https://doi.org/10.1038/s40494-025-01701-2

Cross modal networks for point cloud semantic segmentation of Chinese ancient buildings

Check for updates

Zhiying Xie[1,2], Hanxin Liu[1] ✉, Yuanrong He[1,2] ✉, Yuanhao Shi[1], Peng Yu[1,2], Junqiang Ai[1] & Liang Zhong[1,2]

Point clouds have become essential for digital heritage preservation, aiding in the identification and classification of complex structural elements. However, most datasets rely on single-modal data, limiting their ability to describe real-world scenarios comprehensively. This paper introduces the Real-World Multi-modal Ancient Architecture Point Cloud Semantic Segmentation Dataset (RW-MAPCSD), which includes multi-modal data such as point clouds, line drawings, color, and depth projections, enabling multi-modal analysis of ancient buildings. To address data imbalance, we propose a novel segmentation network, Mask2former-KNN 3D Network (MK3DNet). The network projects point clouds into images while preserving point indices, using image segmentation techniques for initial segmentation. Results are then mapped back to the point cloud and refined with the K-nearest neighbors (KNN) algorithm. Experimental results show significant improvements, with mIoU and OA of 77.47% and 90.85%, respectively, surpassing the Point Transformer network by 21.94% and 5.87%.

As an integral component of Chinese cultural heritage[1], many ancient buildings have gradually deteriorated over time, with eroded surface textures and threatened structural integrity, even facing the risk of disappearance[2]. The digital preservation[3–6] of point cloud data has increasingly become one of the essential methods for protecting ancient architecture. However, significant technical challenges remain in the semantic segmentation of point cloud data[7]. The disorderly and unstructured nature of point cloud data complicates processing. In semantic segmentation tasks, models must identify structural details and functional components from vast amounts of three-dimensional point data, demanding high technical accuracy. Nevertheless, existing point cloud segmentation methods are insufficient for recognizing and refining complex architectural structures[8]. In particular, for irregularly shaped components and closely structured buildings, the segmentation results tend to be coarse, making it difficult to identify and delineate various details, which fails to meet the practical requirements for preservation and restoration. Moreover, most current point cloud semantic segmentation models predominantly rely on single-modal data[9–11]. This data often lacks a nuanced representation of real-world scenes, leading to poor adaptability and insufficient segmentation accuracy of the trained models in authentic ancient architectural environments.

The construction and utilization of datasets are central to the intelligent preservation and restoration of ancient architecture. With ongoing advancements in artificial intelligence, an increasing number of ancient building datasets have been developed, serving diverse purposes such as classification, segmentation, completion, and generation. These datasets exist in various formats, including images, point clouds, and multi-view data, addressing a wide range of requirements from building digitization to 3D reconstruction. Image datasets[12], for instance, are primarily used to document the appearance, intricate details, and localized features of buildings. The Versailles-FP dataset, proposed by Swaileh et al. [13], contains planar images of the Palace of Versailles along with annotated ground and wall features, facilitating studies on architectural evolution and historical development. Similarly, Barz et al. [14] developed a dataset comprising 9,485 church images, providing fine-grained classification labels and bounding box annotations for 631 architectural features, enabling detailed analysis of subtle visual differences in architectural style recognition. Multi-view datasets[15,16], by capturing information from multiple perspectives, provide comprehensive data for 3D reconstruction. Meanwhile, point cloud datasets[17,18], generated through laser scanning and other sensors, offer high-precision digital representations of structural and ornamental details. Zhou et al. [19] introduced a multi-source data fusion approach to address challenges related to data insufficiency and low quality in point cloud datasets of traditional Chinese wooden architectural elements, such as the complex brackets used in official-style structures.

The construction of ancient building datasets is influenced not only by building types and conservation needs but also closely related to the coverage and accuracy requirements of these datasets. Western-style buildings,

[1]School of Computing and Information Engineering, Xiamen University of Technology, Xiamen, China. [2]Big Data Institution of Natural Hazards Monitoring for Digital Fujian, Xiamen University of Technology, Xiamen, China. ✉e-mail: l492115081@gmail.com; 2012112001@xmut.edu.cn

such as churches[10,14] and palaces[13,20], generally feature complex appearances and structures, whereas Eastern wooden structures, particularly ancient Chinese wooden architecture[19,21], exhibit intricate craftsmanship and ornate decoration. Furthermore, datasets for different types of buildings vary in scale and coverage. Some datasets focus on high-precision documentation of individual buildings[22,23], while others encompass multiple structures within a specific region or city[24,25], thereby facilitating cultural heritage conservation efforts on a broader scale. Certain large datasets[26,27] even extend their coverage to the national level, offering more extensive resources for preservation and restoration.

Although existing datasets of ancient buildings have played a significant role in the digital protection of cultural heritage, there are still challenges, such as a limited number of datasets and insufficient model training. Computer-generated synthetic datasets[28,29], such as Building3D[30], although they have certain application value when field scanning cannot be performed, the high precision and authenticity of field scanning datasets remain optimal for accurately representing the current condition of buildings. Furthermore, many existing datasets concentrate on a single modality, lacking multi-modal data fusion, which hampers the comprehensive representation of the complex structural characteristics of buildings, particularly in the digital protection of ancient Chinese timber structures.

With the continuous advancements in technology, the evolution from traditional geometric analysis to machine learning, and eventually to deep learning applications, has progressively driven the intelligent and efficient preservation of ancient architecture through artificial intelligence. Traditional methods primarily rely on geometric modeling techniques, utilizing predefined mathematical models and statistical analyses to reconstruct and segment the shapes, structures, and details of ancient buildings[31–33]. These approaches often employ basic geometric decomposition strategies, such as slicing[34], to stably generate three-dimensional models[35] or identify architectural features[36]. However, their efficiency and adaptability are limited when dealing with large-scale or complex structures. Moreover, the lack of semantic analysis capabilities hinders their application in higher-level building information modeling.

The introduction of machine learning marked a paradigm shift in the preservation of ancient architecture, transitioning from rule-based to data-driven approaches. These methods leverage techniques such as feature selection[37,38] and clustering algorithms[39,40] to achieve structured data processing, thereby minimizing the need for manual intervention. The primary strengths of machine learning lie in its high degree of automation and adaptive learning capabilities, enabling it to address diverse application scenarios. However, its performance heavily depends on data quality and the availability of annotated labels, leading to limitations in scenarios with scarce labeled data.

The rapid advancement of deep learning has ushered ancient architecture preservation into a new era of multidimensional data processing and complex scenario analysis. Techniques such as convolutional neural networks (CNNs)[41] and graph neural networks (GNNs)[42] enable the extraction of rich geometric and semantic features from data sources like images and point clouds, effectively overcoming the limitations of traditional methods and machine learning. For instance, Xiong et al. [43] proposed a novel method based on deep transfer learning by integrating ResNet50 and YOLO v2 networks, which was successfully applied to the detection of Hakka Walled Houses (HWHs). Zhou et al. [44] introduced the Mix Pooling Dynamic Graph Convolutional Neural Network (MP-DGCNN), which redefined edge features and incorporated an internal feature adjustment mechanism alongside a learnable mix pooling operation, enabling efficient learning of local graph features in point cloud topologies. Deep learning methods have demonstrated exceptional performance in tasks such as semantic segmentation, structural recognition, and defect repair by leveraging encoder-decoder architectures[30,45] and multi-scale feature aggregation[46]. Additionally, the incorporation of attention mechanisms like Transformers[47,48] has further enhanced the understanding of semantic relationships among complex architectural components. Compared to traditional methods and machine learning, deep learning significantly improves both accuracy and generalization capabilities, reduces dependency on predefined rules and manual intervention, and provides efficient solutions for large-scale ancient architectural data processing.

As the demands for preserving ancient architecture continue to evolve, single-source data processing methods have become inadequate for meeting the requirements of increasingly complex tasks. Li et al.[49] addressed the limitations of diffusion models in depicting ancient Chinese architecture by introducing a multimodal dataset. This dataset encompasses architectural styles from the Tang to Yuan dynasties, providing diverse data forms such as images, text, and videos, while pioneering the use of pinyin annotations for unique terminology, thereby filling a significant gap in the field. Duan et al. [50] proposed an innovative Multimodal Multi-task Restoration Model (MMRM) that integrates contextual and residual visual information, effectively restoring ancient ideographic scripts. The diversity, complexity, and large-scale nature of architectural data present challenges that single-modality approaches fail to address comprehensively. Consequently, integrating multimodal data for holistic analysis and processing has emerged as a pivotal direction in the research of ancient architectural preservation.

Establishing a high-precision, multimodal point cloud segmentation dataset based on real-world scenarios offers a viable solution to these issues. The main contributions of this paper are outlined in the following sections:

(1) A multi-modal point cloud segmentation dataset for ancient buildings in real-world scenarios, named RW-MAPCD, has been constructed. This dataset integrates rich multimodal data sources, such as point clouds, line drawings, color projections, and depth projections. Point cloud data provides fine spatial geometric information, line drawings illustrate the detailed features of the building structure, while color projections and depth projections supplement the color and depth information of the objects. This multi-level and multi-perspective data combination provides comprehensive contextual information for the model, enhancing its ability to understand and analyze the complex structures of ancient buildings. Figure 1 provides an overview of the RW-MAPCD.

(2) A cross-modal point cloud semantic segmentation network, MK3DNet, is proposed. Addressing the significant imbalance of different types of sample data in the ancient building dataset, the recognition and segmentation ability of complex structures is enhanced by combining 3D point cloud data with 2D image segmentation techniques.

(3) In the RW-MAPCD segmentation task, the performance of MK3DNet significantly surpasses traditional point cloud segmentation methods, showing marked improvements across multiple metrics. Notably, the results are particularly pronounced for components with irregular shapes and complex hierarchies.

The organization of this paper is structured as follows: the first section presents the background and current status of the digital preservation of ancient architecture, along with an analysis of the significance and challenges associated with point cloud segmentation in this domain; the second section reviews the related work on ancient building datasets and the applications of preservation efforts; the third section provides a detailed description of the construction process of RW-MAPCD, the network architecture of MK3DNet, and the evaluation metrics for image and point cloud segmentation tasks; the fourth section presents the experimental results and analysis; finally, the contributions of this paper are discussed and summarized, along with an overview of potential future research directions.

