# Deep Learning Based Semantic Segmentation for BIM Model Generation from RGB-D Sensors

Ishraq Rached[1], Rafika Hajji[2], Tania Landes[3], Rashid Haffadi[4]

[1] College of Geomatic Sciences and Surveying Engineering, Institute of Agronomy and Veterinary Medicine, Rabat 6202, Morocco - Ishraq.rach@gmail.com
[2] College of Geomatic Sciences and Surveying Engineering, Institute of Agronomy and Veterinary Medicine, Rabat 6202, Morocco - r.hajji@iav.ac.ma
[3] ICube Laboratory UMR 7357, Photogrammetry and Geomatics Group, National Institute of Applied Sciences (INSA Strasbourg), 24, Boulevard de la Victoire, 67084 Strasbourg, France - tania.landes@insa-strasbourg.fr
[4] GEOPTIMA, B4, Med El Amraoui Street, Corner of Sebou Street, Office 4, Kénitra, Morocco - rashidhaffadi@gmail.com

**Keywords:** RGB-D Camera, Semantic Segmentation, Deep Learning, As-built BIM.

**Abstract**

RGB-D sensors offer a low-cost and promising solution to streamline the generation of BIM models. This paper introduces a framework designed to automate the creation of detailed and semantically rich BIM models from RGB-D data in indoor environments. The framework leverages advanced computer vision and deep learning techniques to overcome the challenges associated with traditional, labour-intensive BIM modeling methods. The results show that the proposed method is robust and accurate, compared to the high-quality statistic laser scanning TLS. Indeed, 58% of the distances measured between the calculated and the reference point cloud produced by TLS were under 5 cm, and 82% of distances were smaller than 7 cm. Furthermore, the framework achieves 100% accuracy in element extraction. Beyond its accuracy, the proposed framework significantly enhances efficiency in both data acquisition and processing. In contrast to the time-consuming process associated with TLS, our approach remarkably reduces the data collection and processing time by factor of height .This highlights the framework's substantial improvements in accuracy and efficiency throughout the BIM generation workflows, making it a streamlined and time-effective solution.

## 1. Introduction

BIM has emerged as a cornerstone in architecture and construction ((Cheng et al., 2020)), providing a digital representation of physical structures and their associated characteristics. However, the creation and updating of BIM models are essentially manual, time-consuming, error-prone and expensive ((Volk et al., 2014)), especially for existing buildings without digital models. This challenges the widespread adoption of BIM, consequently hindering the realization of its full potential and expected benefits.

One popular approach for BIM modeling is to use TLS (Terrestrial Laser Scanner) to capture detailed 3D point clouds of buildings or structures. However, this method can be time-consuming and costly, especially when dealing with large or complex structures ((Chen et al., 2018); (Tang et al., 2019)). In addition, the accuracy of the resulting BIM models can be affected by many factors such as scan resolution, noise levels, and registration errors. In contrast, Simultaneous Localization and Mapping (SLAM) systems offer a more cost-effective alternative to TLS systems and allow for a comprehensive scanning of the scene with centimeter-level precision ((Li et al., 2020)). However, the cost of certain SLAM systems can reach several thousand dollars, making them unaffordable for some users. Additionally, most of them require an external battery, which is cumbersome and not user-friendly for data collection. Researchers further explore sensor fusion, like LiDAR (Light Detection and Ranging) and photogrammetry combinations, to improve the accuracy and efficiency of BIM modeling. For example, a study by ((Zhang et al., 2022)) combined LiDAR and photogrammetry data to create detailed BIM models of buildings. The authors demonstrated that this hybrid approach could improve the accuracy and completeness of BIM models compared to using a single sensor technology. A more recent development in BIM modeling is the use of panoramic images taken from ground-based or aerial platforms ((Lu and Lee, 2017)). Within this area, researchers have proposed various methods to extract 3D information from 2D images using computer vision techniques like SfM (Structure from Motion) ((Konolige and Agrawal, 2008); (Westoby et al., 2012)) and ((Ortiz et al., 2018)). However, to extract 3D information from two-dimensional (2D) images, an extensive post-processing is needed, including image matching and pose estimation, which are time-consuming, and especially suffer from dark environments, poorly textured areas, and motion blurs ((Lee et al., 2023)).

A recent advancement in BIM modeling is the use of affordable low-cost RGB-D sensors that provide synchronized color and depth information. Various researches have shown the promise of RGB-D sensors in as-built BIM modeling ((Wang et al., 2012); (Henry et al., 2012)). However, these approaches have exhibited limitations in terms of scalability, robustness, and handling dynamic scenes. A more recent research by ((Henry et al., 2012)) proposed an automatic framework to generate as-built BIM model from RGB-D sensor, however this approach remains applicable to regular and small-scale scenes.

Expanding on recent developments in BIM modeling, a further evolution involves the integration of Deep Learning (DL) methods to enhance and automate the generation of BIM models, though it remains in its early stages ((Zabin et al., 2022)). The

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-271-2024  

utilization of deep neural networks and Convolutional Neural Networks (CNNs) has shown promise in various aspects of BIM, such as semantic segmentation, object recognition, and scene reconstruction.

