3D Building Model Segmentation using GNN and ViT

Hanis Rashidan [1], Ivin Amri Musliman [1], Alias Abdul Rahman [1], Gurcan Buyuksalih [2]

[1] 3D GIS Research Lab, Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia, Johor Bahru, Malaysia - mhanis27@gmail.com, ivinamri@utm.my, alias@utm.my
[2] Institute of Marine Science and Management, Istanbul University, Turkiye - gurcanb@istanbul.edu.tr

Keywords: Semantic segmentation, 3D building models, Graph Neural Networks, Vision Transformers.

Abstract

Reliable semantics in 3D building models support practical urban tasks such as planning, asset inventory, and maintenance. This paper presents an approach that pairs graph-based geometry (GNN) with image-based appearance (ViT) to improve component segmentation. A Graph Neural Network (GNN) is first applied to the building mesh to capture structural cues and produce initial labels. Multi-view 2D projections (orthographic and perspective) are then rendered and processed with a Vision Transformer (ViT) to recover visual patterns related to windows, doors, roofs, and walls. The two streams are reconciled through a simple consensus fusion that projects ViT predictions back onto the 3D geometry and refines the labels. In experiments, the proposed pipeline improves accuracy and class-wise consistency over a GNN baseline, with clearer gains on small or visually ambiguous elements.

1. Introduction

The ongoing rapid urbanization and population growth across the globe have triggered significant challenges in urban planning, infrastructure management, and resource allocation. As urban environments evolve into highly complex spaces, city planners, engineers, and decision-makers increasingly rely on accurate and detailed 3D representations of built structures for informed decision-making. A critical aspect of generating useful 3D building models involves semantic segmentation – assigning meaningful and precise labels to various architectural components such as windows, doors, roofs, and walls. These labels enhance spatial analyses, infrastructure evaluation, disaster mitigation strategies, energy efficiency assessments, and various other urban management tasks (Stoter et al., 2020; Biljecki et al., 2015).

Over recent years, substantial research has been dedicated to developing automated approaches to accurately segment and label 3D architectural components. Traditionally, this segmentation has been conducted manually or semi-automatically, involving labor-intensive workflows and substantial reliance on expert human intervention (Alexander & Ben, 2015). However, manual labeling often results in inconsistent outcomes due to inherent subjectivity and human errors. Moreover, as urban datasets grow larger and more diverse, manual approaches are becoming increasingly impractical due to their costs, scalability limitations, and time-consuming nature (Rook et al., 2016).

Machine learning advancements have significantly reshaped this landscape, providing automated methods to handle segmentation tasks more effectively. Graph Neural Networks (GNNs), for example, have emerged as highly promising techniques, leveraging geometric and spatial relationships inherent in 3D mesh data (Wu et al., 2020; Qi et al., 2017). BuildingGNN, a specialized GNN framework specifically tailored for 3D building models, has demonstrated notable accuracy improvements over traditional techniques (Selvaraju et al., 2021; Rashidan et al., 2023). Despite their promising outcomes, GNN-based approaches continue to face limitations. Specifically, these models can struggle to segment semantically ambiguous structures effectively, such as doors and windows, due to their complexity, variety, geometric similarity, and overlapping or unclear spatial boundaries (Kundu et al., 2020).

At the same time, significant strides have been made in computer vision techniques – particularly with the introduction of Vision Transformers (ViTs) which have rapidly gained popularity due to their performance in image classification, detection, and semantic recognition tasks (Hanocka et al., 2019; Dosovitskiy et al., 2021). Trained on large-scale datasets comprising millions of images, ViTs leverage powerful global context capabilities enabled by attention mechanisms, which allow them to accurately recognize intricate visual patterns, subtle textures, and complex structures (Radford et al., 2021; Khan et al., 2022). Although ViTs have primarily excelled within purely 2D domains, their potential applicability to improving 3D semantic segmentation remains a compelling yet largely unexplored of research.

Therefore, this study introduces a hybrid approach that combines Graph Neural Networks (GNNs) and Vision Transformers (ViTs) to address challenges in semantic segmentation of 3D building models. Integrating the geometric reasoning strengths of GNNs with the visual representation capabilities of ViTs, thus, the proposed method aims to improve segmentation robustness and precision. The process begins with initial segmentation using GNNs, followed by the generation of multi-view 2D projections from the segmented 3D meshes. These projections are then refined using pre-trained ViT models, and the enhanced semantic labels are mapped back onto the original 3D geometry through a fusion process. This method is designed to overcome limitations commonly observed in purely geometric or purely visual segmentation approaches.

