

Deep Learning Approaches for Building Extraction and Height Estimation from Satellite Imagery

Hande Ascioglu¹, Tolga Bakirman¹, Bulent Bayram¹

¹ Department of Geomatics Engineering, Yildiz Technical University, Davutpasa Campus, 34220 Esenler, Istanbul, Turkey, hande.ascioglu@std.yildiz.edu.tr, bakirman@yildiz.edu.tr, bayram@yildiz.edu.tr

Abstract

This study investigates a deep learning-based multi-task framework for simultaneous building extraction and building height estimation from satellite imagery. Accurate building footprint delineation and height information are essential for urban monitoring, disaster management, 3D city modelling, and spatial planning. A shared-encoder, dual-decoder architecture with a ResNet34 backbone was implemented and evaluated on the DFC2023 dataset, which includes RGB imagery, normalized DSM data, and building masks from 11 cities. The model was trained for binary building segmentation and pixel-wise height regression within a single framework. The results showed strong segmentation performance, with an Intersection over Union (IoU) of 0.8953 and a Dice score of 0.9227. For height estimation, the model achieved a Mean Absolute Error (MAE) of 4.36 m and an R^2 score of 0.46. While the segmentation results are promising, the height estimation performance is moderate and indicates room for improvement. These findings suggest that multi-task deep learning can effectively learn shared spatial representations, but further work is needed to enhance height prediction accuracy. The proposed approach offers a practical solution for urban remote sensing applications requiring both building outlines and height-related information. As future work, the framework will be refined and evaluated on the ISPRS Vaihingen dataset to assess and potentially improve its generalization capability across different urban environments. The study is positioned as a preliminary CNN-based multi-task baseline rather than a state-of-the-art claim; therefore, particular attention is given to explaining the limitations of the height estimation branch and the factors that may affect its performance across different urban morphologies.

Keywords: building extraction, height estimation, multi-task learning, deep learning, satellite imagery

Introduction

Three-dimensional modelling of urban areas and automatic retrieval of building height information are important for many applications, including urban planning, disaster management, energy analysis, and environmental assessment. Traditional approaches for deriving building heights often rely on LiDAR data or stereo aerial imagery, which can provide accurate results but are costly and not always available for large-scale analysis. In this context, satellite imagery offers a more scalable and cost-effective alternative for extracting building-related information over wide urban areas (Liu et al., 2020; Cao and Huang, 2021; Cai et al., 2023). However, building extraction and height estimation from satellite images remain challenging because of shadows, varying illumination, diverse building forms, and the structural complexity of urban scenes.

Recent advances in deep learning have significantly improved the analysis of remote sensing imagery. Convolutional neural networks have shown strong performance in semantic segmentation and building-related mapping tasks, while recent studies have also demonstrated the potential of deep learning for monocular or multi-source building height estimation (Liu et al., 2022; Ma et al., 2024; Yadav et al., 2024; Zhang et al., 2025). In addition, multi-task learning has emerged as an effective strategy for jointly modelling related outputs through shared feature representations, especially for tasks that require both semantic and geometric

understanding of urban structures (Meng et al., 2023; Feng et al., 2023). Since building footprint extraction and height estimation are strongly related urban analysis tasks, a multi-task framework provides a practical way to learn both semantic and geometric information within a single architecture.

More recently, transformer-based architectures have also attracted attention in remote sensing because of their ability to capture long-range dependencies and global contextual information. For example, SegFormer has been proposed as a simple and efficient transformer-based segmentation architecture, while transformer-based building extraction studies have reported promising results for large-scale urban scenes (Xie et al., 2021; Zhang et al., 2024). Nevertheless, convolution-based models remain strong and efficient baselines, particularly when integrated into task-specific learning frameworks.

The present study does not claim superiority over recent transformer-based or multi-source height estimation approaches. Instead, it provides a controlled ResNet34-based CNN baseline within a multi-task setting on DFC2023. Direct numerical comparison with the literature should be interpreted with caution because existing studies differ in dataset, sensor type, spatial resolution, height reference, and evaluation protocol. Nevertheless, the comparison with recent CNN, transformer, and fusion-based studies provides the methodological context for the next stage of the research, where stronger backbones, attention mechanisms, and explicit height-aware learning strategies will be tested.

