A Refinement Reconstruction Method for Indoor Structures Based on 3D Point Cloud Template Matching

Benhe Cai[1], Shengjun Tang[2] *, Weixi Wang[2], Linfu Xie[2], Xiaoming Li[2], Renzhong Guo[2]

[1] School of Resource and Environmental Sciences, Wuhan University, Wuhan 430072, China - benhe cai@whu.edu.cn
[2] School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China - (shengjuntang,wangwx,linfuxie, lxm, guorz)@szu.edu.cn

Keywords: Point Clouds, Indoor Scene Reconstruction, Model-Driven Approach, 3D Modeling, Multi-modal Feature.

Abstract

Indoor 3D reconstruction is a significant research topic in computer vision and computer graphics, focusing on the construction of complete and accurate models of indoor scenes from 3D point cloud data. Traditional data-driven methods often demonstrate poor robustness, low efficiency, and insufficient semantic information when addressing complex indoor environments. To address these challenges, this paper proposes a variable template matching-based method for indoor 3D scene reconstruction, which reframes the complex reconstruction problem as a matching problem. By adjusting and reconstructing library models according to the original instance parameters of the scene, the proposed method facilitates the fine-grained reconstruction of various complex elements within indoor spaces. Utilizing predefined geometric models and contextual constraints, this approach enhances the precision of indoor scene reconstruction, effectively overcoming the limitations associated with traditional data-driven techniques. Extensive experimental validation confirms the effectiveness of the proposed method, demonstrating its ability to alleviate issues such as point cloud noise, data loss, and occlusions, thereby improving both reconstruction accuracy and efficiency. Furthermore, by enriching the reconstructed models with semantic information, this method provides a more comprehensive data foundation for subsequent applications.

1. Introduction

Indoor scene 3D reconstruction technology aims to recover the three-dimensional structure of a scene from two-dimensional images or point cloud data. This technology is widely applied in various fields, including urban planning, architectural design, cultural heritage preservation (Yang and Zhu, 2021), virtual reality, indoor navigation, and Building Information Modeling (BIM). With the rapid advancement of 3D laser scanning technology and depth cameras, the acquisition of high-precision and high-density indoor point cloud data has become increasingly accessible, significantly propelling the development and application of existing 3D reconstruction methods in indoor environments (Zhang, 2023) (Mabrouk and Zagrouba, 2018) (Zhang et al., 2022). Existing 3D reconstruction methods can be broadly categorized into data-driven and model-driven approaches. Data-driven methods primarily allow the data to 'speak for itself,' learning patterns and features from large datasets without the need for predefined rules or models. Their advantages include adaptability to complex data patterns, the ability to uncover hidden correlations, and a high degree of flexibility. In recent years, deep learning has made significant strides in 3D reconstruction, exemplified by point cloud completion methods based on convolutional neural networks (CNNs) (Charles et al., 2017) (Yuan et al., 2018)and 3D shape understanding techniques utilizing Transformers (Pan et al., 2021). However, data-driven methods often face challenges such as poor model interpretability, reliance on substantial amounts of high-quality training data, and reduced robustness when confronted with noise, missing data, and complex structures. In contrast, model-driven methods leverage prior knowledge and predefined models to guide data analysis and problem-solving. These methods employ geometric model constraints to direct the reconstruction process, decomposing objects into a series of fundamental models and extracting features from data to fit these model parameters. For instance, structural modeling can be conducted using basic geometric primitives such as planes, cylinders, and cuboids (Nan et al., 2010), or through high-precision modeling that integrates semantic segmentation with geometric priors. By constraining the reconstruction process with prior knowledge, model-driven methods significantly enhance reconstruction accuracy, efficiency, and robustness. Furthermore, they offer greater interpretability and are better suited to manage data noise and missing information, making them particularly advantageous for indoor 3D modeling (Dai et al., 2017). Existing research on indoor 3D reconstruction has primarily concentrated on data-driven methods, while the application of model-driven approaches remains relatively limited. The diversity of indoor environments, particularly the challenges posed by complex structures, severe occlusions, and missing data, complicates the ability of traditional geometric models to accurately represent indoor scenes (Mura et al., 2014). For example, indoor environments frequently contain numerous furniture items and decorative objects, and data acquisition may be obstructed by viewpoint limitations and object occlusions, resulting in incomplete point cloud data (Armeni et al., 2019). Furthermore, indoor scenarios necessitate higher reconstruction accuracy and efficiency compared to outdoor environments. Applications such as virtual reality and indoor navigation require more precise and real-time reconstruction results. Consequently, model-driven fine-grained 3D reconstruction methods for indoor environments present significant research value and promising application prospects.

* Corresponding author

1.1 Related work

Data-driven methods primarily rely on models learned from extensive labeled datasets, utilizing deep learning and other techniques to automatically generate 3D models (Charles et al., 2017). Traditional data-driven approaches depend on the geometric features of point cloud data, employing techniques such as point cloud segmentation, plane extraction, and surface reconstruction (Guo et al., 2020). While these methods have shown promising results in specific scenarios, they require substantial labeled data for training and often lack robustness when confronted with noise, missing data, and complex structures (Han et al., 2017). This limitation is particularly pronounced in indoor environments, where occlusions and gaps significantly impair the performance of data-driven methods, complicating the construction of accurate models with semantic information (Kang et al., 2020). To address these limitations, researchers have begun to integrate prior knowledge and model constraints, leading to the emergence of model-driven 3D reconstruction methods, which have gradually become a focal point in the field of 3D reconstruction (Schnabel et al., 2007).

