

A Novel Swin Transformer Based Deep Learning Model for Building Extraction from Very High-Resolution Images

Kavzoglu T.* and Yilmaz E.O.

Department of Geomatics Engineering, Gebze Technical University, Kocaeli, Turkey

*kavzoglu@gtu.edu.tr

Abstract: *The fundamental process of automated building footprint extraction from very high resolution (VHR) images enables geospatial data production that supports urban planning, cartography, and three-dimensional city modeling. This specific task requires precise vectorization of roof outlines for smart city applications. Recent advances in deep learning models (esp. convolutional neural networks (CNNs) and transformers) have significantly contributed to the improvement in the achieved accuracy level. These models learn to identify man-made structures by analyzing image data, which enables them to detect complex patterns inherent in the dataset. Despite these advantages, the task remains challenging, especially in dense urban environments. Heterogeneous building forms, tight spatial proximities, and frequent occlusion by vegetation create complications for the delineation of footprint information. Shadows from high-rise structures also result in misclassification by obscuring boundaries and producing incomplete footprints. These limitations clearly underline the need for robust solutions. In this study, a novel Swin Transformer-based model is proposed to extract building footprint information using the Massachusetts Building Dataset. It aims to precisely identify building footprint boundaries by capturing local textural and global contextual information through a window-based attention mechanism. To assess its effectiveness, the performance of the model was benchmarked against state-of-the-art (SOTA) architectures (i.e., DeepLabV3+, SegFormer, UPerNet, PAN, FPN, and PSPNet). All training and evaluation processes were completed under the same hyperparameter settings for objective evaluation of the proposed deep learning model. Quantitative results revealed that the introduced model showed superior performance based on all accuracy metrics considered. The proposed model demonstrated 87.98% precision, 86.03% recall, 77.94% IoU, 86.96% F-score, and 92.54% overall accuracy. On the other hand, the SOTA models showed varying performances, in that IoUs for SegFormer, UPerNet, DeepLabV3+, PAN, FPN and PSPNet were estimated as 75.41, 75.66, 73.36, 69.78, 71.51, 72.03%, respectively. The study reveals that the proposed model attains greater boundary precision, especially in compact urban settings, and maintains robust generalization performance.*

Keywords: *Building footprint segmentation, Deep learning, Massachusetts building dataset, Swin transformer, VHR imagery.*

Introduction

Building footprint extraction from aerial and VHR satellite imagery is a fundamental geospatial process, which is of critical importance for understanding and managing cities and their development. This automated delineation of human settlements transforms pixels into precise vector data, creating an essential digital inventory of the built environment. This application is essential for many important tasks. It is particularly useful for city planning, tracking population changes, and assessing disaster risks. The data turns unprocessed images

into useful information to be used in decision-making processes. This intelligence provides a solid foundation for making smart decisions, which helps us create a sustainable and tough future. The remote sensing field has recently witnessed substantial interest in detecting building footprints through VHR images (Bittner et al., 2017; Rastogi et al., 2020; Rajamani et al., 2022). The accurate delineation of building footprints or borders is an essential component of several urban applications, which subsequently help decision-making related to sustainable development (Chen et al., 2020; Patil et al., 2025; Zhang et al., 2022). As a result of rapid urbanization, the growing importance of the sustainable city concept promotes the value of building footprint data. It also has a central role in disaster monitoring, including rapid damage assessment and post-disaster recovery operations (Döş et al., 2025; El Ghoul et al., 2025; Feng et al., 2023; Rahnemoonfar et al., 2023). The extracted data can be considered as the foundation of the “smart city”, enabling technologies that improve urban life. As some studies underlined (e.g., Gupta et al., 2025), this specific data is vital for various research, and remote sensing is an indispensable tool, allowing planners and others to work with accurate, real-world information (Huang & Yang, 2024).

The improved quality and resolution of remotely sensed imagery allow for the acquisition of more detailed information about buildings (Debella-Gilo, 2025; Pang et al., 2024). Nonetheless, this can cause noise problems, such as the “salt and pepper” effect due to the technical characteristics of the image. It could make it difficult to distinguish the building from the background. Besides, there are other challenges, including buildings of different scales and complex building structures (Jin et al., 2023). In VHR imagery, building shadows appear more distinct, which complicates the task of separating adjacent structures (Dong et al., 2023). As a result, precise and effective building extraction utilizing VHR remote sensing imagery serves as a significant research domain.

