developing DL systems for object detection and change recognition of buildings, roads, and other features is a major subject of recent research. Multiple neural network-based strategies have been recently introduced for image analysis, such as CNN, FCN, SegNet, DeepLabv3+, and PSPNet (Choure and Prajapat, 2024). U-Net is a kind of convolutional neural network mainly used for image interpretation. Originally, U-Net was used in medical imaging as it works well even with very little labelled data (Ronneberger, Fischer et al., 2015). U-Net is also widely applied for object extraction in manmade environments, like building or road detection. Among various architectures, U-Net has proven to be highly effective for image semantic segmentation due to its encoder-decoder design and skip connections, which preserve spatial context (Fawzy and Barsi, 2024). The U-Net model has already been successfully employed in urban feature extraction using

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-673-2026

multi-spectral and LiDAR data. Building on prior research on urban land cover classification and road detection (Patel and Koringa, 2023, Fawzy, Juhasz et al., 2025-a, Fawzy, Juhasz et al., 2025-b), the current work evaluates the synergistic value of combining spectral and elevation properties to delineate building footprints accurately in complex urban environments. The approach aligns with broader trends in the literature emphasising multimodal data fusion to enhance feature extraction reliability (Sirko, Kashubin et al., 2021, Feng, Xu et al., 2023, Yan, Jiang et al., 2024). This study presents an integrated approach for automatic building map generation using high-resolution RGBI orthophotos and DSM data using a U-Net-based convolutional neural network. The given work contributes to the current state of research by (1) providing a controlled comparison between RGBI orthophoto-based and RGBI+DSM-based building detection using high-resolution airborne data, (2) analyzing the impact of elevation data across multiple urban configurations with varying complexity, and (3) presenting a detailed investigation of misclassification sources such as vegetation proximity, shadow effects, and elevation-induced ambiguities. Unlike many large-scale studies focusing on satellite imagery, this work emphasizes high-resolution orthophoto applications and evaluates the practical benefits and limitations of multimodal data fusion in urban environments.

2. Literature Review

Remote sensing development and increasing access to enhanced-resolution data enable large-scale urban investigations. Focusing on sophisticated satellites, numerous free and commercial datasets are released with submeter resolution images. Besides satellite imagery, orthophotos are commonly used for feature detection in urban settings where higher resolution of airborne images enables more precise building identification (Fan, Ding et al., 2021). Over the past decade, Deep learning (DL) technologies have made significant advances for remote sensing data analysis applications, involving scene classification, image semantic segmentation, and object detection. Different deep learning techniques were developed for building detection, including transfer learning, multi-task learning and transformer/attention mechanism. The National Land Survey of Finland (NLS) started the Advanced Technology for National Topographic Map Updating (ATMU) project with the goals of developing deep learning tools for object detection and change recognition, reducing manual work in GIS production, improving the accuracy and timeliness of geospatial data, supporting map production innovation, and providing high-quality training datasets for society (Zhu, Raninen et al., 2024). The project's results demonstrated a 97.9% accuracy in detecting building changes (Lingli Zhu, 2023). (Sirko, Kashubin et al., 2021) used approximately 100k images with 50 cm resolution to develop an improved U-Net model by investigating the variations in architecture, loss functions, regularisation, pre-training, self-training and post-processing. The resulting pipeline was applied for generating an open buildings dataset of 516M Africa-wide detected building footprints. (Patel and Koringa, 2023) described a training process of U-Net and SegNet with satellite images, which was tuned with various hyperparameters such as learning rate, batch size and epochs. (Aghayari, Hadavand et al., 2023) presented research, where deep network architectures, U-Net and Inception ResNet U-Net (IRNUN), were implemented and evaluated in automatic building detection from aerial imagery. In comparison, the IRNUN could detect buildings of different shapes, structures, textures, and colours in images in almost all regions, though the U-Net couldn't detect very large buildings. That is because the IRNUN model is wide and deep with few variations in the number of parameters. Segmentation processes of urban features can leverage not only images but also elevation data as effective supplementary information. (Jin, Li et al., 2025) introduced a novel approach for extracting building outlines from Digital Surface Models (DSMs) and orthophotographs. Distance-Constrained Clustering (DCC) and Gradient-based Optimal Contour Selection (G-OCS) algorithms were described to cluster contour lines and to select building outlines. The DSM provided initial contour lines, while the orthophotograph indicates where vegetation is obstructing the building regions. (Saito and Aoki, 2015) used Convolutional Neural Networks (CNN) to learn mapping from raw pixel values in aerial imagery to three object labels (buildings, roads, and others). A patch-based semantic segmentation approach was taken, so firstly a large aerial imagery was divided into small patches and then the CNN was trained with those patches and corresponding three-channel map patches. Using satellite data, (Maurya, Mittal et al., 2023) applied a deep learning-based segmentation model for urban images captured from satellites and produced five categories of cultivated land, uncultivated land, residences, water, and forest. Furthermore, several researchers have recognised the potential of using free but lower-resolution satellite images. Many of them deal with the production and use of so-called super-resolution images in this field. (Feng, Xu et al., 2023) proposed an improved model for feature super-resolution semantic segmentation based on a neural network (EDSR_NASUnet) to show that high-resolution building footprints can be extracted from medium-resolution satellite (Sentinel-2) images. The model was used to create the first nation-scaled building extraction in China with more than 86 million buildings. Therefore, both satellite and airborne images can be used as inputs for this semantic segmentation process. Recent advances in state-of-the-art building detection increasingly rely on transformer-based architectures and attention mechanisms, which outperform conventional CNNs in capturing long-range spatial dependencies (Yan, Zhao et al., 2024, Yan, Cao et al., 2025). Additionally, domain generalization and cross-area validation have become essential to assess model robustness beyond training regions. While U-Net remains competitive due to its efficiency and lower computational demand, it may struggle to generalize across diverse urban morphologies without additional training strategies such as transfer learning or domain adaptation.