Model-driven approaches leverage prior knowledge and constraints, integrating models from fields such as architecture and interior design to guide and optimize the reconstruction process (Mura et al., 2014). Compared to traditional methods, model-driven 3D reconstruction demonstrates enhanced adaptability to complex indoor environments and structures through model fitting and optimization. This leads to improvements in accuracy, completeness, and efficiency, while also providing greater robustness and interpretability (Oesau et al., 2013). Common model-driven techniques include the RANSAC algorithm, least squares method, constrained least squares method, and graph matching techniques. The RANSAC algorithm is frequently employed to extract features from noisy data and to fit model parameters (Fischler and Bolles, 1981). The least squares method optimizes model parameters by minimizing errors between the model and the data (Hartley and Zisserman, 2003). The constrained least squares method extends this approach by incorporating constraints such as symmetry, parallelism, and orthogonality to enhance model accuracy (Furukawa and Ponce, 2009). Graph matching techniques represent point cloud data and models as graph structures, utilizing graph matching algorithms for effective model fitting (Zhou et al., 2018). These model-driven methods have found widespread application in areas such as building reconstruction and tunnel reconstruction. For example, researchers have successfully reconstructed roof structures using LiDAR point cloud data by leveraging prior knowledge of building roofs, including planar, gabled, and hipped roof models (Oesau et al., 2013). In tunnel reconstruction, researchers decompose tunnel structures into fundamental components and employ constrained least squares methods to accurately reconstruct 3D tunnel models.

In summary, the successful application of model-driven methods in outdoor environments, such as building and tunnel reconstruction, further validates their effectiveness and robustness. However, these existing model-driven methods are inadequate for addressing the challenges associated with complex indoor 3D reconstruction. Firstly, the complexity and diversity of indoor scenes hinder traditional geometric models from fully capturing their structures (Armeni et al., 2019). Current model libraries lack comprehensive support for objects such as furniture and decorations, complicating model selection and parameter estimation. Different scenes necessitate different modeling strategies, thereby increasing the adaptability requirements of the methods. Secondly, model-driven approaches heavily depend on prior knowledge and predefined models, which may result in suboptimal reconstruction outcomes when the environment deviates significantly from these predefined models (Dai et al., 2017). Moreover, occlusions and noise continue to pose challenges that affect the quality of point cloud data and reconstruction accuracy (Wu et al., 2011). The computational demands of processing high-precision and high-density point cloud data present another significant hurdle, as high computational complexity can reduce efficiency. The generalization capability of models may also be limited when confronted with complex or unique environments, and the optimization and adjustment processes often require substantial manual intervention, which hinders automation. Finally, although some semantic information has been integrated into model-driven reconstruction, its effectiveness remains suboptimal. Efficiently incorporating and utilizing semantic information to enhance reconstruction quality is still an unresolved issue.

To address these challenges, we propose a variable-template matching-based indoor 3D reconstruction method that effectively tackles the point cloud reconstruction challenges in indoor scenes and demonstrates superior performance compared to traditional techniques. The contributions are as follows.

- **Multi-modal Feature Extraction:** We propose a method that integrates geometric features from point clouds and visual features from multi-view depth maps. PointNet extracts initial point cloud features, refined via a Transformer encoder, while the CLIP model extracts visual features from depth projections. A Transformer-based fusion mechanism generates highly discriminative global features for model retrieval.
- **Optimized Model Retrieval:** We introduce a similarity retrieval approach that iteratively optimizes multiple similarity metrics and dynamically adjusts thresholds. This method efficiently selects the top-K most similar models, improving retrieval precision and robustness.
- **Dynamic Model Matching:** We develop a high-precision matching strategy combining rough alignment via RANSAC and fine alignment using ICP. Preprocessing steps, including outlier removal and coordinate normalization, ensure accurate transformation estimation and robust alignment.

2. Methodology

Our model utilizes an instance point cloud Q, which is obtained through indoor scene semantic instance segmentation, along with a predefined model library S. The reconstruction process is structured into three stages: multi-modal feature extraction, model retrieval, and model matching. Figure 1 shows the overall flow of our research method.

3. Data Processing

The quality of input data and model libraries is one of the key factors to ensure the accuracy of feature extraction, model retrieval, and fitting. In the following, we will introduce the relevant works of data preprocessing in this study.

