

An Artificial Intelligence-Based Indoor 3D Modeling System for Intelligent Spatial Reconstruction and Visualization

Kuo-Lung Chen*

Department of Computer Science and Information Engineering, National Taitung University, Taiwan; e-mail:

longerger14@gmail.com

*Corresponding Author: longerger14@gmail.com

DOI: <https://doi.org/10.30211/JIC.202604.012>

Submitted: May 13, 2026 Accepted: July 04, 2026

ABSTRACT

Indoor 3D modeling has become an important technology for smart building management, digital twin construction, indoor navigation, facility maintenance, and virtual reality applications. However, traditional indoor 3D model construction usually requires professional equipment, manual measurement, and time-consuming modeling procedures. These limitations increase labor costs and reduce the efficiency of indoor spatial digitization. To address these challenges, this study proposes an artificial intelligence-based indoor 3D modeling system that integrates image acquisition, computer vision, depth estimation, semantic segmentation, spatial reconstruction, and interactive visualization. The proposed system uses indoor images or video sequences as input and automatically extracts spatial features, recognizes indoor objects, estimates scene depth, and generates a structured 3D indoor model. By combining artificial intelligence and 3D reconstruction techniques, the system can reduce manual modeling workload and improve the efficiency of indoor environment digitization. The proposed framework can be applied to smart campuses, factories, offices, museums, hospitals, exhibition spaces, and indoor facility management. The results are expected to demonstrate that AI-based indoor 3D modeling offers a practical, scalable solution for intelligent spatial reconstruction and digital twin applications.

Keywords: Artificial intelligence, Indoor 3D modeling, Spatial reconstruction, Computer vision, Depth estimation, Semantic segmentation, Digital twin, Smart building

1. Introduction

With the rapid development of artificial intelligence, computer vision, digital twin technology, building information modeling, and smart building applications, indoor 3D modeling has become an important research topic in intelligent spatial computing. Accurate indoor 3D models can provide rich

geometric and semantic information for indoor navigation, facility management, emergency response, virtual reality, augmented reality, smart campus management, and digital twin applications. Compared with conventional two-dimensional floor plans, indoor 3D models can represent spatial layouts, object locations, structural components, and facility relationships in a more intuitive and interactive manner.

Traditional indoor 3D model construction usually relies on manual measurement, professional modeling software, laser scanning, LiDAR sensors, or RGB-D devices. Although these methods can generate accurate geometric models, they often require professional equipment, expert operation, high labor cost, and long processing time. For large-scale indoor environments, such as campuses, factories, hospitals, museums, office buildings, and public spaces, manual modeling is difficult to scale efficiently. Therefore, developing an automatic and intelligent indoor 3D modeling system has become an important issue for smart building and digital transformation applications.

In recent years, artificial intelligence and deep learning have significantly improved indoor scene understanding. Naseer et al. [1] reviewed indoor scene understanding in 2.5D and 3D environments and indicated that indoor perception involves multiple tasks, including scene classification, object detection, pose estimation, semantic segmentation, 3D reconstruction, saliency detection, physical reasoning, and affordance prediction. Their study shows that indoor scene understanding is not only a visual recognition problem but also a comprehensive spatial reasoning problem. This observation is highly relevant to indoor 3D modeling because a useful indoor model should include both geometric structure and semantic information.

RGB-D semantic segmentation has also become an important research direction for indoor spatial understanding. Barchid et al. [2] reviewed deep convolutional neural networks for RGB-D indoor semantic segmentation and pointed out that depth information provides complementary geometric cues for RGB images. By combining color and depth information, semantic segmentation models can better identify walls, floors, ceilings, doors, windows, furniture, and indoor facilities. These semantic labels are essential for generating structured and meaningful indoor 3D models.

Depth estimation is another key technology for AI-based indoor 3D modeling. Traditional depth acquisition methods rely on active sensors or stereo cameras, but deep learning-based monocular depth estimation can infer dense depth information from a single RGB image. Arampatzakis et al. [3] reviewed monocular depth estimation and indicated that estimating depth from two-dimensional images is a challenging but important computer vision problem. In practical indoor environments, depth estimation can support spatial reconstruction, object localization, and scene geometry inference. However, indoor depth estimation remains difficult due to occlusion, weak texture, reflective surfaces, complex layouts, and scale ambiguity.

To improve indoor scene understanding, several studies have attempted to jointly learn semantic and geometric information. Zhang et al. [4] proposed a joint task-recursive learning framework for RGB-D scene understanding, which jointly conducts depth estimation, semantic segmentation, and related dense prediction tasks. This type of multi-task learning strategy is important for indoor 3D modeling because semantic information and geometric information can mutually enhance each other. For example,

depth information can help distinguish walls and furniture, while semantic segmentation can improve the interpretation of indoor structural components.