Methods

Data acquisition and preprocessing

The ancient buildings in RW-MAPCD are all located in Longyan City, Fujian Province, and cover multiple historical periods. This includes the Chaojing Public Ancestral Hall (PAH), Liang PAH, and Rongnan PAH prior to the Qing Dynasty, the Zhongxian Mansion (Ma) during the Qing Dynasty, and the Chen Ancestral Shrine (AS) during the Republic of China. A multi-source data fusion approach[21] was employed to acquire and process high-precision 3D point cloud data of the historical architecture, providing

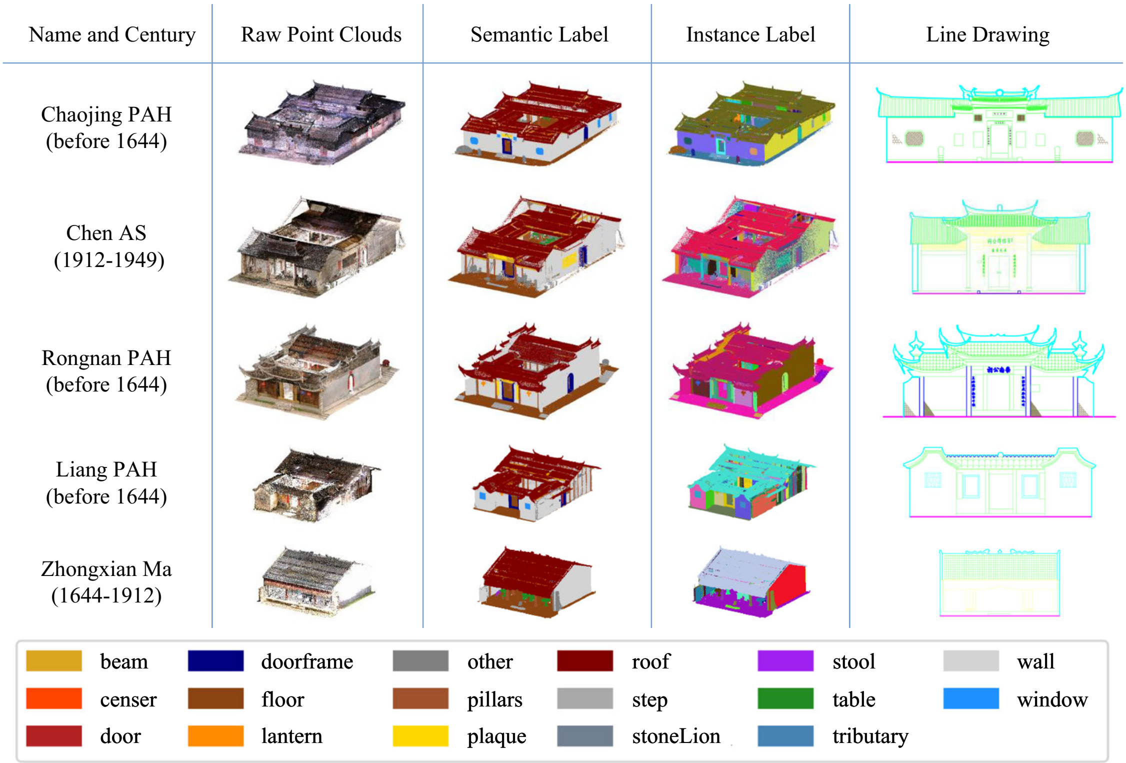
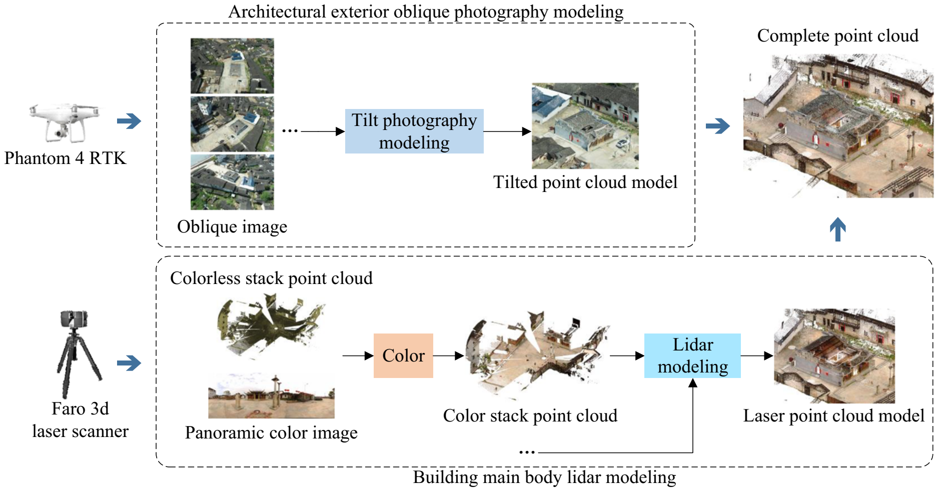


Fig. 1 | Overview of the RW-MAPCD.

Fig. 2 | Data preprocessing process.



significant support for digital preservation and analysis. The data acquisition process is illustrated in Fig. 2.

Firstly, an unmanned aerial vehicle (UAV) is used to acquire oblique imagery, capturing the external information of the structure, which is then processed through software to generate a high-resolution oblique point cloud model. In this process, the image resolution captured by the UAV is 1.5 cm, and the point cloud density is approximately 30 thousand points per square meter. Concurrently, a laser scanner is used to perform a comprehensive indoor and outdoor scan of the building structure, and the data is stitched together to obtain indoor and outdoor laser point cloud models. The laser scan has a point cloud density of 1 million points per square meter. Finally, the fusion and registration of oblique point clouds and laser point clouds are performed to generate a complete indoor and outdoor colored point cloud model of the historical architecture.

The preprocessing of the point cloud data from the historical architecture is a critical step in its digital representation, encompassing several important steps such as region clipping, point cloud filtering, point cloud down-sampling, data normalization, and coordinate system unification. Region clipping is performed to remove irrelevant portions, thereby reducing redundant data and ensuring the preservation of architectural details; point cloud filtering is used to remove noise and unnecessary points, reducing errors in the processed data. This process helps to improve the quality of the point cloud data, thereby enhancing the accuracy and computational efficiency of the model; uniform down-sampling is applied to decrease computational load and enhance recognition capability; centroid translation normalization is used to unify the scale, eliminating discrepancies between data sources; coordinate system unification is achieved through covariance matrix analysis to ensure consistency in subsequent processing.

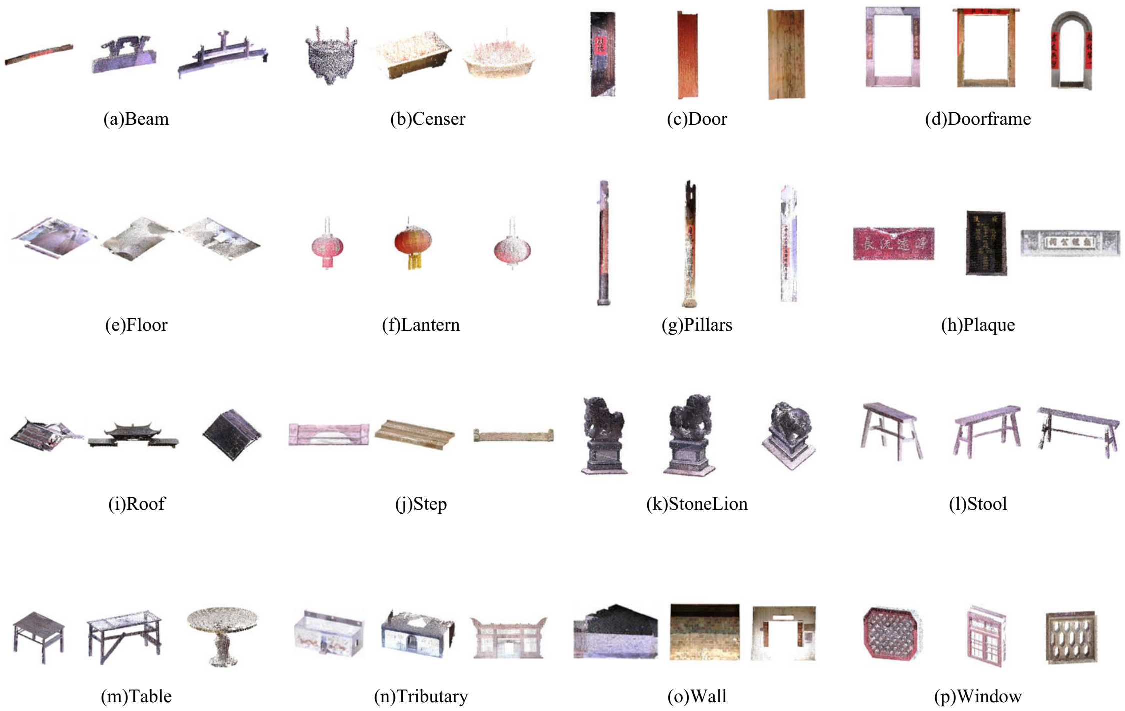


Fig. 3 | Ancient building point cloud category.

Data annotation

In this paper, five representative point cloud datasets of Chinese ancient buildings, including ancestral halls, temples, and ancestral homes, are manually labeled to reflect distinct architectural styles and structural features. These datasets provide diverse research samples for the comprehensive analysis and classification of typical elements in ancient buildings. Furthermore, architectural line drawings were generated based on the fused point cloud data, as illustrated on the right side of Fig. 1. These drawings offer precise and intuitive data support for the structural analysis of ancient buildings. They not only preserve the geometric features and detailed information of the buildings but also effectively visualize their core components and overall layout.

During the annotation process, the architectural elements were categorized into 17 distinct classes based on their structural functions and visual characteristics. Noise was assigned to the Other (or) category, which improved the robustness of the noise-trained network. The primary categories include Beam (bm), Censer (cn), Door (dr), Doorframe (df), Floor (fl), Lantern (lt), Other (or), Pillar (pl), Plaque (pq), Roof (rf), Step (sp), Stone Lion (sn), Stool (sl), Table (tb), Tributary (tr), Wall (wl), and Window (wd), as summarized in Fig. 3. The annotations for each category strictly adhere to relevant architectural standards[51,52], ensuring scientific rigor, standardization, and accuracy.

To qualitatively describe the morphological coverage of the data, we conducted a detailed analysis of the collected point cloud data. Each category of architectural elements is represented in fine detail through the point cloud data, which accurately reflects the structural details of the buildings and closely aligns with the segmentation results of the target elements. For instance, the roof is represented with high-density point cloud data that captures its curved shape and layers, while the pillars, due to their vertical structure, exhibit higher density and clarity in the point cloud data. These data not only faithfully restore the morphological features of ancient buildings but also ensure that the subsequent segmentation algorithms can effectively identify and process the boundaries and details of each element.

Construction of RW-MAPCD

The indoor and outdoor point cloud model of a single complete ancient building contains a large volume of data, which is unsuitable for deep learning model processing. Based on the high-precision point cloud data obtained and its detailed annotation information, the entire ancient building is segmented into regions and further decomposed into several scenes with specific semantics. Point cloud data for each scene is rotated and translated to generate projection images from 24 different perspectives, including color and depth projection images, along with semantic and instance labels. RW-MAPCD encompasses five ancient buildings, categorized into 17 classes and 52 scenes, comprising 627.24 million point clouds, 1248 color projection images, 1248 depth maps, and 20 line drawings. Figure 4 illustrates the construction process of the multimodal ancient building point cloud segmentation dataset.

Defining surrounding boxes. The column serves as the vertical support for the building, usually located at the four corners of the room or at critical internal positions. The space formed between the columns typically defines the fundamental boundary of the room. Beams are horizontal structural members that connect columns to create the ceiling structure of the room. The arrangement of beams plays a crucial role in determining the dimensions and configuration of the room. In ancient architecture, columns are usually arranged according to specific rules and proportions to create a “column grid”. The layout of the column grid determines how the rooms are divided. Each “grid” or “block” generally corresponds to a distinct room or area. The space between columns can be further subdivided by walls, screens, windows, among others, forming independent rooms or areas. A longitudinally-oriented room is formed between the front and rear columns, while a transversely-oriented room is formed between the left and right columns. This paper employs the clipping box mode of Trimble RealWorks, a software for point cloud processing, to generate a point cloud bounding box, as shown in the bounding box depicted in the center of Fig. 4. There are some repeated areas between the adjacent boundaries of each bounding box, to facilitate the acquisition of a complete point cloud scene.