The structure of the paper is as follows: Section 2 reviews related work, summarizing existing methods and their respective advantages and constraints. Section 3 outlines the proposed method GNN+ViT. Section 4 presents the experimental evaluation, highlighting improvements in segmentation performance. Section 5 concludes the study and outlines potential directions for future research.

2. Related Work

2.1 Semantic Segmentation of 3D Building Models

Semantic segmentation is a crucial process in GIS and urban modeling, assigning meaningful labels to discrete building components, thereby improving data interoperability, visualization, and analytical capabilities (Biljecki & Ohori, 2015; Rook et al., 2016). Traditional manual and semi-automatic methods have long dominated this domain, primarily relying on user-assisted labeling procedures (Demir et al., 2015; Alexander & Ben, 2015). However, the labor-intensive nature and lack of scalability of these approaches are increasingly recognized as significant bottlenecks, especially with growing urban datasets and complex architectural styles (Kalogerakis et al., 2010).

Automated methodologies leveraging computer vision and machine learning techniques have rapidly gained momentum to overcome these challenges. Neural network-based approaches such as PointNet++ (Qi et al., 2017) and MeshCNN (Hanocka et al., 2019) have significantly improved semantic labeling accuracy by effectively capturing complex spatial and geometric relationships inherent in 3D models. However, challenges remain, particularly in accurately segmenting complex architectural features, dealing with large-scale datasets, and managing varying data quality (Hu et al., 2021).

2.2 Graph Neural Networks (GNNs) in 3D Segmentation

GNN-based approaches have specifically shown promising results in addressing segmentation challenges for 3D architectural models (Wu et al., 2020; Selvaraju et al., 2021). GNNs leverage graph structures representing 3D models, incorporating both local and global spatial relationships through neural message-passing mechanisms (Zhou et al., 2020). Methods such as BuildingGNN have demonstrated high accuracy, particularly in segmenting common building components like roofs and walls (Selvaraju et al., 2021). However, limitations persist, particularly in segmenting ambiguous and geometrically similar components such as doors and windows as experimented by Rashidan et al., (2023).

2.3 Vision Transformers and Image-based Segmentation

Vision Transformers (ViTs) have rapidly transformed computer vision tasks, demonstrating superior accuracy in image recognition, detection, and semantic segmentation tasks (Dosovitskiy et al., 2021). Unlike traditional CNNs, ViTs utilize attention mechanisms, allowing them to model long-range contextual relationships effectively, critical in accurate visual recognition (Radford et al., 2021; Khan et al., 2022). The success of ViTs in identifying subtle visual features and handling complex visual datasets strongly suggests their potential utility in refining geometric segmentation results visually.

2.4 Multi-View Projection Techniques and Semantic Fusion

Multi-view projection techniques have successfully enhanced semantic understanding by representing 3D structures across multiple 2D views, providing robust coverage against occlusion, ambiguity, and viewpoint variation (Kundu et al., 2020). View-fusion algorithms combine multiple semantic predictions across views, significantly improving segmentation reliability (Zhang et al., 2019). Despite these successes, limited research has explicitly combined these multi-view methods with GNN-based segmentation results to capitalize on the strengths of both visual and geometric segmentation methods.

Recognizing the complementary strengths and limits of prior methods, this study introduces an integrated framework that combines a GNN with a ViT. The design exploits the GNN's spatial and structural sensitivity and the ViT's capacity for global visual context, yielding improved segmentation accuracy - especially for challenging semantic classes. The approach offers a practical refinement over existing techniques, addressing common sources of error and opening avenues for detailed urban analysis and future research.

3. Methodology

This research presents an integrated methodology aimed at improving semantic segmentation in 3D building models. The proposed multi-stage framework combines the geometric reasoning capabilities of Graph Neural Networks (GNNs) with the contextual visual recognition strengths of Vision Transformers (ViTs). The workflow comprises three main stages: (1) initial segmentation performed using a GNN model, (2) generation of multi-view projections and semantic refinement through ViT-based visual inference, and (3) projection of refined semantic labels back onto the 3D geometry using a multi-view fusion algorithm. Each stage is described in detail below, including relevant mathematical formulations, underlying process rationale, and implementation considerations.

3.1 GNN-based Initial Segmentation

Initially, the raw 3D building models, represented in mesh format, are semantically segmented using a Graph Neural Network framework. Specifically, the BuildingGNN approach introduced by Selvaraju et al. (2021) is employed due to its proven efficacy in handling complex architectural structures.