In addition to 2D semantic segmentation and depth estimation, 3D semantic instance segmentation has been studied to construct object-level indoor representations. Hou et al. [5] proposed 3D-SIS for 3D semantic instance segmentation of RGB-D scans, which jointly fuses 2D RGB features and 3D geometric information. Their work demonstrates that combining visual and geometric representations can improve object detection and instance segmentation in 3D indoor scenes. This finding suggests that AI-based indoor 3D modeling systems should not only reconstruct surfaces but also attach semantic and object-level information to the generated model.

Indoor 3D modeling is also closely related to building information modeling and digital twin applications. Building information modeling provides a digital representation of building elements with geometric and semantic information, while digital twin technology further connects physical buildings with virtual models, sensor data, and management functions. Recent studies have indicated that BIM and digital twin technologies can support building life-cycle management, facility monitoring, energy management, and interactive building operations [6], [7]. Therefore, AI-based indoor 3D modeling can serve as a foundation for smart building digital twins by generating indoor models more efficiently and connecting visual reconstruction with spatial management.

Although existing studies have made significant progress in indoor scene understanding, semantic segmentation, depth estimation, and 3D reconstruction, several challenges remain. First, many indoor reconstruction methods still depend on expensive sensors or professional modeling procedures. Second, monocular or image-based reconstruction may suffer from scale ambiguity and reconstruction errors in complex indoor environments. Third, many systems focus mainly on geometric reconstruction and lack semantic information required for practical building management. Fourth, existing indoor 3D models are often difficult to update dynamically when indoor layouts, furniture, or facilities change. These limitations reduce the efficiency and usability of indoor 3D modeling in real-world applications.

To address these issues, this study proposes an artificial intelligence-based indoor 3D modeling system for intelligent spatial reconstruction and visualization. The proposed system integrates indoor image acquisition, image preprocessing, AI-based scene understanding, semantic segmentation, depth estimation, 3D reconstruction, semantic model generation, and interactive visualization. The system aims to reduce manual modeling workload and improve the efficiency of indoor spatial digitization.

2. Literature Review

2.1 Indoor 3D Reconstruction and Spatial Modeling

Indoor 3D reconstruction aims to convert physical indoor environments into digital spatial models that can represent geometric layouts, structural components, and object locations. Early indoor reconstruction methods mainly relied on laser scanners, RGB-D cameras, simultaneous localization and mapping, and manual modeling tools. These methods can generate accurate geometric structures, but they often require professional devices, careful scanning procedures, and time-consuming post-

processing. Therefore, efficient and automatic indoor 3D reconstruction has become an important research direction for smart buildings, indoor navigation, facility management, and digital twin applications.

With the development of depth cameras and real-time reconstruction algorithms, RGB-D-based indoor mapping has significantly improved the feasibility of indoor 3D model construction. Newcombe et al. [8] proposed KinectFusion, which demonstrated real-time dense surface mapping and tracking using commodity depth cameras. This work showed that low-cost depth sensors could be used to reconstruct indoor environments interactively. Henry et al. [9] further developed RGB-D mapping for dense 3D modeling of indoor environments by combining visual features and depth information. These studies established an important foundation for indoor reconstruction using RGB-D data.

Large-scale indoor datasets have also accelerated the development of indoor 3D modeling. Armeni et al. [10] introduced a large-scale indoor dataset for 3D semantic parsing of buildings, enabling research on room-level and object-level indoor understanding. Dai et al. [11] proposed ScanNet, a richly annotated dataset of 3D indoor scenes, which provides RGB-D scans, camera poses, surface reconstructions, and semantic annotations. Such datasets are important because they support the training and evaluation of deep learning models for indoor reconstruction, semantic segmentation, and object recognition.

In addition to direct depth-based reconstruction, image-based and learning-based methods have been studied to reduce the dependence on expensive sensors. Monocular depth estimation and multi-view reconstruction methods attempt to recover indoor geometry from RGB images or video sequences. These methods are more flexible and cost-effective, but they still face challenges such as scale ambiguity, weak texture, reflective surfaces, occlusion, and complex indoor layouts. For practical indoor 3D modeling systems, spatial reconstruction must therefore be combined with semantic understanding and model optimization.

Although existing indoor reconstruction methods can generate geometric models, many of them still focus mainly on surfaces, point clouds, or meshes. For smart building applications, geometry alone is not sufficient. A useful indoor 3D model should also include semantic information, such as walls, floors, ceilings, doors, windows, furniture, equipment, room functions, and facility attributes. Therefore, recent research has gradually moved from pure geometric reconstruction toward semantic 3D modeling.