Scene partitioning strategy. Based on the characteristics of the beam-column structure, ancient buildings can be categorized into various scenes, such as gate, foyer, corridor, lobby, room, and courtyard. These scenes not only represent the functional aspects of ancient architecture but also reflect the logic and aesthetic principles underlying its spatial

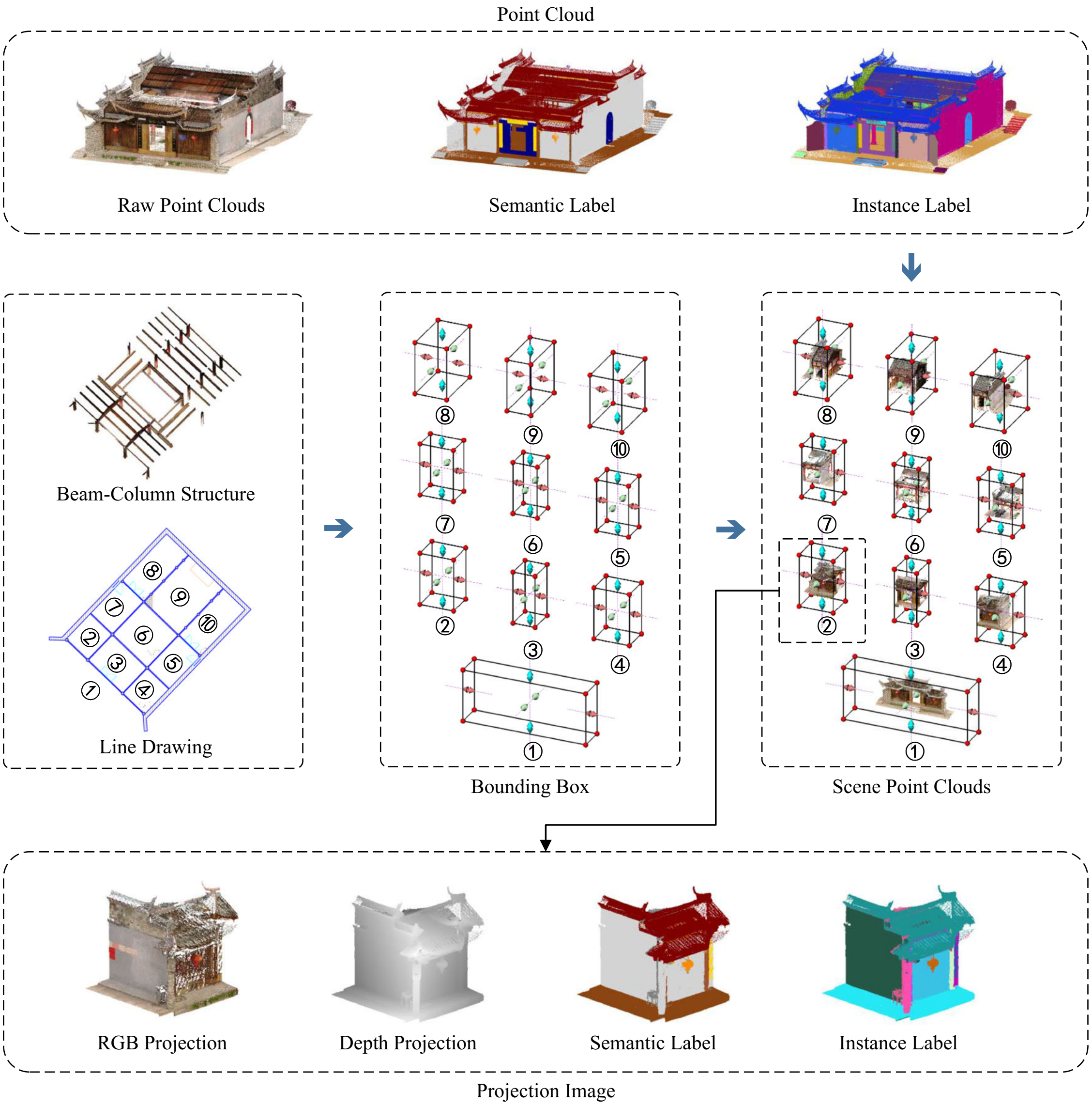


Fig. 4 | Construction process of RW-MAPCD.

organization. The scene classification of ancient buildings in RW-MAPCD is illustrated in Fig. 5. Specifically, 1 denotes the gate, 2 the foyer, 3 the corridor, 4 the courtyard, 5 the room, and 6 the lobby.

The gate, often situated at the building's entrance, serves as a symbolic architectural feature that embodies authority and sanctity. Comprising elements such as gate towers, pillars, and eaves, it requires detailed identification of its structural hierarchy and intricate details during digital segmentation and reconstruction. The foyer, a fundamental component of interior spaces, typically functions as the central area for activities in households, ancestral halls, or temples. Its beam-and-pillar framework is characterized by a symmetrical layout and notable spatial depth, which can be accurately reconstructed through the extraction and analysis of pillar and beam features. Corridors, functioning as connecting passages between different areas, are readily identifiable in point cloud segmentation due to their elongated, linear structure. The lobby, designed for significant ceremonies, showcases a grand architectural style with intricate multi-layered roofing and decorative carvings that necessitate focused attention during segmentation and reconstruction. Rooms, the most prevalent spaces for living or storage in ancient architecture, are distinguishable in point cloud segmentation by their enclosed walls and regular geometric configurations. The courtyard, serving as an open space within the building, acts as a central hub connecting various rooms and spaces. Its digital reconstruction emphasizes accurately mapping the spatial relationships between the courtyard and its surrounding structures. As depicted in Fig. 4, the architecture is divided into 10 labeled scenes (①-⑩), including one gate, three foyers, two corridors, one courtyard, two rooms, and one lobby.

The detailed data distribution of point clouds following the division of scenes is presented in Table 1. It illustrates the intricate details and spatial arrangements across the different scenes.

Coordinate transformation. The point cloud data P consists of the spatial coordinates and various attributes of each point, and can be represented by Eq. (1).

$$P = \{(x_i, y_i, z_i, r_i, g_i, b_i, s_i, i_i, idx_i) | i = 1, 2, \ldots, N\} \tag{1}$$

Here, (x_i, y_i, z_i) represents the three-dimensional coordinates, (r_i, g_i, b_i) denotes the color information, s_i is the corresponding semantic label, i_i represents the instance label, and idx_i refers to the point position index.

In order to capture the details of the point cloud more comprehensively, this method generates 24 different perspective projection images by rotating and translating the point cloud around the X, Y, and Z axes. The rotation operation in three-dimensional space is implemented using Euler angles and rotation matrices.

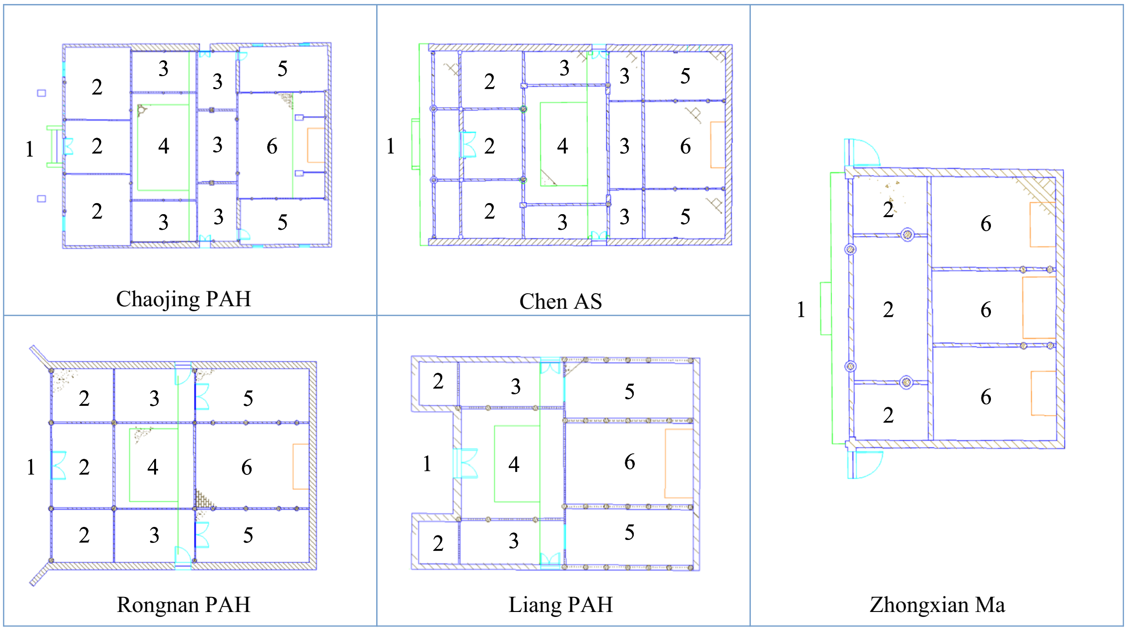


Fig. 5 | Point cloud scene delineation based on beam-column structure.

Table 1 | Detailed point cloud data distribution of ancient buildings after dividing the scene

Data source	Chaojing PAH		Chen AS		Rongnan PAH		Liang PAH		Zhongxian Ma	
Number	13 scenes		13 scenes		10 scenes		9 scenes		7 scenes	
Name	Points (million)	Percent	Points (million)	Percent	Points (million)	Percent	Points (million)	Percent	Points (million)	Percent
Beam	7.94	7.67%	9.02	10.04%	21.50	8.65%	13.25	10.02%	4.53	8.52%
Censer	0.02	0.02%	0.01	0.01%	0.19	0.08%	0.10	0.08%	0.16	0.31%
Door	0.99	0.96%	2.42	2.70%	8.45	3.40%	2.74	2.07%	0	0%
Doorframe	1.63	1.57%	2.82	3.13%	7.82	3.14%	4.63	3.50%	0	0%
Floor	20.33	19.65%	16.78	18.67%	47.32	19.04%	19.63	14.84%	10.65	20.06%
Lantern	0.57	0.55%	0	0%	0.05	0.02%	0.09	0.07%	0.27	0.51%
Other	7.92	7.66%	5.58	6.20%	8.45	3.40%	3.56	2.69%	1.55	2.91%
Pillars	5.71	5.52%	4.98	5.55%	10.53	4.24%	7.33	5.54%	1.89	3.57%
Plaque	0.26	0.25%	0.72	0.81%	0.41	0.17%	1.08	0.81%	1.24	2.34%
Roof	17.41	16.82%	21.42	23.83%	60.88	24.50%	32.900	24.88%	15.10	28.44%
Step	0.55	0.54%	0.11	0.13%	1.68	0.67%	0	0%	0.09	0.17%
StoneLion	0.10	0.09%	0	0%	0	0%	0	0%	0	0%
Stool	0.34	0.33%	0	0%	0.17	0.07%	0	0%	0.14	0.26%
Table	0.83	0.80%	1.33	1.47%	0.46	0.18%	1.59	1.21%	2.03	3.82%
Tributary	0.30	0.29%	0.50	0.55%	0.27	0.11%	0.45	0.34%	0.70	1.31%
Wall	37.61	36.34%	24.19	26.91%	80.34	32.33%	44.45	33.61%	14.76	27.79%
Window	0.97	0.94%	0	0%	0	0%	0.45	0.34%	0	0%
Total	103.48	100%	89.88	100%	248.52	100%	132.25	100%	53.11	100%
All points(million)										627.24

Given the Euler angles θ_x, θ_y and θ_z, the corresponding rotation matrix R can be expressed as Eq. (2).

$$R = R_x(\theta_x) \cdot R_y(\theta_y) \cdot R_z(\theta_z) \tag{2}$$

For the point $v = \left(x_i, y_i, z_i\right)$, transforming the point cloud from 3D space to the target perspective $v\prime = (x_i', y_i', z_i')$ can be expressed as Eq. (3).

$$v' = R \cdot v - \mathrm{T} \tag{3}$$

This transformation enables the point cloud to exhibit different geometric features across different views, which is crucial for extracting comprehensive information regarding the point cloud.

Multi-view projections. The projection module adopts an algorithm based on perspective projection and pixel selection[53]. The resolution of the projected image is set to 512 × 512, i.e., $H = W = 512$. The depth offset parameter b is set to 0.3, and the number of depth layers D is set to 112. The rotated point cloud is projected onto a three-dimensional grid, where

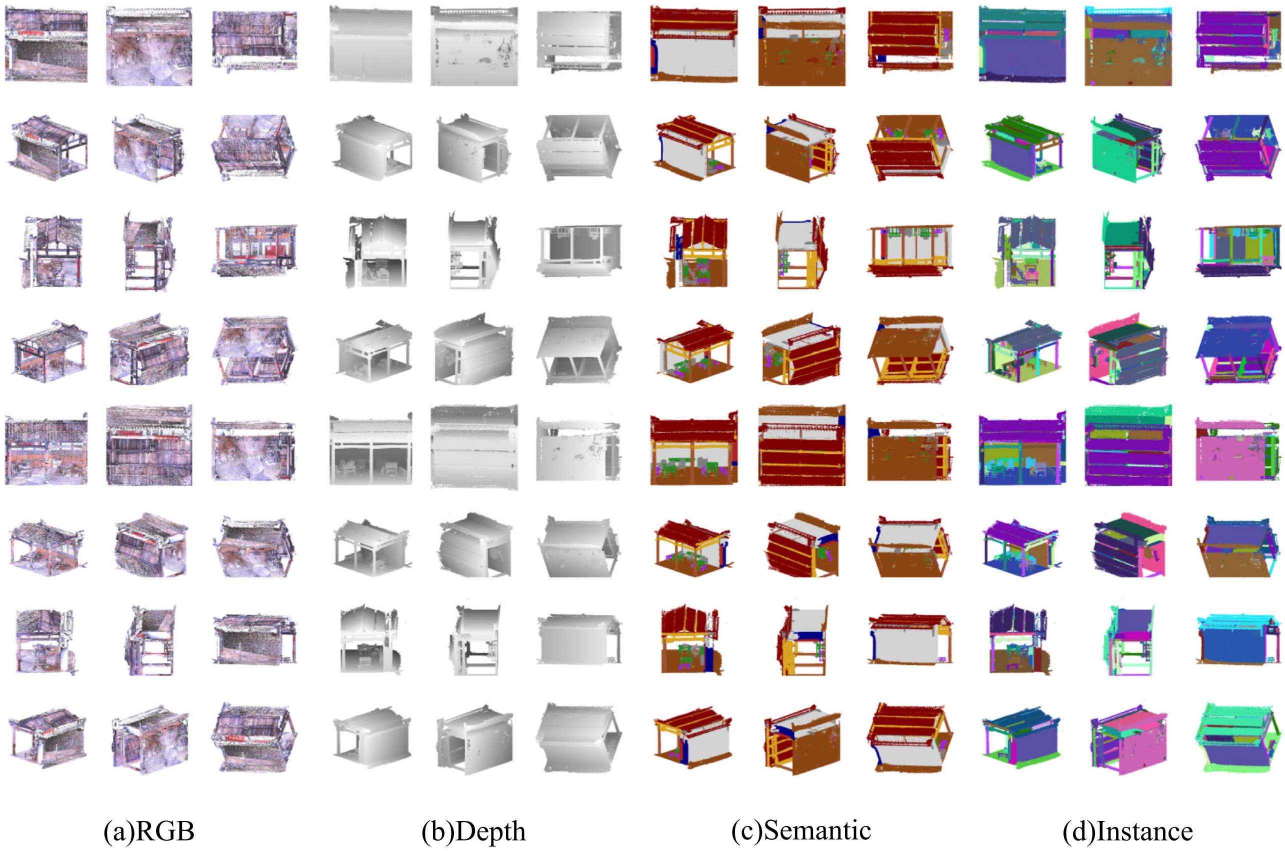


Fig. 6 | Data and labeling of multi-view projected images.

the grid coordinates are computed using Eq. (4) and Eq. (5), and the depth dimension index is determined using Eq. (6).

$$u_i = \left\lceil \frac{(x_i' + 1)}{2} \cdot H \right\rceil \tag{4}$$

$$v_i = \left\lceil \frac{(y_i' + 1)}{2} \cdot W \right\rceil \tag{5}$$

$$d_i = \left\lceil \frac{(z_i' + 1)/2 + b}{1 + b} \cdot (\mathrm{D} - 2) \right\rceil \tag{6}$$

Here, ⌈⌉ denotes the ceiling operation applied to round values upward.

For each pixel in a grid cell, the maximum depth value is selected, representing the surface closest to the observer, thereby generating a depth map, as expressed in Eq. (7).

$$\mathrm{I}_{\mathrm{depth}}(x_i, y_i) = \max_{d_i \in grid} d_i \tag{7}$$

To generate the color map, the index values from the depth map are used to retrieve the corresponding color information from the three-dimensional RGB grid, following the process defined in Eq. (8).

$$\mathrm{I}_{rgb}(x_i, y_i) = RGB\left(x_i, y_i, \mathrm{I}_{\mathrm{depth}}(x_i, y_i)\right) \tag{8}$$

Similarly, the semantic and instance information is extracted from semantic and instance grids using the index values provided by the depth map. The semantic map identifies the semantic category for each pixel, whereas the instance map indicates the object instance to which each pixel belongs.

Figure 6 provides an example of the multi-view projected images and the corresponding label data derived from the point cloud.

Storage point index. By executing the aforementioned calculations, we can store the corresponding point cloud indices for each pixel of the projected image within an index matrix. This matrix documents the relationship between the positions of each pixel in the projected image and the corresponding points in the point cloud. This approach ensures that an accurate mapping relationship is established between the pixels in each projected image and the specific points within the point clouds, thereby providing necessary indexing information for subsequent tasks such as semantic segmentation and label mapping.

MK3DNet

MK3DNet, the Cross-modal Point Cloud Semantic Segmentation Network, enhances the accuracy and efficiency of point cloud segmentation by integrating multi-view projection, point cloud indexing, image segmentation, remapping, and KNN[54] interpolation. Traditional point cloud segmentation models process three-dimensional (3D) data directly, which involves significant computational overhead and presents challenges when handling sparse data. In contrast, image segmentation models demonstrate superior performance in managing intricate details and complex scenes. To address the limitations of point cloud segmentation, this paper proposes projecting the point cloud into multi-view two-dimensional images while recording the correspondence between each projected image pixel and the point cloud data. By employing image segmentation techniques to generate segmentation results, these results are subsequently mapped back to the 3D point cloud according to the point indices, yielding sparse initial segmentation outcomes. Finally, the KNN algorithm is utilized to interpolate and complete the sparse segmentation results, resulting in comprehensive point cloud segmentation outcomes. This method significantly improves the accuracy and completeness of point cloud segmentation.

Figure 7 illustrates the MK3DNet network structure.

Image segmentation. In the projection image segmentation section, the Mask2Former[55] network is employed. Its masked attention mechanism dynamically focuses on the target regions within the image while disregarding irrelevant backgrounds, thereby enhancing both segmentation accuracy and efficiency. Furthermore, the Transformer decoder, by integrating query features, is capable of capturing the global relationships among targets, adapting effectively to complex scenes and irregularly-shaped targets. This makes it particularly well-suited for handling multi-scale and overlapping targets that may arise in projected images. The Masked Attention module is represented by Eq. (9):

$$X_l = softmax\left(M_{l-1} + Q_l K_l^T\right) V_l + X_{l-1} \tag{9}$$

Here, l refers to the index of the current layer. $X_l \in \mathbb{R}^{N \times C}$ represents the N query features, each with C dimensions, in the l-th layer. Meanwhile, $Q_l = f_Q(X_{l-1}) \in \mathbb{R}^{N \times C}$. $K_l, V_l \in \mathbb{R}^{H_l \times W_l \times C}$ denote the transformed image features obtained using $f_K(\cdot)$ and $f_V(\cdot)$, where H_l and W_l represent the spatial resolution of the image features. The value at position (x, y) in M_{l-1} is computed according to Eq. (10).

$$M_{l-1}(x, y) = \begin{cases} 0\ if M_{l-1}(x, y) = 1 \\ -\infty\ otherwise \end{cases} \tag{10}$$

In this context, $M_{l-1} \in \{0, 1\}^{N \times H_l W_l}$ represents the binarized output of the mask prediction, which has been resized by the preceding Transformer decoder layer.

The middle section of Fig. 7 illustrates the process of the Mask2-Former network in processing projection image data. The input image has dimensions of 512 × 512 × 3, and multi-scale features are extracted using the Swin Transformer. The feature dimensions are 16 × 16 × 768, 32 × 32 × 384, 64 × 64 × 192, and 128 × 128 × 96, respectively. These features are then forwarded to the Pixel Decoder, where the spatial dimensions of the output feature map remain unchanged, while the number of channels is uniformly transformed to 256. Subsequently, these processed features are fed into the Transformer Decoder for semantic refinement, which ultimately outputs the mask and category predictions for the semantic segmentation task.

Remapping. After image segmentation, each pixel is assigned a prediction result $\hat{I}_{sem}$. Utilizing the point index mapping matrix M, the segmentation results from the image are remapped back to the 3D point cloud data. For each mapping point i, its semantic label within the point cloud is determined by Eq. (11), thus generating a sparse initial point cloud segmentation result, where only the point clouds corresponding to the image pixels are assigned semantic labels.

$$\hat{P}_{sem}^{sparse}[M(i)] = \hat{I}_{sem} \tag{11}$$

Sparse completion. Due to the occlusion problem during the projection process, the segmentation results obtained through direct mapping are often sparse. To achieve complete segmentation results, this method incorporates the KNN algorithm, which is employed to interpolate and complete the sparse segmentation outcomes. For each unlabeled point i, interpolation is performed using the labels of its K nearest neighbors, where the average label of these neighbors is assigned as the semantic label for the current point. The formula is given by Eq. (12).

$$\hat{P}_{sem}^{complete}(i) = \frac{1}{|N(i)|} \sum_{j\epsilon N(i)} \hat{P}_{sem}^{sparse}(j) \tag{12}$$

Here, $N(i)$ represents the set of K nearest neighbor points of point i, $\hat{P}_{sem}^{sparse}(j)$ represents the semantic label of a neighboring point j, and $|N(i)|$ specifies the number of neighboring points, equal to K. By addressing label loss due to occlusion, this approach enhances the accuracy and ensures spatial consistency in point cloud segmentation results.

Evaluation metrics

To compare the performance of different models, this paper evaluates intersection over union (IoU) and mean intersection over union (mIoU) for both projection image segmentation and point cloud segmentation. IoU, defined as the ratio of the intersection to the union between predicted and ground truth segmentations, quantifies the degree of overlap. The specific formula is shown in Eq. (13).

$$\mathrm{Io}U = \frac{\mathrm{T}P}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

mIoU, the mean IoU across all categories, serves as a metric for overall segmentation accuracy. The specific formula is shown in Eq. (14).

$$\mathrm{mIoU} = \frac{1}{N} \sum_{\mathrm{i}=1}^{N} \mathrm{IoU_i} \tag{14}$$

In addition, this paper calculates the average category accuracy (aAcc) of image segmentation and the overall accuracy (OA) of point cloud segmentation. The aAcc reflects the model's segmentation performance across different categories by averaging the accuracy of each category. The specific formula is shown in Eq. (15).

$$aAcc = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{T}P_i}{\mathrm{TP_i} + FN_i} \tag{15}$$

OA, reflecting the model's overall performance, is calculated as the ratio of correctly predicted samples to the total number of samples. The specific formula is shown in Eq. (16).

$$OA = \frac{\mathrm{T}P + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{16}$$

Here, true positives (TP) and true negatives (TN) indicate the number of correctly classified positive and negative samples, respectively. False positives (FP) and false negatives (FN) represent the misclassified samples. N denotes the total number of categories to be classified.

Results

Experimental setup

The training and testing were conducted on a server equipped with a GeForce RTX 4070 SUPER, running on Ubuntu Linux 22.04.4 LTS. The code was developed in Python using the PyTorch framework and operates within an Anaconda virtual environment. This environment utilizes PyTorch 2.0.1, CUDA 11.7, and Python 3.8.17. A variety of data augmentation and training strategies were implemented to enhance the model's generalization ability and robustness. Specifically, the data augmentation strategies include random rotation, flipping, scaling, and translation to increase data diversity. Additionally, color adjustment, the addition of Gaussian noise, and random cropping were performed to improve the model's adaptability and robustness. Furthermore, training strategies such as learning rate scheduling, early stopping, batch normalization, and regularization were employed to enhance training stability and efficiency. Through these meticulous adjustments and optimizations, the aim is to achieve higher accuracy and better performance in complex tasks.

Dataset partitioning

The dataset was partitioned using a domain-specific strategy to ensure comprehensive training across diverse samples and robust generalization

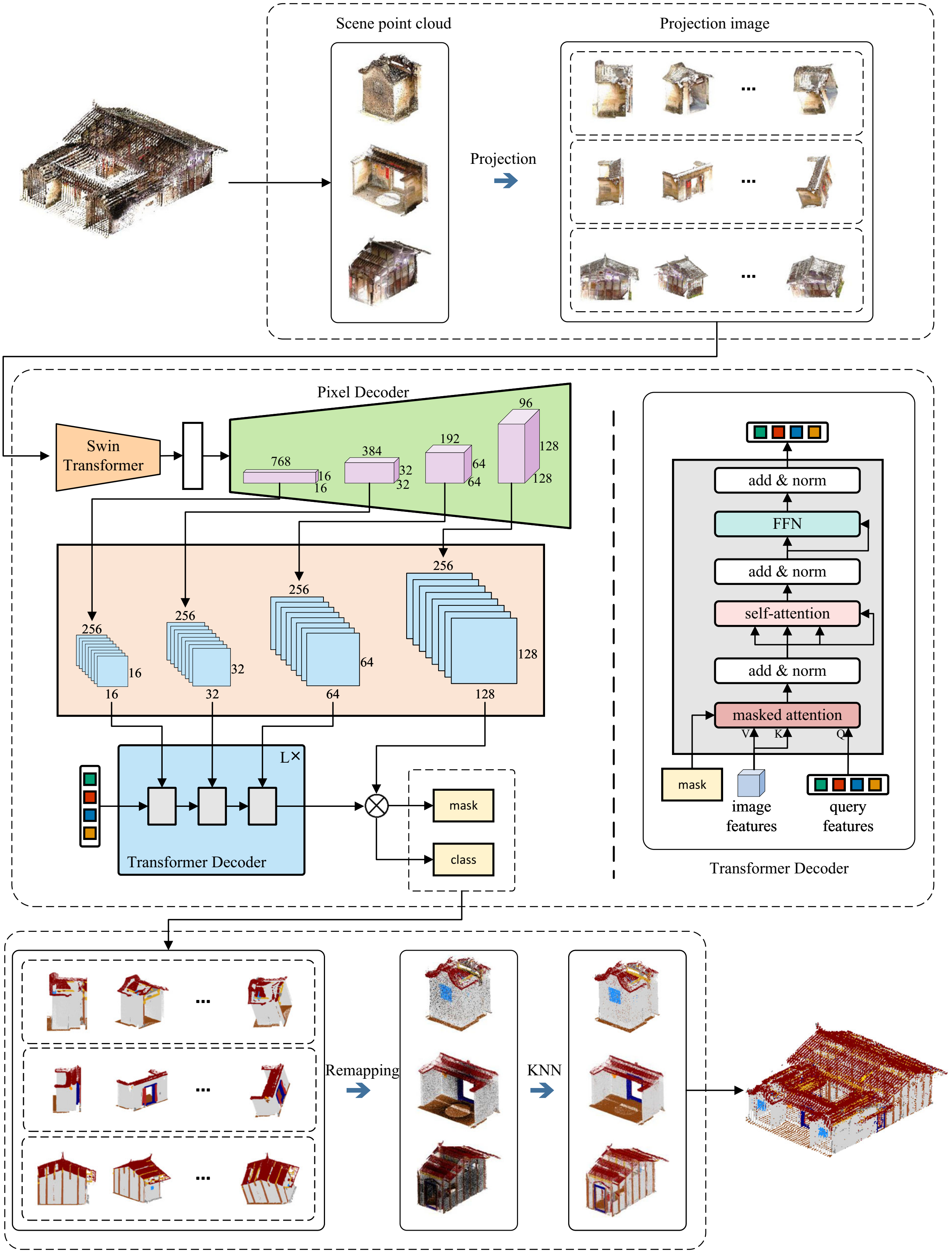


Fig. 7 | Network structure diagram of MK3DNet.

across varying environments. Representative ancient architectural datasets, including Chaojing PAH, Chen AS, Rongnan PAH, and Zhongxian Ma, were allocated to the training set, while Liang PAH was reserved for the test set. The training set spans 17 categories, whereas the test set contains 14, excluding the categories step, stoneLion, and stool.

This partitioning serves two primary purposes. First, incorporating diverse samples in the training set enhances the model's ability to capture generalizable features across ancient architectural structures. Second, the test set, featuring Liang Public Ancestral Hall with its distinct architectural style and intricate details, provides a rigorous benchmark to evaluate the model's robustness and precision in identifying specific architectural characteristics.

Results

In the projection image dataset of RW-MAPCD, several advanced image semantic segmentation algorithms have been employed for evaluation, including DeepLabV3, CCNet, FCN, DeepLabV3 + , GCNet, SegFormer, and Mask2Former. Table 2 presents the semantic segmentation results of different models.

The results demonstrated that Mask2Former exhibited the highest performance, achieving a mIoU of 74.31% and an aAcc of 91.30%. The model outperformed all other algorithms across all categories. This finding indicates that Mask2Former exhibits exceptional capability in handling diverse scenes and complex structures. In contrast, other models displayed comparatively weaker performance in certain categories.

Specifically, the IoU for DeepLabV3, CCNet, FCN, and DeepLabV3+ in the censer and plaque categories was nearly zero, highlighting the limitations of these models in addressing these specific classes. The superior performance of Mask2Former further underscores its adaptability and robustness in complex scenarios. Some of the network segmentation results are visualized in Fig. 8.

Point cloud segmentation represents a fundamental objective of this paper, aiming to achieve precise semantic segmentation of various architectural elements within 3D point cloud data. In the context of digital preservation of ancient structures, traditional point cloud segmentation models often encounter challenges such as segmentation inaccuracies and excessive computational costs when addressing large-scale and intricately shaped buildings. To assess the performance of contemporary point cloud segmentation models, this paper evaluates several classical segmentation models, which were trained and tested on the point cloud dataset in RW-MAPCD. Comparative experiments were conducted to analyze the performance of each model in segmenting different architectural features. The results of the semantic segmentation of the point cloud are presented in Table 3.

Among all models, MK3DNet (Ours) achieved the highest mIoU and OA, reaching 77.47% and 90.85%, respectively, and significantly outperformed other point cloud segmentation methods. Notably, in categories such as plaque and window, the IoU values for most models are close to zero. This indicates that MK3DNet demonstrates significant advantages in addressing these difficult-to-segment categories and effectively addresses the prevalent low IoU issue in existing methods. Although its performance in the floor, roof, and table categories is not as high as that of other models, with values of only 87.61%, 85.42%, and 90.00%, respectively, MK3DNet remains far ahead in most categories, compensating for its relative weaknesses in a few categories. Figure 9 presents the visualization results of projected image segmentation for some models.

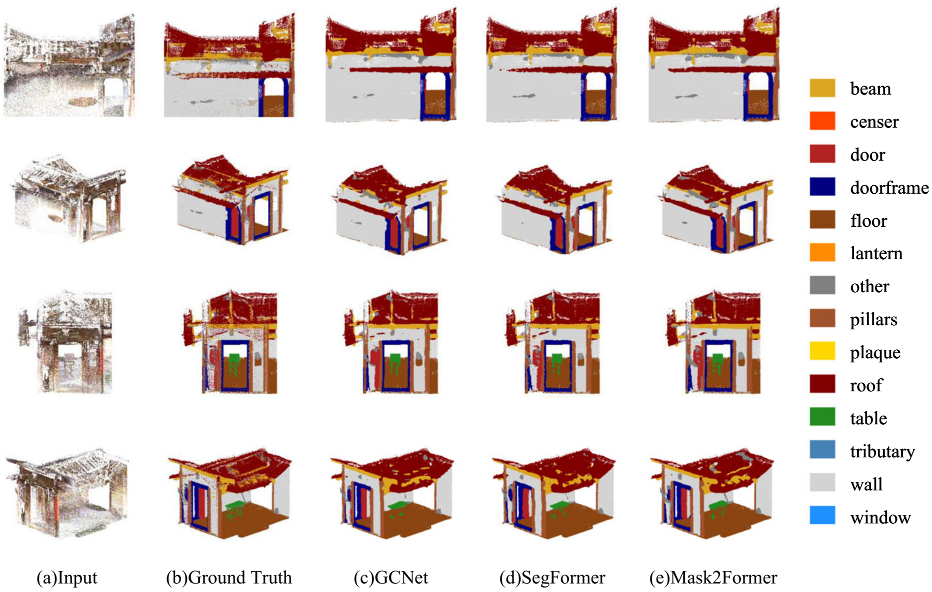
Analysis

In the point cloud segmentation task of ancient building components, MK3DNet demonstrates excellent performance and significant advantages. From Fig. 10, it can be observed that the model performs well across multiple categories, exhibiting robust generalization capabilities and effectively

Table 2 | Detailed semantic segmentation results for projected image datasets

Methods	mIou (%)	aAcc (%)	Per Class IoU(%)													
			bm	cn	dr	df	fl	lt	or	pl	pq	rf	tb	tr	wl	wd
DeepLabV3[56]	22.18	69.63	24.0	0	19.34	12.83	63.24	16.65	2.91	37.35	0	62.6	9.21	3.45	57.43	1.45
CCNet[57]	22.94	70.02	25.91	0	25.55	9.02	62.98	19.45	1.71	36.47	0	63.77	9.46	6.51	59.48	0.86
FCN[58]	23.63	71.6	26.01	1.23	24.34	13.35	65.86	18.69	2.61	38.76	0	64.64	9.6	4.23	60.67	0.78
DeepLabV3 + [59]	24.91	72.17	28.51	0	24.97	16.38	69.63	17.72	3.56	39.1	0	65.73	11.74	0.71	61.43	9.32
GCNet[60]	69.89	90.55	65.08	72.44	75.74	77.94	89.7	62.57	41.07	72.8	55.18	84.19	55.83	70.4	87.81	67.67
SegFormer[61]	70.88	90.71	65.51	71.71	75.30	77.44	89.67	64.45	40.46	73.46	63.83	84.60	57.48	71.40	88.1	68.9
Mask2Former[55]	74.31	91.30	68.59	79.80	79.44	81.59	90.10	68.86	45.84	76.63	69.54	85.11	61.91	73.63	88.73	70.60

Fig. 8 | Projected image segmentation results.



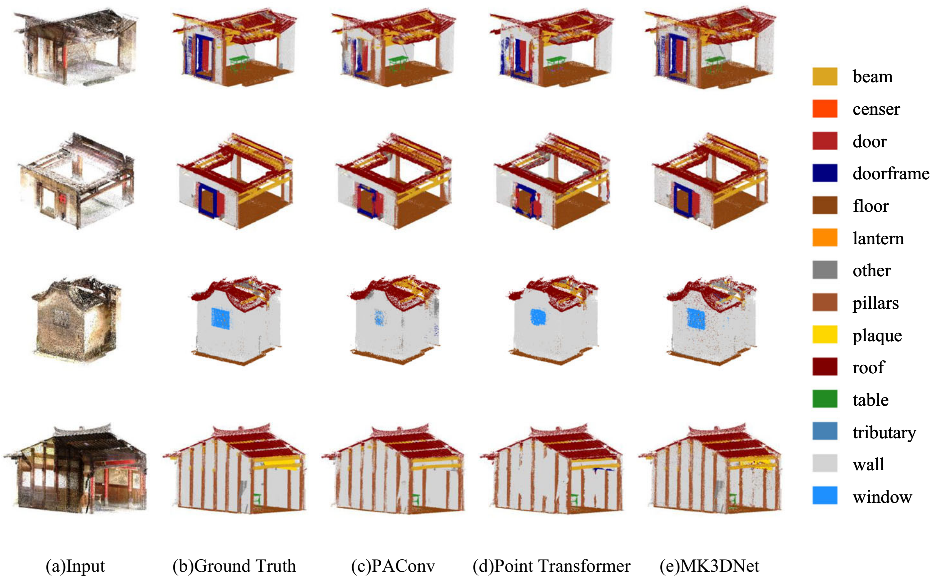
(a)Input (b)Ground Truth (c)GCNet (d)SegFormer (e)Mask2Former

Table 3 | Detailed semantic segmentation results for ancient architecture point cloud dataset

Methods	mIou (%)	OA (%)	Per Class IoU(%)													
			bm	cn	dr	df	fl	lt	or	pl	pq	rf	tb	tr	wl	wd
RandLA-Net[62]	38.65	76.08	50.17	0.86	11.89	14.55	94.22	25.19	6.76	30.47	0.00	82.81	82.32	74.47	67.34	0.00
PointNet + +[63]	44.11	83.60	69.81	0.12	13.74	10.91	96.02	66.82	8.16	55.33	0.00	90.53	98.36	31.31	76.18	0.31
DGCNN[64]	45.13	84.60	59.63	0.08	29.74	11.31	95.81	22.01	7.78	58.21	0.00	87.78	99.43	75.14	84.70	0.15
PVCNN[65]	52.22	82.79	59.84	54.35	32.64	32.20	96.51	69.39	7.95	48.17	0.00	87.27	98.16	63.85	80.74	0.00
PAConv[66]	55.30	87.86	73.35	7.35	48.32	37.35	**96.75**	70.61	9.89	62.02	0.01	**94.25**	**99.59**	77.51	83.55	13.66
Point Transformer[67]	55.53	84.98	60.09	57.72	42.44	49.34	92.14	8.67	16.00	54.55	0.00	86.49	91.28	74.04	82.21	62.42
MK3DNet	**77.47**	**90.85**	**75.39**	**79.60**	**83.09**	**83.38**	87.61	**81.05**	**44.40**	**75.49**	**66.86**	85.42	90.00	**79.33**	**88.08**	**64.85**

Bold represents the highest indicator value.

Fig. 9 | Point cloud segmentation results.



(a)Input (b)Ground Truth (c)PAConv (d)Point Transformer (e)MK3DNet

handling the variations in complex shapes, textures, and spatial structures. Notably, in the categories of beam, floor, roof, and wall, the model's performance is exceptionally remarkable, with the number of correctly classified points being 1346448, 2301264, 3497049, and 5792430, respectively. Moreover, for the censer and lantern categories with few samples, although the overall number of correctly classified points is relatively low, the model can still sustain a reasonable classification performance in the context of limited data, demonstrating its adaptability in few-shot scenarios.

According to the confusion matrix, precision, recall, and F1 score for each category can be calculated, as illustrated in Table 4. The table shows that the model exhibits high precision and recall across most categories, indicating its ability to accurately predict the target categories while effectively covering the actual target points. The precision of the floor category reaches 0.98, with a recall of 0.89, demonstrating both high classification accuracy and recall in segmentation tasks. In contrast, the roof category achieves a recall of 0.88 and a precision of 0.96, reflecting the model's outstanding performance on this key component and its suitability for handling complex large-scale building data.

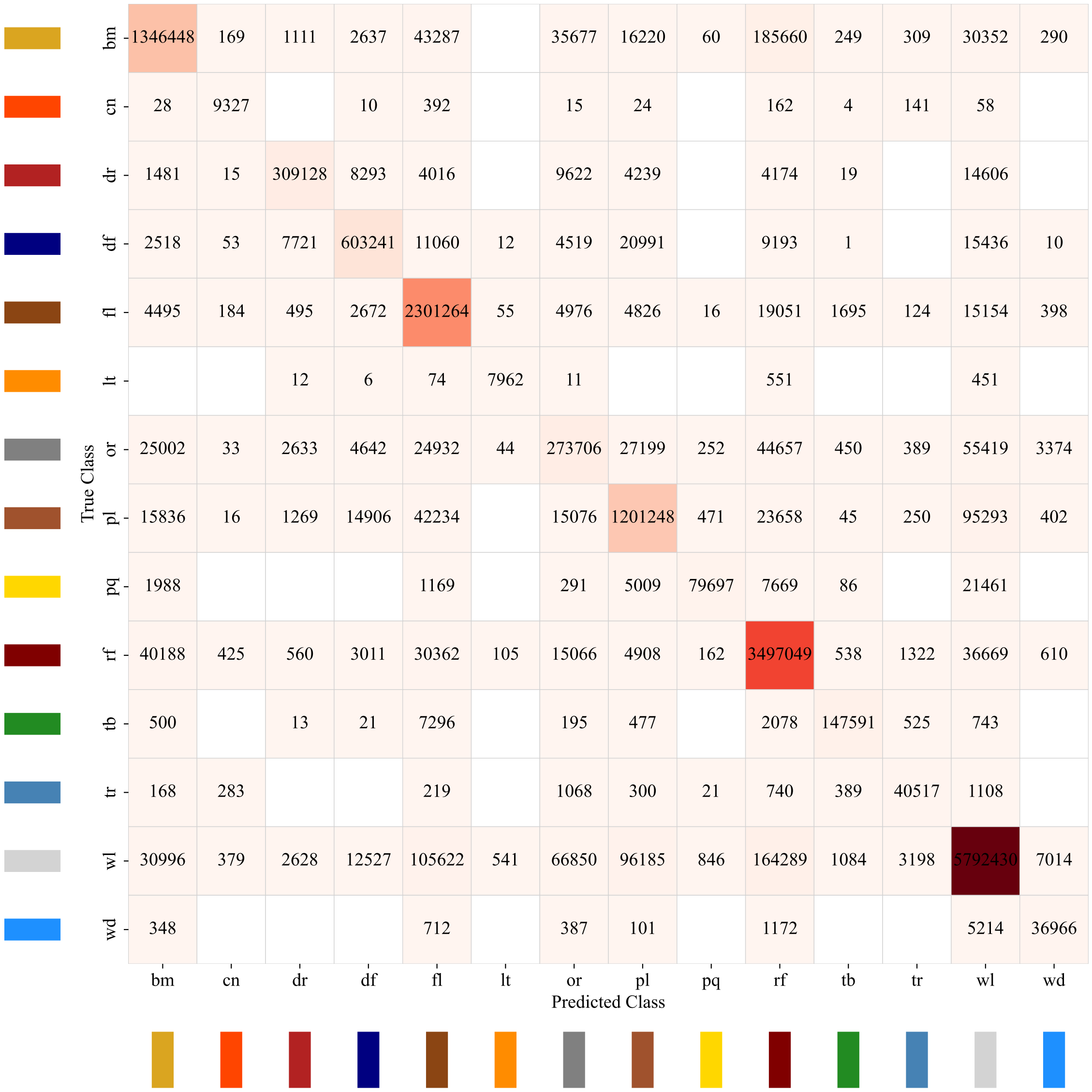
However, the performance of certain categories in the classification model is relatively low. For example, the precision of the 'other' category is 0.59, the recall is 0.64, and the F1 score is only 0.61. This may be attributed to the scattered distribution of samples within this category and the high diversity of features, leading to confusion when distinguishing the 'other' category from the other categories. Furthermore, although the recall of the plaque category is as high as 0.98, its precision is relatively low, at only 0.68, indicating a notable degree of misclassification within this category, which necessitates further optimization.

Overall, the model demonstrates excellent performance in the segmentation tasks of primary categories such as beam, floor, roof, and wall, with F1 scores exceeding 0.86. Notably, the F1 score for the wall category reaches 0.94, highlighting the model's efficiency and reliability for key components.

Figure 11 illustrates the IoU performance of different models across various categories. Overall, our model demonstrates outstanding performance in several categories, with IoU values exceeding 60%, and the curve reflects relatively smooth trends, indicating its stable performance in these categories. In contrast, the IoU values of other models in certain categories, such as censer, other, plaque, and window, are notably low, with some reaching as low as 0%, indicating significant recognition difficulties and instability for these models in these categories. Although Point Transformer performs well in some categories, such as floor, roof, and table, it exhibits poor performance in censer, door, doorframe, and plaque categories, failing to achieve stable IoU above 60%.

Discussion

The existing ancient building datasets and network models are primarily based on single-modal features[43,44], which usually focus on the separate processing of 3D point clouds and 2D images[18,19]. This paper explores the application potential of multimodal data and advanced segmentation models in improving the semantic segmentation accuracy of ancient



True Class \ Predicted Class	bm	cn	dr	df	fl	lt	or	pl	pq	rf	tb	tr	wl	wd
bm	1346448	169	1111	2637	43287		35677	16220	60	185660	249	309	30352	290
cn	28	9327		10	392		15	24		162	4	141	58	
dr	1481	15	309128	8293	4016		9622	4239		4174	19		14606	
df	2518	53	7721	603241	11060	12	4519	20991		9193	1		15436	10
fl	4495	184	495	2672	2301264	55	4976	4826	16	19051	1695	124	15154	398
lt			12	6	74	7962	11			551			451	
or	25002	33	2633	4642	24932	44	273706	27199	252	44657	450	389	55419	3374
pl	15836	16	1269	14906	42234		15076	1201248	471	23658	45	250	95293	402
pq	1988				1169		291	5009	79697	7669	86		21461	
rf	40188	425	560	3011	30362	105	15066	4908	162	3497049	538	1322	36669	610
tb	500		13	21	7296		195	477		2078	147591	525	743	
tr	168	283			219		1068	300	21	740	389	40517	1108	
wl	30996	379	2628	12527	105622	541	66850	96185	846	164289	1084	3198	5792430	7014
wd	348				712		387	101		1172			5214	36966

Fig. 10 | Segmentation confusion matrix for MK3DNet.

Table 4 | Summary of segmentation performance indicators by category

Category	bm	cn	dr	df	fl	lt	or	pl	pq	rf	tb	tr	wl	wd
Precision	0.81	0.92	0.87	0.89	**0.98**	0.88	0.59	0.85	0.68	0.96	0.93	0.90	0.92	0.82
Recall	0.92	0.86	0.95	0.93	0.89	0.91	0.64	0.87	**0.98**	0.88	0.97	0.87	0.95	0.75
F1-Score	0.86	0.89	0.91	0.91	0.93	0.90	0.61	0.86	0.80	0.92	0.95	0.88	**0.94**	0.79

Bold marks the top Precision, Recall, and F1-Score for each category.

building point cloud data and proposes solutions to address the shortcomings of traditional point cloud segmentation methods in capturing complex architectural details.

However, it is important to recognize that digital surveys in the architectural field are often affected by factors such as shadow cones, noise, and reflections, which may occlude or distort the surveyed 3D data. These factors can impact the completeness and accuracy of the point clouds, which in turn affects the performance of the segmentation model. Shadow cones may cause occlusions in poorly illuminated areas, while reflections or noise from nearby surfaces can lead to inaccurate point cloud data. These challenges can result in misrepresentations of architectural elements, particularly those with intricate or subtle details, which are crucial for accurate segmentation. In this context, the multimodal features provided by the RW-MAPCD dataset represent a significant advancement in this field. The proposed MK3DNet segmentation method integrates multiple modalities, which helps reduce the impact of such interferences.

In the RW-MAPCD, there exists a significant imbalance in the distribution of samples across different categories, which critically affects the performance of the single-modal point cloud segmentation model. As illustrated in Table 5, the floor and roof categories constitute 18.11% and 23.32% of the dataset, respectively, and possess distinct geometric features that facilitate the network's ability to learn the segmentation patterns for these categories effectively; thus, they exhibit superior performance in other networks. Although the category of table accounts for only 0.99% of the

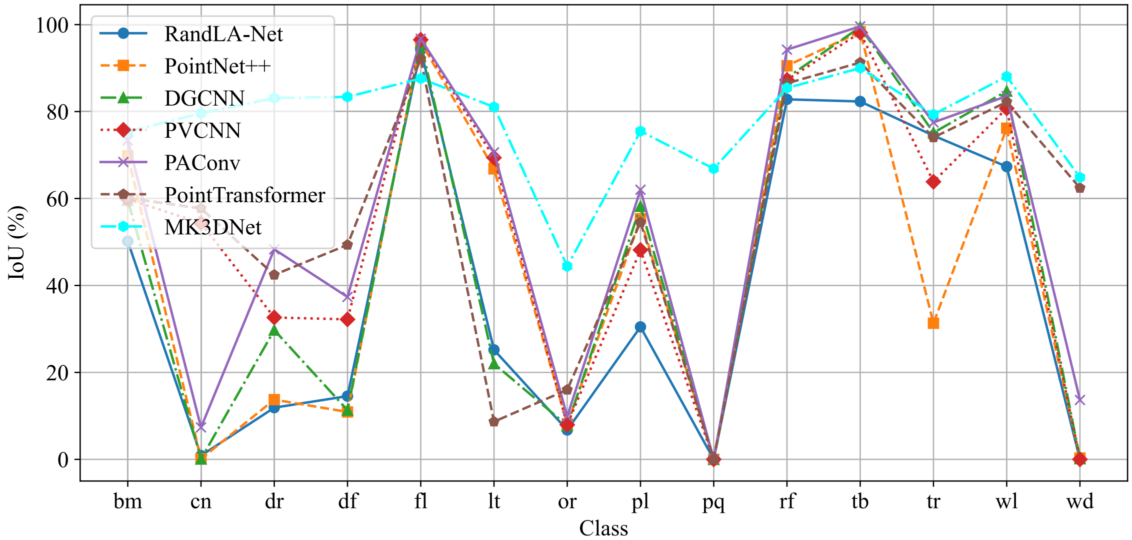


Fig. 11 | Comparison of the effect of point cloud semantic segmentation network.

Table 5 | Sample distribution and segmentation IoU relationships

	bm	cn	dr	df	fl	lt	or	pl	pq	rf	tb	tr	wl	wd
Points Sum (million)	56.24	0.48	14.60	16.90	**114.71**	0.98	27.06	30.44	3.71	**147.71**	**6.24**	2.22	201.35	1.42
Percent Sum(%)	8.87	0.08	2.30	2.67	**18.11**	0.15	4.27	4.81	0.59	**23.32**	**0.99**	0.35	31.77	0.22
PointNet + +	69.81	0.12	13.74	10.91	**96.02**	66.82	8.16	55.33	0.00	**90.53**	**98.36**	31.31	76.18	0.31
Point Transformer	60.09	57.72	42.44	49.34	**92.14**	8.67	16.00	54.55	0.00	**86.49**	**91.28**	74.04	82.21	62.42
MK3DNet	75.39	79.60	83.09	83.38	**87.61**	81.05	44.40	75.49	66.86	**85.42**	**90.00**	79.33	88.08	64.85

dataset, it is relatively common in most point cloud datasets, allowing it to demonstrate satisfactory performance in other networks as well. Despite the substantial sample imbalance observed in the ancient building data, the proposed method exhibits superior performance compared to other existing point cloud segmentation networks.

These research findings align with the perspectives presented in the literature[49,50] regarding the role of multimodal learning in the protection of cultural heritage, indicating that rich contextual information can significantly enhance segmentation accuracy. The outcomes of this study are promising; however, some limitations still exist. First, the coverage of the RW-MAPCD dataset needs to be broadened to accommodate a wider variety of architectural styles and diverse environmental contexts. Secondly, the reliance of MK3DNet on KNN for segmentation and completion introduces significant computational complexity. Future investigations could focus on optimizing computational efficiency and exploring more effective methodologies. Furthermore, while the current work primarily utilizes color projection and point cloud data, the potential value of other modalities, such as depth maps and line drawings, within the dataset has not been thoroughly explored. Future research should delve deeper into the contributions of these modalities, particularly concerning the identification and classification of complex structural elements, to enhance the overall model's expressiveness and applicability.

In terms of application for conservation and digital heritage services, the proposed approach opens up new possibilities for accurately capturing and documenting heritage sites. With the ability to segment intricate architectural elements, the model can support the creation of highly detailed 3D models for use in preservation projects. These models can assist in the assessment of structural integrity, guide restoration efforts, and enable virtual tours for educational or cultural engagement purposes. Additionally, digital archives can be developed that allow future generations to explore and study these buildings without compromising their physical preservation. The study's findings lay a strong foundation for future advances in the application of multimodal segmentation models to support the preservation of cultural heritage in both physical and digital forms.

Data availability

The dataset generated and analyzed in the current research is available from the corresponding author upon request.

Code availability

The code used in the current research is available from the corresponding author upon request.

Received: 13 December 2024; Accepted: 10 April 2025;
Published online: 03 May 2025

References

1. Liu, J. & Wu, Z. K. Rule-based generation of ancient Chinese architecture from the song dynasty. *J. Comput. Cultural Herit. (JOCCH)* **9**, 1–22, https://doi.org/10.1145/2835495 (2015).
2. Hu, Q. et al. Fine surveying and 3D modeling approach for wooden ancient architecture via multiple laser scanner integration. *Remote Sens.* **8**, 270, https://doi.org/10.3390/rs8040270 (2016).
3. Biryukova, M. V. & Nikonova, A. A. The role of digital technologies in the preservation of cultural heritage. *Muzeol.ógia a kult.úrne dedičstvo* **5**, 1 (2017).
4. Adane, A., Chekole, A. & Gedamu, G. Cultural heritage digitization: Challenges and opportunities. *Int. J. Computer Appl.* **178**, 1–5, https://doi.org/10.5120/ijca2019919180 (2019).
5. Hu, Y. et al. Measurement and analysis of facial features of terracotta warriors based on high-precision 3D point clouds. *Herit. Sci.* **10**, 40, https://doi.org/10.1186/s40494-022-00662-0 (2022).
6. Lague, D., Brodu, N. & Leroux, J. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei

canyon (N-Z). *ISPRS J. Photogramm. Remote Sens.* **82**, 10–26, https://doi.org/10.1016/j.isprsjprs.2013.04.009 (2013).

7. Guo, Y. et al. Deep learning for 3D point clouds: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4338–4364, https://doi.org/10.1109/TPAMI.2020.3005434 (2020).
8. Haznedar, B., Bayraktar, R., Ozturk, A. E. & Arayici, Y. Implementing PointNet for point cloud segmentation in the heritage context. *Herit. Sci.* **11**, 2, https://doi.org/10.1186/s40494-022-00844-w (2023).
9. Armeni, I. et al. 3d semantic parsing of large-scale indoor spaces. Proceedings of the IEEE conference on computer vision and pattern recognition;1534-1543. https://doi.org/10.1109/CVPR.2016.170. (2016)
10. Hackel, T. et al. Semantic3D.net: A new Large-scale Point Cloud Classification Benchmark. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. https://doi.org/10.5194/isprs-annals-IV-1-W1-91-2017 (2017).
11. Chen, M. et al. An Aerial Photogrammetry Benchmark Dataset for Point Cloud Segmentation and Style Translation. *Remote Sens*. **16**, 22 (2024).
12. Djelliout, T. & Aliane, H. Building and evaluation of an Algerian Cultural Heritage dataset using convolutional neural networks. 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS);1-7. https://doi.org/10.1109/PAIS56586.2022.9946666 (2022).
13. Swaileh, W. et al. Versailles-FP Dataset: Wall Detection in Ancient Floor Plans. Springer, Cham. https://doi.org/10.1007/978-3-030-86549-8_3 (2021).
14. Barz, B. & Denzler, J. Wikichurches: A fine-grained dataset of architectural styles with real-world challenges. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). https://doi.org/10.5281/zenodo.5166986 (2021).
15. Robot Vision Group. National Laboratory of Pattern Recogntion. Institute of Automation, Chinese Academy of Sciences. 3D Reconstruction Dataset. http://vision.ia.ac.cn/data.
16. Gabara, G. & Sawicki, P. CRBeDaSet: A benchmark dataset for high accuracy close range 3D object reconstruction. *Remote Sens.* **15**, 1116, https://doi.org/10.3390/rs15041116 (2023).
17. Matrone, F. et al. A benchmark for large-scale heritage point cloud semantic segmentation. *Int. Arch. Photogrammetr. Remote Sens. Spatial Inform. Sci*. **43**, 1419–1426, https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1419-2020 (2020).
18. Yang, B. S., Han, X. & Dong, Z. Point cloud benchmark dataset WHU-TLS and WHU-MLS for deep learning. *Natl Remote Sens. Bull.* **25**, 231–240, https://doi.org/10.11834/jrs.20210542 (2021).
19. Zhou, C., Dong, Y. & Hou, M. DGPCD: a benchmark for typical official-style Dougong in ancient Chinese wooden architecture. *Herit. Sci.* **12**, 201, https://doi.org/10.1186/s40494-024-01293-3 (2024).
20. Grilli, E. & Remondino, F. Machine learning generalisation across different 3D architectural heritage. *ISPRS Int. J. Geo-Inf.* **9**, 379, https://doi.org/10.3390/ijgi9060379 (2020).
21. Gao, X. et al. Ancient Chinese architecture 3D preservation by merging ground and aerial point clouds. *ISPRS J. Photogramm. Remote Sens.* **143**, 72–84, https://doi.org/10.1016/j.isprsjprs.2018.04.023 (2018).
22. Liu, Y., Gao, W. & Hu, Z. A large-scale dataset for indoor visual localization with high-precision ground truth. *Int. J. Robot. Res.* **41**, 129–135, https://doi.org/10.1177/02783649211052064 (2022).
23. Walczak J. InLUT3D: Challenging real indoor dataset for point cloud analysis. arXiv preprint. 2024; arXiv:2408.03338. https://doi.org/10.48550/arXiv.2408.03338.
24. Hossain, M. et al. Building Indoor Point Cloud Datasets with Object Annotation for Public Safety. SMARTGREENS; 45–56. https://doi.org/10.5220/0010454400450056 (2021).
25. Ibrahim, M., Akhtar, N., Wise, M. & Mian, A. Annotation tool and urban dataset for 3d point cloud semantic segmentation. *IEEE Access* **9**, 35984–35996, https://doi.org/10.1109/ACCESS.2021.3062547 (2021).
26. Wang R., Huang S., Yang H. Building3d: A. urban-scale dataset and benchmarks for learning roof structures from point clouds. Proceedings of the IEEE/CVF International Conference on Computer Vision; 20076-20086. https://doi.org/10.1109/ICCV51070.2023.01837 (2023).
27. Krapf, S., Mayer, K. & Fischer, M. Points for energy renovation (PointER): A point cloud dataset of a million buildings linked to energy features. *Sci. Data* **10**, 639, https://doi.org/10.1038/s41597-023-02544-x (2023).
28. Pierdicca, R., Mameli, M., Malinverni, E. S., Paolanti, M. & Frontoni, E. Automatic generation of point cloud synthetic dataset for historical building representation. Augmented Reality, Virtual Reality, and Computer Graphics: 6th International Conference, AVR 2019, Santa Maria al Bagno, Italy, June 24–27, 2019, Proceedings, Part I 6. Springer International Publishing; 203–219. https://doi.org/10.1007/978-3-030-25965-5_16 (2019).
29. Battini, C. et al. Automatic generation of synthetic heritage point clouds: Analysis and segmentation based on shape grammar for historical vaults. *J. Cultural Herit.* **66**, 37–47, https://doi.org/10.1016/j.culher.2023.10.003 (2024).
30. Morbidoni, C., Pierdicca, R., Paolanti, M., Quattrini, R. & Mammoli, R. Learning from synthetic point cloud data for historical buildings semantic segmentation. *J. Comput. Cultural Herit. (JOCCH)* **13**, 1–16, https://doi.org/10.1145/340926 (2020).
31. Albano, R. Investigation on roof segmentation for 3D building reconstruction from aerial LIDAR point clouds. *Appl. Sci.* **9**, 4674, https://doi.org/10.3390/app9214674 (2019).
32. Hamid-Lakzaeian, F. Structural-based point cloud segmentation of highly ornate building façades for computational modelling. *Autom. Constr.* **108**, 102892, https://doi.org/10.1016/j.autcon.2019.102892 (2019).
33. Salamanca, S., Merchán, P., Espacio, A., Pérez, E. & Merchán, M. J. Segmentation of 3D Point Clouds of Heritage Buildings using Edge Detection and Supervoxel-Based Topology. *Sens. (Basel)* **24**, 4390, https://doi.org/10.3390/s24134390 (2024).
34. Zolanvari, S. M. I., Laefer, D. F. & Natanzi, A. S. Three-dimensional building façade segmentation and opening area detection from point clouds. *ISPRS J. Photogramm. Remote Sens.* **143**, 134–149, https://doi.org/10.1016/j.isprsjprs.2018.04.004 (2018).
35. Presti, N. L. et al. Streamlining FE and BIM Modeling for Historic Buildings with Point Cloud Transformation. International Journal of Architectural Heritage; 1-14. https://doi.org/10.1080/15583058.2024.2330942 (2024).
36. Tao, W., Xiao, Y., Wang, R., Lu, T. & Xu, S. A Fast Registration Method for Building Point Clouds Obtained by Terrestrial Laser Scanner via 2D Feature Points. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. https://doi.org/10.1109/JSTARS.2024.3392927 (2024).
37. Dong, Y., Hou, M., Xu, B., Li, Y. & Ji, Y. Ming and Qing Dynasty Official-Style Architecture Roof Types Classification Based on the 3D Point Cloud. *ISPRS Int. J. Geo-Inf.* **10**, 650, https://doi.org/10.3390/ijgi10100650 (2021).
38. Croce, V. et al. From the semantic point cloud to heritage-building information modeling: A semiautomatic approach exploiting machine learning. *Remote Sens.* **13**, 461, https://doi.org/10.3390/rs13030461 (2021).
39. Stober, D. & Raguz-Lucic, N. Trends, Problems, and Solutions from Point Cloud via Non-Uniform Rational Basis Spline to Building Information Modelling: Bibliometric and Systematic Study. *Buildings* **14**, 564, https://doi.org/10.3390/buildings14030564 (2024).
40. Wang, H., Shi, Y., Yuan, Q. & Li, M. Crack Detection and Feature Extraction of Heritage Buildings via Point Clouds: A Case Study of Zhonghua Gate Castle in Nanjing. *Buildings* **14**, 2278, https://doi.org/10.3390/buildings14082278 (2024).

41. Zou, Z., Zhao, P. & Zhao, X. Automatic segmentation, inpainting, and classification of defective patterns on ancient architecture using multiple deep learning algorithms. *Struct. Control Health Monit.* **28**, e2742, https://doi.org/10.1002/stc.2742 (2021).
42. Cao, Y. & Scaioni, M. 3DLEB-Net: Label-efficient deep learning-based semantic segmentation of building point clouds at LoD3 level. *Appl. Sci.* **11**, 8996, https://doi.org/10.3390/app11198996 (2021).
43. Xiong, Y., Chen, Q., Zhu, M., Zhang, Y. & Huang, K. Accurate detection of historical buildings using aerial photographs and deep transfer learning. IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium;1592-1595. https://doi.org/10.1109/IGARSS39084.2020.9323541 (2020).
44. Zhou, C., Dong, Y., Hou, M., Ji, Y. & Wen, C. MP-DGCNN for the semantic segmentation of Chinese ancient building point clouds. *Herit. Sci.* **12**, 249, https://doi.org/10.1186/s40494-024-01289-z (2024).
45. Zhao, J. et al. DSC-Net: learning discriminative spatial contextual features for semantic segmentation of large-scale ancient architecture point clouds. *Herit. Sci.* **12**, 274, https://doi.org/10.1186/s40494-024-01367-2 (2024).
46. Zhang, R. et al. MSFA-Net: A Multiscale Feature Aggregation Network for Semantic Segmentation of Historical Building Point Clouds. *Buildings* **14**, 1285, https://doi.org/10.3390/buildings14051285 (2024).
47. Zhao, J. et al. Semantic segmentation of point clouds of ancient buildings based on weak supervision. *Herit. Sci.* **12**, 232, https://doi.org/10.1186/s40494-024-01353-8 (2024).
48. Xiao, C., Chen, Y., Sun, C., You, L. & Li, R. AM-ESRGAN: Super-Resolution Reconstruction of Ancient Murals Based on Attention Mechanism and Multi-Level Residual Network. *Electronics* **13**, 3142, https://doi.org/10.3390/electronics13163142 (2024).
49. Li, B. et al. Building a Chinese ancient architecture multimodal dataset combining image, annotation and style-model. *Sci. Data* **11**, 1137, https://doi.org/10.1038/s41597-024-03946-1 (2024).
50. Duan, S., Wang, J. & Su, Q. Restoring Ancient Ideograph: A Multimodal Multitask Neural Network Approach. arXiv preprint. 2024; arXiv:2403.06682. https://doi.org/10.48550/arXiv.2403.06682.
51. Guo, Q. Yingzao Fashi: Twelfth-century chinese building manual. *Architectural Hist.* **41**, 1–13, https://doi.org/10.2307/1568644 (1998).
52. Li, S. Reconstituting Chinese building tradition: the Yingzao fashi in the early twentieth centur. *J. Soc. Architectural Historians* **62**, 470–489, https://doi.org/10.2307/3592498 (2003).
53. Zhu, X. et al. PointCLIP V2: Prompting CLIP and GPT for Powerful 3D Open-world Learning. Proceedings of the IEEE/CVF International Conference on Computer Vision;2639-2650. http://arxiv.org/abs/2211.11682 (2023).
54. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27, https://doi.org/10.1109/TIT.1967.1053964 (1967).
55. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R. Masked-attention mask transformer for universal image segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition;1290-1299. http://arxiv.org/abs/2112.01527 (2022).
56. Chen, L. C., Papandreou, G., Schroff, F. & Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. CoRR abs/1706.05587 (2017).
57. Huang, Z. et al. Ccnet: Criss-cross attention for semantic segmentation.Proceedings of the IEEE/CVF international conference on computer vision;03-612. https://doi.org/10.48550/arXiv.1811.11721 (2019).
58. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation.oceedings of the IEEE conference on computer vision and pattern recognition;431-3440. https://doi.org/10.1109/CVPR.2015.7298965 (2015).
59. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV);801-818. http://arxiv.org/abs/1802.02611 (2018).
60. Lin, X., Guo, Y. & Wang, J. Global Correlation Network: End-to-End Joint Multi-Object Detection and Tracking. CoRR abs/2103.12511 (2021).
61. Xie, E. et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **34**, 12077–12090 (2021).
62. Hu, Q. et al. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR);11108–11117. https://doi.org/10.1109/CVPR42600.2020.01112 (2020).
63. Qi, C. R., Yi, L., Su, H. & Guibas, L. J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems;30. http://arxiv.org/abs/1706.02413 (2017).
64. Wang, L., Huang, Y., Hou, Y., Zhang, S. & Shan, J. Graph attention convolution for point cloud semantic segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition;10296–10305. https://doi.org/10.1109/CVPR.2019.01054 (2019).
65. Liu, Z., Tang, H., Lin, Y. & Han, S. Point-voxel cnn for efficient 3d deep learning. Advances in neural information processing systems;32 (2019).
66. Xu, M., Ding, R., Zhao, H. & Qi, X. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds.Proceedings of the IEEE/CVF conference on computer vision and pattern recognition;3173-3182. https://doi.org/10.1109/CVPR46437.2021.00319 (2021).
67. Engel, N., Belagiannis, V. & Dietmayer, K. Point Transformer. *IEEE Access* **9**, 134826–134840, https://doi.org/10.1109/ACCESS.2021.3116304 (2021).

Acknowledgements

This study was supported by the Natural Science Foundation of Fujian Province, China (Grant no. 2023J011432), the Natural Science Foundation of Xiamen, China (Grant No.3502Z202373063), the National Natural Science Foundation of China (Grant No. 42401422), the Construction Science and Technology Research and Development Project of Fujian Province, China (Grant no. 2022-K-85,2023-K-27), the Social Science Plan Project of Fujian Province, China (Grant no. FJ2024BF071), and the Longyan City Industry University Research Joint Innovation Project (2022LYF18001).

Author contributions

Conceptualization, H.L. and Z.X.; Methodology, Z.X. and Y.H.; Formal analysis, Y.S.; writing—original draft preparation, H.L.; writing—review and editing, H.L., Z.X. and Y.H.; supervision, H.L.; project administration, P.Y., J.A., and L.Z.; funding acquisition, Z.X., Y.H., P.Y., and L.Z. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Hanxin Liu or Yuanrong He.