3.1.1 Data Preparation

An essential aspect of developing the semantic segmentation model involves acquiring labelled dataset. The dataset chosen for the model training comes from the BuildingNet dataset, and accessible at buildingnet.org. This dataset functions as a comprehensive repository of 3D building models, each uniformly labelled with exterior annotations for various architectural components. BuildingNet exhibits diversity, encompassing a range of architectural styles, sizes, and complexities.

3.1.2 Semantic Label Prediction

The semantic labeling process using BuildingGNN involves a structured, three-step pipeline aimed at capturing both geometric and relational features from 3D building mesh data as shown in the Figure 1. The process begins with node initialization, where each subgroup within the mesh is treated as a distinct node. These nodes are assigned initial representations derived from the subgroup's intrinsic geometric attributes such as orientation, surface area, and centroid position. This representation serves as the foundational feature vector that is refined throughout subsequent processing stages.

BuildingGNN

Input: 3D building mesh with subgroup
Output: Labels per subgroup
Step 1: Node Initialization
- Create nodes representing each subgroup in the mesh.
- Initialize node representations based on subgroup attributes.

Step 2: Establish edges to capture relationships
- Proximity (based on spatial distance).
- Supportive relations (upright axis correlation).
- Similarity (detecting symmetric arrangements).
- Containment (subgroup enclosure).

Step 3: GNN Processing
- Update node and edge representations.
- Employ neural message passing and MLPs.
- Decode node to assign labels through softmax.

End

Figure 1. Overview of the main stages involved in labeling a 3D building mesh using BuildingGNN.

Following initialization, graph construction is performed by establishing edges that encode meaningful spatial and structural relationships between nodes. Four types of relationships are considered critical for semantic interpretation: (1) Proximity, which captures closeness between subgroups based on Euclidean distance; (2) Supportive relations, which assess vertical alignment and upright axis correlation – useful for detecting vertical structures like walls and doors; (3) Similarity, which measures visual or geometric resemblance between components, such as repeated window patterns; and (4) Containment, which checks whether one subgroup encloses another – an important clue for detecting nested or hierarchical components.

In the final stage, GNN processing is executed through iterative message passing. Both node and edge representations are updated across multiple layers, with information propagated between connected nodes. Each node aggregates feature information from its neighbours, allowing the model to capture complex spatial dependencies. This is followed by a multi-layer perceptron (MLP) and softmax classifier, which decode the enriched node representations into semantic labels such as wall, roof, window, or door. Through this combination of geometric initialization and relational learning, BuildingGNN provides a foundation for semantic segmentation in 3D building modeling.

3.2 Multi-view Rendering and Semantic Refinement Using Vision Transformer

The output from GNN segmentation serves as an input to the second stage, where the segmented meshes are refined using Vision Transformers.

3.2.1 Multi-view 2D Rendering

To leverage visual semantic recognition, each 3D mesh is projected into multiple 2D images, including orthographic (top, front, side) and perspective views. The rendering process is

- Orthographic projection transforms 3D coordinates (x, y, z) to 2D coordinates (x', y'):

  $(x', y') = (x, y),\ z' = 0$

- Perspective projection is computed as:

  $x' = x / (f \cdot z),\ y' = y / (f \cdot z)$

  where f is the focal length determining perspective intensity.

Multiple projections from varying viewpoints ensure larger semantic coverage, reducing the risk of missing semantic information due to occlusions or complex geometries.

3.2.2 Semantic Inference via Vision Transformer

Rendered images are input to a ViT model (e.g., CLIP by OpenAI), chosen due to its extensive knowledge of visual semantics, obtained from large-scale image datasets. ViTs rely on self-attention mechanisms to understand complex visual structures.

3.3 Multi-view Semantic Fusion and Mapping to 3D Geometry

In the last stage, semantic labels predicted from multiple 2D views are aggregated and mapped onto the original 3D mesh geometry. Given semantic predictions from multiple views, a consensus-based fusion approach is implemented. For each mesh face, semantic labels from all views are aggregated, and the final semantic label is computed using weighted voting.

Final labels are mapped onto the corresponding 3D faces of the original mesh model. We use ray-casting or inverse projection algorithms to determine correspondence between 2D pixel predictions and 3D mesh faces. This mapping ensures precise spatial alignment between 2D semantic predictions and 3D geometric elements.
Post-processing techniques are applied to further enhance label accuracy and consistency:

- Spatial label smoothing – labels are spatially smoothed across connected mesh elements to reduce noise and isolated misclassifications.
- Geometric consistency check ensures that labels assigned to geometrically similar faces remain consistent across neighbouring regions.