3. Methodology

Compared to previous works focused on VHR satellite imagery (Fawzy, Juhasz et al., 2025-a), the use of airborne orthophotos provides finer spatial detail, enabling the identification of small and complex building structures; in addition, DSMs add elevation information for better scene analysis. Given the research objectives, the presented methodology (Figure 1) starts with a preprocessing phase to present the orthophotos and DSMs in an appropriate format to be combined as inputs for the model. The orthophoto and the DSM were co-registered as part of the preprocessing to ensure spatial alignment. Both datasets were merged into a five-band composite (R, G, B, NIR, and elevation). The combined data is used to create a training dataset with the relevant classes and labels. Four land cover classes: buildings, roads, vegetation, and others, were manually labelled using the MATLAB Image Labeler tool. The collected training dataset is used to train, assess, and validate a U-Net model for image classification in urban environments. The developed U-Net model adheres to a standard encoder-decoder layout with skip connections between corresponding layers (Ronneberger, Fischer et al., 2015). Following the model development and training, the entire scene is classified into four land cover classes using the single RGBI photo and the combined DSM with RGBI data. The classified images are eventually binarised, where building features are detected and assessed in order to compare the single orthophoto findings to the RGBI orthophoto integrated with DSM. It should be noted that the training, validation, and testing samples are derived from the same region. Although different sub-areas were used, this may limit the evaluation of the model's generalization capability across unseen urban environments where relevant training data is needed.

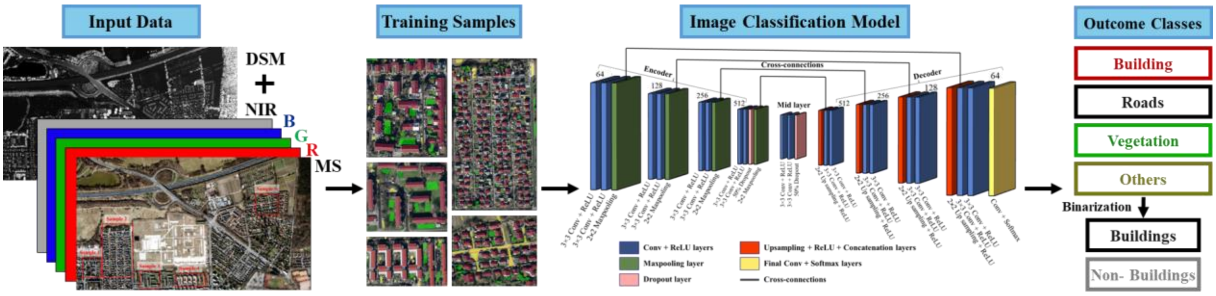

Figure 1. Procedures of the presented methodology.

4. Experimental Works

4.1. Study Area and Data Used

To apply the developed model, an appropriate study area is selected from Germany, Hannover, encompassing a densely populated area with a variety of urban surface materials and building structures. The State Office for Geoinformation and Land Surveying of Lower Saxony provided the orthophotos and the DSM, along with other common and useful spatial data sources like digital terrain models, 3D building models (LoD1 and LoD2), house data, land use, topographic, cadastral, and overview maps of the German state. To guarantee high-quality input data, the RGBI orthophoto is sourced with a 20 cm ground pixel size, in addition to a DSM with 1 m spatial resolution produced from aerial LiDAR data. Both datasets have the same Coordinate Reference System (CRS) and spatial details to facilitate the direct integration (LGLN, 2025). Combining DSMs with optical data has been observed as a successful strategy for enhancing urban mapping quality (Jahan, Zhou et al., 2018, Al-Najjar, Kalantar et al., 2019). Multispectral orthophotos and DSMs significantly enhance land cover semantic segmentation, particularly in distinguishing raised buildings from plain surfaces (Yiğit and Uysal, 2020). The study area is located near the Bothfeld area in the north-east of Hannover, Germany (52°25'29.34" N, 9°47'8.46" E). Five areas of interest, representing different levels of built-up complexity, are selected (Samples 1-5) to apply the model and evaluate its robustness. The scene shows 4 main classes: buildings, roads, vegetation, and others (Figure 2). The geometric accuracy of the orthophoto dataset is within sub-pixel level, consistent with its 20 cm ground sampling distance. The dataset is generated through aerial photogrammetric processing with accurate ground control points, ensuring high positional accuracy suitable for detailed urban mapping applications. This level of geometric precision supports reliable alignment with DSM data and reference building layers. The dataset was divided into training, validation, and testing subsets using spatially distinct patches to minimize data leakage; however, all samples originate from the same urban region.

