Engineering and Technology Journal
Journal homepage: https://etj.uotechnology.edu.iq

A segmentation of buildings from multi-scene drone images using deep learning models: case study- a car multi-story garage in the camp of the university of technology

Akram H. Jalil*, Imzahim A. Alwan

Civil Engineering Dept., University of Technology-Iraq, Alsina'a street, 10066 Baghdad, Iraq.
*Corresponding author Email: bce.23.30@grad.uotechnology.edu.iq

HIGHLIGHTS

- A novel multi-view deep learning framework was developed using MobileNetV2, ResNet50, VGG16, and InceptionV3.
- The framework achieved a segmentation accuracy of 93%, surpassing conventional building segmentation methods.
- Features were integrated with a Vision Transformer to enhance segmentation performance.
- The approach was validated on UAV imagery of a multi-story garage at the University of Technology.
- Strong potential for urban planning, drone mapping, and image analysis applications.

Keywords:

Segmentation
Deep learning
Pre-trained models
Vision transformer
Multi-views
Drone images

ABSTRACT

Accurate building segmentation is essential for urban planning, monitoring, and mapping. Most deep learning approaches rely on single-view images, limiting segmentation accuracy due to the loss of spatial context. This re-search proposes a multi-view deep learning framework that integrates features extracted from four pre-trained CNN models—MobileNetV2, Res-Net50, VGG16, and InceptionV3—to capture multi-angle and multi-scale details. A vision transformer is employed to fuse and refine these features, enhancing global context and boundary precision. The proposed method was evaluated using UAV imagery of a multi-story garage at the University of Technology, captured by a DJI Mavic 2 Pro with a high-resolution Hasselblad L1D-20c camera. Experiments on the augmented building dataset achieved a segmentation accuracy of 93%, with notable improvements in Intersection over Union (IoU) and F1-score compared to standard CNN-based approaches. These results demonstrate the robustness of the model under varying lighting and occlusion conditions and highlight its potential for high-precision urban building segmentation.

1. Introduction

Semantic segmentation, which involves assigning a class label to every pixel in an image, is a fundamental task in computer vision crucial for scene under-standing [1]. In remote sensing, semantic segmentation is especially important for analyzing geospatial data with applications in land use classification, environ-mental monitoring, infrastructure planning, and urban development [2,3]. As cities expand and change, the need for accurate, automated extraction of building footprints and urban features from high-resolution imagery has become increasingly vital [4].

Remote sensing platforms—including satellites, drones, and aircraft—capture extensive imagery with high spatial and spectral resolution, enabling detailed observation of Earth's surface [2]. However, these images are complex due to factors like varying lighting, shadows, occlusions, and the diverse, dense arrangement of urban objects such as buildings, roads, vegetation, and vehicles [3,5]. These complexities make precise segmentation challenging, as similar textures and close proximities often cause overlap and ambiguity [6].

Building segmentation from aerial or satellite images is particularly critical for urban applications such as growth modeling, disaster response (e.g., post-earthquake damage assessment), population estimation, infrastructure management, and cadastral mapping [4]. However, this task is complicated by overlap-ping structures, diverse roof shapes, perspective distortions, and

http://doi.org/10.30684/etj.2025.158830.1931
Received 15 April 2025; Received in revised form 26 July 2025; Accepted 28 July 2025; Available online 30 September 2025
2412-0758/University of Technology-Iraq, Baghdad, Iraq

background clutter [6]. Traditional image processing techniques typically struggle with these complexities due to their limited ability to extract robust features and contextual information [5].

The advent of deep learning, especially convolutional neural networks (CNNs), has revolutionized semantic segmentation [7]. Encoder-decoder architectures like U-Net [3], SegNet [8], and RedNet [9] learn hierarchical image features and have proven successful across domains such as medical imaging, road scene understanding, and agriculture. In remote sensing, CNNs are now the standard for pixel-wise classification [5]. Nevertheless, significant challenges remain, particularly in complex urban environments [5]. A key limitation of conventional CNN-based segmentation models is their reliance on single-view images. While computationally efficient, single-view inputs often miss important spatial and structural details due to occlusion, lighting vari-ation, or angle, leading to incomplete segmentation [6]. Moreover, CNNs' fixed receptive fields limit their capacity to model long-range dependencies and global context, critical for accurately segmenting buildings of varying size and shape in high-resolution imagery [7]. Although encoder-decoder networks have introduced innovations like skip connections (U-Net [3]) and residual learning (RedNet [9]) to preserve spatial details better, performance often degrades in scenes with structural complexity or multi-scale objects [6]. Some models, such as SegNet [8], use pooling indices to retain spatial information, but may still fail to capture fine boundaries. Others like RedNet leverage RGB-D and residual links but are tailored to indoor scenes and do not generalize well to outdoor aerial imagery with broader context [9].

To overcome these issues, researchers have integrated techniques like spatial pyramid pooling and attention mechanisms into encoder-decoder networks to improve multi-scale context awareness. For example, DeepLabv3+ employs atrous spatial pyramid pooling, and PSPNet uses pyramid pooling to aggregate global context [10]. However, these models still mostly rely on single-view in-puts, limiting the spatial diversity and completeness of segmentation [6].

Despite advances from prior work—ranging from U-Net [3] and SegNet [8] to attention-based models like DeepLabv3+ and PSPNet [10]—most approaches inadequately capture the full spatial complexity of urban scenes due to their dependence on single-view imagery and constrained context modeling. This re-search extends these foundations by combining hierarchical feature extraction, spatial detail preservation, and multi-scale context aggregation within a unified multi-view deep learning framework. It addresses critical gaps such as insufficient structural representation, occlusion robustness, and scale variation han-dling [6].

The proposed method integrates multiple pre-trained CNN backbones—MobileNetV2, ResNet50, VGG16, and InceptionV3—selected for their complementary strengths in feature extraction, scale sensitivity, and spatial resolution [6, 11]. This multi-backbone, multi-view approach extracts a richer and more diverse set of features compared to single-architecture models. A vision trans-former module is then used to fuse these features, capturing global context and long-range dependencies via self-attention, which CNNs alone cannot effectively model [6].

Applied to real-world UAV imagery of a multi-story garage under diverse lighting, scale, and occlusion conditions, the framework achieves a segmentation ac-curacy of 93%, outperforming standard models such as U-Net [3] and SegNet [8] in metrics like Intersection over Union (IoU), F1-score, and boundary precision. Data augmentation techniques further improve generalization across building shapes and environmental scenarios [6].

2. Contributions

This study introduces several key innovations advancing remote sensing-based building segmentation:

1. **Multi-view learning integration:** Incorporates multiple image perspectives to overcome limitations of single-view methods, improving completeness and structural detail capture [6,11].
2. **Fusion of multiple CNN backbones**: Leverages four complementary pre-trained networks to extract diverse, representative features at multiple scales and depths [6,11].
3. **Transformer-based feature fusion:** Employs a vision transformer to model long-range dependencies and global context, enhancing segmentation accuracy in complex urban layouts.
4. **Application to real-world UAV imagery:** Demonstrates practical applicability and robustness in real urban scenarios rather than synthetic or benchmark datasets.
5. **Superior segmentation performance:** Achieves significantly higher ac-curacy and boundary precision than conventional CNN-based models, showing robustness to scale, occlusion, and lighting variations.

This research presents a unified framework that addresses the core limitations of existing segmentation methods by combining multi-view inputs, diverse CNN feature extraction, and transformer-based fusion. It offers a scalable, accurate, and robust solution for urban building segmentation from aerial imagery, con-tributing to smarter urban analysis and infrastructure development. The approach lays the groundwork for future exploration of hybrid transformer-CNN architectures and real-time aerial segmentation applications [6]. The conclusions of this study suggest that our strategy is superior to the existing one.

3. Methodology

Over 50% of the world's population lives in urban areas, and this trend is predicted to continue growing. Rapid urbanization necessitates advanced techniques for addressing urban planning issues with unique properties, including reduced scanning time, cost, and accuracy [12]. By using multi-view imagery, this study aims to improve building segmentation accuracy in drone images. Convolutional Neural Networks (CNNs) have played a major role in these developments. Their novel structures and layers—ReLU, Dropout, Batch Normalization, early halting, and previously trained deep models—have considerably improved object recognition, notably for buildings. CNNs have significantly enhanced object detection [3].

We provide a multi-view image approach that takes into account several important aspects to accomplish precise building segmentation. First, it is critical to take into account each object's segmentation accuracy simultaneously when objects from various viewing angles are included in the images. Furthermore, pictures with many perplexing geo-objects often affect accurate segmentation [13]. Using multi-view fusion, this work combines the advantages of four well-known pre-trained CNN systems: MobileNetV2 [13], ResNet50 [14], VGG16 [15], and InceptionV3 [16]. By efficiently using multiple perspectives, this technique enhances the semantic segmentation of buildings and the efficiency of current segmentation models.

This method gathers supplementary information by considering multiple dataset perspectives, enabling a more thorough search for items or structures. Our technique combines multiple views to enhance remote sensing image segmentation models and address the limitations of single-view analysis. The benefits of our suggested multi-view segmentation method over conventional single-view techniques are highlighted in this paper. To illustrate the efficacy of our method, we conduct extensive testing and assessments on an actual dataset. The segmentation accuracy is much improved by our method, which combines characteristics from all four models. Furthermore, our study supports environmental surveillance, building maintenance, and durability. In addition to improving segmentation accuracy and ability, the suggested multi-view technique collects complementary information from many dataset perspectives, resulting in a broader view of the urban area. Developing urban areas sustainably requires precise building segmentation to improve resource usage and ecological design. Accurately identifying building structures improves maintenance methods, enabling targeted improvements and minimizing the effect on the environment. Consequently, our study not only addresses the current challenges in building segmentation technology but also advances the goals of sustainability and promotes reliable built environment management. To achieve precise building segmentation in remote sensing data, our work aims to develop and utilize a novel multi-view pre-trained model and Vision Transformer (ViT).

3.1 Dataset

Limited testing and adverse effects on the findings might come from poor data quality [17]. The research uses 10,000 multi-view drone photos with 512 × 512-pixel resolution and associated masks to build the model as illustrate in Figure 1. Using 20% for testing, 20% is for validation, and 60% is for training [18,19]. The collection comprises high-resolution RGB images of drone-captured scenes, showing houses, roads, plants, and other structures.

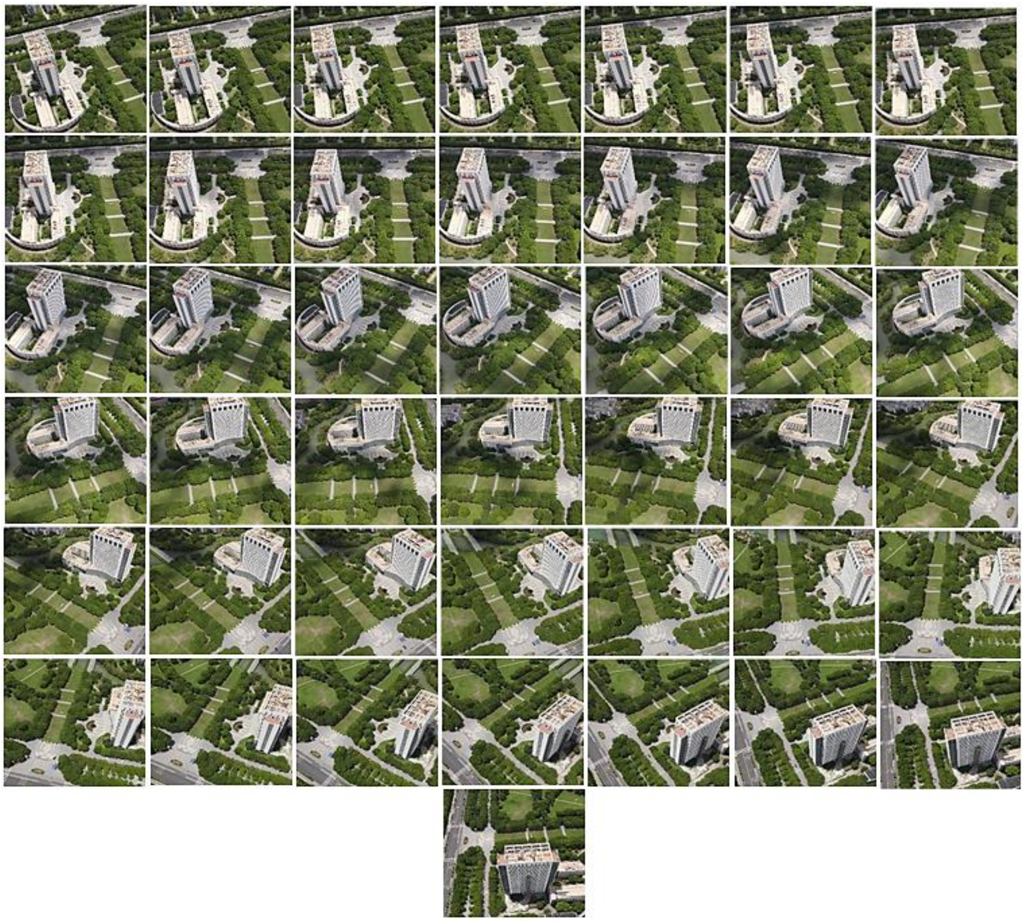


Figure 1: The SUES-200 dataset, used for the model training, validation, and testing [20]

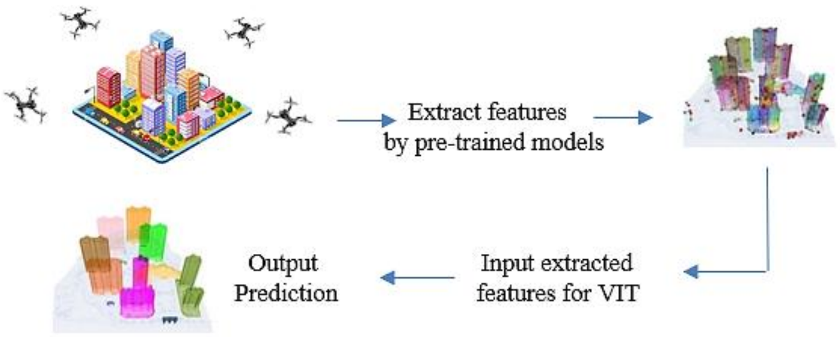
3.2 Proposed approach

Despite recent progress in building segmentation, many existing approaches rely on single-view images, which limit the model's ability to capture complex spatial structures and diverse perspectives of buildings. These limitations result in lower segmentation accuracy and reduced generalization, particularly in environments with architectural diversity and occlusion. To address these gaps, this study presents a unique approach that incorporates multi-view building imaging for accurate building segmentation. Our method utilizes multiple pre-trained models for multi-view analysis, extracting characteristics from images to achieve a common goal. After that, a Vision Transformer takes these features as inputs and uses them to segment buildings. We take 100 images of the garage in a circular layout, designated as view1, view2, ..., view100. Applying pre-trained base models, each view processes all angles to extract characteristics from drone input photos individually. The models include MobileNetV2, ResNet50, VGG16, and InceptionV3. These models were chosen as pre-trained bases because they can capture certain properties at different spatial resolutions. Multi-view fusion combines information from diverse angles to improve our model's segmentation. This multi-view fusion strategy directly addresses the gap in the literature regarding limited spatial awareness in conventional segmentation models. By integrating different viewpoints, the model can overcome occlusion issues and capture subtle structural differences that are often missed in single-perspective methods.

This approach enhances the model's ability to recognize complex structures and provides a more comprehensive understanding of details. The model's segmentation accuracy is enhanced by incorporating feature maps that encompass both local and global information through the use of multi-views. Detailed architectural details and different perspectives improve object segmentation in this multi-view model. It enables the model to preserve the spatial and channel-level properties required for effective segmentation while utilizing both local and global data. By using extra data from many views on the same data, rather than depending on just one, the approach allows for a more comprehensive evaluation of the dataset. The multi-view method provides additional information from views 1, 2..., 100 of our datasets. This enables the model to analyze the data more quickly, which could help it understand it better and create more accurate segments.

The second step uses the Vision Transformer (ViT) architecture to incorporate pre-trained model characteristics into the final target, improving segmentation. Because of its flexibility, the model may change how it behaves in response to the features and complexity of the incoming data. Multi-view integration and ViT architecture feature abstraction enable our model to obtain special data or high-level semantic information, depending on the building segmentation requirements. Our technique enables the gradual improvement of characteristics by creating a segmentation map that combines accurate data from lower abstraction levels with more generic or high-level data, thereby enhancing understanding. The approach creates a segmentation map by improving features and retrieving more sophisticated data from different angles. Low-level abstraction characteristics are coarser and capture unique visual patterns. These characteristics capture local image information, which is essential for recognizing small objects and fine details. The characteristics get more complicated as we reach higher abstraction levels. They encode information that is larger and more generic, with a particular emphasis on the overall shape, structure, or semantic importance of the image. These higher-level characteristics identify key aspects and give context by showing how image portions relate.

Using a new multi-view fusion strategy, our deep multi-view segmentation solution draws on the Vision Transformer (ViT) architecture and numerous pre-trained models. In the practical part of the research, 100 views of a car multi-story garage at the University of Technology were used and then processed by pre-trained models, such as MobileNetV2, ResNet50, VGG16, and InceptionV3, to make strong segmentation assessments across borders. These models were chosen because they capture distinct properties at different spatial resolutions, improving both local and global feature extraction. This technique collects local and global data, improving the model's segmentation. The main contribution of this work is a novel integration of multi-view image fusion with the Vision Transformer (ViT) architecture, supported by multiple pre-trained CNN backbones, to enhance building segmentation performance. This study aims to improve segmentation accuracy by leveraging comprehensive scene understanding gained from diverse spatial perspectives, offering a robust and scalable solution for real-world applications in urban mapping and 3D reconstruction. Figure 2 illustrates the structure of the proposed multi-view segmentation.



-

Figure 2: Process structure for the proposed multi-view segmentation

3.3 Why using vision transformer (ViT)?

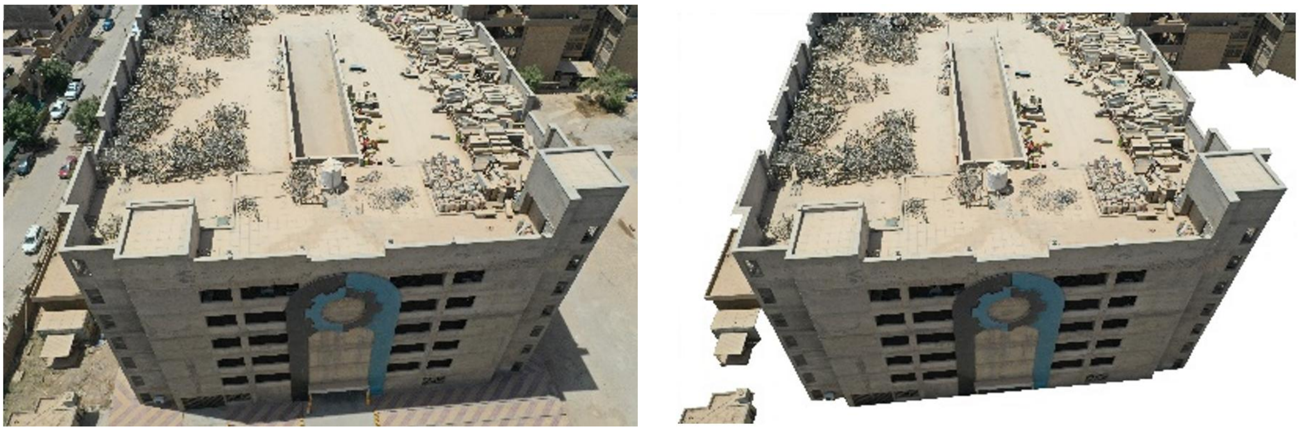
Within the boundaries of semantic segmentation through the utilization of Vision Transformers (ViTs), it has been demonstrated that the incorporation of multi-view features derived from pre-trained models may greatly enhance the classification accuracy [21]. The purpose of this strategy is to collect both local and global contextual information by utilizing the capabilities of both transformers and convolutional neural networks (CNNs). When it comes to segmentation tasks, the model can successfully manage complicated structures and fine details since it incorporates hierarchical spatial characteristics from pre-trained models with the ability of ViTs to represent long-range relationships.

The combination of these properties improves the model's capacity to generalize over a wide range of datasets and demanding settings, such as occlusions or different lighting conditions. Furthermore, the utilization of pre-trained models reduces the need for large task-specific training data, thereby accelerating convergence and decreasing the likelihood of overfitting. In the context of urban building segmentation, where a wide range of architectural styles and environmental conditions provide considerable challenges, the integration of multi-view characteristics is very useful [22].

4. Assessment of performance

4.1 Materials for the study

The dataset we used to implementation in our suggested method for building segmentation is made up of 100 images taken by using a DJI Mavic 2 Pro UAV, built with a Hasselblad L1D-20c compact camera containing a 1-inch CMOS sensor (13.2 mm width × 8.8 mm height), focal length of 10.26 mm, resolution of 20 megapixels, and image size of 5472 × 3648 pixels for car multi-story garage of camp of University of Technology [12]. As seen in Figure 3, these masks match the images' spatial aspects.



Original image **Buildings mask**

Figure 3: Original Image and Its Building Mask

To evaluate the efficacy of our suggested approach in comparison to current methods, we employ four common metrics: Accuracy (ACC) as in Equation 1, Precision (Prec.) as in Equation 2. Recall (Rec) as in Equation 3 and F1-score as in Equation 4. Higher metrics indicate greater model performance. Overall performance is measured by accuracy, whereas precision, recall, and F1-score assess a model's capacity to handle unbalanced datasets and categorize complicated categories [24]. The following equations are used to compute these indicators:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{F1} = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} \tag{4}$$

- **Accuracy (Acc.):** The percentage of predictions that were right, including true positives and negatives.
- **Precision:** Is defined as the ratio of the number of accurately predicted positive observations (true positives) to the total number of predicted positive observations (predicted positives plus false positives). The number of selected things that are relevant is shown.
- **Recall (Rec.):** The percentage of real positives, which includes both true positives and false negatives, in comparison to the percentage of accurate positive observations that were anticipated (also known as "true positives").
- **F1-score:** Precision-recall harmonic mean. It balances accuracy and recalls with one metric [3].

Below, we compare our approach to other popular segmentation methods:

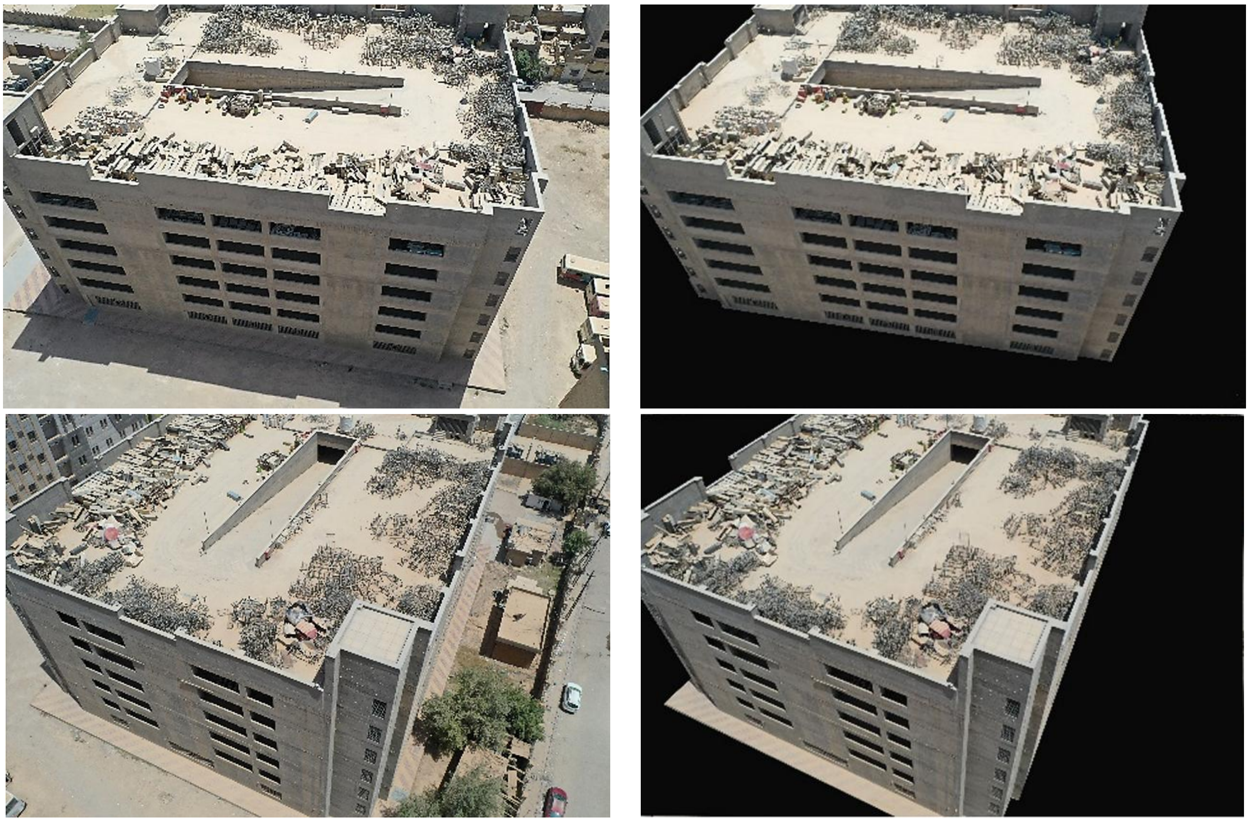
- Deep learning with satellite imagery and geospatial data (U-Net model) [3]. Modified U-Net model for Image segmentation [3].

- Segmenting Buildings model in Satellite Images (seg-model) [3].
- Support Vector Machine Classifier (SVM) [23].

The True Positive (TP) and True Negative (TN) rates evaluate the test's ability to identify positive and negative cases, respectively. Positive instances wrongly categorized as negative are called False Negative (FN), whereas negative cases incorrectly labeled as positive are called False Positive (FP). These rates reveal the model's ability and accuracy in recognizing real-world scenarios using actual data [3].

4.2 Implementation approach

Several pre-trained models, including CNN-based architectures and Vision Transformers (ViTs), can extract features. Before being supplied to the Vision Transformer for additional processing, these feature maps are fused or integrated. The ViT can identify intricate buildings and structures more successfully thanks to the expanded feature set, which improves segmentation outcomes. The final segmentation results employing a Vision Transformer, numerous pre-trained models, and multi-view images taken by drones are shown in Figure 4 [25,26].



Original image **Our building segmentation**

Figure 4: Last results of building segmentation

4.3 Practical application in the context of building

Our results have great promise for building applications. The multi-view model's distinctive properties, which are enhanced by collecting several target object viewpoints, are relevant to many building-related domains. Our technology mostly identifies buildings precisely. A thorough comprehension of building components is achieved by combining MobileNetV2, ResNet50, VGG16, and InceptionV3 characteristics. Urban planning relies on correct building boundaries for zoning and land-use choices. The enhanced dataset improves building comprehension, which advances digital twin technology. Remote sensing and GIS are essential for addressing building segmentation challenges, improving spatial analysis, and enabling informed urban planning decisions by authorities and policymakers [27,28,29]. Additionally, our segmentation method helps build management by revealing the building's spatial layout. This helps building managers optimize maintenance schedules, diagnose structural challenges, and plan interventions more precisely.

Our method also helps urban planners allocate resources for construction and infrastructure projects. Accurate building segmentation helps municipalities and urban planners manage resources for sustainable and planned urban growth. Construction-related environmental impact studies need detailed building segmentation. Our technique helps identify and analyze building footprints, enhancing environmental impact assessments, and encouraging sustainable development. Our flexible technique offers real-time building change monitoring, which is essential for tracking modifications, detecting illegal changes, and complying with building laws. Additionally, our segmentation results may be readily linked with BIM systems. This connection expands BIM model data, giving engineers, architects, and building professionals complete knowledge of building structures.

4.4 Experimental results

By utilizing the whole dataset, we were able to analyze the efficacy of our suggested strategy in comparison to other competitor approaches. Results are provided in Table 1, where the values that are the greatest (which indicate the best possible performance) are underlined in bold characters.

The research's results show that our approach works better than current modern segmentation techniques, exhibiting more accuracy when compared to comparable methods. Accuracy (ACC), Precision (Prec.), Recall (Rec.), and F1-score are some of the essential metrics that are included in Table 1. These metrics provide a complete assessment of the performance of the model in the building segmentation work. But it's important to note that IoU directly from accuracy, precision, recall, or F1-score isn't precise, because IoU depends on the actual overlap of predicted vs. true regions.

Table 1: Comparison of the classification strengths of different strategies

Model	Acc.	Prec.	Rec.	F1-score	IoU
SVM [23]	49%	51%	52%	46%	31%
U-Net [3]	81%	79%	78%	76%	62%
Modified U-Net [3]	87%	81%	81%	81%	69%
Seg-model [3]	57%	60%	65%	53%	36%
Multi-view U-Net [2]	91%	86%	88%	87%	75%
Ours	93%	88%	90%	89%	85%

The simple Support Vector Machine (SVM) [30,31] has 49% accuracy, 51% precision, 52% recall, and 46% F1-score. These results are significantly lower than our deep learning-based multi-view pre-trained models. SVM, a non-deep learning algorithm, fails to capture complicated spatial connections and fine features in remote sensing images, especially when buildings change in size, form, and context. However, deep learning methods like the multi-view pre-trained model in our study excel at segmentation tasks. They are better at this type of task because they can independently learn hierarchical representations, recognize complex structures, and conform to different spatial environments.

In U-Net [3], the deep learning model improves across all measures, with 81% accuracy, 79% precision, 78% recall, and 76% F1-score. U-Net's encoder-decoder structure collects contextual and geographical information, improving performance above older techniques. The Modified U-Net [3] enhances accuracy, precision, recall, and F1-score to 87%, 81%, and 81%, respectively. Advancements like skip connections let the model maintain fine-grained characteristics during segmentation. The Seg-model [3] has 57% accuracy, 60% precision, 65% recall, and 53% F1-score. Although the Seg-model performs worse than the SVM, it outperforms it, demonstrating the benefits of deep learning in segmentation.

The multi-view U-Net outperforms competitors with 91% accuracy, 86% precision, 88% recall, and 87% F1-score. This deep model uses multi-view fusion to merge MobileNetV2 and ResNet50 data. These complex architectural designs and varied views improve the model's remote sensing building segmentation accuracy.

In contrast to previous studies, our model not only performs well-known designs like U-Net and multi-view U-Net, but it also introduces a novel approach for integrating feature richness from multiple angles. Multiple pre-trained models help the system learn high-level abstract representations and low-level spatial features, improving generalization and segmentation.

These findings advance previous work on multi-view and ensemble techniques for remote sensing segmentation. Our technique demonstrates that increasing the number of views and model diversity can enhance segmentation accuracy and reliability, unlike previous techniques that relied on a limited number of views or less varied feature fusion.

In conclusion, our suggested multi-view, pre-trained combined model improves on both conventional and cutting-edge techniques in terms of performance and flexibility. The model is an offering option for future applications in developing segmentation and more general remote sensing tasks because of its capacity to include many views and extract richer feature representations.

5. Conclusion

This study proposed a novel multi-view approach to urban building segmentation, aiming to address the limitations of single-view segmentation methods commonly used in remote sensing data. As expected, incorporating images captured from multiple angles significantly enhanced the model's understanding of complex urban structures. Compared to traditional single-view methods, our multi-view strategy achieved a notable improvement in segmentation accuracy, reaching 93%—a result that outperforms widely-used models such as SVM, U-Net, Modified U-Net, and Seg-model, particularly in terms of precision, recall, and F1-score. The experimental results confirm our hypothesis that integrating complementary features from several pre-trained CNN architectures—MobileNetV2, ResNet50, VGG16, and Inception_v3—would enhance model robustness and accuracy. This combination allowed the model to leverage the strengths of different architectures in processing spatial details from diverse perspectives. This novelty distinguishes this research from earlier work that typically relies on single-architecture or single-view inputs.

By outperforming established baselines on real-world datasets, this work advances the current body of knowledge in remote sensing and semantic segmentation. It demonstrates that a multi-view deep learning framework not only increases segmentation performance but also introduces a scalable and adaptable solution for urban analysis tasks. The improved accuracy directly contributes to better mapping, infrastructure monitoring, and urban planning, aligning with sustainability goals through more informed environmental and resource management strategies.In addition, the study lays a foundation for broader applications beyond building segmentation. The framework shows potential for adaptation to other tasks, including smart city development, disaster management, land use classification, and environmental monitoring. Future work could extend the model to larger and more heterogeneous datasets to assess generalizability across different geographies. Incorporating recent deep learning innovations and alternative data sources, such as LiDAR or hyperspectral imagery, may further increase the model's versatility and performance.

This research not only confirms the value of multi-view learning in improving segmentation tasks but also contributes a novel, hybrid architectural framework that can inform both academic and applied efforts in remote sensing data and urban analytics.

Author contributions

Conceptualization, **A. Jalil,** and **I. Alwan**; data curation, **A. Jalil**; formal analysis, **A. Jalil**; investigation, **A. Jalil,** and **I. Alwan**; methodology, **A. Jalil,** and **I. Alwan**; project administration, **I. Alwan**, resources, **A. Jalil,** and **I. Alwan**; software, **A. Jalil**; supervision, **I. Alwan**; validation, **A. Jalil,** and **I. Alwan**; visualization, **I. Alwan**; writing—original draft preparation, **A. Jalil**; writing—review and editing, **A. Jalil,** and **I. Alwan**. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

[1] A. Noori, S. Shaker, R. A. Azeez, Street Scene understanding via Semantic Segmentation Using Deep Learning, Eng. Technol. J., 40 (2022) 588-594. http://doi.org/10.30684/etj.v40i4.2120

[2] M. P. Barbato, F. Piccoli, P. Napoletano, Ticino: A multi-modal remote sensing dataset for semantic segmentation, Expert. Syst. Appl., 249 (2014) 123600. http://doi.org/10.1016/j.eswa.2024.123600

[3] S. El Hajjar, H. Kassem, F. Abdallah, H. Omrani, Enhancing building segmentation by deep multiview classification for advancing sustainable urban development, J. Build. Eng., 83 (2024) 108421. http://doi.org/10.1016/j.jobe.2023.108421

[4] S. Chen, Y. Ogawa, C. Zhao, Y. Sekimoto, Large-scale individual building extraction from open-source satellite imagery via super-resolution-based in-stance segmentation approach, ISPRS J. Photogramm. Remote Sens., 195 (2023) 129-152. http://doi.org/10.1016/j.isprsjprs.2022.11.006

[5] I. Kassar Akeab, Improved Image Segmentation Algorithm Using Graph-Edges, Eng. Technol. J., 28 (2010) 2247-2258. https://doi.org/10.30684/etj.28.11.14

[6] W. Boulila, H. Ghandorh, S. Masood, A. Alzahem, A. Koubaa, F. Ahmed, Z. Khan, J. Ahmad, A transformer-based approach empowered by a self-attention technique for semantic segmentation in remote sensing, Heliyon, 10 (2024) e29396. https://doi.org/10.1016/j.heliyon. 2024.e29396

[7] K. O'Shea, R. Nash, An Introduction to Convolutional Neural Networks, arXiv:1511.08458v2 [cs.NE], (2015)1-11. https://doi.org/10.48550/arXiv.1511.08458

[8] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, arXiv:1511.00561v3 [cs.CV], (2015)1-14. https://doi.org/10.48550/arXiv.1511.00561