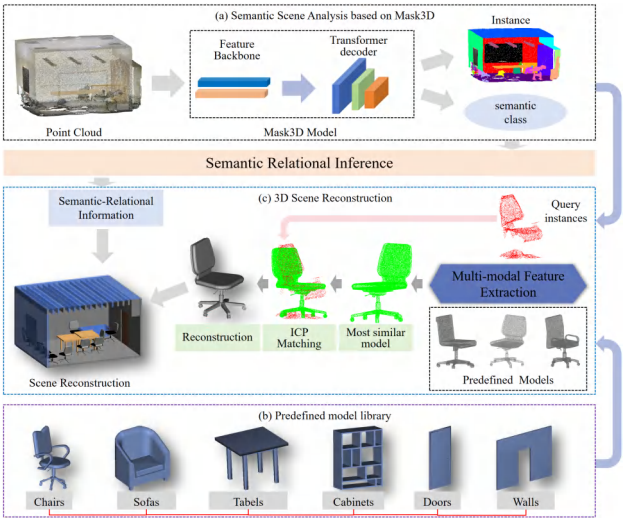


Figure 1. An overview of our approach. With an input (a) indoor scene instance and (b) predefined model library, the output 3D scene is reconstructed with a three-stage process (c): Multi-modal feature extraction, model retrieval, and model matching.

3.1 Public Datasets

To verify the feasibility and reliability of the method, we conduct experimental analyses using the large public dataset of indoor scenes known as S3DIS, which is generated from the Matterport 3D laser scanner. The S3DIS dataset encompasses six indoor regions, containing over 215 million points, 70,496 standard RGB images, 1,413 panoramic RGB images, and 272 indoor scenes with instance-level semantic annotations. The total area covered exceeds 6,000 square meters and includes 13 categories. Each point in the dataset is characterized by attributes such as surface normals, coordinates, and semantic annotations, which provide rich data information that ensures the reliability of the experimental results.

3.2 Predefined Model Library

In the field of model-driven 3D reconstruction, a predefined model library is the basis of the research, and it is essential for achieving efficient and accurate 3D modeling. In our study, we designed and constructed a predefined model library specifically for indoor environments, which is organized to manage various structural entities and movable accessories within buildings. The basic construction classification framework illustrated is illustrated in Figure 2, we adopted a multi-scale modular design to develop the library, ranging from single buildings to individual rooms and specific example models. This approach allows for rapid customization and extension of building models while maintaining consistency and accuracy, enabling users to select appropriate building components based on varying needs and application scenarios. For instance, when considering a single room, the predefined model library decomposes the building structure into structural entities (e.g., ceilings, floors) and accessory structural entities (e.g., windows, doors), as well as movable parts (e.g., furniture such as chairs, sofas, tables, cabinets) and other miscellaneous categories. This hierarchical categorization approach effectively breaks down complex interior structures into smaller, more manageable components, thereby enhancing the organization, storage, and management of predefined models, while also improving the usability, flexibility, and scalability of the model library.

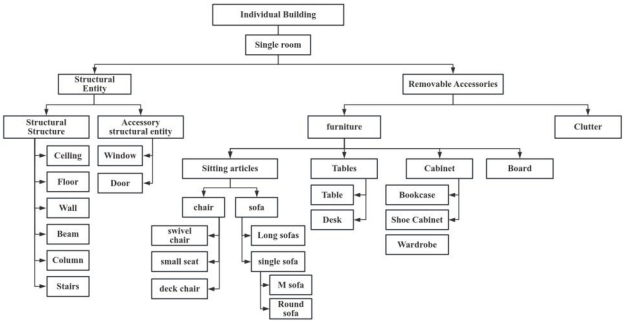


Figure 2. The basic classification framework of the predefined model library.

In this study, in order to realize effective 3D reconstruction, we established a predefined model library containing more than 300 different kinds of styles by means of Revit software and network data retrieval. This library includes structural entities such as 10 types of ceilings, 10 types of floors, and 30 types of walls. Additionally, the accessory structural entities consist of 60 types of windows and 45 types of doors. The movable parts library features 50 types of sofas, 60 types of chairs, 35 types of tables, and 30 types of coffee tables, among others. Some of the predefined models are shown in Figure 3.

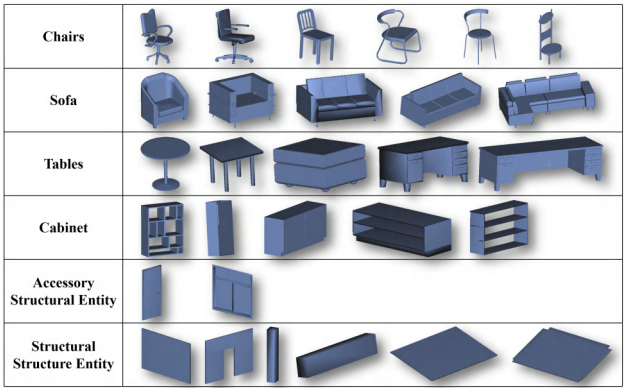


Figure 3. Partial Display of Predefined Models.

4. Method Implementation

In this section, we detail the implementation of our proposed method for indoor 3D scene reconstruction based on variable template matching, accompanied by a comprehensive explanation of the key techniques employed.

4.1 Input Data Preparation

Based on the public S3DIS dataset (Armeni et al., 2016), in order to obtain the instance information that needs to be reconstructed from the disordered point cloud. We firstly adopt the Mask3D model (Schult et al., 2023) to perform instance-semantic segmentation on the original point cloud to realize the fine-grained extraction and classification of the complex indoor scene, in order to recognize the key constituent parts of the indoor scene, such as the furniture, the walls and the floors, and obtain the corresponding information such as point cloud data

and semantics of the components as input for query instances, as shown in Figure 4(a).

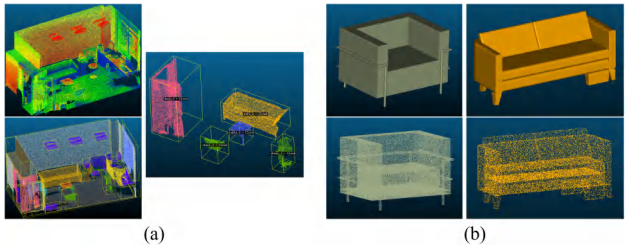
For the candidate models in the constructed predefined model library, we have designed a highly robust BIM model loading and conversion method.The core of this method is its ability to achieve efficient data conversion and processing through adaptive sampling and support for multiple formats. In this method, we introduce a dynamic threshold adjustment mechanism that selects the appropriate voxel size for downsampling based on the density distribution of the point cloud, thereby optimizing processing efficiency. This method not only accommodates BIM models in various formats but also enhances the efficiency and quality of data conversion through adaptive sampling and density-aware preprocessing. Specifically, for the models in the BIM model library, the appropriate loading method is selected based on the file type. For mesh models (e.g., OBJ, STL), we utilize the Open3D library to convert them into point clouds and perform adaptive sampling (as shown in equation(1).), results as shown in Figure 4(b).

$$N_{\text{sample}} = N_{\text{base}} + \alpha \cdot \frac{N_{\text{effective}}}{D_{\text{local}}} \tag{1}$$

where N_{sample} = number of sampled points
N_{base} = base number of sampled points
$N_{\text{effective}}$ = effective number of points
α = adaptive factor controlling the impact of density
D_{local} = local density factor

Based on equation (1), the sampling process is dynamically adjusted according to the characteristics of the point cloud, ensuring balanced sampling across different regions. The base sample count (N_{base}) ensures a minimum resolution, while the effective points address noise and outliers. The adaptive factor (α) controls the sampling sensitivity to density variations, ensuring appropriate sampling of high-density areas without over- or under-sampling.

Usually, the query point cloud and the predefined model may differ in the coordinate system, scale, and resolution, which can significantly affect feature extraction and model retrieval. To mitigate this issue, we first compute the bounding box of both the query and model point clouds to determine their center points and extents. Then, we scale the point clouds based on the maximum range to fit them within a unit cube. Finally, a translation is applied to move the point cloud's center of mass to the origin, ensuring consistency in scale and position for subsequent processing.



(a) (b)

Figure 4. Input data preprocessing. (a) 3D scene semantic instance segmentations. (b) Adaptive sampling based on BIM model.

4.2 Multi-modal Feature Extraction

Feature extraction from point cloud data is a critical step in reconstruction, as it significantly influences the accuracy of subsequent model retrieval. We propose a multimodal feature fusion-based feature extraction method that integrates the global feature extraction capabilities of Transformers with the depth information enhancement provided by multi-view projections. This approach harnesses both the geometric information inherent in point clouds and the visual information derived from multi-view depth maps, thereby enhancing robustness and distinguishing ability.

In this study, we employ high-fidelity projection techniques to convert sparse 3D point cloud data into multi-view depth maps, this enhancing spatial information capture and structural perception in three-dimensional space, thereby boosting model recognition performance. Specifically, as shown in equation (2), we define N fixed viewpoints, which can be dynamically adjusted depending on the application scenario. Each viewpoint represents a unique camera position and orientation, enabling the capture of the point cloud from various angles to generate multiple-depth maps. A perspective projection transformation is applied to convert the 3D point cloud data into 2D depth maps, which are then smoothed using Gaussian filtering. As a result, we can get the depth map of N views, and each depth map reflects the depth of the point cloud from a specific camera viewpoint.

$$\mathbf{p}_2^{(i)} = \mathbf{K} \cdot \mathbf{R}^{(i)} \cdot \mathbf{p}_3 + \mathbf{t}^{(i)} \tag{2}$$

where $\mathbf{p}_3$ = 3D point in the point cloud, $[X, Y, Z, 1]^T$
$\mathbf{p}_2$ = 2D projection point on the image plane, $[x, y, 1]^T$
$\mathbf{K}$ = camera intrinsic matrix
$\mathbf{R}$ = rotation matrix
i = current viewpoint (total N viewpoints)

For resulting N multi-view depth maps, we use a pre-trained CLIP visual Transformer (ViT-B/16) to extract depth features, as shown in equation (3), which are encoded into feature vectors for subsequent fusion. For the point cloud, we initially apply the PointNet method to compute local feature vectors, and these vectors are subsequently processed using a Transformer Encoder Layer, which utilizes an 8-head Multi-Head Attention mechanism to learn the global dependencies among the points. and aggregates the features into a comprehensive descriptor vector by global max pooling, resulting in a 128-dimensional feature representation through a fully connected layer and obtain global feature finally,as shown in equation (4). . After obtaining the depth map features and point cloud features, the feature information from both modalities is effectively integrated using a fully connected layer and a Transformer encoder, ultimately resulting in a highly discriminative global feature representation, as shown in equation (5).