This study evaluates a deep learning-based multi-task framework for simultaneous building extraction and building height estimation from satellite imagery. The proposed approach uses a shared encoder and dual decoder structure with a ResNet34 backbone and is assessed on the DFC2023 dataset (IEEE GRSS, 2023). The main objective is to examine whether a unified architecture can provide accurate and efficient performance for both segmentation and height regression within the same model. In response to the preliminary nature of the experiments, the contribution is defined as an initial baseline analysis that identifies the strengths of the segmentation task and the current limitations of the height regression task. The paper therefore focuses not only on the reported metrics, but also on error sources, city-level variability, and methodological design choices affecting height estimation.

Data and Method

The experiments were conducted on the DFC2023 dataset prepared for the 2023 IEEE GRSS Data Fusion Contest Track 2 (IEEE GRSS, 2023). The dataset consists of 1,773 image tiles of size 512×512 pixels collected from 11 cities: Berlin, Brasilia, Copenhagen, Portsmouth, New Delhi, New York, San Diego, Rio, Barcelona, Sydney, and Sao Luis. The data were divided into training, validation, and test sets with a 70/15/15 split, corresponding to 1,232, 270, and 271 images, respectively. Each sample includes RGB optical imagery, normalized Digital Surface Model (nDSM) height maps, and semantic building segmentation masks. In the implemented setting, RGB optical imagery was used as the input to the shared encoder, while the building mask and the nDSM-derived height map were used as supervised targets for the segmentation and height regression branches, respectively. Thus, the nDSM layer was not treated as an additional image input channel; it served as the reference height information for training and evaluation of the height decoder.

A multi-task learning framework with a shared encoder and two task-specific decoders was implemented for simultaneous building extraction and building height estimation. As illustrated in Fig. 1, the encoder was used to learn common feature representations from the input RGB images. The first decoder generated binary building masks using sigmoid activation, whereas the second decoder produced continuous height maps for pixel-wise

regression. In the implemented configuration, ResNet34 pre-trained on ImageNet was used as the backbone encoder. The height estimation decoder follows the same encoder feature hierarchy but replaces the final classification-oriented output with a regression head. Multi-scale features from the encoder are progressively upsampled and combined in the decoder to recover spatial detail, and the final layer uses linear activation to estimate continuous height values at pixel level. This design allows the segmentation branch to focus on semantic building boundaries, while the height branch learns geometric variation inside the building regions. However, because the current baseline does not include explicit height-aware attention, boundary-constrained regression, or long-tailed height re-weighting, the height decoder remains more sensitive to tall buildings and heterogeneous urban forms.

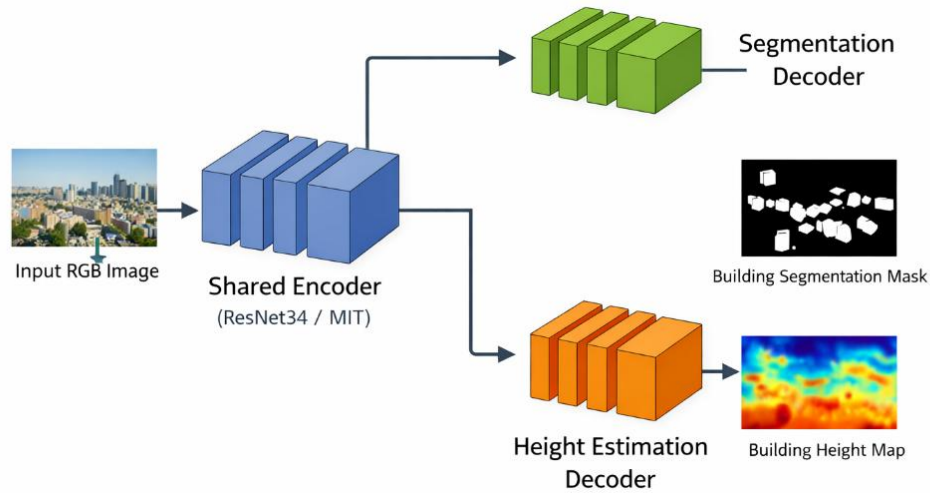


Fig. 1. Multi-Task Learning Architecture (ResNet34 Encoder - Dual Decoder).

Loss Functions. For the segmentation task, a combined Binary Cross-Entropy and Dice Loss formulation was used to improve class balance and boundary quality. For height estimation, regression loss functions were applied to minimize the difference between predicted and reference height values. The total training loss was defined as a weighted combination of segmentation and height estimation losses. More specifically, the height branch was optimized with MSE and Smooth L1 components in order to penalize continuous height differences while reducing the sensitivity to extreme outliers. The total objective can be expressed as

$$L_{\text{total}} = w1L_{\text{seg}} + w2L_{\text{height}} \quad (1)$$

Since segmentation and height regression have different numerical scales and convergence behaviour, the choice of task weights is important. In this baseline, fixed task weights were used; adaptive loss balancing will be investigated in future experiments to reduce the dominance of the easier segmentation task and improve height prediction.