The integrated method involves substantial computational steps, requiring careful optimization to ensure efficiency. To address this, parallel computation is leveraged throughout key stages, including rendering, inference, and fusion. Efficient GPU implementation, along with the use of optimized data structures such as sparse tensors and spatial indexing, enables the system to handle large-scale urban datasets effectively.

3.4 Evaluation and Validation

The effectiveness of the proposed GNN+ViT method was assessed using quantitative evaluation method. The primary evaluation metric used in this study is the Intersection-over-Union (IoU), a standard metric widely applied in semantic segmentation tasks to measure the accuracy of predicted labels against ground truth labels. Additionally, visual analysis was conducted to assess the quality of segmentation outputs and to identify specific areas of improvement.

Intersection-over-Union (IoU) calculates the overlap between the predicted labels, and the ground truth labels for each semantic class. The IoU for each class is defined as the ratio of the intersection to the union of the predicted and ground truth sets. The formula used for IoU calculation is as follows:

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|}$$

where A represents the set of predicted labels for a specific class, such as Roof, Wall, Window, or Door, and B denotes the set of ground truth labels for the same class. The numerator $|A \cap B|$ refers to the number of correctly predicted labels, representing the intersection between the prediction and ground truth sets. The denominator $|A \cup B|$ refers to the total number of elements in both sets, representing the union of the prediction and ground truth labels. IoU values range from 0 to 1, where a value of 1 indicates a perfect overlap, and a value of 0 indicates no overlap.

In this study, IoU values were calculated separately for each semantic class to assess the segmentation performance of both the GNN-only model and the GNN+ViT model. The calculation process involved two distinct stages. First, the initial segmentation was performed using the GNN model, where semantic labels were assigned based solely on geometric features extracted from the 3D mesh. For each class, the number of correctly predicted labels (intersection) and the total number of labels (union) were recorded, and the IoU values were calculated based on the above formula.

In the second stage, the ViT model was employed to refine the segmentation output generated by the GNN. The GNN-labeled output was rendered into multiple 2D views, including front and side views. The ViT model processed each view to refine the semantic labels by leveraging visual information. The multiple view predictions were then combined using a consensus-based voting mechanism to determine the final label for each mesh face. Following this refinement process, IoU values were recalculated for each class, considering the updated labels generated by the ViT.

4. Results and Discussion

To evaluate the effectiveness of the proposed method Graph Neural Network (GNN) and Vision Transformer (ViT) approach for semantic segmentation of 3D building models, a comprehensive set of experiments was conducted using a diverse dataset comprising residential buildings sourced from buildingnet.org. The dataset includes variations in architectural styles, structural complexity, and component detailing, providing a robust basis for performance assessment.

The evaluation employs metrics such as Intersection-over-Union (IoU), focusing on four semantic classes: Roof, Wall, Window, and Door. Visualizations are included to support the analysis of improvements in segment consistency and semantic fidelity.

Figure 2 illustrates examples of segmentation outputs from both the GNN-only pipeline (left) and the proposed GNN + ViT model (right). Noticeable improvements in the roof and dormer regions, where the ViT-enhanced model produced more coherent roof plane boundaries and reduced class fragmentation. These refinements suggest that the Vision Transformer's global context awareness allows the model to capture structural patterns beyond local geometric relationships, which are often insufficient when using GNNs alone.

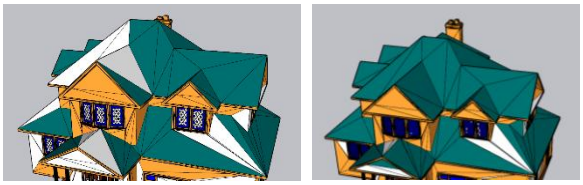

Figure 2. Segmentation comparison between the GNN (left) and GNN + ViT (right) models, showing smoother roof planes and improved consistency.

Similarly, Figure 3 presents close-up views of the wall and window regions, comparing the GNN and GNN + ViT output. The proposed method demonstrates improved delineation between window frames and wall surfaces, minimizing misclassification and edge noise. This enhancement indicates that the multi-view fusion strategy employed in the ViT component effectively leverages visual redundancy, facilitating better feature alignment and error correction.

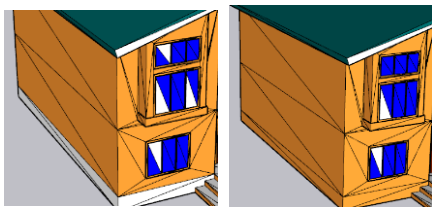

Figure 3. Close-up of wall and window segmentation, where the proposed model provides fewer misclassifications.