$$\mathbf{f}_{\text{depth}} = \mathrm{F}\left(\left\{ \frac{\mathrm{CLIP}(\mathrm{RP}(\mathbf{pc}, i))}{\|\mathrm{CLIP}(\mathrm{RP}(\mathbf{pc}, i))\|} \right\}_{i=1}^{N}, \mathrm{dim} = 0 \right) \tag{3}$$

where $\mathbf{pc}$ = point cloud data
$\mathrm{RP}(\mathbf{pc}, i)$ = projection from the i-th viewpoint
$\mathrm{CLIP}(\cdot)$ = CLIP model for feature extraction
$\|\mathrm{CLIP}(\mathrm{RP}(\mathbf{pc}, i))\|$ = normalization of the feature
$\mathbf{f}_{\text{depth}}$ = concatenated feature of all N views

$$\mathbf{f}_{\text{point}} = \mathbf{W}_3 \left(\max\left(\mathrm{T}\left(\mathbf{X}_2 + \mathrm{P}(\mathbf{X}_2)\right), \dim = 0\right)\right) + \mathbf{b}_3 \quad (4)$$

where $\mathbf{X}_2$ = local feature
$\mathbf{X}$input = input point cloud data
$\mathrm{P}(\mathbf{X}_2)$ = position encoding
$\mathrm{T}(\cdot)$ = Transformer Encoder
$\mathbf{X}_{\text{pooled}}$ = global feature
$\mathbf{W}_3$ = weight matrix for the final fully connected layer
$\mathbf{b}_3$ = bias for the final fully connected layer
$\mathbf{f}_{\text{point}}$ = final global feature

$$\mathbf{f}_{final} = \text{Transformer}(\mathbf{f}_{\text{point}} \oplus \mathbf{f}_{\text{depth}}) \quad (5)$$

Through this series of operations, we obtain a highly discriminative global feature that effectively integrates the geometric information of the point cloud with the visual information from the depth map. This feature not only yields a high-precision representation of the point cloud data but also captures information from various viewpoints, thereby providing more accurate and robust feature representations for subsequent tasks, such as model retrieval and 3D reconstruction.

4.3 Model Retrieval

In the model retrieval phase, in order to improve the accuracy and robustness of model retrieval, we design a similarity retrieval method that incorporates multi-method iterative optimization and introduces a dynamic threshold adjustment mechanism to adapt to different data distributions and matching requirements, aiming at efficiently identifying the model that is most similar to the query point cloud from a set of candidate models. The method dynamically adjusts the similarity threshold through continuous iterative computation and dynamically selects the appropriate similarity computation method to achieve the purpose of obtaining the optimal solution. The essence of this method lies in the integration of multiple similarity computation techniques, which greatly enhances the robustness of model detection. Specifically, based on the multi-modal fused features extracted from the query point cloud and candidate model in Section 4.2, we calculate the similarity between the query feature and candidate feature as follows equation (6):

$$\text{sim}_m(f_q, f_k) \text{ for each method } m \quad (6)$$

where f_q = feature vector of the query instance
f_k = feature vector of the candidate model
sim_m = similarity score calculated for method m

To improve the robustness of similarity calculation, we perform a weighted fusion of the results from different similarity computation methods, obtaining a weighted similarity score:

$$\text{sim}_{fused}(f_q, f_k) = \sum_{m=1}^{M} w_m \cdot \text{sim}_m(f_q, f_k) \quad (7)$$

where M = total number of similarity calculation methods
w_m = weight for each method m

In the retrieval process, we adopt a dynamic threshold adjustment mechanism to update the similarity threshold based on the current similarity calculation results. The threshold update formula is as follows:

$$\text{thresh}_{t+1} = \text{thresh}_t + \alpha \left(\text{sim}_{fused}(f_q, f_k) - \text{thresh}_t\right) \quad (8)$$

where α = learning rate,
controlling the speed of threshold update.

In each iteration, we dynamically select the most suitable similarity calculation method m^* based on the current similarity values:

$$m^* = \arg\max_{m \in M} \text{sim}_m(f_q, f_k) \quad (9)$$

Finally, after multiple iterations of calculation and optimization, we rank the candidates according to the weighted similarity scores and output the top K most similar candidate models to the query point cloud, ensuring the accuracy and robustness of the retrieval results:

$$\text{Top-K}(f_q) = \arg\max_{f_k} \text{sim}_{\text{fused}}(f_q, f_k) \quad (10)$$

Following these calculations, we obtain a similarity ranking of the image object in relation to all models in the library, as illustrated in Figure 5(b). For reference, this study presents the top three retrieval results.

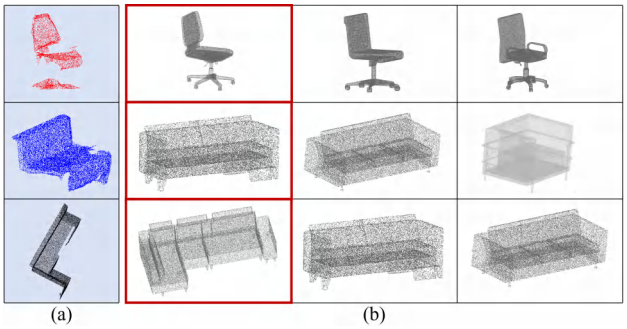


Figure 5. Model retrieval result for the query instance. (a) Input query instance. (b) Top 3 retrieval results in candidate predefined models (Models are sorted in descending order based on similarity scores).