The performance of semantic segmentation algorithms has notably advanced due to the adoption of deep neural networks and extensive RGB-D datasets. In enhancing segmentation accuracy, these models utilize depth information from scene depth sensors alongside the conventional RGB image, as highlighted by ((Barchid et al., 2021)). While traditional CNNs have demonstrated success in RGB-D semantic segmentation ((Long et al., 2015); (Eftekhar et al., 2021); (Song et al., 2015); (Gupta et al., 2014); (Su and Wang, 2016); (Wang et al., 2016); (Kamran and Sabbir, 2018) and (Chen et al., 2017)), their computational needs can be daunting for resource-constrained platforms like mobile devices.

Recent advancements in As-Built BIM generation face challenges related to scene dimension and complexity when using RGB-D sensors. In this area, the computational and time demands associated with DL techniques like FCN and CNN, underscore the need for alternative approaches. Exploring lightweight architectures, transfer learning, and pre-trained models can offer efficient alternatives that balance accuracy while minimizing computational overhead, when dealing with complex indoor scenes.

In this research, an automatic system for indoor As-Built BIM generation has been developed. This system autonomously carries out four distinct steps: acquisition and preprocessing of RGB-D data, semantic segmentation, 3D reconstruction, and BIM generation. The proposed framework is designed to not only overcome the limitations of traditional methods but also address the challenges posed by irregular scenes. By leveraging innovative techniques such as lightweight deep learning architectures and transfer learning, computational efficiency has been optimized and accuracy has been enhanced, making this system a robust solution for indoor As-Built BIM generation.

The experimental results underscore the efficacy of this approach. Utilizing an RGB-D camera, the system demonstrates commendable accuracy in handling noisy data. This accuracy, when combined with a significantly reduced processing time compared to TLS (16 minutes as opposed to 134 minutes for TLS), emphasizes the practicality and efficiency of this method. This makes it a viable option for real-world applications where both accuracy and processing speed are of paramount importance. However, it is acknowledged that this approach requires further enhancements, particularly in handling highly complex scenes with numerous occlusions.

The paper is organized as follows: Section 2 presents related work in RGB-D semantic segmentation with deep learning and 3D reconstruction. Section 3 provides a comprehensive description of the methodology. In Section 4, we present experimental findings, and Section 5 concludes with insightful discussions and outlines future development prospects.

## 2. Related Work

This section discusses research related to automating BIM generation using RGB-D data. Our proposed framework incorporates two key steps: semantic segmentation and 3D reconstruction. Therefore, we will explore relevant work in both areas.

### 2.1 Semantic Segmentation with RGB-D Data

Semantic Segmentation involves analyzing each pixel in the RGB image and assigning it a semantic label, such as ”wall”, ”floor,” ”window,” etc. The performance of semantic segmentation algorithms has notably advanced due to the adoption of DL networks and extensive RGB-D datasets (Barchid et al., 2021). While traditional CNNs have succeeded in RGB-D semantic segmentation, their computational demands are challenging for devices like mobile phones. In contrast, FCNs are considered as standard models in deep learning, allowing for flexible input and output sizes.

However, FCN's semantic segmentation result isn't detailed enough, even when combining information from its high and low layers (Li et al., 2020). To address these challenges, research has focused on lightweight and efficient architectures. (Chollet, 2017) proposed depth-wise separable convolutions for building lightweight encoders. (Yin et al., 2023) further explored this concept by introducing D-Former, a pre-training framework utilizing a Transformer-based encoder with depth-wise convolutions specifically designed for RGB-D data. This approach achieves state-of-the-art results while maintaining computational efficiency.

### 2.2 3D Reconstruction from RGB-D Data

3D reconstruction is the process of creating a three-dimensional representation or model of an object or scene; it is a crucial step in as-built BIM creation using low cost RGB-D sensors. This process typically involves several steps, including feature detection, feature matching, camera pose estimation, global and refined registration, and point cloud colorization.

The first step in creating a point cloud is to detect features in the RGB-D images. Features are distinctive points or areas in an image that can be reliably and robustly detected. Many algorithms have been used for this purpose, such as SIFT (Lowe, 2004), SURF (Bay et al., 2006), and ORB (Rublee et al., 2011). Once features have been detected in the images, the next step is to match these features across different images. This involves finding pairs of features that correspond to the same point in the scene. Feature matching can be done using various methods, such as brute-force matching, FLANN-based matching (Muja and Lowe, 2009).

After features have been matched, the next step is to estimate the pose of the camera for each image. This involves determining the position and orientation of the camera relative to the scene. There are several methods for pose estimation, such as (Perspective-n-Point) PnP (Pan and Wang, 2021) and its variants.