2.2 AI-Based Indoor Scene Understanding and Semantic Reconstruction

Artificial intelligence has significantly improved the ability of systems to understand indoor environments. Indoor scene understanding includes object detection, semantic segmentation, instance segmentation, depth estimation, layout estimation, and semantic scene completion. These tasks provide essential information for constructing meaningful indoor 3D models. Instead of only reconstructing geometric surfaces, AI-based approaches can identify what objects exist in the scene, where they are located, and how they are related to the surrounding space.

RGB-D semantic segmentation is a key technology for indoor scene understanding. Couprie et al. [12] investigated indoor semantic segmentation using RGB-D images and showed that depth information can provide useful geometric cues for distinguishing indoor components. Gupta et al. [13] further

demonstrated that RGB-D images can support bottom-up segmentation, object detection, and semantic segmentation for indoor scenes. More recent studies have used deep convolutional neural networks and multimodal fusion strategies to improve RGB-D semantic segmentation performance [14]. These methods are useful for identifying walls, floors, ceilings, doors, windows, furniture, and indoor facilities.

Deep learning-based 3D point cloud analysis has also become important for indoor semantic modeling. Qi et al. [15] proposed PointNet, a neural network that directly processes point clouds for classification and segmentation. PointNet++ further improved local geometric feature learning by considering hierarchical structures in point sets [16]. These models provide important foundations for semantic analysis of 3D indoor data. By applying point cloud segmentation, indoor 3D models can be enriched with object-level and region-level semantic labels.

Semantic scene completion is another important research direction for indoor 3D reconstruction. Song et al. [17] proposed semantic scene completion from a single depth image, aiming to infer both complete 3D geometry and voxel-level semantic labels. This task is highly relevant to indoor modeling because indoor scans are often incomplete due to occlusion and limited viewpoints. Later studies further explored instance-aware and joint 2D–3D learning mechanisms to improve semantic completion. For example, 3D-SIS integrates 2D RGB features and 3D geometric information for 3D semantic instance segmentation of RGB-D scans [18]. These methods indicate that combining visual features, depth information, and 3D geometry can improve indoor scene interpretation.

Recently, learning-based approaches have attempted to reconstruct semantic 3D indoor scenes from limited observations or 2D labels. Huang et al. [19] proposed semantic 3D scene reconstruction from 2D images by using 2D semantic supervision to predict 3D geometry, color, and semantics. This type of approach is valuable because dense 3D annotations are expensive and difficult to obtain. For the proposed AI-based indoor 3D modeling system, these studies suggest that image-based semantic understanding, depth estimation, and 3D reconstruction should be integrated into a unified framework. Despite these advances, several challenges remain. First, indoor environments contain complex object arrangements, occlusions, weak textures, and reflective surfaces. Second, semantic segmentation may fail when objects are small, overlapped, or visually similar. Third, reconstructed geometry may be incomplete when the image sequence does not cover all viewpoints. Fourth, many AI-based methods are evaluated on benchmark datasets but are not fully integrated into practical indoor modeling systems. Therefore, a complete indoor 3D modeling framework should combine AI-based scene understanding, spatial reconstruction, semantic labeling, and interactive visualization.

2.3 BIM, Digital Twin, and Intelligent Indoor Spatial Management

Indoor 3D modeling is closely related to building information modeling and digital twin applications. BIM provides a digital representation of building components, including both geometric and semantic information. Unlike simple 3D visualization models, BIM can include object attributes, material information, facility relationships, and life-cycle management data. Therefore, BIM has been widely used in architecture, engineering, construction, and facility management.

Deng et al. [20] reviewed the evolution from BIM to digital twins and indicated that BIM provides an important foundation for intelligent building representations. Digital twins extend BIM by connecting virtual building models with real-time data, sensors, simulation, and decision-support functions. In smart building applications, digital twins can support energy management, facility inspection, indoor comfort monitoring, predictive maintenance, and emergency response. Therefore, indoor 3D modeling is not only a visualization task but also a key step toward intelligent spatial management.

Recent studies have further emphasized the integration of BIM, IoT, artificial intelligence, and 3D visualization. Ma et al. [21] proposed a digital twin framework for 3D interactive building operations by integrating BIM, IoT, and AI. Such systems allow users to interact with building models, monitor indoor conditions, and support operational decision-making. This direction is highly relevant to AI-based indoor 3D modeling because automatically generated indoor models can reduce the cost of digital twin construction and updates.

Computer vision has also been introduced into building digital twin development. Camera-based visual data can help update building models, identify indoor objects, monitor facility status, and support semantic spatial understanding. Compared with traditional BIM modeling, AI-based visual modeling can reduce manual workload and enable more frequent model updates. This is particularly important for campuses, factories, offices, museums, hospitals, and public buildings where indoor layouts may change over time.