4.4 Model Matching

After obtaining the most similar predefined model, it is essential to align and adjust the retrieved 3D model with the objects in the input image. This alignment must ensure that the model matches the input object in terms of spatial position, scale, and orientation, thereby achieving the final 3D reconstruction. To attain high-precision alignment between the query instance point cloud and the target point cloud, we have designed a dynamic matching method. This method progressively enhances registration accuracy and robustness through a series of optimization steps, ensuring both efficiency and reliability in complex scenes.

First, based on the most similar predefined model retrieved, we obtain the original query point cloud and the predefined model

point cloud for matching. Preprocessing operations are conducted on both point clouds, including outlier removal, unit normalization, and coordinate system alignment, to mitigate the effects of data noise and discrepancies on the registration results. Following these operations, we acquire a set of partially overlapping, clean point clouds. Subsequently, the RANSAC algorithm is employed for coarse matching, allowing for the rapid estimation of an approximate alignment transformation matrix, as shown in equation (11).

$$E(T) = \sum_{i=1}^{N} ||T(\mathbf{x}_i) - \mathbf{y}_i||^2 \tag{11}$$

where $\mathbf{x}_i$ = source point of query instance, $[X, Y, Z]^T$
$\mathbf{y}_i$ = corresponding point of target, $[x, y, z]^T$
$\mathbf{T}$ = transformation matrix
N = number of matching points
$E(T)$ = error metric to minimize
$T(\mathbf{x}_i)$ = the transformation matrix

Building on this, the ICP algorithm is utilized to iteratively identify the closest point pairs and refine the transformation matrix, thereby performing fine matching to enhance alignment accuracy, as shown in equation (12)(13)(14). To increase the algorithm's adaptability, we introduced a dynamic threshold adjustment mechanism during matching. This mechanism dynamically modifies the maximum corresponding point distance threshold in the ICP algorithm based on the feature range of the point cloud, facilitating an adaptive adjustment of matching precision according to the local features of the point cloud and circumventing the limitations associated with fixed thresholds.

$$E(T) = \sum_{i=1}^{N} ||T(\mathbf{x}_i) - \mathbf{y}_i||^2 \tag{12}$$

$$R, t = \text{argmin} \sum_{i=1}^{N} ||T(\mathbf{x}_i) - \mathbf{y}_i||^2 \tag{13}$$

$$T' = T \cdot \begin{pmatrix} R & \mathbf{t} \\ 0 & 1 \end{pmatrix} \tag{14}$$

where $\mathbf{x}_i$ = source point of query instance, $[X, Y, Z]^T$
$\mathbf{y}_i$ = corresponding point of target, $[x, y, z]^T$
$\mathbf{T}$ = transformation matrix
R = rotation matrix
$\mathbf{t}$ = translation vector
N = number of matching points
$E(T)$ = error metric to minimize

Furthermore, we conducted post-processing on the final alignment results, which included centroid and angle optimization to further enhance alignment accuracy. A validation mechanism was implemented to ensure the reliability of the alignment results, ultimately achieving the outcome depicted in Figure 6(b).

5. Experiment and Analysis

To evaluate the effectiveness of our method, we conduct tests on various indoor scene point clouds sourced from public datasets and analyze the results.Based on the query point cloud of

Figure 6. The match results between the query instance and the predefined model. (a) Input indoor scene query instance point cloud. (b) Matching result visualization. (c) Output final indoor scene query instance reconstruction result.

semantic strength segmentation and the predefined model library constructed, we validate the effectiveness of the method by controlling the difference of its missing degree (as shown in Figure 7) based on query instances of different types and scales (as shown in Figure 8) as well as the same query instance point cloud, respectively.

For query instance point clouds of different scales, our method demonstrates good adaptability and accuracy, as shown in Figure 8. However, the effectiveness of model retrieval is influenced to varying degrees by the presence of missing data within the input query instance point cloud. Specifically, as depicted in Figure 7, when the input query instance point cloud is relatively complete, our method accurately matches the correct model. When there is missing data, if the missing data is not situated at critical locations, our method can still successfully identify the correct model. However, when the missing data occurs at critical locations, the matching outcomes may become biased. For instance, as shown in Figure 7, if an office chair is missing essential wheels, or if a right-angled sofa lacks critical right-angle information, the matching result tends to favor the model most similar to the current state of the point cloud with missing components, rather than the expected correct model. This shows that our method achieves good reconstruction work in that it can, to a certain extent, overcome the problem of missing data due to occlusion, but further optimization is needed to improve its robustness and accuracy in the face of missing critical information.