These steps result in a set of 3D points, known as a point cloud, which represents the 3D structure of the scene. The process of point cloud optimization begins with global registration, which provides a rough alignment of the RGB-D images. This is typically achieved using feature-based methods that identify and match distinctive points in the images. This involves the use of RANSAC (Fischler and Bolles, 1981), a robust estimation technique that can handle a significant proportion of outliers. RANSAC iteratively estimates the parameters of a mathematical model from a set of observed data points in a way that maximizes the number of inliers (Zhou et al., 2018). The ex-

pected number of iterations in RANSAC can be expressed as:

$$E(k) = \sum_{i=1}^{\infty} i \times p(i) = \sum_{i=1}^{\infty} i \times a^{i-1} \times b = b \times \sum_{i=1}^{\infty} i \times a^{i-1}$$

Where $a$ is the probability that a point is an inlier and $b$ is the probability that at least one of the randomly selected points is an outlier.

However, global registration alone is often insufficient for accurate 3D reconstruction due to noise and other inaccuracies. To refine the alignment, a secondary process known as refine registration is employed. The ICP algorithm is used to minimize the difference between two clouds of points (Chen and Medioni, 1992). ICP iteratively revises the transformation (rotation and translation) needed to align the points in one cloud with corresponding points in the other cloud. This results in a fine-grained alignment of the RGB-D images.

## 3. Methodology

This section describes the methodology followed to generate an As-Built BIM model from indoor scenes captured with an RGB-D sensor. The workflow proposed as shown in Figure 1 comprises four key stages: in the first stage, we capture indoor scenes with Kinect Azure RGB-D sensor then we proceed to depth image processing. The second stage consists of semantic segmentation which involves classification of the scene into distinct objects or surfaces. This is crucial for identifying walls, floors, ceilings, furniture, and other elements within the scene. To do this, a DL model is employed to perform semantic segmentation on the preprocessed data. In the third stage, a point cloud is generated by fusing the depth data with the RGB information, providing a rich representation of the scene's geometry. This point cloud is fused with predictions resulting from the second stage to give a semantically rich point cloud which acts as the foundation for the final stage, where BIM model generation takes place. Finally, the predictions are compared to the ground truth in order to assess the results

### 3.1 Data Acquisition and Preprocessing

The first step involves capturing indoor scenes using a RGB-D camera. The device used in this paper is Kinect azure, an RGB-D sensor offering high-resolution (1920x1080) synchron-ized RGB and depth images at 30 fps. Its portability and ability to be mounted on tripods or robots makes it ideal for capturing diverse indoor environments.

This approach is particularly valuable in the context of RGBD data, as it allows for the reconstruction of depth information in areas where it may be incomplete or unavailable. By inferring depth values based on surrounding color cues, the algorithm contributes to creating a more comprehensive and detailed dataset. Figure 2 shows the difference between the raw depth and the processed depth map. As a result, the filled depth image seems to be more comprehensive and refined, providing a more seamless and accurate representation of the indoor environment. This enhancement is valuable for subsequent stages of the BIM generation framework.

### 3.2 2D Semantic Segmentation

In order to optimize the performance of D-Former, a two-step training process was employed. Initially, the model underwent

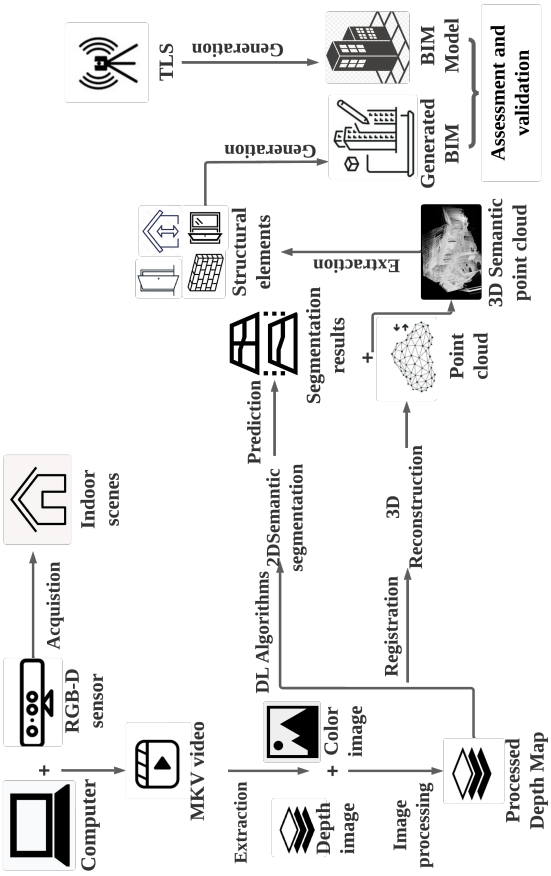


Figure 1. The Methodological workflow.

