

A Comparative Study of U-Net Variants and Transformer-Based Models for Building Segmentation

Danang Budi Susetyo^{1,2}, Tolga Bakirman^{1,3}

¹ Department of Geomatic Engineering, Yildiz Technical University, Istanbul, 34220, Türkiye, danang.susetyo@std.yildiz.edu.tr; bakirman@yildiz.edu.tr

² National Research and Innovation Agency (BRIN), Jakarta, 10340, Indonesia

³ Geography and Environmental Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom

Abstract

Buildings are key geospatial features that support various applications, including urban planning, disaster management, and 3D city modelling. Accurate building extraction from remote sensing imagery remains challenging due to complex shapes and spatial variability. Recent advances in deep learning have shown promising potential to address this issue. This study investigates the effect of different deep learning models on building segmentation performance. Various U-Net-based architectures were examined, including the Simple U-Net, U-Net with a ResNet50 backbone, U-Net with a ResNet50 + Atrous Spatial Pyramid Pooling (ASPP), and UNet++, together with SegFormer as a Transformers-based segmentation model. Two widely used open-source datasets, the Massachusetts Buildings Dataset and the WHU Building Dataset, are employed for evaluation. The experimental results reveal that U-Net with ResNet50 + ASPP performs best on the Massachusetts Buildings Dataset, whereas SegFormer achieves the highest performance on the WHU Building Dataset. This suggests that larger training datasets with higher spatial resolution are more suitable for Transformers-based models, resulting in enhanced segmentation performance.

Keywords: Building segmentation, Deep learning, U-Net, UNet++, SegFormer

Introduction

Buildings are essential geospatial objects, supporting a wide range of applications such as spatial planning, disaster mitigation, and 3D city modelling. Accurately and efficiently extracting building information from geospatial datasets is therefore crucial, as it directly impacts the effectiveness of these applications. Images acquired from satellite and aerial platforms are well suited for this purpose, as recent sensor advancements have significantly improved spatial and temporal resolution (Bakirman et al., 2022). This need has driven extensive research on building extraction, which has evolved alongside advances in image processing and computer vision techniques.

Over the years, researchers have employed various automated approaches for building extraction, including image segmentation-based methods (Jiang et al., 2008; Lee et al., 2008; Liu et al., 2008), building indices (Huang & Zhang, 2011, 2012; Pesaresi et al., 2008), and machine learning techniques (Du et al., 2015; Guo & Du, 2017; Huang et al., 2018; Turker & Koc-San, 2015). However, due to the complexity of building objects, automating building mapping is not a straightforward task. Several factors contribute to this challenge, including the complexity and heterogeneity of building appearances (Benchabana et al., 2023; Hossain & Chen, 2024; Qiu et al., 2023; Raghavan et al., 2022; Y. Wang et al., 2024; Zeng et al., 2022; B. Zhu et al., 2025), visual obstructions in the scene (Sun et al., 2023; S. Wang et al., 2025; X. Wang et al., 2024), difficulties in preserving clear boundaries (Dong et al., 2023; He & Jiang, 2021), and the subtle tonal differences that often make buildings difficult to distinguish from their surroundings (Hu et al., 2021). This challenge also persists in deep learning-based building extraction, making it a topic that continues to receive active research attention.

Various studies on deep learning-based building extraction have attempted to address these challenges. Techniques such as multi-scale feature learning (Feng et al., 2020; D. Zhou et al.,

2020; Q. Zhu et al., 2020), boundary-aware learning (Lin et al., 2023; You et al., 2022; Zeng et al., 2022), and post-processing strategies (Feng et al., 2020; Y. Xie et al., 2020) have been widely adopted. Other studies have explored alternative approaches, including incorporating prior knowledge of buildings (Hao et al., 2024; M. Wang et al., 2024), integrating deep learning with traditional machine learning methods (Qu & Wang, 2025), and applying transfer learning through fine-tuning (Neupane et al., 2025). In addition, semi-supervised learning has been used to mitigate the limited availability of labelled training datasets (Protopapadakis et al., 2021; Xia et al., 2021). However, the effectiveness of these approaches varies, and a comprehensive comparison of different deep learning architectures for building extraction is still needed.

Among these techniques, multi-scale approaches are among the most commonly used, as building heterogeneity occurs in many regions worldwide and represents one of the main challenges addressed in previous studies. Therefore, this study also aims to tackle this issue by combining the U-Net architecture with a ResNet50 backbone and Atrous Spatial Pyramid Pooling (ASPP). U-Net is chosen because it remains widely used in deep learning-based building extraction (Rastogi et al., 2022; Temenos et al., 2022; H. Wang & Miao, 2022), while ASPP is included to better handle the complexity and diversity of building objects. This variation also includes UNet++, which has been proven to achieve higher average performance than the standard U-Net (Z. Zhou et al., 2018). Moreover, the development of Transformer-based segmentation has opened new opportunities for building segmentation tasks, as it offers high accuracy. Accordingly, SegFormer, a simple and efficient Transformer-based architecture for semantic segmentation (E. Xie et al., 2021), is also employed in this study. The main contributions of this study are as follows: (1) a systematic evaluation of U-Net variants and a Transformer-based model under consistent experimental settings; (2) an investigation of the impact of architectural modifications, including backbone replacement and multi-scale feature enhancement using ASPP; and (3) an empirical analysis of model performance across two benchmark datasets with different spatial resolutions and dataset sizes. Overall, this work aims to provide practical insights into the selection of appropriate segmentation architectures for building extraction tasks under varying dataset characteristics.

Data and Method

This study utilizes two widely used open-source datasets for deep learning-based building segmentation: the Massachusetts Buildings Dataset and the WHU Building Dataset. The Massachusetts Buildings Dataset, which covers the Boston metropolitan area in the United States, spanning approximately 340 square kilometers. The images have an 8-bit radiometric resolution with three spectral bands, while the labels are provided as binary masks representing background and building areas. The dataset consists of 151 aerial images, each with a spatial dimension of 1500×1500 pixels, corresponding to an area of 2.25 square kilometers per image. The spatial resolution of the imagery is 1 meter. The dataset is divided into 137 training images, 10 test images, and 4 validation images. The corresponding labels were generated by rasterizing building footprint data obtained from OpenStreetMap. Visual illustrations of the Massachusetts Buildings Dataset, including the aerial images and their corresponding building labels, are provided in Fig. 1.



Fig. 1. Example images and ground-truth labels from the Massachusetts Building Dataset.

Another dataset used in this study is the WHU Building Dataset, which provides aerial imagery along with corresponding building labels, covering the area of Christchurch, New Zealand. Similar to the Massachusetts Buildings Dataset, the WHU Building Dataset consists of three bands images with 8-bit radiometric resolution, accompanied by binary mask labels that distinguish between background and building regions. The dataset is divided into training, validation, and test subsets. In this study, we employed the down-sampled version (including 187,000 buildings) with a ground resolution of 0.3 m, which had already been converted into patch format. Each patch has a spatial size of 512×512 pixels. With this configuration, the training set consists of 5,732 patches, while the validation and test sets each contain 1,228 patches. Examples of the WHU Building Dataset images and their corresponding labels are presented in Fig. 2.



Fig. 2. Example images and ground-truth labels from the WHU Building Dataset.

This study examined five models to observe the effect of architectural modifications on segmentation quality. The first model is Simple U-Net, which adopts the standard U-Net architecture consisting of four encoder blocks and four decoder blocks, along with one bottleneck layer at the center of the network. Each encoder and decoder block comprises two convolutional layers, resulting in a relatively deep network that is capable of extracting hierarchical spatial features at multiple levels of abstraction.

The second model is a U-Net combined with ResNet50 as the backbone. In this model, ResNet50 is employed as the encoder, which consists of 50 layers organized into residual blocks, making the network significantly deeper than the Simple U-Net. Meanwhile, in the decoder, the feature reconstruction is performed through a series of upsampling operations followed by convolutional layers, with skip connections linking the decoder to corresponding encoder feature maps.

The third model is an extension of the second model by incorporating an Atrous Spatial Pyramid Pooling (ASPP) module. In this architecture, ResNet50 is retained as the encoder. The architectural complexity is increased at the bottleneck layer, where the ASPP module is introduced to enhance multiscale feature representation. The ASPP module consists of multiple convolutional layers with different dilation rates, along with a global average pooling branch.

The decoder follows the standard U-Net upsampling scheme with skip connections from the encoder.

Another model used in this study is UNet++. Unlike the three previous models implemented using TensorFlow, this model is implemented within the PyTorch framework using the segmentation-models-pytorch library (Iakubovskii, 2019). UNet++ is configured with a ResNet34 backbone initialized with ImageNet pretrained weights. The architecture of UNet++ incorporates restructured skip pathways, dense skip connections, and integrated deep supervision (Aslantas et al., 2025), enabling more effective multi-scale feature fusion and improved segmentation performance compared to the conventional U-Net.

The final model examined in this study is SegFormer, which integrates Transformer-based encoders with lightweight multilayer perceptron (MLP) decoders (E. Xie et al., 2021). Similar to UNet++, SegFormer is implemented within the PyTorch framework using the segmentation-models-pytorch library and initialized with ImageNet pretrained weights. In this study, the SegFormer-B2 variant is employed as the encoder.

All models were trained for 30 epochs using a batch size of 4 and an input image size of 512×512 pixels. For TensorFlow-based models (Simple U-Net, U-Net with ResNet50, and U-Net with ResNet50-ASPP), pixel values were normalized to the range $[0, 1]$. Meanwhile, for PyTorch-based models (UNet++ and SegFormer), ImageNet normalization with mean values of (0.485, 0.456, 0.406) and standard deviations of (0.229, 0.224, 0.225) was applied. Data augmentation was performed using random horizontal flipping with a probability of 0.5 during training to improve model generalization. For the TensorFlow-based models, the Adam optimizer with binary cross-entropy loss was used. The UNet++ model, implemented in PyTorch, was trained using the Adam optimizer with a learning rate of 1×10^{-4} and Dice loss, while the SegFormer model employed the AdamW optimizer with a learning rate of 6×10^{-5} and Dice loss.

Evaluation was conducted in two ways. First, the trained model was used to predict the entire test dataset. The Massachusetts Buildings Dataset contains 10 images, while the WHU Building Dataset consists of 1,228 image patches. Second, to assess the visualization of building segmentation, evaluations were also performed on selected test images. From the Massachusetts Buildings Dataset, this study uses the test image with ID 23429080_15 for single-image evaluation. Meanwhile, for the WHU Building Dataset, four test images were selected: test_0002, test_0027, test_0033, and test_0034.

Results

The results of the five models were evaluated using five accuracy metrics: overall accuracy (OA), precision, recall, F1-score, and Intersection over Union (IoU). This combination of metrics was employed to provide a comprehensive evaluation of model performance. Overall accuracy was used to assess the general performance of the models, while precision and recall represent the models' ability to accurately detect the target objects. The F1-score was used to measure the balance between precision and recall. Furthermore, IoU was applied to evaluate the spatial agreement between the predicted results and the reference data.

The segmentation results for the Massachusetts Buildings Dataset on the entire test set are presented in Table 1. Overall, incorporating the ASPP module into the ResNet50-based U-Net leads to notable performance improvements, indicating an enhanced capability to capture multi-scale contextual information. This is evidenced by the model achieving the best values across almost all evaluation metrics, demonstrating that the combination of a ResNet50 backbone with an ASPP module not only improves overall accuracy but also enhances the spatial quality of the segmentation results. In contrast, UNet++ and SegFormer exhibit comparatively poorer performance.

Table 1. Accuracy evaluation of building segmentation models on the Massachusetts Buildings Dataset using the entire test set.

Model	OA	Precision	Recall	F1-Score	IoU
Simple U-Net	0.848	0.806	0.244	0.375	0.231
U-Net with ResNet50	0.905	0.848	0.595	0.700	0.538
U-Net with ResNet50 + ASPP	0.910	0.841	0.638	0.726	0.569
U-Net++	0.886	0.696	0.690	0.693	0.530
SegFormer	0.840	0.553	0.738	0.632	0.462
: Best values for each evaluation metric					

The evaluation results on the selected test image for the Massachusetts Buildings Dataset are presented in Table 2. Overall, these results are consistent with those obtained on the entire test set, where U-Net with ResNet50 + ASPP achieves the highest performance. Based on the segmentation visualization, the Simple U-Net produces a large number of false negative segments (Fig. 3), which significantly reduces its recall. Conversely, SegFormer shows the opposite behavior, generating many false positive segments, resulting in lower precision. This performance degradation may be attributed to differences in network architecture and feature representation, where the Transformers-based SegFormer model may require larger training datasets and longer training schedules to effectively learn fine-grained building boundaries.

Table 2. Accuracy evaluation of building segmentation models on selected test image (ID 23429080_15) from the Massachusetts Buildings Dataset.