After the above coarse-to-fine matching, we can get a predefined model that is basically consistent with the query instance. After the above coarse-to-fine matching, we can get a predefined model that is basically consistent with the query instance, and then we combine the semantic, color, and other

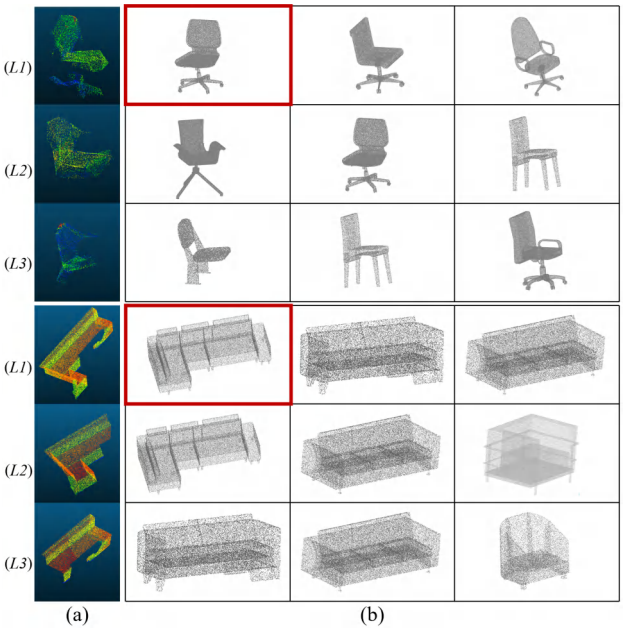


Figure 7. Retrieval results for one query instance point cloud using different degrees of points missingness. (a)Input query point cloud, while L1, L2, L3 correspond to different levels of points missingness. (b) Top 3 retrieval results in candidate predefined models. (With the models within the red boxes indicating the true model.)

information obtained from the semantic-instance segmentation in Section 4.1 to adjust the matched predefined model. Finally, we can get a model with the same location, orientation finally, semantics, color, and relationship of the query instance. Based on this information, we put the final model into the 3D scene that needs to be reconstructed. Utilizing this information, we incorporate the final model into the 3D scene that requires reconstruction. Thus, we have successfully achieved the reconstruction process through matching for 3D scene reconstruction. Figure 9 illustrates our final 3D interior scene.

6. Conclusions

This study presents a method for indoor 3D scene reconstruction based on variable template matching, effectively transforming the complex reconstruction problem into a matching task. The method directly addresses the challenges faced by traditional data-driven approaches in managing intricate indoor environments, including poor robustness, low efficiency, and insufficient semantic information. By constructing a predefined model library and utilizing instance point clouds obtained from semantic instance segmentation as input, the method facilitates the effective retrieval of models from the predefined library. It subsequently matches and adjusts the retrieved models according to the parameters of the query instance. Furthermore, this approach integrates semantic information derived from semantic segmentation with 3D point cloud models, enabling fine-grained reconstruction of diverse movable entities within indoor scenes. This integration not only enhances the accuracy and efficiency of reconstruction but also enriches the semantic content of the model, providing more comprehensive data support for subsequent applications. Experimental results demonstrate that the proposed method, when confronted with challenges such as point cloud noise, missing data, or occlu-

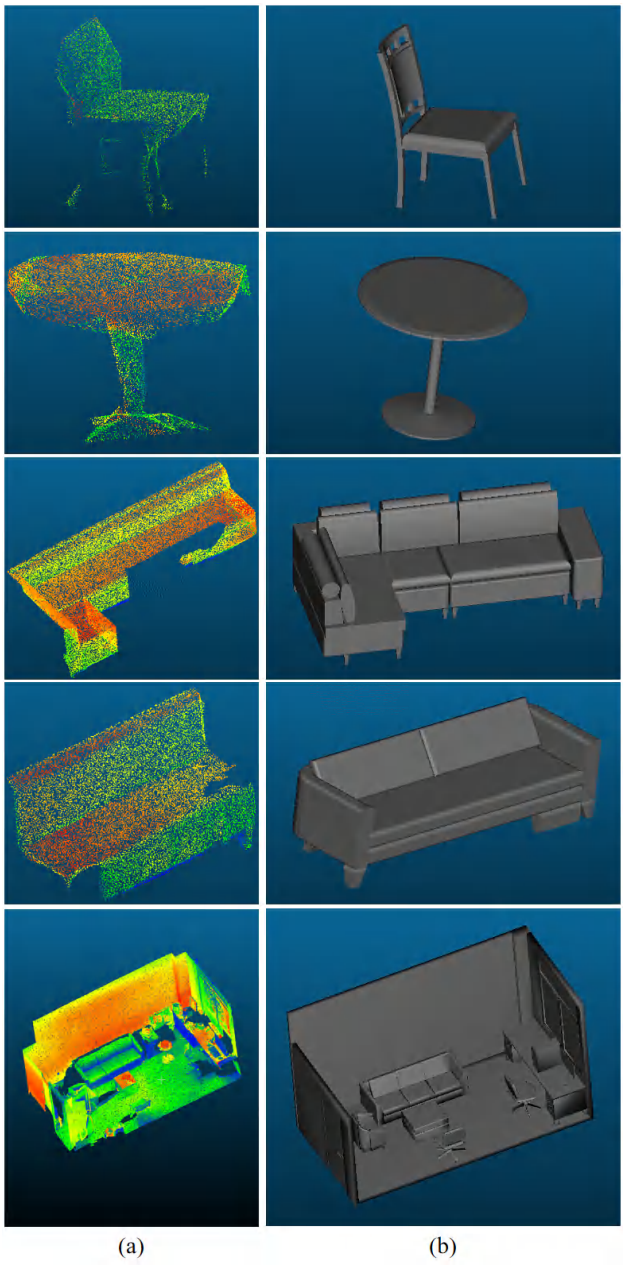


Figure 8. Visualization of different instances and scale reconstruction results. (a) Input indoor scene or query instance point cloud. (b) reconstructed 3D scene or single instance.