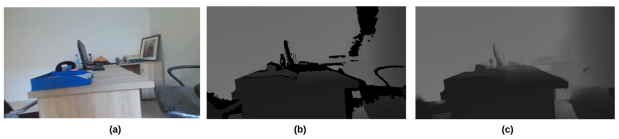


Figure 2. Example of (a) RGB image, (b) Raw depth, (c) filled depth

training on the NYU depth v2 dataset (Silberman and Fergus, 2011), a widely used dataset for RGB-D semantic segmentation tasks. This initial training phase allowed D-Former to learn fundamental features and patterns from a diverse set of depth information.

The second stage consists of semantic segmentation wich is performed using a D-Former model (Yin et al., 2023). In order to optimize the performance of D-Former, a two-step training process was employed. Initially, the model underwent training on the NYU depth v2 dataset (Silberman and Fergus, 2011), a widely used dataset for RGB-D semantic segmentation tasks. This initial training phase allowed DFormer to learn fundamental features and patterns from a diverse set of depth information.

Following the pretraining on NYU depth v2, the model was fine-tuned using our preprocessed dataset. This dataset consists of 167 RGB-D images captured using a Kinect sensor from a variety of diverse architectural styles, encompassing both modern and historical structures from different cities. We are currently working on making this dataset publicly accessible for

broader use. The depth maps were preprocessed using colorization algorithm (Levin et al., 2004). The images are annotated with labels using LabelMe (Russell et al., 2008) and split into 70% for training and 30% for testing.

Results are depicted in Figure 3 and show an increase in Mean Intersection over Union (mIoU) and a decrease in training loss after the training of D-Former on our preprocessed dataset. These metrics are representative of the model's segmentation performance and learning convergence. Both indicators are described below.

- **Mean Intersection over Union (mIoU)**: The rise in mIoU indicates that the model has become more adept at accurately segmenting and assigning semantic labels to pixels in the RGB images. mIoU is a metric that measures the intersection between predicted and ground truth segmentation masks, normalized by the union of these masks. A higher mIoU reflects a better alignment between the predicted and actual segmentation, signifying enhanced overall semantic segmentation accuracy.
- **Training Loss**: The reduction in training loss signifies that the model has successfully minimized the discrepancy between its predicted outputs and the ground truth labels during the training process. As the model iteratively refines its parameters, the loss decreases, indicating improved convergence and alignment with the training data.

This enhanced performance is crucial for the subsequent steps in the BIM framework, where precise semantic understanding of structural elements is essential for generating detailed and semantically rich 3D models. Examples of segmented images illustrating these improvements are presented in Figure 4.

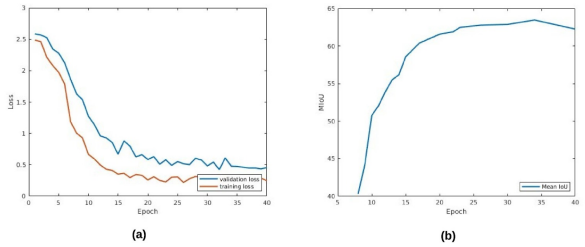


Figure 3. Training Loss during training (a) and MIoU during training Progress (b).


Figure 4. Example of segmentation results.

### 3.3 3D Reconstruction

In 3D reconstruction stage, we employed the Open3D library. Initially, feature detection was conducted to identify distinctive points in the captured images. Subsequently, feature matching was performed to establish correspondences between points in different images. Through camera pose estimation, the relative positions and orientations of the cameras capturing the scene were determined. The global registration between the different fragments is performed by the RANSAC algorithm, followed by refined registration to further improve the alignment accuracy. Following the alignment stage, the point cloud data was obtained through a process of triangulation. This involved projecting the features identified in the images onto their corresponding positions in 3D space, using the camera parameters estimated during the camera pose estimation step. The triangulated points were then merged into a single point cloud representation of the scene.

Finally, the process continues with the superposition of the results from segmentation and the point cloud to achieve a colorized point cloud representation of the scene. Figure 5 bellow show an example of 3D point cloud and colorized point cloud.

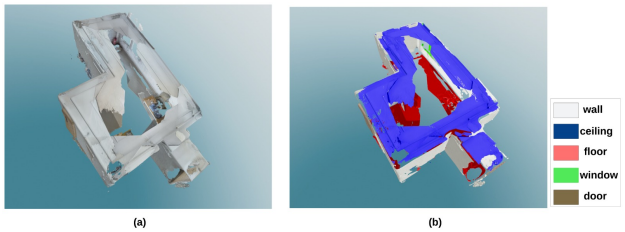


Figure 5. Example of (a) 3D Point Cloud, and (b) 3D semantic point cloud

### 3.4 BIM Generation

The final stage is the generation of the BIM model from the semantic 3D point cloud, guided by semantic labels. This process involves transforming the enriched point cloud, where each point carries semantic information, into a comprehensive and detailed BIM.

In this step, we dissect the point cloud into its constituent structural elements. However, point cloud data often suffer from noise, incompleteness, and irregularity, which pose challenges for accurate and efficient boundary extraction.

Several methods have been proposed to address this problem, which can be broadly classified into two categories: image-based methods and feature-based methods.