Model	OA	Precision	Recall	F1-Score	IoU
Simple U-Net	0.867	0.823	0.427	0.562	0.391
U-Net with ResNet50	0.913	0.844	0.696	0.763	0.617
U-Net with ResNet50 + ASPP	0.921	0.818	0.778	0.798	0.663
U-Net++	0.860	0.642	0.682	0.661	0.494
SegFormer	0.793	0.489	0.721	0.582	0.411
: Best values for each evaluation metric					

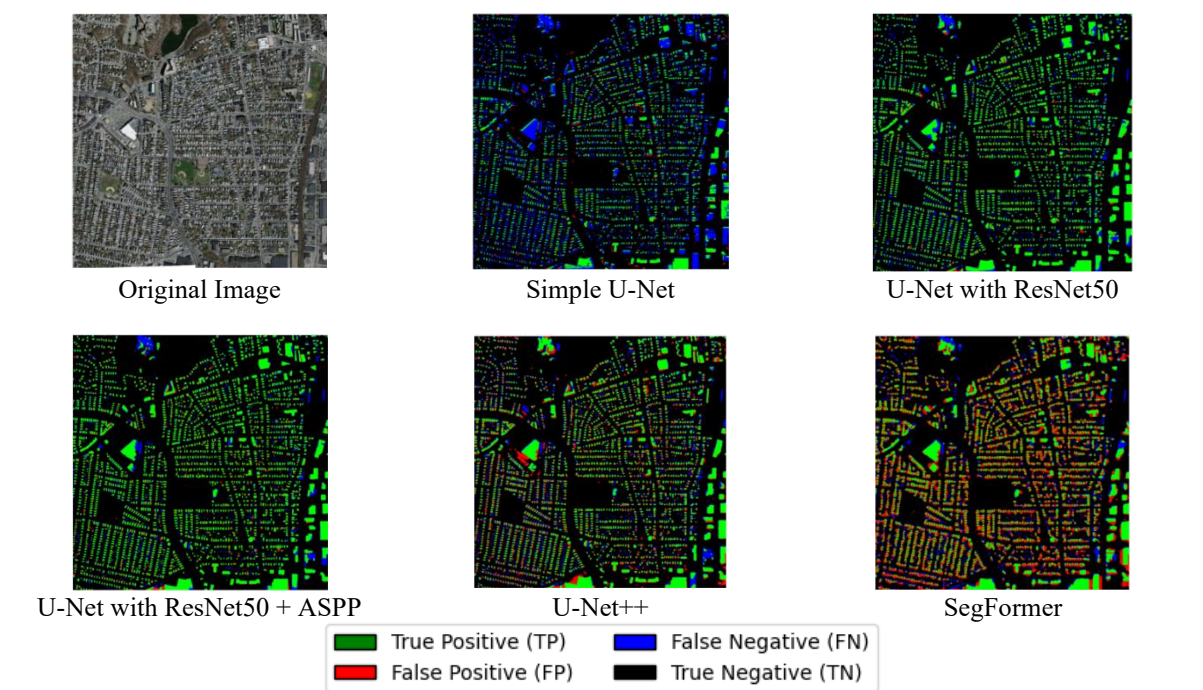


Fig. 3. Visual comparison of building segmentation results from different model architectures on the Massachusetts Buildings Dataset.

The segmentation results for the WHU Building Dataset on the entire test set are presented in Table 3. Overall, this dataset demonstrates better performance than the Massachusetts Buildings Dataset, with overall accuracy (OA) consistently above 0.90 and IoU values mostly exceeding 0.80. In contrast to the Massachusetts Buildings Dataset, SegFormer achieves the best performance among the models on this dataset. The larger training dataset, higher spatial resolution, and more clearly defined building boundaries are likely the main factors contributing to this performance improvement, particularly for Transformers-based models such as SegFormer.

Table 3. Accuracy evaluation of building segmentation models on the WHU Building Dataset using the entire test set.