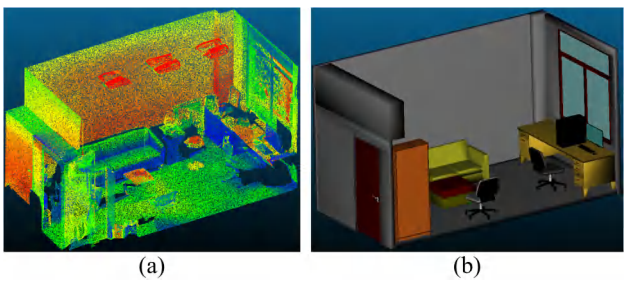


Figure 9. Indoor scene reconstruction result. (a) Input indoor scene point cloud data. (b) Final indoor scene reconstruction result.

sion, significantly improves accuracy, efficiency, and robustness, thereby offering an effective solution for fine-grained 3D reconstruction of indoor scenes.

Despite the demonstrated advantages, several key issues and challenges remain to be explored and resolved in future research. Building a comprehensive and diverse indoor object model library, developing more efficient model-matching algorithms, and better integrating semantic information into the reconstruction process are crucial directions for future studies in this field. These research avenues will not only propel the development of indoor 3D reconstruction technology but also provide robust technical support and innovation potential for related application areas, such as virtual reality, augmented reality, and building information modeling (BIM).In summary, this study offers a valuable perspective and technical approach in the realm of indoor 3D scene reconstruction. Through a model-driven method, it effectively enhances the accuracy and efficiency of reconstruction while also providing guidance for future research efforts. With ongoing technological advancements and deeper investigations, this approach is anticipated to yield broader applications and further optimizations in the future.

7. Acknowledgements

This work was supported in part by Natural Science Foundation of China(Project Nos.42471442), Research Project of Natural Science Foundation of Guangdong Province (Project No. 2024A1515030061), Research Project of Shenzhen S and T Innovation Committee (Project No. KJZD20230923115508017) and Research project of State Key Laboratory of Subtropical Building and Urban Science(Project No. 2023ZB18)

References

Armeni, I., He, Z. Y., Gwak, J. Y., 2019. 3d scene graph: A structure for unified semantics, 3d space, and camera. *Proceedings of the IEEE/CVF international conference on computer vision*, 44, 5664–5673.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., 2016. 3D Semantic Parsing of LargeScale Indoor Spaces. *In IEEE Conference on Computer Vision and Pattern Recognition*.

Charles, R. Q., Su, H., Mo, K., Guibas, L. J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Dai, A., Nießner, M., Zollhöfer, M., 2017. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36, 1.

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.

Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8), 1362–1376.

Guo, Y., Wang, H., Hu, Q., 2020. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12), 4338-4364.

Han, X., Li, Z., Huang, H., 2017. High-resolution shape completion using deep neural networks for global structure and local geometry inference. *Proceedings of the IEEE international conference on computer vision*, 85-93.

Hartley, R., Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Kang, Z., Yang, J., Yang, Z., 2020. A review of techniques for 3d reconstruction of indoor environments. *ISPRS International Journal of Geo-Information*, 9(5), 330.

Mabrouk, A. B., Zagrouba, E., 2018. Abnormal behavior recognition for intelligent video surveillance systems: A review. *IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, 91, 480–491.

Mura, C., Mattausch, O., Villanueva, A. J., 2014. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts. *Computers Graphics*, 44, 20–32.

Nan, L., Sharf, A., Zhang, H., 2010. Smartboxes for interactive urban reconstruction. *ACM Siggraph 2010 Papers*, 1–10.

Oesau, S., Lafarge, F., Alliez, P., 2013. Indoor scene reconstruction using primitive-driven space partitioning and graph-cut. *Eurographics workshop on urban data modelling and visualisation*.

Pan, X., Xia, Z., Song, S., 2021. 3d object detection with pointformer. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7463–7472.

Schnabel, R., Wahl, R., Klein, R., 2007. Efficient RANSAC for point-cloud shape detection. *Computer graphics forum. Oxford, UK: Blackwell Publishing Ltd*, 26(2), 214–226.

Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., 2023. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. *IEEE International Conference on Robotics and Automation (ICRA)*, 8216–8223.

Wu, J., Zhang, C., Xue, T., 2011. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29.

Yang, X., Zhu, S. N., 2021. Application of 3D Laser Scanner in Digitization of Movable Cultural Relics. *IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, 0–0. doi:10.1109/icpeca51329.2021.9362575.

Yuan, W., Khot, T., Held, D., 2018. Pcn: Point completion network. *international conference on 3D vision (3DV)*, 728–737.

Zhang, J., Li, S., Zhao, Z., Gao, Y., 2022. Highly sensitive three-dimensional scanning triboelectric sensor for digital twin applications. *Nano Energy*, 97, 107198–107198. doi:10.1016/j.nanoen.2022.107198.

Zhang, M., 2023. 3D scene image surface reconstruction method based on virtual reality technology. *International Conference on Mathematics, Modeling, and Computer Science (MMCS2022).*, 12625, 733–739.

Zhou, Q. Y., Park, J., Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv preprint*, 1801, 09847.