Image-based methods convert the point cloud data into 2D images and apply edge detection algorithms to find the boundary points. For example, (Xi et al., 2016) proposed a method that divides the point cloud data into different patches based on the coplanarity condition, and then converts each patch into a 2D image according to the depth dimension. An improved Laplace image edge detection method is then applied to each image to extract the boundary points. This method is fast and robust, but it may lose some 3D information due to the projection and discretization of the point cloud data.

Feature-based methods use geometric or topological features of the point cloud data to classify the boundary points. For example, (Dey et al., 2021) proposed a method that uses Delaunay triangulation and distance from the mean point of the neighborhood to extract both inner and outer boundary points of the building point cloud. This method can preserve the 3D information and detect both concave and convex boundaries, but it may

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-271-2024 | © Author(s) 2024. CC BY 4.0 License.

be computationally expensive and sensitive to the point density and distribution.

To generate automatic as-built BIMs from low cost RGB-D sensor data (Li et al., 2020) uses an iterative plane detection algorithm to detect the planes from the point cloud, and then computes the normal vector and distance to the fitted planes for each point. The points that have a large normal vector difference or distance difference with their neighboring points are considered as boundary points. This method can generate accurate and efficient BIMs from low-quality point cloud data, but it may not be suitable for complex indoor environments with curved or non-planar walls.

The low-quality point cloud generated by RGB-D sensors can be a challenge for extracting structural elements such as walls. For identifying walls, we use a wall boundary extraction method proposed by (Li et al., 2020). Figure 6 displays the result obtained after extracting the walls. It can be observed that the result is not refined, and the polygon is not closed. To address this issue, we employed the algorithm suggested in the same article, called "wall boundary refinement." To extract windows and doors it is sufficient to provide the center, width, and the height of the element, the width and height are extracted from the projection on the floor and wall planes, then the center is extracted from the same projections. These information are subsequently integrated into Revit, a BIM software facilitating the creation of accurate and detailed BIM models developed by Autodesk.

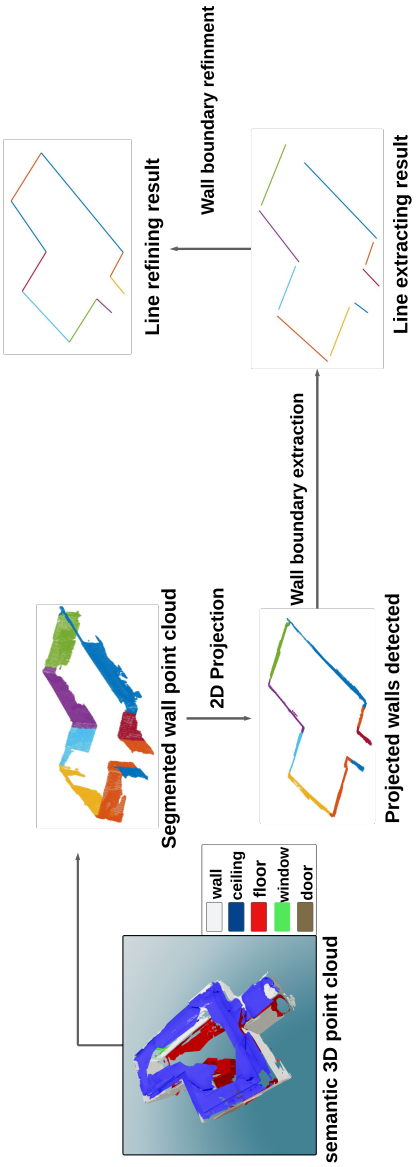


Figure 6. Wall Boundary Extraction Workflow.

## 4. Experimental Tests and Discussion

Two experiments were conducted in distinct scenes: the first involved an irregular room with a small corridor , while the second featured a complex scenario comprising two adjacent rooms, as shown in Figure 7. We captured videos using the Kinect Azure camera, systematically scanning the walls and details by moving in a unidirectional manner and incorporating upward movements to ensure comprehensive coverage of the floor and ceiling.

To assess the performance of our proposed framework, we compared the results with those obtained from the same scenes using a Leica RTC 360 (TLS), which are considered as the ground truth. This evaluation focuses on three key aspects: geometric quality assessment, semantic segmentation and elements extraction accuracy and the efficiency of 3D reconstruction and BIM generation.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-271-2024 | © Author(s) 2024. CC BY 4.0 License.


Figure 7. Captured scenes (a) Irregular room (b) Adjacent rooms

### 4.1 Geometric Quality Assessment

To evaluate the geometric quality of low-cost image-based 3D reconstruction, we compare the results obtained from TLS and RGB-D sensors, which generated two separate point clouds., Table 1 shows details on the dataset gathered from the two experiments using both the TLS and RGB-D Sensor. Following this, a co-registration process was conducted to align and compare the point clouds. The co-registration was performed manually in Cloud Compare.

The point-to-point distances provide a reliable metric for assessing the consistency between these two modeling techniques. The obtained result is illustrated in Figure 8.