Quantitatively, the mean IoU increased from 69.85% in the GNN-only configuration to 78.05% when incorporating the Vision Transformer, representing a performance improvement of approximately 8%. These results confirm that the combination of topological representation learning (through GNNs) and visual context modeling (via ViTs) enhances both spatial reasoning and semantic precision in 3D model interpretation.

5. Conclusion

The results show that combining GNN and ViT improves the semantic segmentation of 3D building models compared to using only GNNs. The combined approach performs better, especially when labelling challenging components like doors and windows, which are typically harder to segment accurately due to their smaller size and visual similarities. These improvements highlight the benefits of using both geometric structure and visual information together.

Although this combined method has advantages, there are also some limitations to consider. One significant issue is the high computational resources needed, especially because more generated images mean greater demands on computing power. This limitation can make the method harder to scale up for large urban datasets or for tasks that require real-time segmentation. Another limitation is related to the range of building designs used in this study. Since the current tests only involved limited types of building structures, the method's ability to handle a wider variety of building styles remains uncertain.

To further advance this research, future studies could focus on addressing these limitations. Reducing computational costs by developing lighter models, simplifying the multi-view approach,

or optimizing the way images are processed could make the method more practical and easier to use. Testing the model on a broader set of building designs (e.g. tropical Asian buildings), including more complex or unusual architectural styles, would help improve its robustness.

In conclusion, the combination of GNN and ViT shows potential in improving the accuracy of building labelling. Nonetheless, further work is needed to enhance computational efficiency and evaluate the method across broader test scenarios to support its practical use in real-world GIS applications.

References

Alexander, T., & Taskar, B. (2015). 3D all the way: Semantic segmentation of urban scenes from start to end in 3D. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1–8).

Biljecki, F., & Ohori, K. A. (2015). Semantic 3D city models in urban planning and analysis. International Journal of Geographical Information Science, 29(3), 408–427.

Demir, I., Aliaga, D. G., & Benes, B. (2015). Coupled segmentation and similarity detection for architectural models. ACM Transactions on Graphics (TOG), 34(4), 1–11.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929

Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., & Cohen-Or, D. (2019). MeshCNN: A network with an edge. ACM Transactions on Graphics (TOG), 38(4), 1–12.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigonis, I., Markham, A., & Trigoni, N. (2021). Learning geometric features for 3D mesh segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 43(12), 4370–4385.

Kalogerakis, E., Hertzmann, A., & Singh, K. (2010). Learning 3D mesh segmentation and labeling. In ACM SIGGRAPH 2010 Papers (pp. 1–12).

Khan, S. A., Shi, Y., Shahzad, M., & Zhu, X. X. (2020). FGCN: Deep feature-based graph convolutional network for semantic segmentation of urban 3D point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 198–199).

Kundu, A., Yin, X., Fathi, A., Ross, D., Brewington, B., Funkhouser, T., & Pantofaru, C. (2020). Virtual multi-view fusion for 3D semantic segmentation. In European Conference on Computer Vision (ECCV) (Vol. 12369, pp. 518–535). Springer International Publishing.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems (NeurIPS), 30, 5099–5108.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML). https://arxiv.org/abs/2103.00020

Rashidan, H., Musliman, I. A., & Rahman, A. A. (2023). Semantic segmentation of building models with deep learning in CityGML. In Proceedings of the 3D GeoInfo Conference.

Rook, M., Biljecki, F., & Diakité, A. A. (2016). Towards automatic semantic labelling of 3D city models. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-4/W1, 23–30. https://doi.org/10.5194/isprs-annals-IV-4-W1-23-2016

Selvaraju, P., Nabail, M., Loizou, M., Maslioukova, M., Averkiou, M., Andreou, A., Drettakis, G., & Kalogerakis, E. (2021). BuildingNet: Learning to label 3D buildings. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 10397–10407). https://doi.org/10.1109/ICCV48922.2021.01024

Stoter, J. E., Arroyo Ohori, K. A., Dukai, B., Labetski, A., Kavisha, K., Vitalis, S., & Ledoux, H. (2020). State of the art in 3D city modelling: Six challenges facing 3D data as a platform. GIM International: The Worldwide Magazine for Geomatics, 34(4), 10–13.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Zhang, Z., Hua, B. S., & Yeung, S. K. (2019). ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1607–1616). https://doi.org/10.1109/ICCV.2019.00170

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., & Song, L. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001