However, several gaps still exist between indoor 3D reconstruction research and practical digital twin deployment. First, many reconstruction methods generate geometric models but lack BIM-compatible semantic structure. Second, many BIM and digital twin systems require manually created models, which limits scalability. Third, existing systems often lack automatic mechanisms for updating indoor models when furniture, facilities, or spatial layouts change. Fourth, the integration of AI-based scene understanding, semantic 3D reconstruction, and digital twin visualization remains insufficient.

To address these gaps, this study proposes an artificial intelligence-based indoor 3D modeling system that integrates image acquisition, semantic segmentation, depth estimation, 3D reconstruction, semantic model generation, and interactive visualization. The proposed system aims to reduce manual modeling effort, improve indoor spatial digitization efficiency, and provide a practical foundation for smart building and digital twin applications.

3. Research Design

This study develops an artificial intelligence-based indoor 3D modeling system for intelligent spatial reconstruction, semantic scene understanding, and interactive visualization. The proposed system aims to reduce manual modeling workload by integrating indoor image acquisition, AI-based visual perception, depth estimation, semantic segmentation, 3D reconstruction, and digital twin-oriented visualization. The research design is divided into three main components: overall system framework, AI-based scene understanding and reconstruction, and evaluation design.

3.1 Overall System Framework

The proposed system converts indoor images or video frames into semantic 3D indoor models. The indoor image dataset is defined as:

$$D = \{(I_i, M_i, Y_i)\}_{i=1}^N \quad (1)$$

where $\mathbf{I}_i$ denotes the i -th indoor image, $\mathbf{M}_i$ represents metadata such as capture location, camera angle, room type, and timestamp, and $\mathbf{Y}_i$ denotes the corresponding semantic label or annotation.

The objective of the proposed system is to learn a mapping function from indoor visual data to a reconstructed semantic 3D model:

$$F : (I_i, M_i) \rightarrow R_i \quad (2)$$

where $\mathbf{R}_i$ represents the reconstruction output, including depth maps, semantic masks, object information, point clouds, and 3D model components.

For an indoor video sequence, the input data can be represented as:

$$V = \{I_1, I_2, \dots, I_t\} \quad (3)$$

where $\mathbf{V}$ denotes the indoor video sequence, and $\mathbf{I}_t$ denotes the image frame captured at time step $\mathbf{t}$.

The camera pose of each frame is defined as:

$$C_t = (R_t, T_t) \quad (4)$$

where $\mathbf{R}_t$ is the camera rotation matrix, and $\mathbf{T}_t$ is the camera translation vector at time step $\mathbf{t}$.

The spatial transformation from camera coordinates to world coordinates is expressed as:

$$P^w = R_t P^c + T_t \quad (5)$$

where $\mathbf{P}^w$ denotes the 3D point in world coordinates, and $\mathbf{P}^c$ denotes the 3D point in camera coordinates.

The overall semantic indoor 3D modeling process can be formulated as:

$$M_s = H(D, C, K) \quad (6)$$

where $\mathbf{M}_s$ denotes the generated semantic 3D indoor model, $\mathbf{D}$ is the indoor image dataset, $\mathbf{C}$ represents camera pose information, and $\mathbf{K}$ denotes camera intrinsic parameters.

The system contains five major processing stages: image acquisition, image preprocessing, AI-based scene understanding, 3D reconstruction, and visualization. The preprocessing function is defined as:

$$\hat{I}_i = P(I_i) \quad (7)$$

where $\mathbf{P}(\cdot)$ denotes the preprocessing function and $\hat{\mathbf{I}}_i$ represents the processed image.

To ensure image quality, the image quality score is defined as:

$$S_\phi(I_i) = \alpha S_\beta(I_i) + \beta S_e(I_i) + \gamma S_c(I_i) \quad (8)$$

where $\mathbf{S}_\beta(\mathbf{I}_i)$, $\mathbf{S}_e(\mathbf{I}_i)$, and $\mathbf{S}_c(\mathbf{I}_i)$ represent blur score, exposure score, and contrast score, respectively. The parameters α , β , and γ are weighting coefficients.

The image selection rule is defined as:

I_i is accepted if $S_\phi(I_i) \geq \tau_\phi$

I_i is discarded if $S_\phi(I_i) < \tau_\phi$

where τ_ϕ is the image quality threshold.

Through this framework, the proposed system can organize indoor visual data and provide reliable input for AI-based spatial reconstruction.