In recent years, researchers have proposed numerous building extraction methods and compared their performances with SOTA models. For instance, Yılmaz et al. (2025) conducted a performance analysis for U-Net and U-Net++ for building footprint extraction and applied explainable AI to comprehend the decision-making process of the deep learning models. Similar to our objective in this study, Li et al. (2024) proposed a new model called HD-Net for building footprint extraction via deeply supervised body and boundary decomposition. Empirical analyses using the Massachusetts, WHU, and Inria datasets demonstrate that HD-Net provides competitive outcomes with a minimal parameter burden. Liu et al. (2024) also suggested a model for building segmentation from satellite and aerial images. The proposed model, STransU2Net, combines a Transformer and CNN to extract

buildings at various sizes. The results in terms of the IoU metric revealed SOTA results with 91.04% and 59.09% for aerial and satellite imagery, respectively.

Building footprints extraction methods may be divided into two broad groups as traditional and deep learning methods. Traditional methods (e.g., support vector machine and random forest) face various challenges (Haidarh et al., 2025; Kavzoglu & Bilucan, 2023). Because high-quality remote sensing images contain a large amount of feature detail, they require prior knowledge. While these algorithms extract feature parameters from the color, contours, and shape of objects and generate decision rules, they are generally limited to specific building shapes. Spectral brightness differences between objects, such as the same object having different spectral appearances, lead to low accuracy and unclear object boundaries (Guo et al., 2022; Zhu et al., 2020). Also, the intricate structures and high density of buildings, compounded by tree-induced occlusions, considerably hinder the extraction of reliable information from images.

With advances in the computer vision domain, deep learning models have become progressively well-recognized (Trigka & Dritsas, 2025; Voulodimos et al., 2018). The capacity of CNNs to obtain discriminative features has become a popular tool in image processing applications. CNN-based architectures such as DeepLabV3+ (Chen et al., 2018) and U-Net (Ronneberger et al., 2015) are recognized for their ability to obtain detailed segmentation information using various datasets. However, CNN-based models may have difficulty in capturing the global context because they focus only on local features (Pereira & Hussain, 2024; Xie et al., 2025).

Originally designed for natural language processing, Transformers employ a self-attention mechanism to analyze both local structures and global context (Vaswani et al., 2017). This approach has contributed to more effective representation of spatial relationships. Swin Transformer has been a key breakthrough in this research field. Its intelligent design employs a hierarchy of windows that shift between layers, which makes it both powerful and computationally efficient (Liu et al., 2021). Thus, it can capture small details in VHR images while still understanding the overall scene. Some recent studies (e.g., Xiao et al., 2022; Yang et al., 2025) have reported that Swin Transformer-based models produce more accurate results than other methods for segmenting images.

This study proposes a novel Swin Transformer-based model for building footprint extraction that can extract various features from VHR imagery. To evaluate and confirm the superior performance of the model, its performance was compared with that of well-known SOTA models, including DeepLabV3+, SegFormer, UperNet, PAN, FPN and PSPNet. It should be

noted that all models were trained under the same experimental conditions for objective comparison and evaluation.

Study Area and Dataset

In this study, the Massachusetts Building Dataset, which is a widely-used dataset in building footprint extraction studies, was employed to evaluate the proposed and SOTA models. It consists of VHR images collected in 2011 over the city of Boston, MA. The dataset includes a variety of urban patterns with high-rise buildings to detached residential structures (Mnih, 2013). This diversity helps to develop a deep learning model with high generalization capabilities. The structure of the data under investigation is illustrated with images and their corresponding ground truth masks (Figure 1). While the left columns illustrate the original images, the right columns display the mask images. Within the masks, buildings are shown in yellow, while the background is represented in dark purple.