[9] J. Jiang, L. Zheng, F. Luo, Z. Zhang, RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation, arXiv:1806.01054v2 [cs.CV], (2018)1-14. https://doi.org/10.48550/arXiv.1806.01054

[10] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, W. Qi, Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encod-er-Decoder with Spatial Pyramid Pooling, IEEE Access, 7 (2019) 128774-128786. https://doi.org/10.1109/ACCESS.2019.2940527

[11] F. Hassan AL Kathy, Digital Video Automatic Segmentation Algorithms Using Edge Detection, Eng. Technol. J., 28 (2010) 2405-2412. https://doi.org/10.30684/etj.28.12.10

[12] R. M. Ridha, I. A. Alwan, H. S. Ismael, Accuracy assessment of 3D model reconstructed from UAV images by the distribution of the ground control points (GCPs), AIP Conf. Proc., 3105, 2024, 050078. https://doi.org/10.1063/5.0212203

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, arXiv:1801.04381v4 [cs.CV], (2018)1-14. https://doi.org/10.48550/arXiv.1801.04381

[14] A. S. B. Reddy, D. S. Juliet, Transfer learning with RESNET-50 for malaria cell-image classification, International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, 945-949. https://doi.org/10.1109/ICCSP.2019.8697909

[15] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556v6 [cs.CV], (2014)1-14. https://doi.org/10.48550/arXiv.1409.1556

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, arXiv:1512.00567v3 [cs.CV], (2015)1-10. https://doi.org/10.48550/arXiv.1512.00567

[17] M. Ressan, R. Hassan, Improving Machine Learning Performance by Eliminating the Influence of Unclean Data, Eng. Technol. J., 40 (2022) 546-539. http://doi.org/10.30684/etj.v40i4.2010

[18] M. Qasim, J. B. Al-Dabbagh, A. N. Abdalla, M. M. Yusoff, G. Hegde, Radial Basis Function Neural Network Model for Optimizing Thermal Annealing Pro-cess Operating Condition, Nano Hybrids, 4 (2013) 21-31. https://doi.org/10.4028/www.scientific.net/NH.4.21

[19] Y. A. Khudhaier, F. S. Kadhim, Y. K. Yousif, Using Artificial Neural Network to Predict Rate of Penetration from Dynamic Elastic Properties in Na-siriya Oil Field, Iraqi J. Chem. Pet. Eng., 21 (2020) 7-14. https://doi.org/10.31699/IJCPE.2020.2.2

[20] R. Zhu, L. Yin, M. Yang, F. Wu, Y. Yang, W. Hu, SUES-200: A Multi-height Multi-scene Cross-view Image Benchmark Across Drone and Satellite, arXiv:2204.10704v2 [cs.CV], (2022)1-16. https://doi.org/10.48550/arXiv.2204.10704

[21] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing, 452 (2021) 48-62. https://doi.org/10.1016/j.neucom.2021.03.091

[22] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H.S. Torr, L. Zhang, Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers, arXiv:2012.15840v3 [cs.CV], (2020)1-12. https://doi.org/10.48550/arXiv.2012.15840

[23] M. H. Khudhur; I. A. Alwan, N. A. Aziz, Comparative study of supervised classification methods of land cover mapping using remote sensing data: A case study in Al-Hawija district/Iraq, AIP Conf. Proc., 3105, 2024, 050070. https://doi.org/10.1063/5.0213746

[24] R. M. Ridha, I. A. Alwan, H. S. Ismael, Accuracy assessment of UAV automated 3D city model for urban planning, AIP Conf. Proc., 2793 (2023) 020004. https://doi.org/10.1063/5.0162664

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, arXiv:1802.02611v3 [cs.CV], (2018)1-18. https://doi.org/10.48550/arXiv.1802.02611

[26] M. Belgiu, L. Drăguţ, Comparing supervised and unsupervised multiresolu-tion segmentation approaches for extracting buildings from very high-resolution imagery, ISPRS J. Photogramm. Remote Sens., 96 (2014) 67-75. https://doi.org/10.1016/j.isprsjprs.2014.07.002

[27] S. A. Mustafa, N. A. Aziz, I. A. Alwan, Geospatial Suitability Mapping for Sustainable Energy Site Selection in Iraq, Eng. Technol. Appl. Sci. Res., 15 (2025) 25192-25198. https://doi.org/10.48084/etasr.11135

[28] R. Chen, X. Li, J. Li, Object-based features for house detection from RGB high-resolution images, Remote Sens. (Basel), 10 (2018) 451. https://doi.org/10.3390/rs10030451

[29] N. A. Aziz, I. A. Alwan, O. E. Agbasi, Integrating remote sensing and GIS techniques for effective watershed management: a case study of Wadi Al-Naft Basins in Diyala Governorate, Iraq, using ALOS PALSAR digital elevation model, Appl. Geomat.,16 (2024) 67-76. https://doi.org/10.1007/s12518-023-00540-9

[30] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey on support vector machine classification Applications, challenges and trends, Neurocomputing, 408 (2020) 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

[31] B. Jasim, O. Jasim, A. AL-Hameedawi, Evaluating Land Use Land Cover Classification Based on Machine Learning Algorithms, Eng. Technol. J., 42 (2024) 557-568. http://doi.org/10.30684/etj.2024.144585.1638