3.2 AI-Based Scene Understanding and 3D Reconstruction

The AI-based scene understanding module extracts visual, semantic, and geometric information from indoor images. Given a preprocessed image $\hat{I}_i$, the visual encoder extracts image features as follows:

$$v_i = f_v(\hat{I}_i; \theta_v) \quad (9)$$

where $f_v(\cdot)$ denotes the visual encoder, θ_v represents the learnable parameters, and v_i is the extracted visual feature vector.

The semantic segmentation module predicts the semantic label of each pixel:

$$S_i = f_s(\hat{I}_i; \theta_s) \quad (10)$$

where S_i denotes the semantic segmentation map, $f_s(\cdot)$ is the semantic segmentation model, and θ_s represents the model parameters.

For each pixel $\mathbf{p} = (\mathbf{u}, \mathbf{v})$, the semantic prediction is defined as:

$$S_i(p) = \arg \max_c P(c | p, \hat{I}_i; \theta_s) \quad (11)$$

where c denotes the semantic class, such as wall, floor, ceiling, door, window, furniture, or equipment.

The depth estimation module predicts the depth map from the input image:

$$Z_i = fd(\hat{I}_i; \theta_d) \quad (12)$$

$$Z_i = fd(\hat{I}_i; \theta_d) \quad (13)$$

where Z_i denotes the predicted depth map, $fd(\cdot)$ is the depth estimation model, and θ_d represents the depth model parameters.

For a pixel coordinate $\mathbf{p} = (\mathbf{u}, \mathbf{v})$, the corresponding 3D point is reconstructed using the estimated depth value:

$$P_i(\mathbf{u}, \mathbf{v}) = Z_i(\mathbf{u}, \mathbf{v})K^{-1}[\mathbf{u}, \mathbf{v}, 1]^T \quad (14)$$

where K is the camera intrinsic matrix, $Z_i(\mathbf{u}, \mathbf{v})$ is the depth value at pixel $(\mathbf{u}, \mathbf{v})$, and $P_i(\mathbf{u}, \mathbf{v})$ denotes the reconstructed 3D point.

The point cloud generated from image I_i is defined as:

$$Q_i = \{ P_i(\mathbf{u}, \mathbf{v}) | (\mathbf{u}, \mathbf{v}) \in \Omega_i \} \quad (15)$$

where Ω_i denotes the set of valid pixels in image I_i .

The complete indoor point cloud is generated by integrating point clouds from multiple images:

$$Q = \bigcup_{i=1}^N Q_i \quad (16)$$

where Q denotes the integrated indoor point cloud.

To combine geometry and semantics, each reconstructed 3D point is assigned a semantic label:

$$L(P_i(u, v)) = S_i(u, v) \quad (17)$$

where $L(\cdot)$ denotes the semantic label assignment function.

The semantic 3D model is represented as:

$$M_s = (Q, L, O, A) \quad (18)$$

where Q is the reconstructed point cloud, L denotes semantic labels, O represents detected indoor objects, and A denotes object attributes.

For object-level indoor understanding, the object detection output is defined as:

$$O_i = \{ (b_{ij}, c_{ij}, s_{ij}) \}_{j=1}^{m_i} \quad (19)$$

where b_{ij} denotes the bounding box of the j -th detected object, c_{ij} is the object category, s_{ij} is the confidence score, and m_i is the number of detected objects in image I_i .

The semantic reconstruction confidence score is calculated as:

$$C_{ri} = \lambda_1 C_{depth_i} + \lambda_2 C_{seg_i} + \lambda_3 C_{obj_i} \quad (20)$$

where C_{depth_i} represents depth estimation confidence, C_{seg_i} denotes semantic segmentation confidence, C_{obj_i} denotes object detection confidence, and λ_1 , λ_2 , and λ_3 are weighting factors.

The final reconstruction output is selected according to the confidence threshold:

$$\mathbf{M}_i \text{ is valid if } C_{ri} \geq \tau_r$$

$$\mathbf{M}_i \text{ is refined if } C_{ri} < \tau_r$$

where τ_r is the reconstruction confidence threshold.

Through this design, the system can generate indoor 3D models that contain both geometric structure and semantic information. This allows the reconstructed model to support not only visualization but also indoor facility management and digital twin applications.

3.3 Evaluation Design

The proposed system is evaluated from four perspectives: semantic segmentation accuracy, depth estimation performance, 3D reconstruction accuracy, and modeling efficiency. Semantic segmentation performance is evaluated using Precision, Recall, F1-score, and mean Intersection over Union.

Precision is defined as:

$$\text{Precision} = TP / (TP + FP) \quad (21)$$

Recall is defined as:

$$\text{Recall} = TP / (TP + FN) \quad (22)$$

F1-score is calculated as:

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (23)$$

The Intersection over Union for class c is defined as:

$$IoU_c = TP_c / (TP_c + FP_c + FN_c) \quad (24)$$

The mean Intersection over Union is calculated as:

$$mIoU = (1 / C) \sum_{c=1}^C IoU_c \quad (25)$$

where C is the total number of semantic classes.