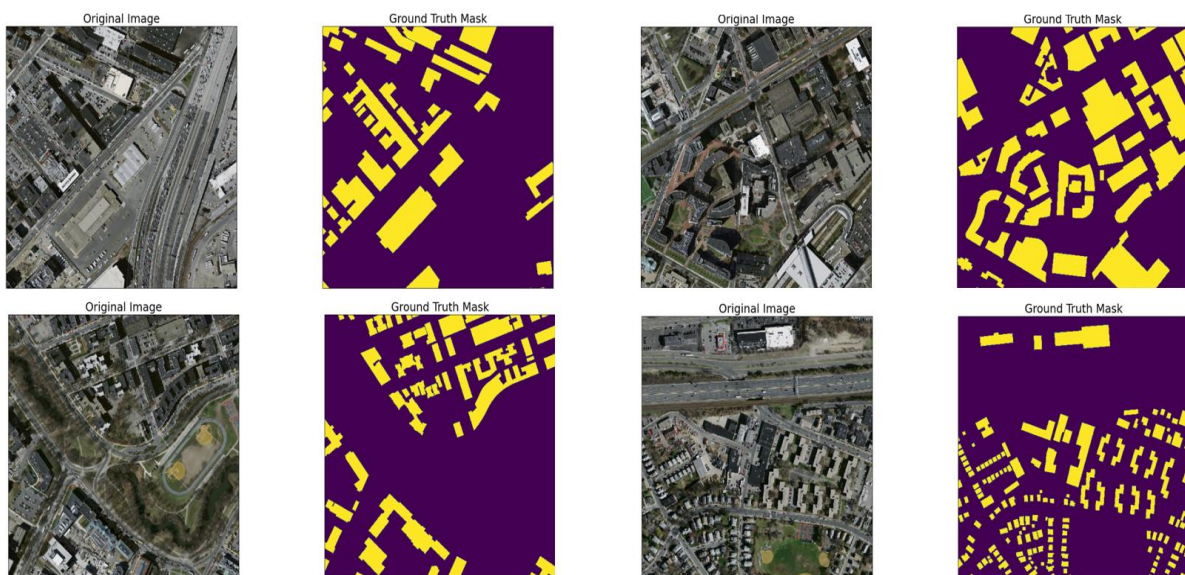


Figure 1: Examples of VHR satellite imagery (left) and the corresponding building ground truth masks (right).

Each image in the dataset is 1500x1500-pixel size, and its ground truth, produced from OpenStreetMap with labels, is provided. To prepare the dataset for deep learning applications, the original aerial images were divided into 512x512 pixels with a half overlap. Furthermore, data augmentation techniques, including random rotation, translation, and normalization, were used to improve generalization performance, so that overfitting was avoided.

Methodology

This research proposes a deep learning model to achieve more accurate building footprint extraction. Leveraging the Swin Transformer architecture and its feature pyramid mechanism, the model effectively refines footprint extraction from VHR satellite images. Its performance was evaluated against SOTA models using accuracy measures of precision, recall, IoU, F-score, and overall accuracy.

a. Proposed Model Architecture

The proposed model features a multi-component structure with a backbone formed by the Swin Transformer. This backbone has been proven to facilitate the capture of complex details by dividing the image into smaller components using a sliding window mechanism (Raghu et al., 2021). It involves the integration of features at various levels for the purpose of preserving details and semantic information. The model was augmented with an auxiliary output layer with the objective of achieving balanced learning. In the final part of the model, a decoder was employed to generate the resulting feature maps. These features enable the optimization of architecture, thereby ensuring the attainment of robust generalization capabilities and dependable performance in the domain of building footprint extraction.

b. SOTA Models Architecture

The study compares the proposed model with several leading SOTA models (i.e., DeepLabV3+, SegFormer, UPerNet, PAN, FPN and PSPNet) for building footprint extraction to validate its success. All SOTA models have been extensively utilized in semantic segmentation studies. It can be stated that each model exhibits unique characteristics. For example, several models are shown to possess CNN-based architecture (e.g., DeepLabV3+, FPN, PSPNet and PAN), whilst others reveal Transformer-based architecture (e.g., SegFormer and UPerNet). The rationale behind the selection of these models is to facilitate a comparative analysis of the performance of CNN and Transformer-based models. The SOTA models utilized in the study are described as follows:

- **DeepLabV3+** model is recognized for its effectiveness in modeling dense convolutional operations and multi-scale contextual features. This architecture successfully captures multi-scale contexts by using atrous convolutional and spatial pyramid pooling (Chen et al., 2018).