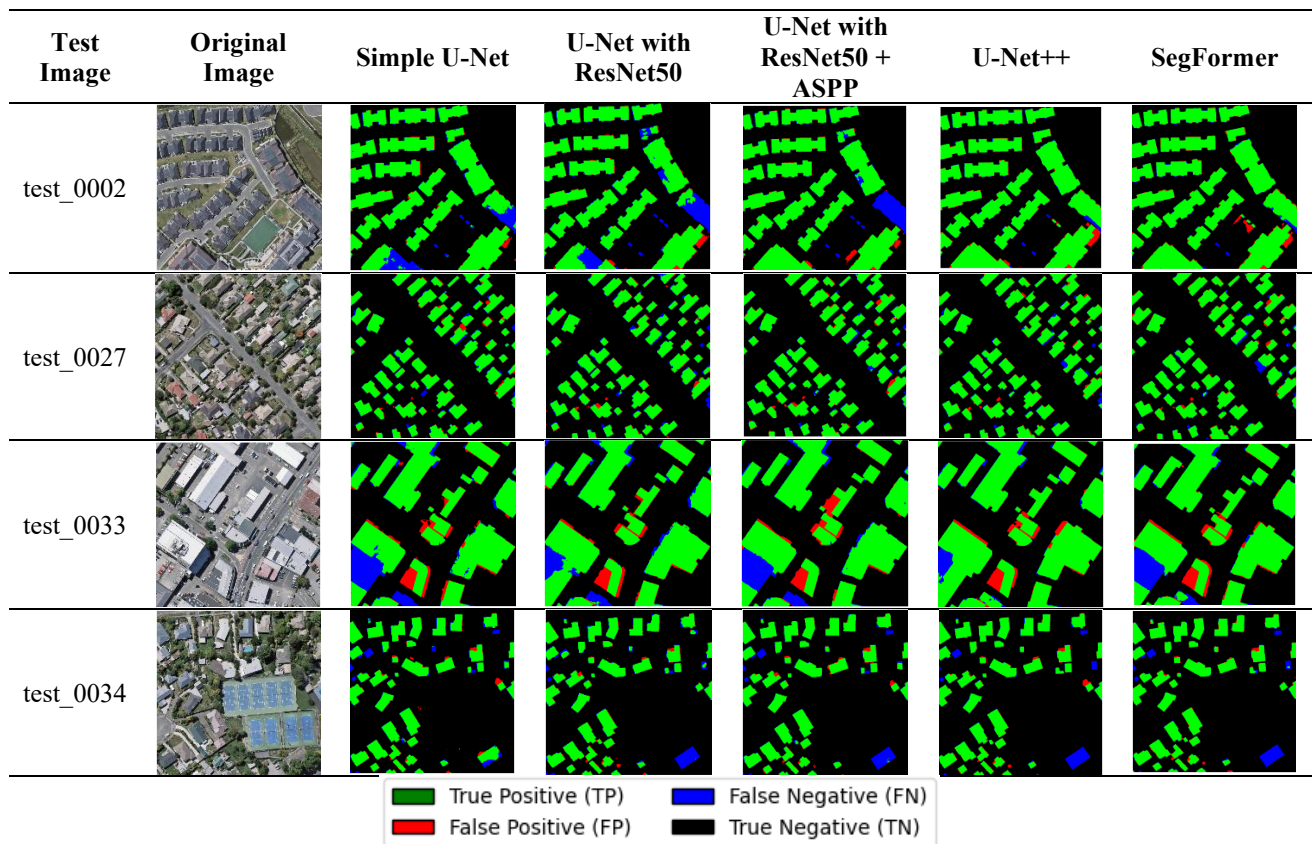
Model	OA	Precision	Recall	F1-Score	IoU
Simple U-Net	0.980	0.948	0.924	0.936	0.879
U-Net with ResNet50	0.978	0.941	0.918	0.930	0.869
U-Net with ResNet50 + ASPP	0.981	0.952	0.924	0.938	0.883
U-Net++	0.983	0.949	0.940	0.945	0.896
SegFormer	0.984	0.951	0.947	0.949	0.903
: Best values for each evaluation metric					

The performance on selected test images is presented in Table 4. Overall, the results are consistent with those obtained on the entire dataset, where SegFormer achieves the best performance among the evaluated models. This superiority is particularly evident in test_0002 and test_0027, where SegFormer demonstrates noticeably better segmentation accuracy. The visual results for this dataset are illustrated in Fig. 4, further confirming the quantitative findings.

Table 4. Accuracy evaluation of building segmentation models on selected test images from the WHU Building Dataset.

Model	Data Test	OA	Precision	Recall	F1-Score	IoU
Simple U-Net	test_0002	0.948	0.963	0.893	0.927	0.864
	test_0027	0.965	0.928	0.928	0.928	0.866
	test_0033	0.906	0.906	0.849	0.876	0.780
	test_0034	0.970	0.934	0.903	0.918	0.849
U-Net with ResNet50	test_0002	0.933	0.953	0.861	0.905	0.826
	test_0027	0.959	0.928	0.906	0.917	0.846
	test_0033	0.928	0.922	0.893	0.907	0.830
	test_0034	0.964	0.936	0.866	0.900	0.818
U-Net with ResNet50 + ASPP	test_0002	0.945	0.956	0.891	0.922	0.856
	test_0027	0.967	0.934	0.932	0.933	0.874
	test_0033	0.911	0.910	0.860	0.884	0.792
	test_0034	0.967	0.942	0.879	0.909	0.833
U-Net++	test_0002	0.970	0.954	0.964	0.959	0.922
	test_0027	0.965	0.931	0.927	0.929	0.868
	test_0033	0.942	0.917	0.938	0.927	0.864
	test_0034	0.966	0.937	0.877	0.906	0.829
SegFormer	test_0002	0.971	0.950	0.972	0.961	0.925
	test_0027	0.968	0.940	0.932	0.936	0.879
	test_0033	0.920	0.912	0.882	0.897	0.813
	test_0034	0.969	0.947	0.882	0.913	0.841

: Best accuracy values for each test image

**Fig. 4.** Visual comparison of building segmentation results from different model architectures on the WHU Building Dataset.

Conclusion

This study compares five different models for building segmentation using the Massachusetts Buildings Dataset and the WHU Building Dataset. Several U-Net-based architectures were evaluated, including U-Net with a ResNet50 backbone, the integration of ASPP, and UNet++, along with SegFormer as a Transformers-based segmentation model. Overall, U-Net with ResNet50 + ASPP achieves the best performance on the Massachusetts Buildings Dataset. In contrast, different results are observed on the WHU Building Dataset, where SegFormer produces the best performance. This indicates that a larger training dataset with higher spatial resolution can better support Transformers-based models, leading to improved segmentation results.

Future work may further investigate the effectiveness of these architectural modifications on more complex and large-scale datasets, as well as explore multi-resolution learning strategies. In addition, cross-dataset generalization experiments could be conducted to assess the robustness of the models when applied to datasets with different geographic and spatial characteristics. Further improvements may also be achieved by incorporating post-processing techniques to refine segmentation boundaries and by exploring semi-supervised learning approaches to reduce dependency on large amounts of labeled data.

References

- Aslantas, N., Bakirman, T., Selbesoğlu, M. O., & Bayram, B. (2025). The Role of Ensemble Deep Learning for Building Extraction from VHR Imagery. *International Journal of Engineering and Geosciences*, 10(3), 352–363. <https://doi.org/10.26833/ijeg.1587798>
- Bakirman, T., Komurcu, I., & Sertel, E. (2022). Comparative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset. *Expert Systems with Applications*, 202. <https://doi.org/10.1016/j.eswa.2022.117346>
- Benchabana, A., Kholadi, M. K., Bensaci, R., & Khaldi, B. (2023). Building Detection in High-Resolution Remote Sensing Images by Enhancing Superpixel Segmentation and Classification Using Deep Learning Approaches. *Buildings*, 13(7). <https://doi.org/10.3390/buildings13071649>
- Dong, D., Ming, D., Weng, Q., Yang, Y., Fang, K., Xu, L., Du, T., Zhang, Y., & Liu, R. (2023). Building Extraction from High Spatial Resolution Remote Sensing Images of Complex Scenes by Combining Region-Line Feature Fusion and OCNN. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 4423–4438. <https://doi.org/10.1109/JSTARS.2023.3273726>
- Du, S., Zhang, F., & Zhang, X. (2015). Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 107–119. <https://doi.org/10.1016/j.isprsjprs.2015.03.011>
- Feng, W., Sui, H., Hua, L., Xu, C., Ma, G., & Huang, W. (2020). Building extraction from VHR remote sensing imagery by combining an improved deep convolutional encoder-decoder architecture and historical land use vector map. *International Journal of Remote Sensing*, 41(17), 6595–6617. <https://doi.org/10.1080/01431161.2020.1742944>
- Guo, Z., & Du, S. (2017). Mining parameter information for building extraction and change detection with very high-resolution imagery and GIS data. *GIScience and Remote Sensing*, 54(1), 38–63. <https://doi.org/10.1080/15481603.2016.1250328>
- Hao, M., Chen, S., Lin, H., Zhang, H., & Zheng, N. (2024). A prior knowledge guided deep learning method for building extraction from high-resolution remote sensing images. *Urban Informatics*, 3. <https://doi.org/10.1007/s44212-024-00038-8>
- He, S., & Jiang, W. (2021). Boundary-assisted learning for building extraction from optical remote sensing imagery. *Remote Sensing*, 13(4), 1–18. <https://doi.org/10.3390/rs13040760>

- Hossain, M. D., & Chen, D. (2024). Target-based building extraction from high-resolution RGB imagery using GEOBIA framework and tabular deep learning model. *Geomatica*, 76(1). <https://doi.org/10.1016/j.geomat.2024.100007>
- Huang, X., Hu, T., Li, J., Wang, Q., & Benediktsson, J. A. (2018). Mapping urban areas in China using multisource data with a novel ensemble SVM method. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8), 4258–4273. <https://doi.org/10.1109/TGRS.2018.2805829>
- Huang, X., & Zhang, L. (2011). A Multidirectional and Multiscale Morphological Index for Automatic Building Extraction from Multispectral GeoEye-1 Imagery. *Photogrammetric Engineering & Remote Sensing*, 77(7), 721–732.
- Huang, X., & Zhang, L. (2012). Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1), 161–172. <https://doi.org/10.1109/JSTARS.2011.2168195>
- Hu, Q., Zhen, L., Mao, Y., Zhou, X., & Zhou, G. (2021). Automated building extraction using satellite remote sensing imagery. *Automation in Construction*, 123. <https://doi.org/10.1016/j.autcon.2020.103509>
- Iakubovskii, P. (2019). *Segmentation Models Pytorch*. GitHub. https://github.com/qubvel/segmentation_models.pytorch
- Jiang, N., Zhang, J. X., Li, H. T., & Lin, X. G. (2008). Semi-Automatic Building Extraction from High Resolution Imagery Based on Segmentation. *International Workshop on Earth Observation and Remote Sensing Applications*.
- Lee, D. H., Lee, K. M., & Lee, S. U. (2008). Fusion of Lidar and Imagery for Reliable Building Extraction. *Photogrammetric Engineering & Remote Sensing*, 74(2), 215–225.
- Lin, H., Hao, M., Luo, W., Yu, H., & Zheng, N. (2023). BEARNet: A Novel Buildings Edge-Aware Refined Network for Building Extraction From High-Resolution Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, 20. <https://doi.org/10.1109/LGRS.2023.3272353>
- Liu, Z., Cui, S., & Yan, Q. (2008). Building Extraction from High Resolution Satellite Imagery Based on Multi-scale Image Segmentation and Model Matching. *International Workshop on Earth Observation and Remote Sensing Applications*.
- Neupane, B., Aryal, J., & Rajabifard, A. (2025). Fine-Tuning-Based Transfer Learning for Building Extraction from Off-Nadir Remote Sensing Images. *Remote Sensing*, 17(7). <https://doi.org/10.3390/rs17071251>
- Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A robust built-up area presence index by anisotropic rotation-invariant textural measure. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1(3), 180–192. <https://doi.org/10.1109/JSTARS.2008.2002869>
- Protopapadakis, E., Doulamis, A., Doulamis, N., & Maltezos, E. (2021). Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery. *Remote Sensing*, 13(3), 1–24. <https://doi.org/10.3390/rs13030371>
- Qiu, W., Gu, L., Gao, F., & Jiang, T. (2023). Building Extraction From Very High-Resolution Remote Sensing Images Using Refine-UNet. *IEEE Geoscience and Remote Sensing Letters*, 20. <https://doi.org/10.1109/LGRS.2023.3243609>
- Qu, Y., & Wang, C. (2025). A self-supervised building extraction method based on multi-modal remote sensing data. *Remote Sensing Letters*, 16(1), 77–88. <https://doi.org/10.1080/2150704X.2024.2440668>
- Raghavan, R., Verma, D. C., Pandey, D., Anand, R., Pandey, B. K., & Singh, H. (2022). Optimized building extraction from high-resolution satellite imagery using deep learning. *Multimedia Tools and Applications*, 81(29), 42309–42323. <https://doi.org/10.1007/s11042-022-13493-9>

- Rastogi, K., Bodani, P., & Sharma, S. A. (2022). Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto International*, 37(5), 1501–1513. <https://doi.org/10.1080/10106049.2020.1778100>
- Sun, Y., Chen, J., Huang, X., & Zhang, H. (2023). Multi-Level Perceptual Network for Urban Building Extraction from High-Resolution Remote Sensing Images. *Photogrammetric Engineering and Remote Sensing*, 89(7), 427–434. <https://doi.org/10.14358/PERS.22-00103R1>
- Temenos, A., Temenos, N., Doulamis, A., & Doulamis, N. (2022). On the Exploration of Automatic Building Extraction from RGB Satellite Images Using Deep Learning Architectures Based on U-Net. *Technologies*, 10(1). <https://doi.org/10.3390/technologies10010019>
- Turker, M., & Koc-San, D. (2015). Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *International Journal of Applied Earth Observation and Geoinformation*, 34, 58–69. <https://doi.org/10.1016/j.jag.2014.06.016>
- Wang, H., & Miao, F. (2022). Building extraction from remote sensing images using deep residual U-Net. *European Journal of Remote Sensing*, 55(1), 71–85. <https://doi.org/10.1080/22797254.2021.2018944>
- Wang, M., Su, L., Yan, C., Xu, S., Yuan, P., Jiang, X., & Zhang, B. (2024). RSBuilding: Towards General Remote Sensing Image Building Extraction and Change Detection with Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62. <https://doi.org/10.1109/TGRS.2024.3439395>
- Wang, S., Guo, M., Wang, D., Cao, W., Huang, Y., & Zhang, J. (2025). Optimisation of building contour extraction in high-resolution remote sensing images: An adaptive cluster segmentation algorithm for improving deep learning models. *Expert Systems with Applications*, 280. <https://doi.org/10.1016/j.eswa.2025.127562>
- Wang, X., Tian, M., Zhang, Z., He, K., Wang, S., Liu, Y., & Dong, Y. (2024). SDSNet: Building Extraction in High-Resolution Remote Sensing Images Using a Deep Convolutional Network with Cross-Layer Feature Information Interaction Filtering. *Remote Sensing*, 16(1). <https://doi.org/10.3390/rs16010169>
- Wang, Y., Wang, S., Yuan, J., Dou, A., & Gu, Z. (2024). A Deep Learning Method for Building Extraction from Remote Sensing Images by Fuzing Local and Global Features. *Journal of Sensors*. <https://doi.org/10.1155/2024/5575787>
- Xia, L., Zhang, X., Zhang, J., Yang, H., & Chen, T. (2021). Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection. *Remote Sensing*, 13(11). <https://doi.org/10.3390/rs13112187>
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *arXiv*. <http://arxiv.org/abs/2105.15203>
- Xie, Y., Zhu, J., Cao, Y., Feng, D., Hu, M., Li, W., Zhang, Y., & Fu, L. (2020). Refined Extraction of Building Outlines from High-Resolution Remote Sensing Imagery Based on a Multifeature Convolutional Neural Network and Morphological Filtering. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1842–1855. <https://doi.org/10.1109/JSTARS.2020.2991391>
- You, D., Wang, S., Wang, F., Zhou, Y., Wang, Z., Wang, J., & Xiong, Y. (2022). EfficientUNet+: A Building Extraction Method for Emergency Shelters Based on Deep Learning. *Remote Sensing*, 14(9). <https://doi.org/10.3390/rs14092207>
- Zeng, Y., Guo, Y., & Li, J. (2022). Recognition and extraction of high-resolution satellite remote sensing image buildings based on deep learning. *Neural Computing and Applications*, 34(4), 2691–2706. <https://doi.org/10.1007/s00521-021-06027-1>

- Zhou, D., Wang, G., He, G., Long, T., Yin, R., Zhang, Z., Chen, S., & Luo, B. (2020). Robust building extraction for high spatial resolution remote sensing images with self-attention network. *Sensors (Switzerland)*, 20(24), 1–19. <https://doi.org/10.3390/s20247241>
- Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *arXiv*. <http://arxiv.org/abs/1807.10165>
- Zhu, B., Yu, D., Xiao, X., Shen, J., Cui, Z., Su, Y., Li, A., & Li, D. (2025). AP-PointRend: An Improved Network for Building Extraction via High-Resolution Remote Sensing Images. *Remote Sensing*, 17(9), 1481. <https://doi.org/10.3390/rs17091481>
- Zhu, Q., Li, Z., Zhang, Y., & Guan, Q. (2020). Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields. *Remote Sensing*, 12(23), 1–18. <https://doi.org/10.3390/rs12233983>