The comparative analysis reveals that our method generates a point cloud comparable to that obtained through laser scanning. Specifically, more than half of the distances (57,7%) measured were under 5 cm, and the majority of distances (82%) were smaller than 7 cm as shown in Figure 8. The attained accuracy, with such a high percentage of distances falling within the 5-8 cm range, is deemed satisfactory for various modeling applications, such as documentation, visualization, virtual reality simulations, augmented reality overlays, and spatial analysis.

| Experiment | Sensor | Stations | Frames | Raw Points (M) | Sampled Points (M) |
|---|---|---|---|---|---|
| Irregular room | TLS (RTC360) | 1 | - | 31.15 | 1.57 |
| | Kinect Azure | - | 3528 | 7.12 | 1.10 |
| Adjacent rooms | TLS (RTC360) | 2 | - | 39.97 | 1.12 |
| | Kinect Azure | - | 6600 | 13.20 | 0.80 |

Table 1. Acquisition details from TLS and RGB-D Sensor.

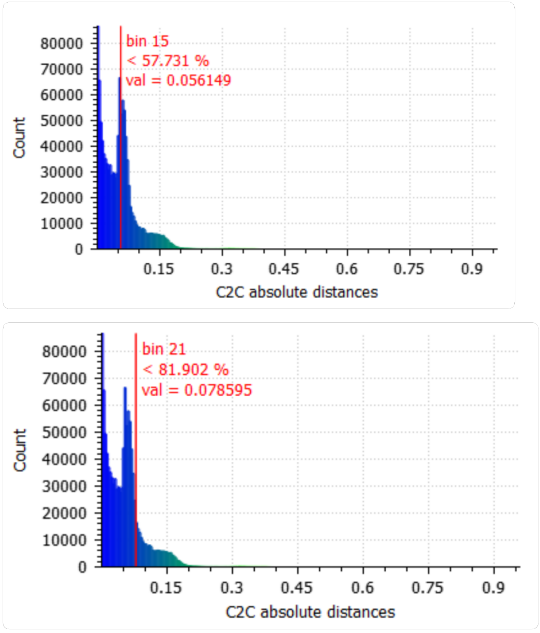


Figure 8. Histogram of distance between reconstructed point cloud and TLS.

In order to evaluate the presence of odometry effect, we overlaid the point clouds obtained from the RGB-D sensor and TLS, as illustrated in Figure 9. The observed errors along the edges of details are likely attributed to the limited capture angles. Unfortunately, capturing objects, especially smaller ones, from all possible angles is impractical using RGB-D sensors.

### 4.2 Semantic Segmentation and Elements Extraction Accuracy

To validate the semantic segmentation and the accuracy of structural element extraction in the proposed approach, a visual assessment has been performed.

In the first experiment, the room has two doors, one window, and ten distinct walls. The second experiment is more complex with a scene that comprises two adjacent rooms, with two windows, three doors, and twelve distinct walls.

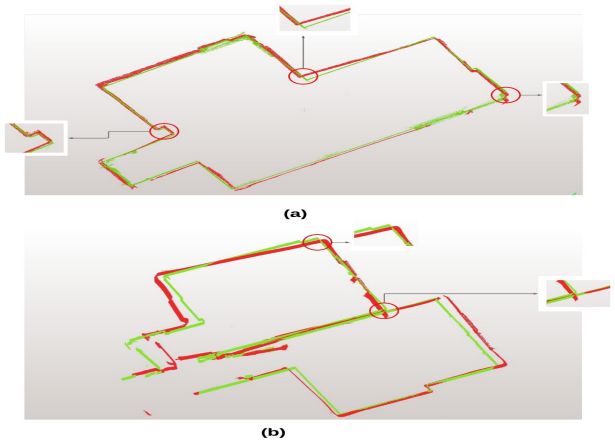


Figure 9. Overlay of point clouds captured by TLS in green and Kinect Azure in red, (a) Experiment 1, (b) Experiment 2.

The findings demonstrate that the proposed method successfully identifies and extracts all existing structural elements in the scenes as shown in table 2. Figures 10 and 11 shows the BIM models generated by our proposed method alongside manually generated ground truth from TLS.

| Experiment | Structural Element | True Number | Extracted Number | Accuracy |
|---|---|---|---|---|
| Irregular room | Wall | 10 | 10 | 100% |
| | Ceiling | 1 | 1 | 100% |
| | Floor | 1 | 1 | 100% |
| | Door | 2 | 2 | 100% |
| | Window | 1 | 1 | 100% |
| Two adjacent room | Wall | 12 | 12 | 100% |
| | Ceiling | 2 | 2 | 100% |
| | Floor | 2 | 2 | 100% |
| | Door | 2 | 2 | 100% |
| | Window | 2 | 2 | 100% |

Table 2. Results for the structural elements extraction compared to the reference.

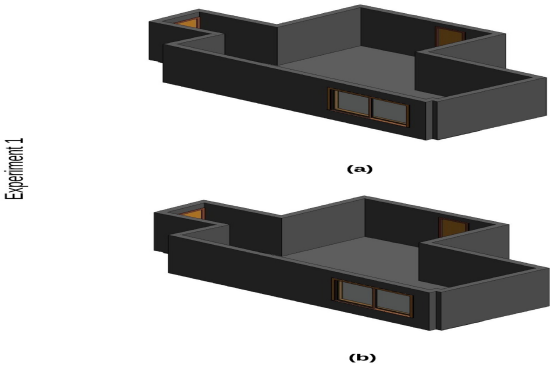


Figure 10. BIM Model Comparison (a) Proposed Method, and (b) Ground Truth.

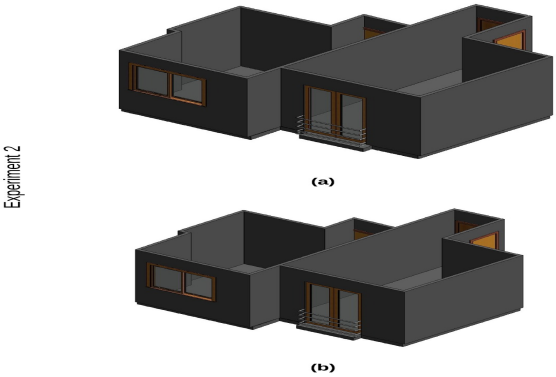


Figure 11. BIM Model Comparison (a) Proposed Method, and (b) Ground Truth.

### 4.3 Efficiency of 3D Reconstruction and BIM Generation

To evaluate the efficiency of our framework, we assess computational resources, processing time, and overall system performance during 3D reconstruction and BIM generation. The objective is to determine the framework's feasibility for real-world applications, considering its computational demands and speed of operation.

As shown in Table 3, the time required for data collection, data processing and BIM model generation is reduced from 134 minutes for two experiments with TLS, to 16 minutes with the RGB-D sensor.

| | TLS | | RGB-D sensor | |
|---|---|---|---|---|
| | Irregular room (s) | Two adjacent room(s) | Irregular room(s) | Two adjacent room(s) |
| Acquisition | ≈ 2400 | 3000 | 178 | 220 |
| Processing | 600 | 750 | 50 | 80 |
| BIM Generation | 600 | 720 | 200 | 250 |
| Total | 134 min | | 16 min | |

Table 3. Time comparison of data collection, processing, and BIM Model generation: TLS vs. RGB-D Sensor.

### 4.4 Discussion

The experiment results indicate that our approach, utilizing an RGB-D camera demonstrates acceptable accuracy for handling noisy data. Over half (57.7%) of the measured distances fall under 5 cm, highlighting its effectiveness for various modeling applications like documentation, visualization, and augmented reality applications. This accuracy, coupled with significantly reduced processing time compared to traditional TLS (134 minutes vs. 16 minutes) emphasize the practicality and efficiency of our method, making it a viable option for real-world applications where both accuracy and processing speed are crucial considerations. However, this approach still requires improvements, particularly in scenes that are highly complex and involve numerous occlusions.

For future endeavors, we strongly recommend:

- Incorporating Sensor Fusion: Combine RGB-D sensors with other systems such as LiDAR or thermal cameras. This integration proves particularly beneficial in overcoming limitations related to the restricted capture angles of RGB-D sensors.
- Investigating Online Learning and Adaptation: explore methods for enabling the framework to continuously learn and adapt to diverse scenes in real-time. This evolution would significantly enhance the versatility and practical applicability of the framework, addressing dynamic environmental conditions effectively.
- Large-Scale Scene Reconstruction: Investigating scalability for handling larger and more complex scenes will push the boundaries of the method's applicability and open doors for broader use cases.

## 5. Conclusion

This paper proposes an automatic framework for cost-effective as-built BIM generation using RGB-D sensors and advanced Deep Learning techniques. The incorporation of D-Former for 2D semantic segmentation, combined with a 3D reconstruction pipeline, enables comprehensive scene understanding and accurate extraction of BIM elements. The results highlight the potential of our approach method for generating detailed, semantically rich models across various applications. While demonstrating effectiveness, challenges may arise in scenes with high clutter, occlusions, and intricate architecture. Nevertheless, ongoing refinements and optimizations are anticipated to make this framework a valuable and flexible tool for BIM generation in different situations.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W5-2024-271-2024 | © Author(s) 2024. CC BY 4.0 License.

## References

Barchid, S., Mennesson, J., Djéraba, C., 2021. Review on indoor rgb-d semantic segmentation with deep convolutional neural networks. *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, 1–4.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. 3951, 404–417.

Chen, C., Yang, B., Song, S., Tian, M., Li, J., Dai, W., Fang, L., 2018. Calibrate multiple consumer RGB-D cameras for low-cost and efficient 3D indoor mapping. *Remote Sensing*, 10(2), 328.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

Chen, Y., Medioni, G., 1992. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3), 145–155.

Cheng, J. C., Chen, K., Chen, W., 2020. State-of-the-art review on mixed reality applications in the AECO industry. *Journal of Construction Engineering and Management*, 146(2), 03119009.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Dey, E. K., Awrangjeb, M., Kurdi, F. T., Stantic, B., 2021. Building boundary extraction from lidar point cloud data. *2021 Digital Image Computing: Techniques and Applications (DICTA)*, 1–6.

Eftekhar, A., Sax, A., Malik, J., Zamir, A., 2021. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10786–10796.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Gupta, S., Girshick, R., Arbeláez, P., Malik, J., 2014. Learning rich features from rgb-d images for object detection and segmentation. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, Springer, 345–360.

Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., 2012. RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments. *The international journal of Robotics Research*, 31(5), 647–663.

Kamran, S. A., Sabbir, A. S., 2018. Efficient yet deep convolutional neural networks for semantic segmentation. *2018 international symposium on advanced intelligent informatics (SAIN)*, IEEE, 123–130.

Konolige, K., Agrawal, M., 2008. FrameSLAM: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5), 1066–1077.

Lee, M. J. L., Wen, W., Au, S. L. M., 2023. Integrating Building Information Modeling and Panoramic Structure-from-Motion for Accurate Camera Pose Estimation.

Levin, A., Lischinski, D., Weiss, Y., 2004. Colorization using optimization. *ACM SIGGRAPH 2004 Papers*, 689–694.

Li, Y., Li, W., Tang, S., Darwish, W., Hu, Y., Chen, W., 2020. Automatic indoor as-built building information models generation by using low-cost RGB-D sensors. *Sensors*, 20(1), 293.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60, 91–110.

Lu, Q., Lee, S., 2017. Image-based technologies for constructing as-is building information models for existing buildings. *Journal of Computing in Civil Engineering*, 31(4), 04017005.

Muja, M., Lowe, D., 2009. Flann-fast library for approximate nearest neighbors user manual. *Computer Science Department, University of British Columbia, Vancouver, BC, Canada*, 5, 6.

Ortiz, L. E., Cabrera, E. V., Gonçalves, L. M., 2018. Depth data error modeling of the ZED 3D vision sensor from stereolabs. *ELCVIA: electronic letters on computer vision and image analysis*, 17(1), 0001–15.

Pan, S., Wang, X., 2021. A survey on perspective-n-point problem. *2021 40th Chinese Control Conference (CCC)*, IEEE, 2396–2401.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. Orb: an efficient alternative to sift or surf. 2564–2571.

Russell, B. C., Torralba, A., Murphy, K. P., Freeman, W. T., 2008. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77, 157–173.

Silberman, N., Fergus, R., 2011. Indoor scene segmentation using a structured light sensor. *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, IEEE, 601–608.

Song, S., Lichtenberg, S. P., Xiao, J., 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Su, W., Wang, Z., 2016. Regularized fully convolutional networks for rgb-d semantic segmentation. *2016 Visual Communications and Image Processing (VCIP)*, IEEE, 1–4.

Tang, S., Zhang, Y., Li, Y., Yuan, Z., Wang, Y., Zhang, X., Li, X., Zhang, Y., Guo, R., Wang, W., 2019. Fast and automatic reconstruction of semantically rich 3D indoor maps from low-quality RGB-D sequences. *Sensors*, 19(3), 533.

Volk, R., Stengel, J., Schultmann, F., 2014. Building Information Modeling (BIM) for existing buildings—Literature review and future needs. *Automation in construction*, 38, 109–127.

Wang, J., Wang, Z., Tao, D., See, S., Wang, G., 2016. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, Springer, 664–679.

Wang, J., Zhang, C., Zhu, W., Zhang, Z., Xiong, Z., Chou, P. A., 2012. 3d scene reconstruction by multiple structured-light based commodity depth cameras. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 5429–5432.

Westoby, M. J., Brasington, J., Glasser, N. F., Hambrey, M. J., Reynolds, J. M., 2012. 'Structure-from-Motion'photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179, 300–314.

Xi, X., Wan, Y., Wang, C., 2016. Building boundaries extraction from points cloud using an image edge detection method. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1270–1273.

Yin, B., Zhang, X., Li, Z., Liu, L., Cheng, M.-M., Hou, Q., 2023. Dformer: Rethinking rgbd representation learning for semantic segmentation. *arXiv preprint arXiv:2309.09668*.

Zabin, A., González, V. A., Zou, Y., Amor, R., 2022. Applications of machine learning to BIM: A systematic literature review. *Advanced Engineering Informatics*, 51, 101474.

Zhang, P., He, H., Wang, Y., Liu, Y., Lin, H., Guo, L., Yang, W., 2022. 3D urban buildings extraction based on airborne lidar and photogrammetric point cloud fusion according to U-Net deep learning model segmentation. *IEEE Access*, 10, 20889–20897.

Zhou, Q.-Y., Park, J., Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*.