- **SegFormer** model provides a balance between efficiency and accuracy due to its hierarchical design. Its structure is distinguished by its lightweight and efficient design. It also incorporates an MLP-segmentation head in the decoder section instead of convolution (Xie et al., 2021).
- **UPerNet** architecture includes the Feature Pyramid Network and Pyramid Pooling module, which offers clear advantages because its encoder-decoder design integrates multi-level features to segment objects of different sizes (Xiao et al., 2018).
- **PAN** model is a specialized architecture, including the Pyramid Attention Module. By weighting features at various scales, the module allows the network to generate richer feature representations. It also incorporates attention mechanisms to highlight informative features during segmentation (Li et al., 2018).
- **FPN** is an architecture developed to effectively capture multi-scale features in CNNs. This model learns both fine details and broader context simultaneously by combining feature maps of different resolutions obtained from the base network (e.g., ResNet). Thus, it improves the detection of small objects while leveraging general scene knowledge (Lin et al. 2017).
- **PSPNet** is a deep learning architecture developed for semantic understanding. The model's fundamental innovation is a structure called the pyramid pooling module. It combines both local details and global context information by performing pooling operations at different scales on the input feature map. Thus, PSPNet enables more accurate and comprehensive classification of objects, especially in complex scenes (Zhao et al., 2017).

c. **Hyperparameter Settings and Accuracy Assessment**

All analyses were conducted with the Massachusetts Building Dataset using a patch-based training strategy. Before the training phase and accuracy assessment, the dataset was divided into training, validation, and test subsets. The same hyperparameter setting (Table 1) was preferred for the models so that performance comparisons can be performed objectively. The models were trained for a total of 200 epochs by employing the AdamW optimization algorithm to update the network parameters. During the training process, a fixed learning rate of 1×10^{-4} was applied to control the step size of the updates, while a batch size of 8 was used to process the training samples in each iteration. This configuration

was selected to ensure stable convergence, efficient utilization of computational resources, and a balanced trade-off between training speed and segmentation accuracy. The loss function employed in this study is a composite formulation that combines two components, namely Cross-Entropy Loss and Dice Loss. Importantly, all encoder weights were initialized randomly, allowing the models to learn dataset-specific features rather than relying on pre-trained parameters.

Table 1: Hyperparameter configurations for the training process.

Epoch	200
Optimizer	AdamW
Learning rate	1×10^{-4}
Batch size	8
Loss Function	Cross-Entropy & Dice

To conduct a thorough assessment of the performance of the proposed and SOTA models, five accuracy metrics were estimated (i.e., precision, recall, IoU, F-score, and overall accuracy). Precision measures the proportion of correctly predicted building pixels, while Recall evaluates the ability to identify all present building pixels. Additionally, IoU is a measurement that quantifies the overlap between predicted and reference building masks. F-score is an evaluation metric that demonstrates a harmonic balance between precision and recall. The final evaluation metric used in the study is overall accuracy, which is the proportion of correctly classified pixels in the building and background classes.

Results and Discussion

The performance analysis was conducted for the proposed model and the SOTA architectures considered in this study, employing the Massachusetts building dataset, which is considered a benchmark in building footprint extraction studies. Under identical conditions, the results accurately reflect the comparative strengths of the models. Results of the accuracy assessment are given in Table 2.

Table 2: Performance comparison of the proposed model and SOTA models on the Massachusetts Building Dataset (%).

Model	Precision	Recall	IoU	F-score	Overall Accuracy
DeepLabV3+	86.14	81.60	73.36	83.60	90.95
SegFormer	87.19	83.42	75.41	85.13	91.70
UPerNet	86.36	84.42	75.66	85.34	91.62
PAN	82.81	77.17	69.78	80.21	92.26
FPN	83.96	80.74	71.51	82.21	90.02
PSPNet	84.70	80.90	72.03	82.60	90.32
Proposed Model	87.98	86.03	77.94	86.96	92.54

The proposed Swin Transformer-based model exhibited the optimal performance in all metrics evaluated. It outperformed SOTA models with a precision of 87.98%, thereby minimizing errors in incorrectly classifying non-building areas as buildings. Remarkably, the model obtained the highest recall rate of 86.03%, accurately identifying a substantial proportion of real buildings. The model achieves an IoU of 77.94% and an F-score of 86.96%, which indicates its capacity to extract building boundaries with both precision and consistency. The overall accuracy rate reached its peak of 92.54%. It specifies consistent success not only in building areas but also across the entire image. Among the other models that were compared, SegFormer and UPerNet provided results that were competitive, but they were inferior to the proposed model. Even though DeepLabV3+, FPN and PSPNet are regarded as a sophisticated CNN architecture, it exhibited substandard performance, especially regarding IoU and F-score metrics. The PAN model, despite achieving a high overall accuracy of 92.26%, revealed significantly low precision, recall, and IoU values. This indicates that while the model was successful in background classification, it was insufficient in building extraction. The results indicate that the proposed model improves accuracy, reliability, and generalizability in building extraction. A comparative analysis of several deep learning-based segmentation approaches applied to VHR satellite imagery is given in Figure 2. The first column displays the original input image, followed by the ground truth in the second column. Columns three through eight show the predictions generated by widely used models from the literature, while the final column depicts the outputs of the proposed method.

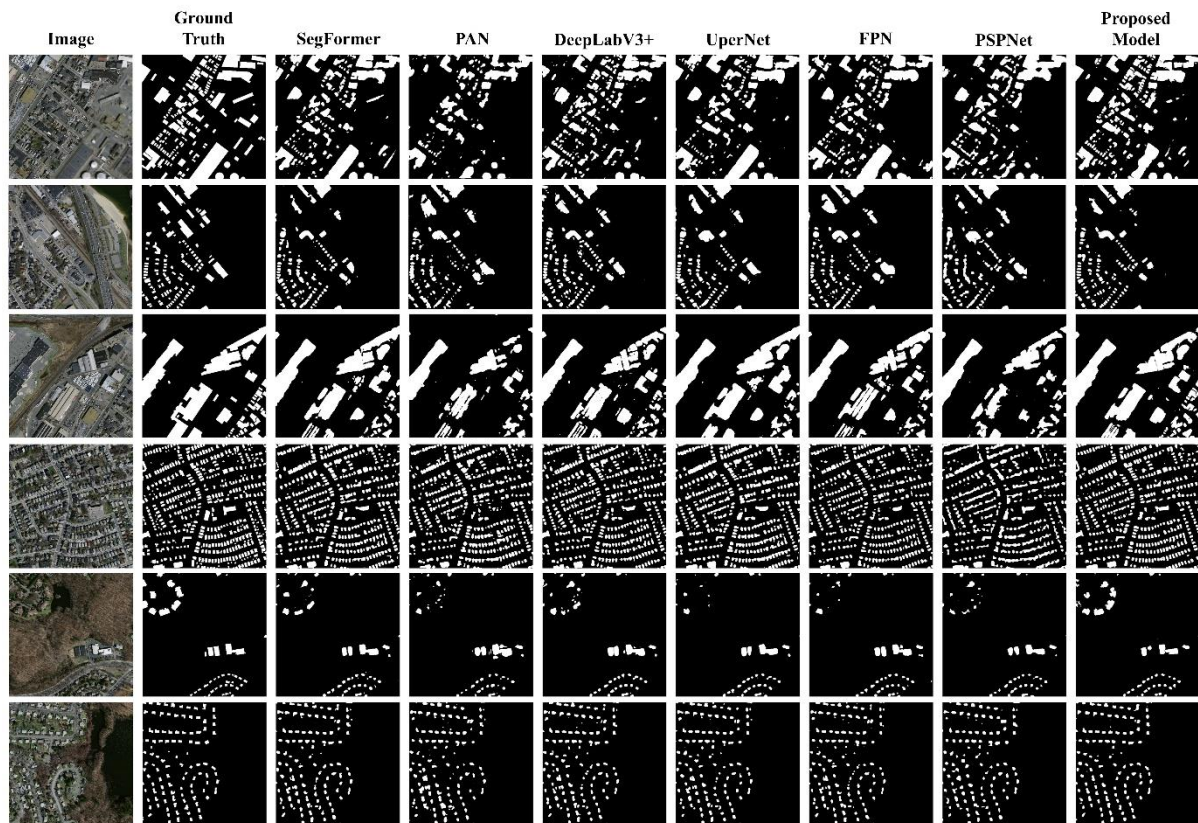


Figure 2: Performance comparison of SOTA and the proposed models.

While buildings are represented by clear, sharp, and compact polygons in ground truth, significant differences were observed between the outputs of the models. SegFormer could extract general building shapes but produced jagged edges and false positives. Also, roads and spaces were partially predicted as buildings in densely populated areas. The PAN model contains more noise, partially correctly predicts large buildings, but fails to distinguish small structures. The DeepLabV3+ model provides more detailed segmentation, but in some areas, broken and fragmented building boundaries are noticeable. While it is partially successful in recognizing small buildings, shape distortions are prominent. While UperNet is relatively successful in capturing large-scale buildings, it misses many small and scattered structures. While the FPN model was observed to be particularly effective at distinguishing small, scattered buildings, it was also realized that the model tends to segment buildings in some areas. Conversely, the PSPNet model offered a satisfactory representation of the overall building density, yet it exhibited a loss of fine detail and the potential for blurring of edge patterns. However, its edge accuracy is more satisfactory than SegFormer and PAN. The proposed model, on the other hand, captures both large and small-scale buildings more completely and accurately. It significantly reduces false positives and more clearly reveals

building boundaries. It is observed that the proposed model provides the most reliable results, closest to the ground truth, especially in complex urban areas.

Conclusion and Recommendation

This study addresses the extraction of building boundaries from VHR satellite imagery. For this purpose, a novel Swin Transformer-based deep learning model was developed and evaluated against SOTA models. Experimental analysis conducted on the Massachusetts Building Dataset revealed that the proposed transformer-based approach outperforms all commonly used architectures, such as DeepLabV3+, SegFormer, UPerNet, PAN, FPN and PSPNet. The proposed model achieved the highest results in terms of key evaluation metrics. Specifically, the IoU values of 77.94% and F-score of 86.96% provided significant improvements over SOTA models. The model proved effective at integrating small-scale details with large-scale context, resulting in more precise and reliable identification of buildings in dense urban environments. Tests showed that the model produces complete and well-defined building outlines. The ability to cleanly separate adjacent buildings is critical in cities, as it ensures that subsequent analyses, such as monitoring land use and preparing for emergencies, are based on trustworthy data. Ultimately, this work confirms the significant promise of transformer models for remote sensing. It offers a solution to a long-standing problem in geospatial sciences. However, among the limitations of the study, relying on a single comparison dataset may be considered a limitation for the generalizability of the model. Furthermore, although it is successful in detecting small buildings, poor performance may be observed in complex urban areas. On the other hand, to eliminate doubts regarding the use of black-box deep learning models, a new paradigm of explainable AI strategies encompassing both global and local explanations (Adadi & Berrada, 2018; Teke & Kavzoglu, 2024) will be investigated to generate more reliable and transparent results. There are various labeling errors in the Massachusetts dataset used in the study, so correcting these errors is of great importance. Another issue is that only RGB images were employed in this study, which is a common practice in the current literature. Other spectral bands (e.g., NIR), contextual, and textual features can be utilized. Therefore, future studies plan to test the proposed model on multiple datasets and use data containing elevation information produced using LiDAR or radar in addition to other spectral bands.

Acknowledgments

The authors would like to note that following the submission of this conference paper, the

extended version of the study entitled “DeepSwinLite: A Swin Transformer-Based Light Deep Learning Model for Building Extraction Using VHR Aerial Imagery” was published as a journal article, which can be accessed through <https://doi.org/10.3390/rs17183146>.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138-52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Bittner, K., Cui, S., & Reinartz, P. (2017). Building extraction from remote sensing data using fully convolutional networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-1/W1*, 481–486. <https://doi.org/10.5194/isprs-archives-XLII-1-W1-481-2017>
- Chen, C., Deng, J., & Lv, N. (2020). Illegal constructions detection in remote sensing images based on multi-scale semantic segmentation. In *2020 IEEE International Conference on Smart Internet of Things (SmartIoT)* (pp. 300–303). IEEE. <https://doi.org/10.1109/SmartIoT49966.2020.00053>
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 801–818). Springer. https://doi.org/10.1007/978-3-030-01234-2_49
- Debella-Gilo, M. (2025). Relative performance of super-resolved Sentinel-2 and Copernicus VHR images in mapping built-up areas and building footprints using deep learning. *European Journal of Remote Sensing*, 58(1). <https://doi.org/10.1080/22797254.2025.2517381>
- Dong, X., Cao, J., & Zhao, W. (2023). A review of research on remote sensing images shadow detection and application to building extraction. *European Journal of Remote Sensing*, 57(1). <https://doi.org/10.1080/22797254.2023.2293163>
- Döş, M. E., Seyrek, E. C., & Uysal, M. (2025). Deep learning based building change detection by integrating building footprint data to pre and post-earthquake VHR satellite images from February 6, 2023, Kahramanmaraş earthquake: a case study for Hatay-Antakya. *International Journal of Remote Sensing*, 1–24. <https://doi.org/10.1080/01431161.2025.2557587>
- El Ghoul, S. S., Tayeh, B. A., Baghdadi, A., Alaloul, W. S., & Abu Aisheh, Y. I. (2025). Key factors shaping post-disaster building damage assessment: insights from the Gaza Strip as a conflict zone. *Journal of Asian Architecture and Building Engineering*, 1–21. <https://doi.org/10.1080/13467581.2025.2483992>
- Feng, L., Xu, P., Tang, H., Liu, Z., & Hou, P. (2023). National-scale mapping of building footprints using feature super-resolution semantic segmentation of Sentinel-2 images. *GIScience & Remote Sensing*, 60(1). <https://doi.org/10.1080/15481603.2023.2196154>

- Guo, H., Du, B., Zhang, L., & Su, X. (2022). A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 240–252. <https://doi.org/10.1016/j.isprsjprs.2021.11.005>
- Gupta, A., Dixit, M., & Mishra, V. K. (2025). Advancing urban analysis through joint rooftop segmentation and building height estimation. *Procedia Computer Science*, 260, 947–956. <https://doi.org/10.1016/j.procs.2025.03.278>
- Haidarh, M., Mu, C., Liu, Y., & He, X. (2025). Exploring traditional, deep learning and hybrid methods for hyperspectral image classification: A review. *Journal of Information and Intelligence*. Advance online publication. <https://doi.org/10.1016/j.jiixd.2025.04.002>
- Huang, J., & Yang, G. (2024). Research on urban renewal based on semantic segmentation and spatial syntax: Taking Wuyishan City as an example. In *Proceedings of the International Conference on Smart Transportation and City Engineering* (pp. 1179–1185). IEEE. <https://doi.org/10.1117/12.3024365>
- Jin, H., Fu, W., Nie, C., Yuan, F., & Chang, X. (2023). Extraction of building from remote sensing imagery base on multi-attention L-CAFSSFM and MFFM. *Frontiers in Earth Science*, 11, 1268628. <https://doi.org/10.3389/feart.2023.1268628>
- Kavzoglu, T. & Bilucan, F. (2023). Effects of auxiliary and ancillary data on LULC classification in a heterogeneous environment using optimized random forest algorithm. *Earth Science Informatics*, 16(2), 415–435. <https://doi.org/10.1007/s12145-022-00874-9>
- Li, Y., Hong, D., Li, C., Yao, J., & Chanussot, J. (2024). HD-Net: High-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 209, 51–65. <https://doi.org/10.1016/j.isprsjprs.2024.01.022>
- Li, H., Xiong, P., An, J., & Wang, L. (2018). Pyramid attention network for semantic segmentation. *arXiv Preprint, arXiv:1805.10180*. <https://doi.org/10.48550/arXiv.1805.10180>
- Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. <https://doi.org/10.1109/CVPR.2017.106>
- Liu, G., Diao, K., Zhu, J., Wang, Q., & Li, M. (2024). STransU2Net: Transformer based hybrid model for building segmentation in detailed satellite imagery. *PLoS One*, 19(9), e0299732. <https://doi.org/10.1371/journal.pone.0299732>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9993–10002). IEEE. <https://doi.org/10.1109/ICCV48922.2021.00986>
- Mnih, V. (2013). *Machine learning for aerial image labeling* (Doctoral dissertation, University of Toronto).

Pang, Z., Hu, R., Zhu, W., Zhu, R., Liao, Y., & Han, X. (2024). A Building Extraction Method for High-Resolution Remote Sensing Images with Multiple Attentions and Parallel Encoders Combining Enhanced Spectral Information. *Sensors*, 24(3), 1006. <https://doi.org/10.3390/s24031006>

Patil, A. N., Kawarkhe, C. S., Bukam, M. M., Divkar, A. M., Cherian, M., & Y. I., J. M. (2025). A review on GIS and machine learning frameworks for building footprint detection. In *2025 International Conference on Emerging Trends in Industry 4.0 Technologies (ICETI4T)* (pp. 1–7). IEEE. <https://doi.org/10.1109/ICETI4T63625.2025.11132195>

Pereira, G. A., & Hussain, M. (2024). A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. *arXiv Preprint, arXiv:2408.15178*. <https://arxiv.org/abs/2408.15178>

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 12116–12128. <https://doi.org/10.48550/arXiv.2108.08810>

Rahnemoonfar, M., Chowdhury, T., & Murphy, R. (2023). RescueNet: A high-resolution UAV semantic segmentation dataset for natural disaster damage assessment. *Scientific Data*, 10(1), 913. <https://doi.org/10.1038/s41597-023-02799-4>

Rajamani, T., Sevugan, P., & Ragupathi, S. (2022). Automatic building footprint extraction and road detection from hyperspectral imagery. *Journal of Electronic Imaging*, 32(1), 011005. <https://doi.org/10.1117/1.JEI.32.1.011005>

Rastogi, K., Bodani, P., & Sharma, S. A. (2020). 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>

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Teke, A., & Kavzoglu, T. (2024). Exploring the decision-making process of ensemble learning algorithms in landslide susceptibility mapping: Insights from local and global explainable AI analyses. *Advances in Space Research*, 74(8), 3765–3785. <https://doi.org/10.1016/j.asr.2024.06.082>

Trigka, M., & Dritsas, E. (2025). A Comprehensive Survey of Deep Learning Approaches in Image Processing. *Sensors*, 25(2), 531. <https://doi.org/10.3390/s25020531>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 5998–6008). Curran Associates. <https://doi.org/10.48550/arXiv.1706.03762>

- Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018, 7068349. <https://doi.org/10.1155/2018/7068349>
- Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 418–434). Springer. https://doi.org/10.1007/978-3-030-01249-6_26
- Xiao, X., Guo, W., Chen, R., Hui, Y., Wang, J., & Zhao, H. (2022). A Swin Transformer-based encoding booster integrated in U-shaped network for building extraction. *Remote Sensing*, 14(11), 2611. <https://doi.org/10.3390/rs14112611>
- 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*, 34, 12077–12090. <https://doi.org/10.48550/arXiv.2105.15203>
- Xie, Y. H., Huang, B. S., & Li, F. (2025). UnetTransCNN: Integrating transformers with convolutional neural networks for enhanced medical image segmentation. *Frontiers in Oncology*, 15, 1467672. <https://doi.org/10.3389/fonc.2025.1467672>
- Yang, D., Gao, X., Yang, Y., Guo, K., & Xu, L. (2025). Advances and future prospects in building extraction from high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18(8), 6994–7016. <https://doi.org/10.1109/JSTARS.2025.3538662>
- Yılmaz, E. Ö., Teke, A., & Kavzoğlu, T. (2025). A performance analysis of U-Net and U-Net++ in building footprint extraction using XAI. *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, Şile, İstanbul, Türkiye, 1–4. <https://doi.org/10.1109/SIU66497.2025.11111828>
- Zhang, H., Zheng, X., Zheng, N., & Shi, W. (2022). A multiscale and multipath network with boundary enhancement for building footprint extraction from remotely sensed imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 8856–8869. <https://doi.org/10.1109/JSTARS.2022.3214485>
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2881–2890. <https://doi.org/10.1109/CVPR.2017.660>
- 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), 3983. <https://doi.org/10.3390/rs12233983>