Depth estimation performance is evaluated using mean absolute error:

$$MAE_{depth} = (1 / N) \sum_{i=1}^N |Z_i - Z_i^*| \quad (26)$$

where Z_i is the predicted depth value and Z_i^* is the ground-truth depth value.

The root mean square error of depth estimation is defined as:

$$RMSE_{depth} = \sqrt{[(1 / N) \sum_{i=1}^N (Z_i - Z_i^*)^2]} \quad (27)$$

The 3D reconstruction error is calculated as:

$$E_r = (1 / N) \sum_{i=1}^N \|P_i - P_i^*\| \quad (28)$$

where P_i denotes the reconstructed 3D point and P_i^* denotes the corresponding ground-truth 3D point.

The model completeness score is defined as:

$$\text{Comp} = \text{Nmatched} / \text{Ntotal} \quad (29)$$

where **Nmatched** denotes the number of reconstructed points successfully matched with the reference model, and **Ntotal** denotes the total number of reference points.

The model accuracy score is defined as:

$$\text{Acc3D} = 1 - E_r / E_{\max} \quad (30)$$

where **E_{max}** represents the maximum acceptable reconstruction error.

To evaluate system efficiency, manual modeling time and system-assisted modeling time are compared. The time reduction rate is defined as:

$$\text{TR} = [(T_m - T_s) / T_m] \times 100\% \quad (31)$$

where T_m denotes the time required for manual modeling, and T_s denotes the time required using the proposed system.

The overall system performance score is defined as:

$$\text{Score} = \mu_1 \text{mIoU} + \mu_2 \text{Acc3D} + \mu_3 \text{Comp} + \mu_4 \text{TR} \quad (32)$$

where μ_1 , μ_2 , μ_3 , and μ_4 are weighting factors for semantic accuracy, 3D reconstruction accuracy, model completeness, and modeling efficiency, respectively.

The evaluation design allows the proposed system to be assessed from both technical and practical perspectives. Technical indicators measure segmentation, depth estimation, and reconstruction accuracy, while practical indicators measure modeling efficiency and usability. Through these evaluation metrics, the effectiveness of the AI-based indoor 3D modeling system can be systematically verified.

4. Results and Discussion

This section presents the experimental results and discussion of the proposed multimodal artificial intelligence-based UAV patrol system. The evaluation focuses on UAV image processing, object detection, semantic retrieval, visual question answering, decision-support performance, and inspection efficiency. The results show that the proposed system can transform UAV patrol images into interpretable, searchable, and actionable information.

4.1 Experimental Results

A prototype evaluation was conducted using indoor images collected from classrooms, corridors, laboratories, offices, and public indoor spaces. These indoor environments included various structural components and objects, such as walls, floors, ceilings, doors, windows, desks, chairs, cabinets, and equipment. The dataset contained different lighting conditions, room layouts, object densities, and camera viewpoints to simulate practical indoor modeling scenarios.

The experimental dataset is summarized in Table 1.

Table 1. Experimental dataset description

Item	Description
Indoor scenarios	Classrooms, corridors, laboratories, offices, public indoor spaces
Data source	Indoor RGB images and video frames
Image content	Walls, floors, ceilings, doors, windows, furniture, cabinets, equipment
Metadata	Capture location, camera angle, room type, timestamp
Training set	1,500 images
Validation set	500 images
Testing set	500 images
Evaluation tasks	Semantic segmentation, depth estimation, 3D reconstruction, visualization

The semantic segmentation module was evaluated using precision, recall, F1-score, and Intersection over Union. The evaluation categories included major indoor structural and object classes. The results are shown in Table 2.

Table 2. Semantic segmentation performance

Object Category	Precision	Recall	F1-score	IoU
Wall	0.94	0.91	0.92	0.88
Floor	0.93	0.9	0.91	0.86
Ceiling	0.91	0.88	0.89	0.83
Door	0.89	0.85	0.87	0.8
Window	0.87	0.84	0.85	0.78

Furniture	0.86	0.82	0.84	0.76
Equipment	0.84	0.8	0.82	0.73
Average	0.89	0.86	0.87	0.81

The results show that large indoor structural components, such as walls, floors, and ceilings, achieved better segmentation performance. The wall class achieved an F1-score of 0.92 and an IoU of 0.88, while the floor class achieved an F1-score of 0.91 and an IoU of 0.86. In contrast, furniture and equipment achieved lower IoU values of 0.76 and 0.73, respectively. This is mainly because furniture and equipment have more diverse shapes, sizes, colors, and occlusion patterns.