Figure 2. The study area samples, Hannover, Germany.

4.2. Reference Data

Reliable information is collected as a benchmark for assessing model performance, allowing for quantitative evaluations of building feature extraction accuracy. Reference building polygons are obtained from the Open Street Map (OSM) database. OSM layers are compatible with web maps, GIS software, analysis tools, and three-dimensional visualisation of footprint shape, height, type, and other attributes (Geofabrik, 2025). The OSM building layers are first obtained in a shapefile format with a height attribute, and afterwards transformed into a geo-raster in order to align the classified building images (Figure 3-a).

4.3. Sample Selection and Data Labelling

Accurate, pure, representative, and uniformly distributed training samples are required for optimal classification outcomes (Fawzy, Khodary et al., 2021). In deep learning, comprehensive training data leads to better model performance and increases the ability to distinguish between various classes (Kim, Baek et al., 2025). For interactive data labelling, the MATLAB Image Labeler tool provides a straightforward window to pick appropriate samples with their related classes (MathWorks, 2025). Five sample images are carefully labelled, ensuring a representative distribution of all classes for training, testing, and validating the developed model (Figure 3-b).

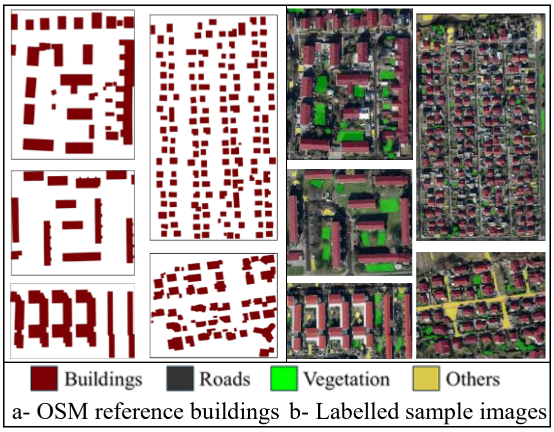

Figure 3. Reference and labeled data.

4.4. U-Net for Image Classification

4.4.1. U-Net Model Architecture

In order to improve the classification quality, the applied U-Net model integrates an encoder-decoder design with skip connections between relevant layers. The linkages between the encoder and decoder blocks improve the segmentation accuracy and preserve spatial information. Each encoder block contains four blocks of two 3×3 convolutional layers with ReLU activations followed by 2×2 max-pooling layers. The decoder mirrors this structure with additional upsampling layers. The bottleneck, in the middle of the U-Net, is made up of two 3×3 convolutional layers with ReLU activations, along with a dropout layer to promote generalisation. The dropout layer (50%) is included in the final encoder and bottleneck layers to mitigate overfitting. The segmented image is generated using a 1×1 convolutional layer with a softmax activation function, giving each pixel in the input layer a class probability.

4.4.2. U-Net Model Training

For accurate predictions, besides the right neural network, the quantity, quality, and variety of training data, as well as the appropriate parameters, must be optimised. A randomly generated distribution is determined for the starting weights and biases of the convolutional layers. Next, the model is trained using Stochastic Gradient Descent with Momentum (SGDM) to accelerate the convergence by incorporating momentum in weight updates with a learning rate of 0.05 and a momentum value of 0.9. The training dataset comprised 16,000 image patches (128×128 pixels) and 2,000 validation patches. The cross-entropy loss function guided optimisation, while accuracy and loss metrics were monitored to preserve a stable optimisation process while enabling effective processing. The model is trained through 150,000 iterations, divided into 30 epochs of 5,000 iterations. The cross-entropy loss function is applied to measure the difference between the ground truth and predicted labels, serving as a guide for parameter adjustments during the training cycles. Two models were trained and applied, through MATLAB R2023b environment, for comparison:

- **Model A:** RGBI orthophoto (4 band input).
- **Model B:** orthophoto + DSM (5 band input).

The models are trained and applied using the following computational capabilities, aiming to bridge the gap between deep learning and systems with limited resources:

- **Processor:** Intel(R) Core (TM) i5-9600K CPU @ 3.70 GHz.
- **Memory:** 16.00 GB RAM @ 2400 MHz.
- **GPU:** NVIDIA GeForce GTX 1650 (4 GB VRAM).

The training progress is monitored by regularly evaluating the accuracy and loss curves for both the training and validation sets (Figures 4, 5). Significant learning efficiency and robustness in semantic segmentation tasks were demonstrated by the trained models, with validation accuracy of **96.31%** for the RGBI orthophoto model through 3,166 minutes and **98.06%** for the RGBI orthophoto integrated with the DSM model through 3,200 minutes.

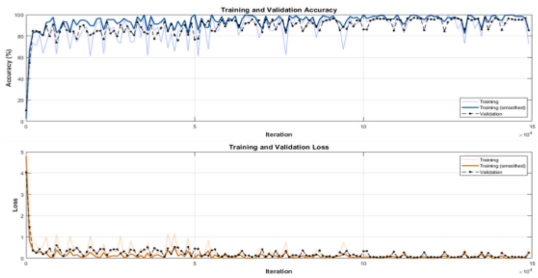

Figure 4. Training accuracy and loss progress for model A: RGBI orthophoto.

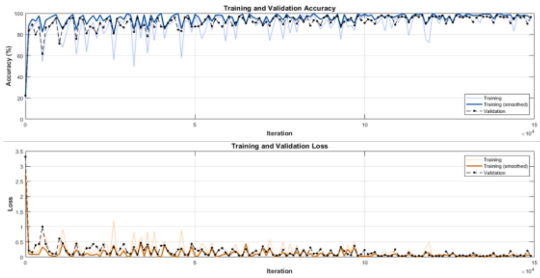

Figure 5. Training accuracy and loss progress for model B: RGBI orthophoto with DSM integration.

5. Results and Discussions

5.1. Image Classification Procedures

Classification outcomes of the two pre-trained U-Net models are obtained using RGBI orthophoto alone and RGBI orthophoto combined with DSM. Scenes are classified into four classes: buildings, roads, vegetation, and others (Figure 6).

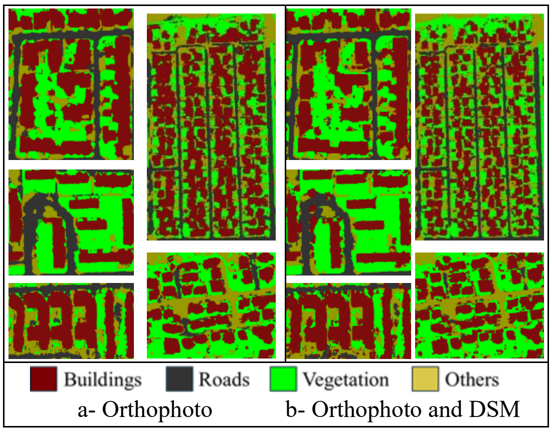

Figure 6. Classified images using the U-Net models.

Building surface classification, however, presents several challenges like the un-referenced building extensions, the concrete grounds within building areas, and the pavement and parking spaces established by building materials, leading to misclassifications. Despite these limitations, the findings provide a reasonable segmentation of urban features due to the integration of DSM, particularly in differentiating between buildings and similar features. The elevation information helps to resolve the spectral confusion between buildings and surrounding impervious features, reducing inaccurate outcomes. The DSM information improves the consistency of the detected building objects in complex urban scenarios where spectral similarities previously hindered accuracy. To enable a reliable accuracy assessment, the classified images are presented as building binary masks (Figure 7) for comparison with the reference information (Figure 3-a).

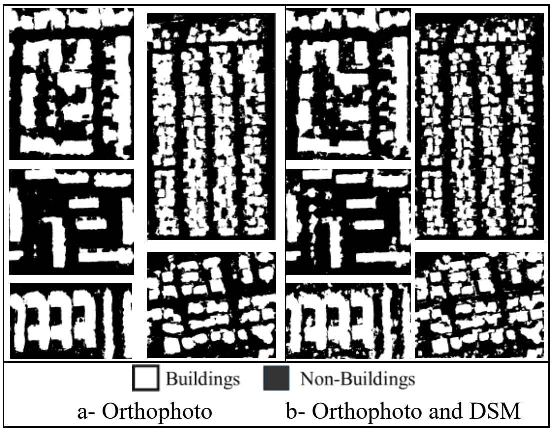

Figure 7. Extracted buildings using the U-Net models.

5.2. Accuracy Assessment

The inclusion of elevation data significantly reduced the classification uncertainty between buildings and other impervious surfaces such as concrete landscapes, soil, roads, and parking areas. This improvement can be illustrated by the clearer delineation of building boundaries, particularly in densely built regions where shadows and tree occlusions hinder segmentation accuracy. Consequently, the U-Net model of RGBI orthophoto with DSM integration exhibited superior segmentation accuracy qualitatively and quantitatively compared to the orthophoto-only configuration. Qualitative evaluation reveals that the identified buildings in the classified orthophoto lack accuracy and experience false detections when non-building surfaces (e.g., road, other classes) are presented as buildings. Meanwhile, integrating the DSM with the orthophoto improves the results with notable increases in accuracy, which is in line with earlier research on building extraction (Fawzy and Barsi, 2024, Fawzy, Dowajy et al., 2024, Fawzy, Juhasz et al., 2025-a). Additionally, incorporating DSMs as a fifth input channel improved model generalisation and robustness against spectral ambiguities caused by roof material and colour variability. The outcomes corroborate studies by (Jin, Li et al., 2025) that more accurate building outlines were obtained by combining orthophotos and DSM-derived contours. Quantitative assessment is employed using a confusion matrix through four standard metrics: completeness (Eq. 1), correctness (Eq. 2), quality (Eq. 3), and overall accuracy (Eq. 4) (Rutzinger, Rottensteiner et al., 2009).

$$\text{Completeness} = TP/(TP+FN) \tag{1}$$

$$\text{Correctness} = TP/(TP+FP) \tag{2}$$

$$\text{Quality} = TP/(TP+FP+FN) \tag{3}$$

$$\text{Overall accuracy} = (TP+TN)/(TP+FP+FN+TN) \tag{4}$$

The confusion matrix is formed based on True Positives (TP), representing the number of building pixels correctly classified; True Negatives (TN), indicating the number of non-building pixels successfully identified; False Positives (FP), showing non-building pixels incorrectly classified as building; and False Negatives (FN), referring to building pixels incorrectly assigned to non-building class. The metrics of the RGBI orthophoto with DSM classification and the RGBI orthophoto-only classification's values are shown in (Table 1), while building differences can be seen in (Figure 8).

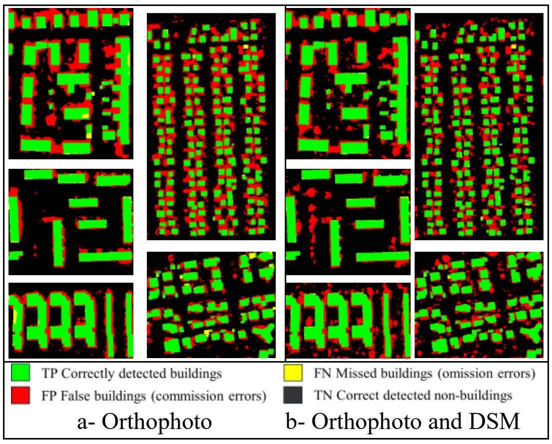

Figure 8. Building differences using the U-Net models.

Table 1: Metrics of the orthophoto with the DSM model compared to the RGBI orthophoto model.

	TP (pixels)	TN (pixels)	FP (pixels)	FN (pixels)	Completeness	Correctness	Quality	Overall accuracy
RGBI&DSM-S1	154126	331028	100565	1357	0.99	0.61	0.60	0.83
RGBI&DSM-S2	553301	1873844	571064	13695	0.98	0.49	0.49	0.81
RGBI&DSM-S3	123665	373363	52362	4212	0.97	0.70	0.69	0.90
RGBI&DSM-S4	168594	224367	99879	1368	0.99	0.63	0.62	0.80
RGBI&DSM -S5	193732	424674	121957	1063	0.99	0.61	0.61	0.83
RGBI-S1	151139	309368	122225	4344	0.97	0.55	0.54	0.78
RGBI-S2	557654	1697926	746982	9342	0.98	0.43	0.42	0.75
RGBI-S3	124640	375730	49995	3237	0.97	0.71	0.70	0.90
RGBI-S4	166197	249248	74998	3765	0.98	0.69	0.68	0.84
RGBI-S5	186687	446791	99840	8108	0.96	0.65	0.63	0.85

Compared to the reference dataset, building objects retrieved from the orthophoto classification only show a few differences, like misclassified road pixels among the building surfaces, that are addressed by the orthophoto integration with DSM (Figure 9).

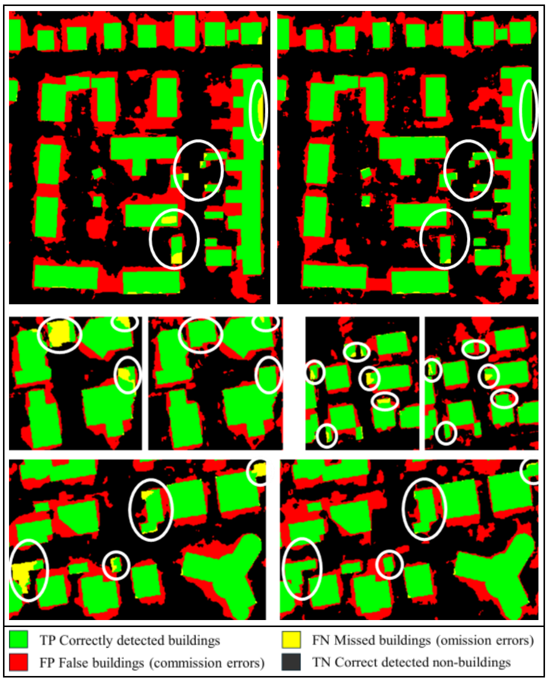

Figure 9. Detected building differences of orthophoto only (left), and orthophoto and DSM (right).

The detected buildings using orthophoto and DSM showed more accurate measurements with fewer disruptions and incorrect classifications (Figure 10). However, misclassifications remain evident and can be grouped into three main categories, as illustrated in Figures 8–10: (1) vegetation-related errors, where trees adjacent to buildings exhibit similar structural characteristics; (2) shadow-induced errors, where low reflectance regions obscure building boundaries and reduce spectral separability; and (3) elevation-induced ambiguities, where DSM values of tall vegetation or irregular terrain resemble building heights. The shadow effect significantly impacts orthophoto-based classification, as shaded building regions often resemble roads or background classes due to reduced spectral contrast (Figure 9). While DSM integration mitigates some of these issues by introducing height information, it may also introduce new errors when vegetation exhibits elevation values similar to buildings. Trees located near buildings often lead to over-segmentation, as their canopy height is interpreted as part of the building structure (Figure 10). Consequently, DSM improves building detection in open and clearly structured urban areas (Samples 1–3) but may affect accuracy in vegetation-dense environments (Samples 4 and 5), where both spectral and elevation features overlap and increase the likelihood of false positives. Despite the challenges of the concrete impervious surfaces, the surrounding trees, and the irregular building extensions, the DSM-enhanced extraction aligns more closely with the reference datasets. The RGBI with DSM model demonstrates the effectiveness of integrating elevation data for complete and accurate building objects (Samples 1-3); however, the model still struggles to be generalised throughout various urban environments with tree occlusions and height inconsistencies close to intricate structures. The accuracy metrics of the two models are presented in (Figure 11), highlighting several keys:

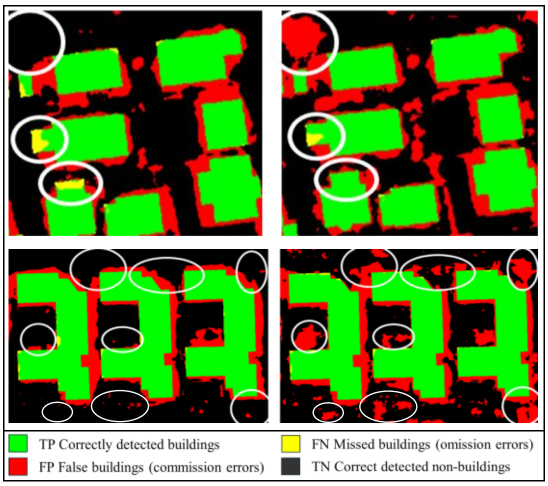

Figure 10. Misclassification caused by vegetation proximity and shadow effects on orthophoto only (left), and orthophoto and DSM (right) outcomes.

Completeness: The DSM, along with the RGBI, strategy achieves higher completeness because the elevation information helps detect most building pixels accurately with fewer missed buildings (lower FN). In contrast, the RGBI-only approach misses more building pixels, as similar spectral signatures can cause certain buildings, especially low or ambiguous ones, to remain undetected.

Correctness: The integrated approach exhibits better correctness, demonstrating that buildings are more likely to be accurate when considering elevation information. Meanwhile, spectral data alone can result in confusion in class differentiation, reducing the correctness, as evidenced by the increasing number of false alarms (non-building pixels mistakenly classified as building) in the RGBI-only classification. However, FP, which is impacted by incorrectly identified nearby trees and un-referenced building extensions, is crucial to correctness and quality. Samples 3-5 show a slight decline in correctness and quality measurements of the integrated orthophoto and DSM scenario, which is not related to the model defect. Even though it could be the consequence of overweighting the elevation data due to the complicated surroundings where building labels dominated the training samples relative to road, vegetation, and other classes. On the other hand, although elevation information improves the detection of actual building pixels, in some cases (Sample 4, 5), it also increases the likelihood of producing false positives, where certain non-building areas are classified as buildings. In several cases, the DSM input appears to be overweighted, particularly in complex urban surroundings where small height variations or local topographic noise resemble building-like elevation patterns. This effect contributes to elevated FP counts in the integrated classification. In addition, the training dataset is dominated by building samples, which may have biased the model toward predicting the building class more frequently. This class imbalance further amplifies false-positive outcomes when ambiguous features – such as narrow roads, vegetation patches, or unreferenced building extensions – display spectral or elevation characteristics similar to buildings. Consequently, the reduced correctness values of the RGBI&DSM model do not necessarily indicate poorer performance but rather highlight the interaction between elevation data sensitivity and class-imbalance effects.

Quality: The quality metric measures the balance between completeness and correctness, reflecting the significant superiority of the DSM and RGBI case when compared to the sole orthophoto for building extraction in built-up areas.

Overall Accuracy: The better overall accuracy of the integrated model versus the model without DSM validates the importance of combining DSM data with orthophotos in challenging classification tasks, particularly for scenarios with buildings of comparable surrounding features.

Although DSM integration consistently improves all evaluation metrics, the observed gains remain moderate. This can be attributed to the already high quality and spatial resolution of the RGBI orthophotos, which provide strong spectral and textural information for building detection. Consequently, the additional elevation information offers only incremental benefits. Furthermore, DSM data may introduce noise and ambiguities, particularly in areas with vegetation or complex roof structures, limiting the magnitude of improvement. This highlights that while multimodal fusion is beneficial, its impact depends on the baseline data quality and scene complexity.

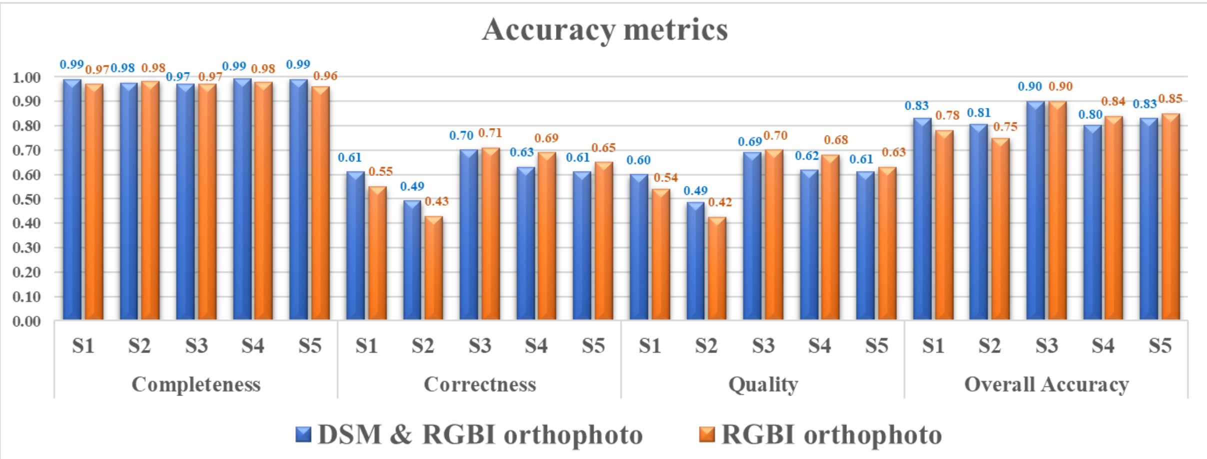

Figure 11. Accuracy metrics of the extracted buildings by RGBI orthophoto with DSM versus the solely RGBI orthophoto.

6. Conclusions and Future Work

The applied approach developed a deep learning-based U-Net semantic segmentation model for urban building detection using high-resolution RGBI orthophotos and DSM data. Findings proved that providing the U-Net framework with high-resolution orthophotos, which are rich in spectral information, and DSMs, which are valuable with elevation data, significantly improved the building footprint extraction. The applied methodology extends previous work on building detection and urban land cover classification by employing the multi-source data fusion paradigm for building mapping. Despite the promising results, there are still challenges with the model generalising over various urban environments with tree occlusions and height irregularities close to complex structures. Future studies will look into optimised deep learning architectures for multimodal data integration and multi-temporal analysis to strengthen urban feature extraction. As well, it is anticipated that normalised difference indices, like the NDVI, NDBI, etc., will enhance the comprehension of urban land cover that requires verification (Fawzy, Mostafa et al., 2020). To increase the model's scalability across different urban environments with distinctive features, it is recommended to incorporate cutting-edge deep learning techniques, such as transformer-based architectures or attention mechanisms to overcome the challenges and limitations of U-Net performance. Further studies will include cross-city and cross-region validation using independent datasets to evaluate the robustness and transferability of the trained models. This will help assess the generalization capability of multimodal approaches under varying urban conditions.

Acknowledgement

The research reported in this paper is part of project ID 2022-2.1.1-NL-2022-00012, National Laboratory of Cooperative Technologies, implemented with the support provided by Ministry of Culture and Innovation through the National Research, Development and Innovation Fund. Mohamed Fawzy is funded by the Stipendium Hungaricum Scholarship under the joint executive programme between Hungary and Egypt.

References

Aghayari, S., A. Hadavand, S. Mohamadnezhad Niazi and M. Omidalizarandi (2023). "Building detection from aerial imagery using inception resnet unet and unet architectures." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 10: 9-17.

Al-Najjar, H. A., B. Kalantar, B. Pradhan, V. Saeidi, A. A. Halin, N. Ueda and S. Mansor (2019). "Land cover classification from fused DSM and UAV images using convolutional neural networks." Remote Sensing 11(12): 1461.

Choure, P. and S. Prajapat (2024). Exploring U-Net, FCN, SegNet, PSPNet, Mask R-CNN and Using DeepLabV3+ for multiclass semantic segmentation on satellite images of western ghats. The International Conference on Computing, Communication, Cybersecurity & AI, Springer.

Dabove, P., M. Daud and L. Olivotto (2024). "Revolutionizing urban mapping: deep learning and data fusion strategies for accurate building footprint segmentation." Scientific Reports 14(1): 13510.

Deren, L., Y. Wenbo and S. Zhenfeng (2021). "Smart city based on digital twins." Computational Urban Science 1(1): 4.

Fan, Y., X. Ding, J. Wu, J. Ge and Y. Li (2021). "High spatial-resolution classification of urban surfaces using a deep learning method." Building and Environment 200: 107949.

Fawzy, M. and A. Barsi (2024). "A U-Net Model for Urban Land Cover Classification Using VHR Satellite Images." Periodica Polytechnica Civil Engineering 69(1): 98–108.

Fawzy, M., M. Dowajy, T. Lovas and A. Barsi (2024). Urban Land Cover Classification Using Deep Neural Networks Based on VHR Multi-Spectral Image and Point Cloud Integration. IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium, IEEE.

Fawzy, M., A. Juhasz and A. Barsi (2025-a). "Urban Land Cover Classification Using Deep Neural Networks Based on VHR MS Image and DSM." Acta Polytechnica Hungarica 22(8): 115-135.

Fawzy, M., A. Juhasz and A. Barsi (2025-b). "High-Quality Road Detection Using U-Net-Based Semantic Segmentation with High-Resolution Orthophotos and DSM Data in Urban Environments." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48: 459-465.

Fawzy, M., F. Khodary and Y. G. Mostafa (2021). "Selecting Training Samples Automatically from VHR Satellite Images for Image Classification." Sohag Engineering Journal 1(1): 71-84.

Fawzy, M., Y. G. Mostafa and F. Khodary (2020). "Automatic Indices Based Classification Method for Map Updating Using VHR Satellite Images." JES. Journal of Engineering Sciences 48(5): 845-868.

Fawzy, M., G. Szabó and A. Barsi (2023). "A Shallow Neural Network Model for Urban Land Cover Classification Using VHR Satellite Image Features." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 10: 57-64.

Feng, L., P. Xu, H. Tang, Z. Liu and P. Hou (2023). "National-scale mapping of building footprints using feature super-resolution semantic segmentation of Sentinel-2 images." GIScience & Remote Sensing 60(1): 2196154.

Geofabrik. (2025). "OpenStreetMap Data Extracts." Geofabrik's free download server Retrieved 20 May, 2025, from https://download.geofabrik.de/.

Huang, J., X. Zhang, Q. Xin, Y. Sun and P. Zhang (2019). "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network." ISPRS journal of photogrammetry and remote sensing 151: 91-105.

Jahan, F., J. Zhou, M. Awrangjeb and Y. Gao (2018). "Fusion of hyperspectral and LiDAR data using discriminant correlation analysis for land cover classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11(10): 3905-3917.

Jin, F., X. Li, Y. Zhu, Z. Ou, Y. Ren, S. Peng, C. Li, W. Liu, E. Li and L. Zhang (2025). "Automated building outline extraction from digital surface models and orthophoto: a novel contour-based approach." International Journal of Geographical Information Science: 1-22.

Kim, J.-I., J.-W. Baek and C.-B. Kim (2025). "Hierarchical image classification using transfer learning to improve deep learning model performance for amazon parrots." Scientific Reports 15(1): 3790.

LGLN. (2025). "State Office for Geoinformation and Land Surveying of Lower Saxony." Retrieved 8-Febraury, 2025, from https://www.lgln.niedersachsen.de/startseite/vertrieb_support/geodaten_marktplatz/opengeodata/opengeodata-220509.html.

Lingli Zhu, J. R., Emilia Hattula (2023) "How the National Land Survey of Finland is exploring AI technology, The ATMU project explained."

MathWorks. (2025). "Get Started with the Image Labeler." Retrieved 10 January, 2025, from https://nl.mathworks.com/help/vision/ug/get-started-with-the-image-labeler.html.

Maurya, A., P. Mittal and R. Kumar (2023). "A modified u-net-based architecture for segmentation of satellite images on a novel dataset." Ecological Informatics 75: 102078.

Patel, M. J. and H. P. Koringa (2023). "Deep Learning Architecture U-Net Based Road Network Detection from Remote Sensing Images." International Journal of Next-Generation Computing 14(3).

Ronneberger, O., P. Fischer and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer.

Rutzinger, M., F. Rottensteiner and N. Pfeifer (2009). "A comparison of evaluation techniques for building extraction from airborne laser scanning." IEEE Journal of selected topics in applied earth observations and remote sensing 2(1): 11-20.

Saito, S. and Y. Aoki (2015). Building and road detection from large aerial imagery. Image Processing: Machine Vision Applications VIII, SPIE.

Sirko, W., S. Kashubin, M. Ritter, A. Annkah, Y. S. E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse and J. Quinn (2021). "Continental-scale building detection from high resolution satellite imagery." arXiv preprint arXiv:2107.12283.

Sun, Y., Y. Hua, L. Mou and X. X. Zhu (2021). "CG-Net: Conditional GIS-aware network for individual building segmentation in VHR SAR images." IEEE Transactions on Geoscience and Remote Sensing 60: 1-15.

Yan, P., J. Zhao, R. Hou, X. Duan, S. Cai and X. Wang (2024). "Clustered remote sensing target distribution detection aided by density-based spatial analysis." International Journal of Applied Earth Observation and Geoinformation 132: 104019.

Yan, W., L. Cao, P. Yan, C. Zhu and M. Wang (2025). "Remote sensing image change detection based on swin transformer and cross-attention mechanism." Earth Science Informatics 18(1): 106.

Yan, X., Z. Jiang, P. Luo, H. Wu, A. Dong, F. Mao, Z. Wang, H. Liu and Y. Yao (2024). "A multimodal data fusion model for accurate and interpretable urban land use mapping with uncertainty analysis." International Journal of Applied Earth Observation and Geoinformation 129: 103805.

Yiğit, A. Y. and M. Uysal (2020). "Automatic road detection from orthophoto images." Mersin Photogrammetry Journal 2(1): 10-17.

Zhu, L., J. Raninen and E. Hattula (2024). "GeoAI for Topographic Data Accuracy Enhancement: the AI4TDB project." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48: 221-228.