The model was trained with a batch size of 8 using the Adam optimizer and an initial learning rate of 1×10^{-4} . Cosine Annealing with Warm Restarts was applied as the learning rate scheduling strategy. Mixed precision training was used to improve computational efficiency. Training was stopped early based on validation performance, and the best-performing model was retained for evaluation. Data augmentation included horizontal and vertical flipping, rotation, colour jitter, and normalization. Height values were normalized to the range $[0, 1]$ before training. The validation strategy was used to select the best epoch and reduce overfitting. Since height values are highly uneven across cities and tall buildings are less frequent, the current augmentation and sampling strategy may still under-represent rare high-rise cases. This limitation is considered one of the possible reasons for the moderate height estimation performance.

Results

Training on the DFC2023 dataset with the ResNet34-based multi-task framework demonstrated strong performance for building segmentation and moderate performance for height estimation. Table 1 summarizes the overall test set results, while Fig. 2 presents example visual outputs of the proposed model. To avoid over-interpreting the results, the evaluation is discussed as a baseline experiment: segmentation performance is considered robust, whereas height estimation is treated as the main limitation requiring further methodological improvement.

Table 1. DFC2023 test set performance metrics for the ResNet34-based model.

Task	Metric	Value
Segmentation	IoU	0.8953 (89.53%)
	F1 score	0.9227 (92.27%)
	Precision	0.9205 (92.05%)
	Recall	0.9177 (91.77%)
Height Estimation	MAE	4.36 m
	RMSE	4.12 m
	R ²	0.46

The segmentation results indicate that the proposed framework was able to delineate building footprints with high accuracy. The balanced Precision and Recall values suggest that the model achieved stable performance with relatively low omission and commission errors. In addition, the F1 and IoU scores show that the shared multi-task structure was effective for learning building-related spatial features from satellite imagery. As illustrated in Fig. 2, the predicted building masks were generally consistent with the ground-truth masks, although some boundary inaccuracies and missed structures were observed in more complex urban scenes.

For height estimation, the model achieved a Mean Absolute Error of 4.36 m and an R² value of 0.46, showing that the network captured some height-related information but with limited accuracy while maintaining strong segmentation performance. Although height regression remained more challenging than segmentation, the results indicate that the framework can jointly model semantic and geometric building characteristics, but further improvements are needed for more reliable height prediction. As shown in Fig. 2, the predicted height maps followed the general spatial distribution of building heights, but local deviations and underestimation errors were visible in some samples. The moderate R² value indicates that the current model explains only part of the height variance in the test set. This can be attributed to several factors: (i) tall buildings are less frequent and create a long-tailed target distribution, which encourages underestimation; (ii) shadows, roof materials, viewing geometry, and dense urban blocks make height cues ambiguous; and (iii) the baseline decoder does not explicitly model vertical structure or city-specific morphology. Therefore, the height results should be interpreted as preliminary and diagnostic rather than final performance.

City-level analysis further revealed differences in model behaviour across urban environments. The IoU value remained above 0.74 for all cities, indicating consistent segmentation performance throughout the dataset. Barcelona produced the highest segmentation accuracy, whereas New York and Rio showed the largest height estimation errors, likely due to their more complex urban morphology and broader building height distributions. In contrast, San Diego yielded the lowest height estimation error, which may be related to its relatively

homogeneous low-rise urban structure. These results suggest that segmentation is mainly affected by roof boundary visibility and building density, whereas height estimation is more strongly influenced by the distribution of building heights and the structural complexity of the city. High-rise and heterogeneous scenes increase the regression difficulty because the model tends to predict smoothed height values, leading to larger errors for very tall or morphologically complex buildings.

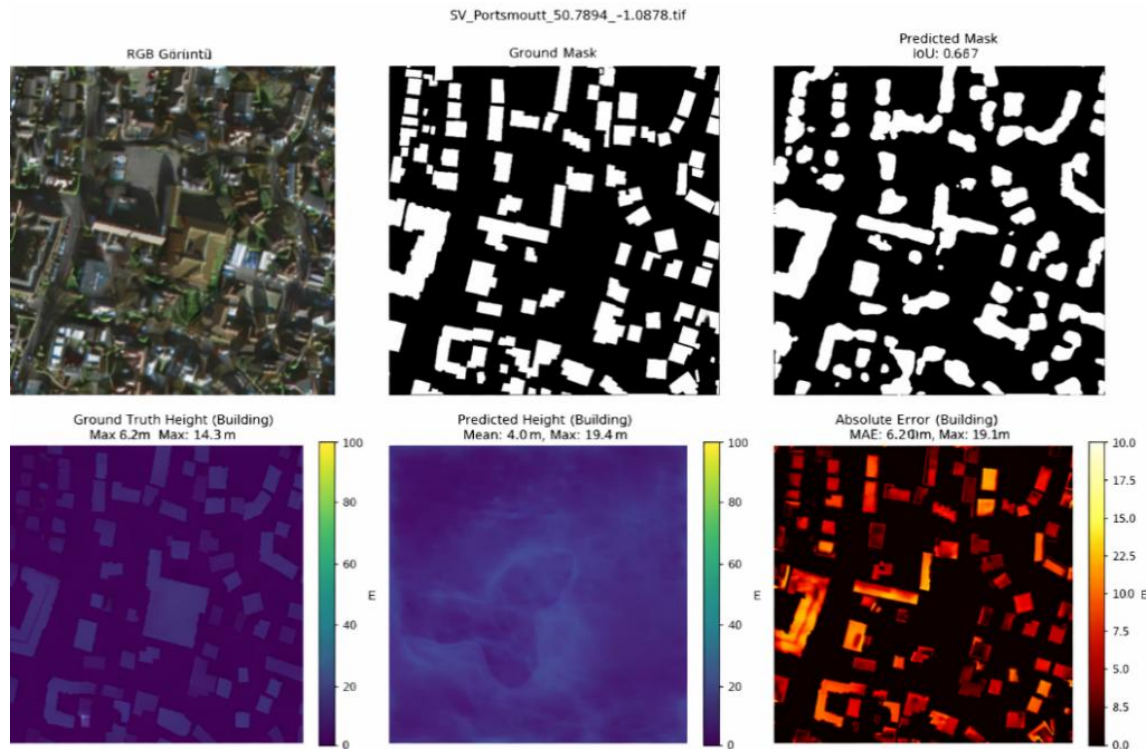


Fig. 2. Visual Results.

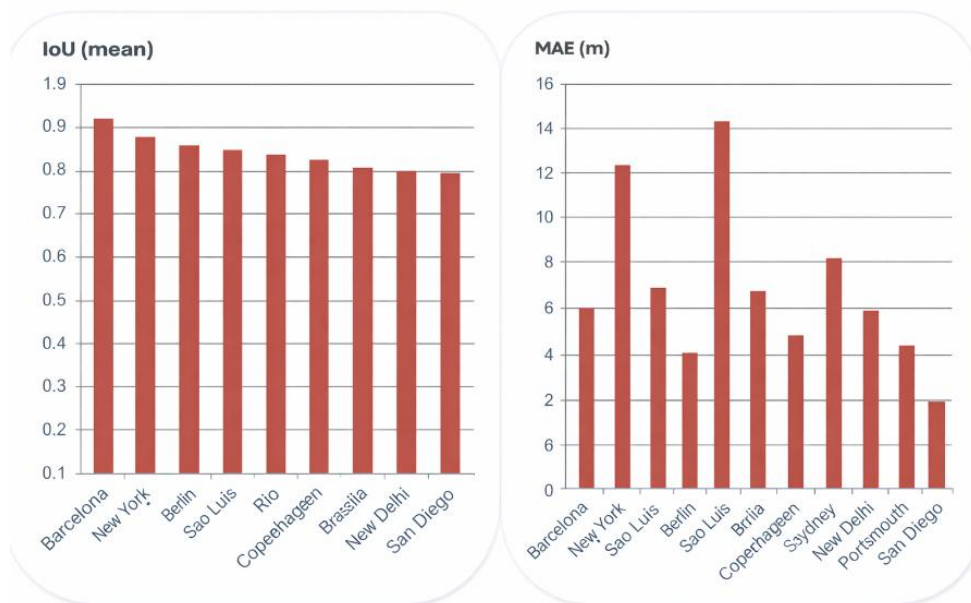


Fig. 3. City-Based Segmentation and Elevation Estimation Performance.

The results show a clear imbalance between the two tasks: building footprint extraction benefits strongly from shared spatial features, while height estimation remains more difficult because

it requires learning continuous geometric information from limited visual cues. The main error sources are expected to be long-tailed height distributions, under-representation of tall buildings, local shadow and illumination effects, and different urban morphologies across cities. In addition, the current ResNet34-based decoder does not include explicit height-aware attention, uncertainty modelling, or city-specific adaptation. For this reason, the current model should be considered a baseline for the ongoing M.Sc. research rather than a final optimized method. Future improvements will include stronger backbones such as SegFormer or other transformer-based encoders, attention-based feature fusion, adaptive task weighting, loss functions designed for long-tailed height regression, and cross-domain testing on ISPRS Vaihingen.

Conclusion

This study demonstrated that a deep learning-based multi-task framework can perform simultaneous building extraction and building height estimation from satellite imagery. The ResNet34-based model achieved strong results on the DFC2023 dataset, with an IoU of 0.8953 and a Dice score of 0.9227 for building segmentation, and an MAE of 4.36 m with an R^2 value of 0.46 for height estimation. While the segmentation results are robust, the height estimation performance is moderate, suggesting that the framework captures some geometric information but requires further refinement. These findings indicate that a unified multi-task architecture can effectively learn semantic building information and partially capture geometric characteristics within a single framework. The main contribution of the study is the establishment and analysis of a preliminary multi-task baseline. The most important limitation is the height estimation branch, which requires improved decoder design, stronger feature fusion, better treatment of long-tailed height values, and more detailed cross-city validation.

The results confirm that shared representation learning is a practical and efficient strategy for joint urban analysis tasks in remote sensing. In particular, the proposed approach provides a useful baseline for applications requiring both building footprints and height-related information, such as urban planning, disaster management, and 3D city modelling. As future work, the framework will be refined and evaluated on the ISPRS Vaihingen dataset to improve height estimation and assess generalization across different urban environments. In addition, future experiments will include comparison with alternative CNN and transformer-based baselines, explicit reporting of height-error distributions, and more detailed analysis of high-rise and low-rise urban categories. These additions are expected to make the methodological contribution clearer and directly address the limitations identified in the review process.

References

- Cai, B., Shao, Z., Huang, X., Zhou, X., Fang, S. (2023) Deep learning-based building height mapping using Sentinel-1 and Sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation*, 122, 103399.
- Cao, Y., Huang, X. (2021) A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas. *Remote Sensing of Environment*, 264, 112590.
- Feng, Y., Sun, X., Diao, W., Li, J., Niu, R., Gao, X., Fu, K. (2023) Height aware understanding of remote sensing images based on cross-task interaction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 233–249.
- IEEE GRSS (2023) 2023 IEEE GRSS Data Fusion Contest – Track 2: Multi-Task Learning of Joint Building Extraction and Height Estimation. <https://ieee-dataport.org/competitions/2023-ieee-grss-data-fusion-contest-track-2>.
- Li, X., Chen, Z., Zhang, L., Jia, D. (2023) HTC-DC Net: Monocular height estimation from single remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–14.

- Liu, C.-J., Krylov, V. A., Kane, P., Kavanagh, G., Dahyot, R. (2020) IM2ELEVATION: Building height estimation from single-view aerial imagery. *Remote Sensing*, 12(17), 2719.
- Liu, W., Sun, X., Zhang, W., Chen, K., Fu, K., Wang, H. (2022) Associatively segmenting semantics and estimating height from monocular remote-sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–17.
- Ma, C., Zhang, Y., Guo, J., Zhou, G., Geng, X. (2024) FusionHeightNet: A multi-level cross-fusion method from multi-source remote sensing images for urban building height estimation. *Remote Sensing*, 16(6), 958.
- Meng, X., Yang, Y., Wang, L., Wang, T., Li, R., Zhang, C., Jia, X. (2023) Multi-task learning of semantic segmentation and height estimation for multi-modal remote sensing images. *Journal of Geodesy and Geoinformation Science*, 6(4), 27–39.
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. (2021) SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 12077–12090.
- Yadav, R., Nascetti, A., Ban, Y. (2024) How high are we? Large-scale building height estimation using a T-SwinUNet model. *Remote Sensing of Environment*, 303, 114015.
- Zhang, H., Sun, M., Li, Q., Liu, Y., Chen, Y. (2025) Structure-aware deep learning network for building height estimation. *International Journal of Applied Earth Observation and Geoinformation*, 126, 103619.
- Zhang, L., Sun, X., Wu, T., Zhang, H. (2024) Transformer-based semantic segmentation for large-scale building footprint extraction from very-high resolution satellite images. *Advances in Space Research*, 73(9), 4341–4356.