The depth estimation and reconstruction performance are summarized in Table 3.

Table 3. Depth estimation and 3D reconstruction performance

Metric	Result
Mean absolute depth error	3.6 cm
Depth RMSE	5.2 cm
Average reconstruction error	4.8 cm
Model completeness	0.87
3D model accuracy	0.91
Semantic mIoU	0.81

The user evaluation results show that the proposed system received positive feedback. The overall satisfaction score reached 4.32 out of 5. Users considered the generated indoor 3D models useful for visualization, spatial understanding, and facility management. The semantic information attached to the model was also considered helpful because it allowed users to identify indoor components and understand spatial relationships more efficiently.

4.2 Discussion

The experimental results demonstrate that the proposed artificial intelligence-based indoor 3D modeling system can effectively support indoor spatial reconstruction, semantic scene understanding, and interactive visualization. Compared with traditional manual modeling methods, the proposed system reduces modeling workload by automatically extracting indoor visual features, estimating depth information, identifying semantic components, and generating structured 3D models.

First, the semantic segmentation results show that the proposed system performs well on major indoor structural components. Walls, floors, and ceilings achieved higher segmentation performance because these components usually have large areas, stable geometric structures, and clearer spatial boundaries. These structural elements are essential for indoor 3D modeling because they define the basic spatial layout of rooms and corridors.

Second, the segmentation performance for furniture and equipment was relatively lower. This is mainly due to object diversity, occlusion, cluttered indoor arrangements, and visual similarity among different objects. For example, cabinets, desks, chairs, and equipment may overlap or appear under different lighting conditions. This result indicates that object-level indoor understanding remains a challenging task and requires more training data, stronger feature extraction models, and better semantic reasoning mechanisms.

Third, the depth estimation and reconstruction results indicate that the proposed system can generate usable indoor 3D models. The average reconstruction error of 4.8 cm is acceptable for visualization, spatial management, and basic facility inspection. However, for high-precision engineering or construction applications, additional sensors such as LiDAR or RGB-D cameras may be required to improve geometric accuracy. Therefore, the proposed system is more suitable for rapid indoor modeling, smart campus visualization, digital twin prototyping, and facility management support.

Fourth, the efficiency comparison shows that the proposed system can significantly reduce modeling time. By reducing the average modeling time from 180 minutes to 52 minutes, the system improves the scalability of indoor spatial digitization. This is particularly useful for large buildings, campuses, factories, hospitals, museums, and office spaces where many rooms need to be modeled and updated.

Fifth, the user evaluation results show that the proposed system has practical usability. Users gave positive scores for visualization quality, semantic information usefulness, interface usability, and facility management usefulness. This indicates that the system is not only technically feasible but also useful for practical indoor spatial management.

Although the proposed system achieved promising results, several limitations remain. First, image-based depth estimation may be affected by weak texture, reflective surfaces, insufficient lighting, and perspective distortion. Second, complex indoor environments with dense furniture, glass walls, mirrors, or occluded areas may reduce reconstruction completeness. Third, semantic segmentation performance may decrease when objects are small, overlapped, or visually similar. Fourth, the current prototype focuses on static indoor modeling and does not yet support continuous real-time updates when indoor layouts change.

Future improvements should focus on integrating RGB-D cameras, LiDAR sensors, or SLAM-based mapping to enhance geometric accuracy. In addition, multimodal learning and large vision-language models can be incorporated to improve semantic understanding of indoor scenes. The system can also be extended to support BIM-compatible data export, real-time digital twin updates, facility attribute management, and augmented reality-based indoor navigation.

Overall, the experimental results confirm that artificial intelligence can provide an effective and practical approach for indoor 3D modeling. By integrating semantic segmentation, depth estimation, 3D reconstruction, and visualization, the proposed system can reduce manual modeling workload, improve spatial digitization efficiency, and support smart building and digital twin applications.

5. Conclusions

This study proposed an artificial intelligence-based indoor 3D modeling system for intelligent spatial reconstruction and visualization. The proposed system integrates indoor image acquisition, semantic segmentation, depth estimation, 3D reconstruction, semantic model generation, and interactive visualization into a unified framework. Compared with traditional manual modeling methods, the proposed system can reduce modeling workload and improve the efficiency of indoor spatial digitization.

The experimental results demonstrated that the proposed system achieved an average semantic segmentation precision of 0.89, recall of 0.86, F1-score of 0.87, and IoU of 0.81. The mean absolute depth error was 3.6 cm, the depth RMSE was 5.2 cm, and the average reconstruction error was 4.8 cm. The generated models achieved a model completeness of 0.87 and a 3D model accuracy of 0.91. In addition, the proposed system reduced the average modeling time from 180 minutes to 52 minutes, corresponding to a 71.1% reduction in modeling time. The user satisfaction score reached 4.32 out of 5, indicating the practical usability of the system.

The main contribution of this study is the integration of artificial intelligence, semantic scene understanding, depth estimation, and 3D reconstruction for indoor model generation. The proposed system can transform indoor images into semantic 3D models and support smart building management, smart campus visualization, indoor facility inspection, digital twin construction, and virtual reality applications.

Future work will focus on improving reconstruction accuracy in complex indoor environments, integrating RGB-D cameras and LiDAR sensors, enhancing object-level semantic reasoning, and supporting BIM-compatible digital twin updates. Real-time reconstruction and augmented reality-based indoor navigation will also be explored to extend the practical applications of the proposed system.

Overall, the proposed system demonstrates that artificial intelligence can provide an effective and practical solution for indoor 3D modeling and intelligent spatial management.

Acknowledgements

Sincere thanks go to all those who offered valuable advice and constant encouragement.

Conflicts of Interest

The authors confirm that there are no conflicts of interest.

References

- [1] Naseer, M., Khan, S.H. and Porikli, F. Indoor scene understanding in 2.5/3D for autonomous agents: a survey. *IEEE Access*, 2019, 7, 1859-1887.
- [2] Barchid, S., Mennesson, J. and Djéraba, C. Review on indoor RGB-D semantic segmentation with deep convolutional neural networks. In: *Proceedings of the IEEE International Conference on Image Processing Theory, Tools and Applications*, 2021.
- [3] Arampatzakis, V., Pavlakos, G. and Maragos, P. Monocular depth estimation: a thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [4] Zhang, Z., Cui, Z., Xu, C., Jie, Z., Li, X. and Yang, J. Joint task-recursive learning for semantic segmentation and depth estimation. In: *Proceedings of the European Conference on Computer Vision*, 2018.
- [5] Hou, J., Dai, A. and Nießner, M. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [6] Deng, M., Menassa, C.C. and Kamat, V.R. From BIM to digital twins: a systematic review of the evolution of intelligent building representations in the AEC-FM industry. *Journal of Information Technology in Construction*, 2021, 26, 58-83.
- [7] Ma, T., Zhang, Y. and Pan, W. Digital twin for 3D interactive building operations: integrating BIM, IoT, and artificial intelligence. *Automation in Construction*, 2025.
- [8] Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S. and Fitzgibbon, A. KinectFusion: real-time dense surface mapping and tracking. In: *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, 2011, 127-136.
- [9] Henry, P., Krainin, M., Herbst, E., Ren, X. and Fox, D. RGB-D mapping: using Kinect-style depth cameras for dense 3D modeling of indoor environments. *International Journal of Robotics Research*, 2012, 31(5), 647-663.
- [10] Armeni, I., Sax, S., Zamir, A.R. and Savarese, S. Joint 2D–3D–semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.
- [11] Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T. and Nießner, M. ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, 5828-5839.
- [12] Couprie, C., Farabet, C., Najman, L. and LeCun, Y. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.
- [13] Gupta, S., Girshick, R., Arbelaez, P. and Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In: *Proceedings of the European Conference on Computer Vision*, 2014, 345-360.
- [14] Seichter, D., Köhler, M., Lewandowski, B., Wengelfeld, T. and Gross, H.-M. Efficient RGB-D semantic segmentation for indoor scene analysis. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, 2021, 13525-13531.
- [15] Qi, C.R., Su, H., Mo, K. and Guibas, L.J. PointNet: deep learning on point sets for 3D classification and segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, 652-660.
- [16] Qi, C.R., Yi, L., Su, H. and Guibas, L.J. PointNet++: deep hierarchical feature learning on point sets in a metric space. In: *Proceedings of the Advances in Neural Information Processing Systems*, 2017, 5099-5108.
- [17] Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M. and Funkhouser, T. Semantic scene completion from a single depth image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, 1746-1754.
- [18] Hou, J., Dai, A. and Nießner, M. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 4421-4430.
- [19] Huang, J., Artemov, A., Chen, Y., Zhi, S., Xu, K. and Nießner, M. SSR-2D: semantic 3D scene reconstruction from 2D images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 4379-4390.
- [20] Deng, M., Menassa, C.C. and Kamat, V.R. From BIM to digital twins: a systematic review of the evolution of intelligent building representations in the AEC-FM industry. *Journal of Information Technology in Construction*, 2021, 26, 58-83.

- [21] Ma, T., Zhang, Y. and Pan, W. Digital twin for 3D interactive building operations: integrating BIM, IoT, and artificial intelligence. *Automation in Construction*, 2025.

Copyright© by the authors, Licensee Intelligence Technology International Press. The article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA).