Zhao *et al. Heritage Science* (2024) 12:274
https://doi.org/10.1186/s40494-024-01367-2

Heritage Science

RESEARCH

Open Access

DSC-Net: learning discriminative spatial contextual features for semantic segmentation of large-scale ancient architecture point clouds

Jianghong Zhao[1,2], Rui Liu[1,2], Xinnan Hua[1,2*], Haiquan Yu[1,2], Jifu Zhao[1,2], Xin Wang[1,2] and Jia Yang[1,2]

Abstract

Semantic segmentation of point cloud data of architectural cultural heritage is of significant importance for HBIM modeling, disease extraction and analysis, and heritage restoration research fields. In the semantic segmentation task of architectural point cloud data, especially for the protection and analysis of architectural cultural heritage, the previous deep learning methods have poor segmentation effects due to the complexity and unevenness of the data, the high geometric feature similarity between different components, and the large scale changes. To this end, this paper proposes a novel encoder-decoder architecture called DSC-Net. It consists of an encoder-decoder structure based on point random sampling and several fully connected layers for semantic segmentation. To overcome the loss of key features caused by random downsampling, DSC-Net has developed two new feature aggregation schemes: the enhanced dual attention pooling module and the global context feature module, to learn discriminative features for the challenging scenes mentioned above. The former fully considers the topology and semantic similarity of neighboring points, generating attention features that can distinguish categories with similar structures. The latter uses spatial location and neighboring volume ratio to provide an overall view of different types of architectural scenes, helping the network understand the spatial relationships and hierarchical structures between different architectural elements. The proposed modules can be easily embedded into various network architectures for point cloud semantic segmentation. We conducted experiments on multiple datasets, including the ancient architecture dataset, the ArCH architectural cultural heritage dataset, and the publicly available architectural segmentation dataset S3DIS. The results show that the mIoU reached 63.56%, 55.84%, and 71.03% respectively. The experimental results prove that our method has the best segmentation effect in dealing with challenging architectural cultural heritage data and also demonstrates its practicality in a wider range of architectural point cloud segmentation applications.

Keywords Ancient architecture, Architectural cultural heritage, Semantic segmentation of point clouds, Attention feature aggregation

Introduction

As an important heritage of human civilization, ancient architecture carries rich historical and cultural information. It is entirely different in structure from modern architecture, being composed of thousands of wooden components like columns, beams, rafters, and tiles assembled in a specific order [1]. Due to the complexity of ancient architectural structures, two-dimensional images face challenges such as single viewing angles, occlusions, and lighting issues, and are insufficient in

*Correspondence:
Xinnan Hua
3051123638@qq.com
[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
[2] Key Laboratory for Architectural Heritage Fine Reconstruction & Health Monitoring, Beijing 100044, China

displaying concave-convex, three-dimensional structures, and decorative details. In contrast, three-dimensional models can more intuitively represent the shape, structure, and construction methods of ancient architectures, supporting digital twin and graphic space interaction applications [2]. With the continuous development of three-dimensional laser scanning technology, its application in the conservation of ancient architectural artifacts is increasing, especially as advancements in deep learning technology in point cloud semantic segmentation have moved beyond traditional methods—which often rely on manually designed features and rules. These methods struggle to adapt to the complex and variable structures of ancient architecture and are inefficient in processing large-scale point cloud data; moreover, machine learning methods based on manual features rely too heavily on feature descriptors, are not suitable for large and complex scenes, and have poor generalization capabilities [3]. End-to-end deep learning methods can assign semantic labels to every point in a scene while balancing algorithm accuracy and complexity well, offering new insights for architectural cultural heritage point cloud segmentation. However, the large volume of point cloud data and its irregular, unstructured, and unordered nature make it difficult to quickly learn discriminative features of large-scale point cloud objects and perform accurate segmentation.

In recent years, many neural network-based methods for semantic segmentation of 3D point clouds have been proposed, mainly divided into three types: projection-based, voxel-based, and point-based methods. When processing large-scale point clouds, projection and voxel-based methods not only increase computational overhead but also require additional operations, such as converting point clouds into other representations and reprojecting intermediate segmentation results back into the point cloud. Unlike these, point-based methods can directly and end-to-end process point cloud data, especially suitable for ancient architectures with complex geometric structures and shapes. This method can flexibly handle various irregular shapes and different resolutions of point cloud data, better preserving the original data information, such as location, color, and normals. Some previous methods have high computational and memory requirements when dealing with large-scale point cloud data such as ancient architectures, and are not suitable for real-time processing of ancient architecture scene data. Recently, a large-scale point cloud semantic segmentation method RandLA-Net [4] has been proposed, known for its efficient downsampling method, which enables it to process point cloud data for large-scale scenes such as ancient architectures. This is particularly important for efficiently solving the segmentation task of cloud data for large-scale scenic spots such as ancient architectures. However, architectural point cloud scenes have complex geometric structures, diverse features such as materials, textures, shapes, and sizes. Especially in traditional Chinese architecture, due to the large and dense number of point clouds in each scene, wooden materials are often used for doors, windows, and columns, while stone materials are used for footings and stone steps. Efficiently distinguishing similar structures between the above categories is more challenging. Therefore, while reducing computational and memory costs, it is also necessary to address the loss of important topological and semantic information in complex geometric structures in ancient architectural scenes, and to learn discriminative features for challenging complex ancient architectural scenes. Inspired by the successful use of attention mechanisms [5–7], and contextual information [8, 9] in many semantic segmentation tasks, we propose the following three questions and propose solutions:

(1) How to efficiently learn highly discriminative local feature aggregation methods from large-scale ancient architecture point cloud data?
(2) How to accurately understand the overall shape and long-distance dependency relationship between different categories of architectural cultural heritage through learning global contextual semantic information?
(3) How to ensure accurate semantic segmentation in large-scale point cloud data of different building types, structures, and complexities?

In response to the above three issues, this article proposes a large-scale point cloud semantic segmentation network architecture for ancient architecture, which consists of a symmetric encoder decoder structure with skip connections. To effectively distinguish the geometric similarity between different categories and comprehensively capture category characteristics, we designed an enhanced dual attention pooling module and a global contextual semantic feature module. The former focuses on the similarity of geometry and appearance, and is applied to each module in the encoder stage to perceive the topological and semantic differences of similar points. The latter learns the global context of each 3D point cloud by utilizing neighborhood position and volume ratio, thereby achieving an understanding of the spatial layout and interrelationships of the entire building scene. The DSC Net we propose can be integrated into various network architectures to handle point cloud

semantic segmentation tasks. Our main contributions are as follows:

(1) We conducted comprehensive experiments and evaluations on our self built ancient architecture dataset, architectural cultural heritage dataset ArCH [10], and public dataset S3DIS [11] on a fully supervised task. The semantic segmentation results demonstrated the robustness and superiority of our method in different styles of architectural scenes. In particular, our method comprehensively considers the complex and similar geometric structures but different appearance categories in various scenes of ancient architecture, providing strong support for the digital analysis and protection of architectural cultural heritage.
(2) We have developed an enhanced dual attention pooling (EDAP) module that can capture more complex and refined local feature information and distinguish the geometric and appearance similarity of adjacent points. This module can be inserted into the feature aggregation of the encoder decoder stage to explore new point cloud segmentation networks.
(3) We have introduced the Global Context Feature (GCF) module, which specifically analyzes the global context of each 3D point from point cloud data. By integrating global context, this module focuses on learning global information from 3D points, which can more effectively handle large-scale spatial changes and complex ancient architecture clusters, thereby improving performance in segmentation tasks.

Related work

In this section, we will thoroughly review point cloud semantic segmentation methods based on deep learning, which can generally be categorized into three types: projection-based methods, voxel-based methods, and point-based methods.

Projection based methods

Inspired by 2D convolutional neural networks, existing work [12] describe a method that involves projecting point clouds onto a two-dimensional plane and then using traditional 2D image segmentation algorithms to process them. Subsequently, the segmentation results are mapped back to the original three-dimensional space to achieve semantic segmentation of the point cloud. In related methods, Tatarchenko [13] project the local surface geometry surrounding each point onto a tangent plane, creating tangent images that can be processed with 2D convolutions. However, these multi-view segmentation methods inevitably introduce a loss of detail due to the projection step, therefore not fully utilizing the underlying geometric and structural information. Despite these methods being able to leverage mature 2D image processing technologies, the projection process inevitably leads to a loss of detail information, and remapping the 2D segmentation results back to three-dimensional space incurs significant computational overhead.

Voxel based methods

Voxel-based point cloud semantic segmentation methods primarily involve converting three-dimensional point cloud data into a voxel format. Specifically, the point cloud is organized into a 3D grid structure of small cubic units, and deep learning techniques are used to predict semantic labels for each voxel, achieving semantic segmentation of the overall point cloud. Given the sparsity of point cloud data and its substantial resource consumption, researchers have proposed various sparse convolution techniques to reduce computational costs [14,15]. Furthermore, to enhance processing performance, researchers have also introduced technologies such as octrees and hash maps [16] to improve efficiency. Due to the high computational demand and significant memory consumption involved in 3D convolutions, research based on this technology has declined in recent years.

Point based methods

Point based methods directly target point clouds for end-to-end operations, which can be divided into the following categories: point convolutional methods, multi-layer perceptron (MLP), graph based methods, RNN based methods, and attention mechanism based methods.

Based on point convolution method

Inspired by the success of convolutional operators in the two-dimensional image domain, several studies have proposed convolutional methods for three-dimensional point clouds [17–20]. These methods mainly extend the traditional concept of image convolution to unordered point cloud data, effectively processing point cloud data through local neighborhood modeling and feature extraction.

The method of multi-layer perceptron (MLP)

The per-point MLP method typically employs shared MLPs as the basic unit. The pioneering PointNet [21], by applying a symmetric function to handle the disorderliness of point clouds, uses MLPs to extract features from each point, followed by a max pooling operation to extract global features of the point clouds across various dimensions. However, PointNet has limitations

in extracting local features. To address this, PointNet++[22] introduced a multi-level feature extraction structure, which effectively enhanced the extraction capability of local and global features. However, it faces issues of excessive computational resource consumption when processing large-scale point clouds. Hu Q [4] proposed an efficient and lightweight network for semantic segmentation of large-scale point clouds, RandLA-Net, which significantly reduces the memory and computational overhead for large-scale point cloud processing by using random point sampling techniques. Based on these findings, researchers proposed a PointNet-based deep learning point cloud segmentation workflow for architectural cultural heritage. Bulent H [23] evaluated the application of the deep learning model PointNet in the segmentation of point clouds of heritage buildings in Gaziantep, Turkey. By analyzing the point cloud data of 28 buildings, it was found that PointNet performs with high accuracy when handling synthetic data, providing a new method for precise classification and segmentation of heritage buildings. To address the problem of low segmentation accuracy caused by the complexity of training scenarios, Literature [24] selected four types of objects: arcs, columns, walls, and windows. Researchers trained the network using annotated point cloud data from field surveys and used the PointNet++method for segmentation, to assess the impact of training data variability on performance.

Graph based and RNN based methods

Graph convolution-based methods extract features by utilizing the topological structure and connections of point cloud data. In contrast, RNN-based methods combine the feature extraction capabilities of CNNs with the temporal information processing ability of RNNs to capture the spatial and temporal correlations in point cloud data, enabling semantic label prediction for each point. DGCNN introduced an EdgeConv module, which generates edge features describing the relationships between a point and its neighbors. RSNet [25] developed a lightweight local dependency module that uses slice pooling layers to transform unordered point cloud features into ordered feature vector sequences. Liu [26] proposed a new method called 3DCNN-DQN-RNN, which integrates three-dimensional convolutional neural networks (CNNs), deep Q networks (DQN), and residual recurrent neural networks (RNNs). Through an "eye window" mechanism, this method effectively locates and segments target class points. However, the complexity of model computation and excessive computational overhead cannot be ignored. 3D CNN and residual RNN together extract robust and discriminative features within the eye window, thus improving the parsing accuracy of point clouds. This method automates the mapping of raw data to classification results, integrating target localization, segmentation, and classification into one. Christian [27] developed and trained an improved DGCNN, the Rad-DGCNN network model, using synthetic point cloud data. This model performed well in real TLS point cloud segmentation, although it still has shortcomings in handling segmentation tasks of similar categories. Literature [28] developed an improved dynamic graph convolutional neural network that uses edge attention convolution technology to reinforce the learning of local features. With the 3DMAX model trained on sampled points, this network can effectively extract the roof structures of ancient architectures from real point cloud data. Roberto Pierdicca [29] and others proposed an improved version of the dynamic graph convolutional neural network (DGCNN), which, by integrating key features such as normals and colors, enhanced the processing capabilities for the newly collected digital cultural heritage dataset ArCH. Francesca Matrone et al. [30] compared the application of machine learning and deep learning in large-scale cultural relic classification, analyzed the advantages and disadvantages of these two technologies, and developed a semantic segmentation architecture DGCNN Mod+3Dfeature that integrates the advantages of these two methods. However, these methods have not fully evaluated the diversity and applicability across different types of datasets and may lead to excessive consumption of network computational resources and low computational efficiency.

Method based on attention mechanism

Point cloud semantic segmentation methods based on attention mechanisms enhance segmentation performance by dynamically adjusting weights to increase focus on key information, considering the relationship between each point's local information and the global context. Yang [31] developed a Point Attention Transformer to simulate interactions between points. Literature [32] introduced a local spatial awareness layer designed to learn spatial distribution weights to capture local geometric structures. Literature [33] built on the structure of 3D Unet [34], designing modules for global feature learning and multi-scale feature fusion. This approach also introduced a sparse tensor-based implementation to reduce unnecessary computations and adapt to the sparsity of 3D point clouds. These methods, by refining the weight adjustments between points, have significantly enhanced the recognition and utilization of key features, greatly improving segmentation performance.

Methodology

In this section, we provide a detailed introduction to a novel semantic segmentation network: DSC Net, which has developed two core modules including an enhanced dual attention pooling module (EDAP) and a global context feature module (GCF). The Enhanced Dual Attention Pooling Module (EDAP) utilizes topology and appearance semantic information, integrates multi-level features, and dynamically adjusts feature weights during the pooling process, effectively improving the network's sensitivity and discriminative ability to local details. The design of this module enables the network to adaptively enhance key features, suppress unimportant information, and more accurately segment the complex and fine similar geometric structures of ancient architectures, distinguishing adjacent point geometric and appearance differences caused by materials and weathering. The Global Context Feature Module (GCF) is responsible for capturing and integrating global information in point cloud data. By analyzing the distribution and structural characteristics of the overall point cloud, this module helps the network grasp the overall semantic context. Not only does it perform well in local areas, but it can also perform effective feature learning and semantic parsing at the global level, further enhancing the model's adaptability and accuracy to complex structures. In addition, the ancient architectural complex spans a large area of space and exhibits significant scale changes. The Global Context Feature Module (GCF) module can effectively handle large-scale spatial changes, and by learning global information, the model can still maintain efficient segmentation performance when facing structures of different scales.

DSC module

We have designed the DSC module to learn discriminative spatial feature aggregation. This section will provide a detailed introduction to the two modules, the Enhanced Dual Attention Pooling Module (EDAP) and the Global Context Feature Module (GCF), and will specifically describe the architecture of the DSC module.

Enhanced dual attention pooling module

This section introduces a method for local feature aggregation called the Enhanced Dual Attention Pooling Module (EDAP), which is designed to differentiate categories that have similar geometric shapes but different appearance structures. Detailed explanations related to this are shown in Fig. 1. The input includes N point clouds, each consisting of three-dimensional coordinates $p_i \in P^{N\times 3}$ and corresponding appearance features $f_i \in P^{N\times d}$. For each point cloud in the set, we use a K-NN algorithm based on Euclidean distance to aggregate its neighboring point set $P_j=\left\{P_j^1,P_j^2,P_j^3,\ldots,P_j^k\right\}$, and obtain the corresponding appearance featuresf_j. The formula for calculating the aggregation of local features to distinguish points with similar properties is defined as follows (Eq. 3–1):

$$F = S\left(F\left(\left[p_i,p_j,f_i,f_j\right]\right)\right) \tag{3-1}$$

Here, S represents a symmetric reduction function, while F is our designed feature aggregation function, which includes per-point multilayer perceptrons (MLP), adaptive weight adjustments, and max pooling operations. The "[]" represents a series of operations covering various manipulations of p_i, p_j, f_i, f_j including raw operations, arithmetic operations (such as addition and subtraction), and data concatenation. Specifically, this enhanced dual attention local feature aggregation method adopts a strategy based on adaptive weights and multi-level feature fusion. This strategy aims to automatically adjust and optimize the weights during the training process to better adapt to the data features and model objectives. This method can fully exploit topological and appearance features, capturing details from coarse to fine levels, enabling the network to dynamically adjust its focus on feature levels, thereby effectively extracting semantic information of different structures and complexities. Below is a detailed explanation of our proposed enhanced dual attention:

(1) Coordinate Position Encoding: Position encoding plays a crucial role in networks based on Transformers and self-attention. For example, in the field of 2D images, the relative position of 2D coordinates is often used for position encoding to enhance image features [35]. However, in 3D space, the absolute coordinates of points may not be suitable for the network to extract high-level features, as the network tends to focus on the relative positions and centroids of points. For N input point clouds, the coordinate encoding of each point can be represented by the centroid coordinates, neighboring point coordinates, relative coordinates, and relative distances. These are processed through a shared multilayer perceptron (MLP) to obtain the encoded features P, which have the same dimensions as the features f_i.

$$P = MLP\left(p_i \oplus p_j \oplus (p_i - p_j) \oplus \left\|p_i,p_j\right\|\right) \tag{3-2}$$

|| || represents the Euclidean distance between two points, and⊕ represents the concatenation operation, which doubles the dimensions after concatenation.

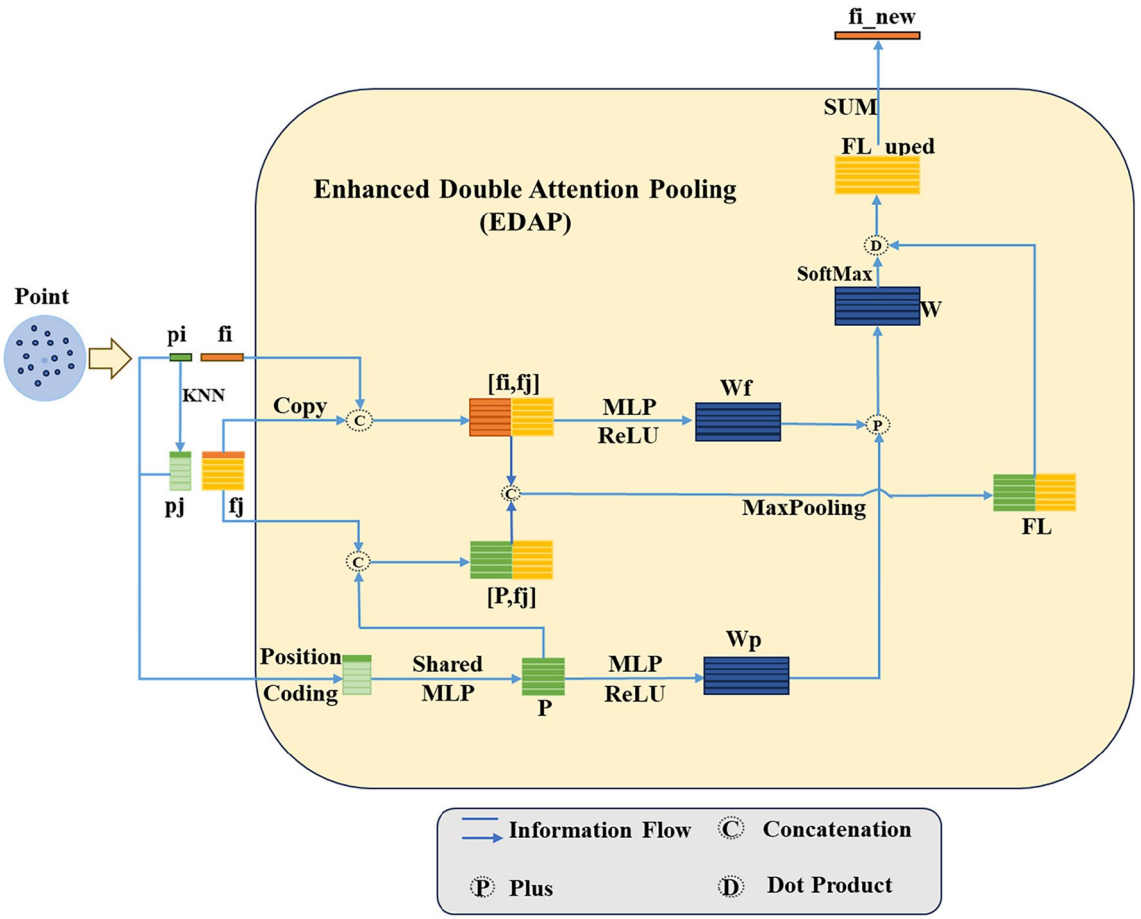


Fig. 1 Schematic diagram of Enhanced Dual Attention Pooling (EDAP) module

Position encoding significantly enhances the model's ability to recognize the positional information of point clouds. By concatenating the encoded features P with the neighboring appearance features f_j, more comprehensive feature information $[P \oplus f_j]$ can be obtained.

(2) Multi-level Feature Fusion: In our method, we first integrate encoded topological information with neighboring appearance features to extract high-level semantic information. Additionally, we pay particular attention to the interactions between local centroid features and their adjacent point features, which can be represented as $[f_i \oplus f_j]$. These features, along with the combination of centroid and neighboring point features $[P \oplus f_j]$, are processed through a shared multilayer perceptron (MLP). Subsequently, the output of the MLP is aggregated through a max pooling operation, as shown in Eq. (3-3), mapping it to a new feature space, thus comprehensively extracting local advanced semantic features FL, further enhancing the semantic expression capabilities of the model.

$$FL = \text{Maxpooling}\left(\text{MLP}\left([f_i \oplus f_j] \oplus [P \oplus f_j]\right)\right) \quad (3\text{-}3)$$

(3) Dual Attention: In this part, our processing procedure is divided into two steps: First, for each point, we balance the topological weights and the computed appearance weights to weight the features $[P \oplus f_j]$:

$$W_p = \text{MLP}(P) \quad (3\text{-}4)$$

Equation (3–4) is crucial for understanding the topological features of the neighborhood, providing advanced topological information. Therefore, after encoding the features, we use a shared multilayer

perceptron (MLP) to learn the topological weights W_p of local points. The coordinate features of local points alone may not be sufficient to distinguish objects of the same class, as differences in texture, color, and shape among objects can make their appearance features difficult to distinguish by the network. Considering that the texture features of the same type of objects are usually similar, we concatenate the features f_i of the centroid and the features f_j of the neighboring points, and use a shared multilayer perceptron (MLP) to perceive appearance features, thereby calculating the local semantic weights W_f

$$W_f = \mathrm{MLP}(f_i \oplus f_j) \tag{3-5}$$

Next, we merge the obtained local geometric topological features and appearance features, and activate both weights using the ReLU function. We use addition to obtain the composite weight. The weight coefficient is denoted by η, and the calculation of the composite weight is shown in Eq. (3–6):

$$W = \mathrm{ReLU}(W_p) + \eta \times \mathrm{ReLU}(W_f) \tag{3-6}$$

(4) Local Feature Aggregation: We calculate the fused attention weights, considering the importance of different positions and appearance features. Using the SoftMax activation function, we perform bilinear weighting on the weights W and the enhanced local neighborhood features FL. Subsequently, we use SUM as the reduction function to aggregate and update the point's features to f_{i_new}, as shown in Eq. (3-7):

$$f_{i\_new} = \mathrm{SUM}(\mathrm{softmax}(W) \odot FL) \tag{3-7}$$

Global context feature aggregation

Local feature aggregation describes the contextual relationships between neighboring points, but for complex structures such as architectural cultural heritage, starting solely from local features is insufficient for global perception. To more effectively express features, we introduce a global context feature module, aimed at enhancing the model's global perception by integrating panoramic scene information, enabling it not only to recognize individual architectural structures but also to effectively handle complex scenes and extensive spatial relationships.

We assume a spherical spatial domain,as shown in Fig. 2, using the position and volume ratio of objects to represent the global context. It is important to note that even objects of the same class may exhibit different styles; their geometric structures are similar, but their positions and orientations vary. Therefore, given the insensitivity of the volume ratio to local and global boundaries, we use this characteristic to recognize subtle geometric deformations of objects within the same class.

$$S_i = \frac{V_g}{V_i} \tag{3-8}$$

In this context, V_i represents the volume of the neighborhood boundary, while V_g represents the global boundary volume. The geometric coordinates X, Y, Z indicate the position of the local neighborhood. Based on this, we have defined the following method for aggregating global context features:

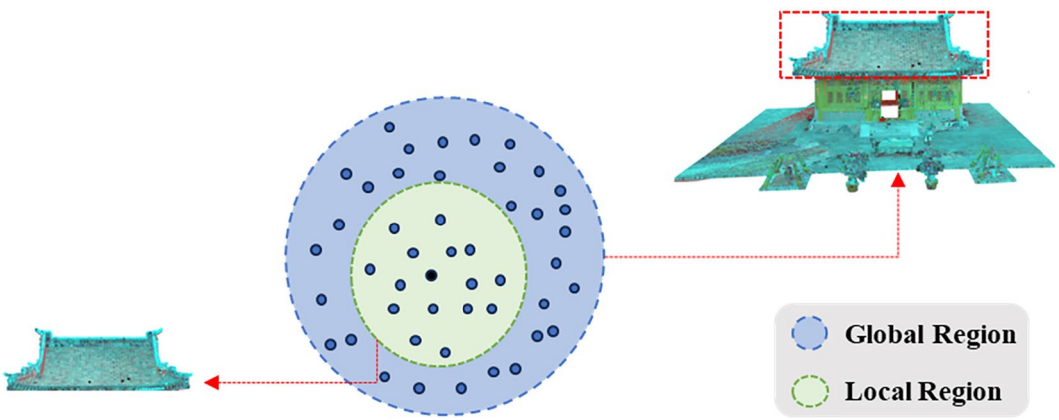


Fig. 2 Display of Global Context Feature (GCF) aggregation module (the roof in the picture belongs to the Hall of Heavenly Gods)

$$F_g = \mathrm{MLP}\left((x_i, y_i, z_i) \oplus S_i\right) \tag{3-9}$$

Here, (x_i, y_i, z_i) represent the coordinates of the point, and "$\oplus$" denotes the concatenation operation.

DSC structure

The structure of the DSC architecture is shown in Fig. 3. This architecture accepts two types of inputs: spatial information and previously learned features. Spatial information is used to learn both local and global semantic features, while previously learned features are specifically used for local feature aggregation. The diagram shows the process of local feature aggregation for points, which are input into the EDAP module for two-level local feature aggregation. Subsequently, the aggregated features are overlaid with the feature map to produce the final local features. Global context information is extracted from spatial information through the GCF module. The output of this module is the learned discriminative spatial features, which are a concatenation of local and global features.

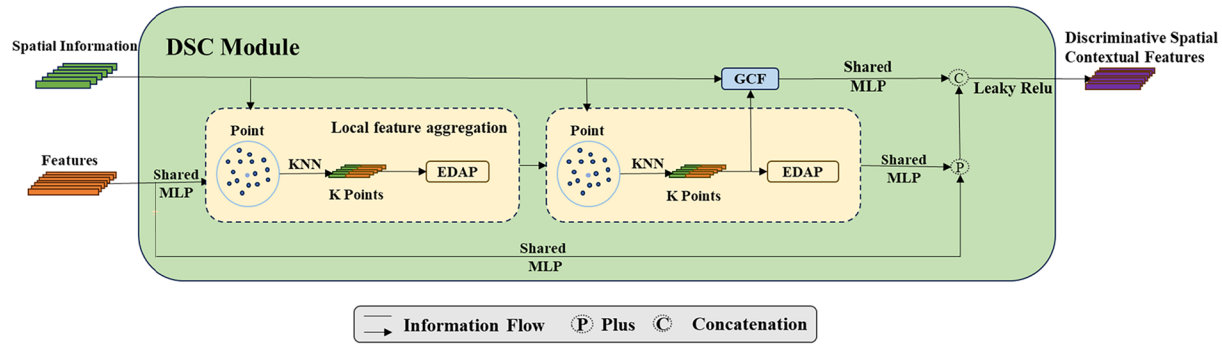


Fig. 3 DSC structure

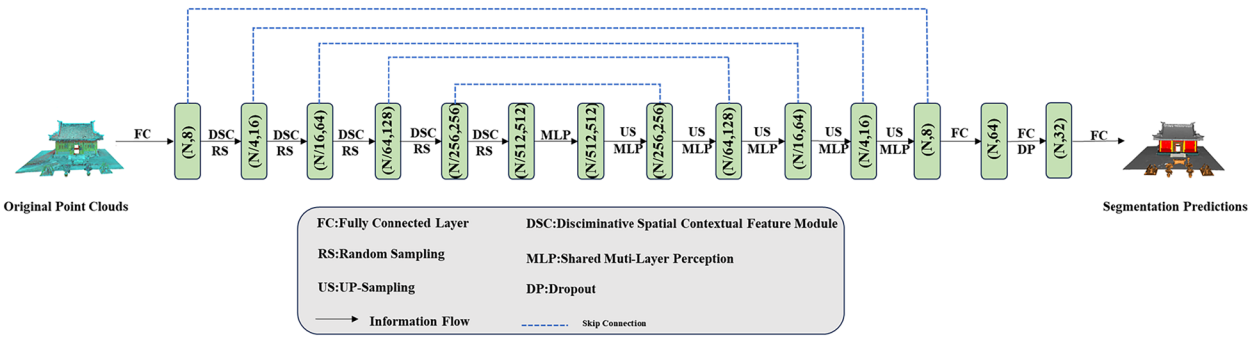


Fig. 4 DSC-Net structure

(a) Beiding Niang Niang Temple and Auxiliary Halls(Area_1) (b) Hall of Heavenly Gods(Area_2) (c) East Side Hall(Area_3)

Fig. 5 Display of buildings in each region of the Ancient architecture dataset

DSC net structure

In this section, we will provide a detailed introduction to our designed network, DSC-Net, which is a symmetric encoder-decoder network architecture. Both the encoder and decoder stages contain the same number of basic blocks, and the workflow is illustrated in Fig. 4. The network input consists of N point clouds, which include coordinate information and features, represented as $P \in R^{N \times 3}$ and $F \in R^{N \times d}$, respectively. Point clouds can be viewed as a collection that integrates topological attributes and appearance features. Features are first input into a shared MLP layer, where dimensions are unified to 8. Subsequently, the encoder, composed of five enhanced attention feature aggregation modules and global context feature modules, progressively encodes features to extract the semantics of multiple features such as color (details in Sect. "Enhanced dual attention pooling module"). After each encoder block, a random point sampling method is used for downsampling. The number of points gradually reduces from N to N/512, and feature dimensions increase from 8 to 512. The next five decoder blocks are used for decoding high-level semantic features. The encoded features are upsampled through nearest neighbor interpolation and connected to the intermediate feature map through skip connections. Finally, three fully connected layers reduce the dimensions of the features to the final output categories, predicting the final semantic labels with the output dimensions of semantic segmentation prediction being $N \times C_{class}$, where C_{class} is the number of categories.

Experiments

Experimental details

In this section, we will comprehensively evaluate our proposed network DSC Net, using datasets including self built ancient architecture dataset, publicly available architectural cultural heritage dataset ArCH, and publicly available dataset S3DIS. Our experimental setup includes virtual CPUs (Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz), a Tesla V100-SXM2 GPU, and all experiments were conducted in a virtual environment equipped with CUDA 11.3 and cuDNN v7 on the TensorFlow 2.6.0 framework. In the experiments with these three datasets, we used an Adam optimizer with an initial learning rate of 10^{-2}. The network underwent 100 training epochs, with the learning rate decreasing by 5% at the end of each epoch, and the number of neighborhood points was set to 16. During the training phase, a fixed number of points (40,960) were sampled from each point cloud. For the testing phase, the entire original point cloud was used, with each point including 3D coordinates and color information.

Datasets

(1) The equipment used in this experiment is the FARO Focus3D X130 3D laser scanner, which can capture 976,000 points per second and scan at a distance of over 130 m. It is equipped with a coaxial high-resolution camera, making the matching of color images and point clouds unbiased. The collected data comes from the architectural heritage of the Niangniang Temple during the Xuande period of the Ming Dynasty in ancient China. It was first built between 1426–1435 during the Ming Dynasty and has a history of over 500 years. The main buildings in the Beiding Niangniang Temple include the Hall of Heavenly Gods, the East Supporting Hall, the Niangniang Hall, the Dongyue Hall, and the Shanmen Hall. The Beiding Niangniang Temple is a typical traditional Chinese wooden architecture, mainly composed of a roof and a pedestal. The roof consists of tiles and roof figures on the roof ridge. Doors and windows are all made of wooden structure, and there are hollow patterns on them. The Niangniang Temple was listed as the seventh batch of municipal level cultural relics protection units in Beijing in 2003. It is one of the "Five Top Temples" in Beijing's history and also a landmark building on the central axis of Beijing.

The point cloud datasets collected in this experiment are the Beiding Niangniang Hall and its auxiliary hall (Area1), the Hall of Heavenly Gods (Area2), and the East Side Hall (Area3). Among them, as for the Niangniang Hall and its auxiliary hall (Area_1), the Niangniang Hall is five rooms wide, with the roof of Xieshan ound ridge roof, green glazed tile and yellow trimmed roof. The auxiliary halls on both sides of Niangniang Hall are respectively gable roof and simple tile roof. The Hall of Heavenly Gods (Area_2) has a width of three rooms and a gable roof with a simple tile roof. The front of the Hall of Heavenly Gods has four five painted wooden doors, with four threshold windows on each side of the doors, and four five painted wooden doors on the back. The East Side Hall (Area_3) is a gable roof with a simple tile roof. Figure 5 shows the appearance characteristics of ancient architectures in three regions. This dataset consists of ten categories, namely Tiebeam, Window,

Table 1 Partial display images for each category in the Ancient architecture dataset

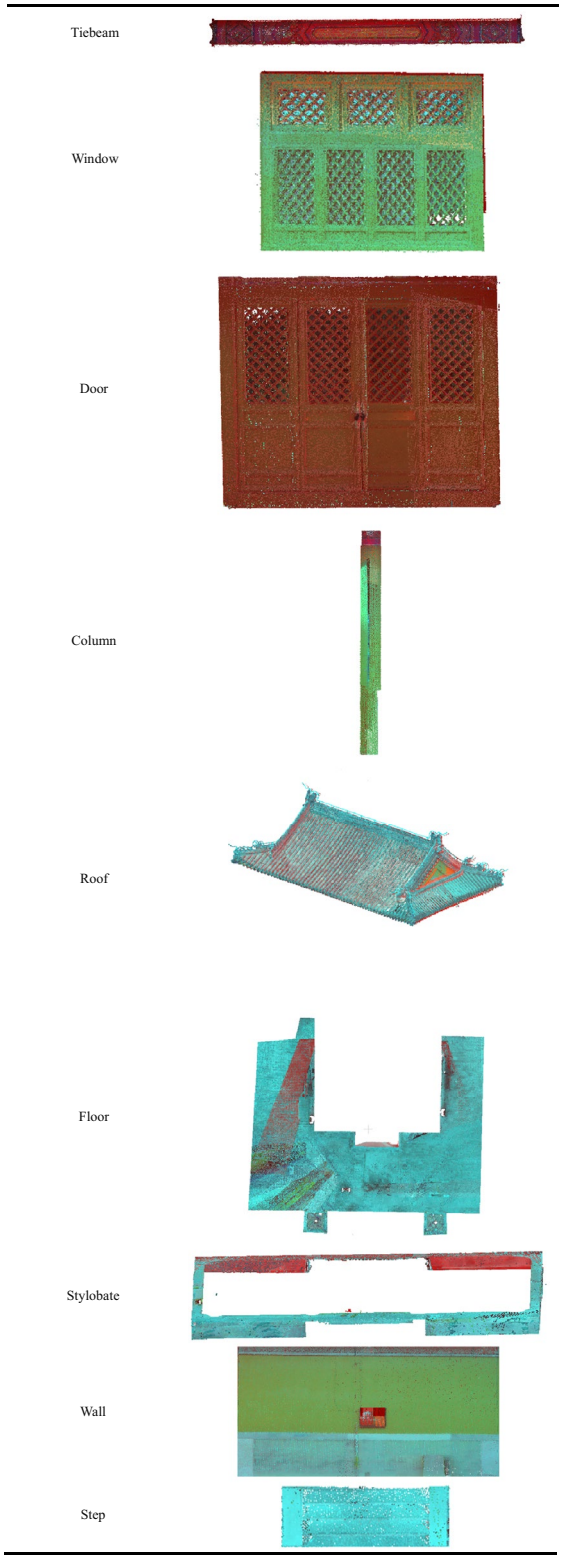


Table 2 Number of point clouds in each area of the Ancient architecture dataset

	Area 1	Area 2	Area 3
Points	76516796	28569973	13882483
Total	118969252		

Door, Column, Roof, Floor, Stylobate, Step, Wall, and Clutter. Table 1 provides a visual representation of the appearance of each category in the dataset. In addition, we also conducted detailed statistics on the number of point clouds in the ancient architecture dataset, and Table 2 shows the number of points in each region and the total number of points. Table 3 shows the number of point clouds corresponding to each category.

(2) ArCH is a large point cloud dataset, jointly released by the University of Turin and other universities and institutions, focusing on the semantic segmentation of point clouds related to historical architectural heritage. The ArCH dataset contains 17 annotated point cloud collections and 10 unannotated collections. These 17 point cloud scenes have been meticulously labeled into 10 categories, including architectural elements such as Vault, Column, Floor, Door,Window, Wall, Moldings, Stair, Arch, and Roof, as shown in Fig. 6.

(3) S3DIS is a 3D indoor space dataset acquired by Stanford University in 2017 through scanning technology. This dataset encompasses six large indoor areas of three different buildings. Each area contains between 20 to 70 rooms, with the number of points in each room ranging from 50,000 to 2.5 million. Each point is labeled with one of thirteen semantic categories. We use only the 3D coordinates, color information, and corresponding labels from the point cloud data to train the network and employ a six-fold cross-validation strategy for evaluation.

Evaluation on Ancient architecture dataset

In order to effectively distinguish the various categories in the ancient architecture dataset, we conducted a comprehensive evaluation of DSC-Net and adopted a K-fold cross validation (K=3) strategy to evaluate the performance of the self built ancient architecture dataset in eight methods. Select one fold as the test set each time, and the other two folds as the training set (where each

Table 3 Number of point clouds for each category in the Ancient architecture dataset

Class	Tiebeam	Window	Door	Column	Roof	Floor	Stylobate	Step	Wall	Clutter
Points	11900299	8924192	17104932	10447809	28082593	12078178	4060321	1824288	21325993	3220647
Total	118969252									

fold corresponds to an area). The detailed results are shown in Table 4 below. We use overall accuracy (OA) and mean intersection to union ratio (mIoU) as standard indicators for evaluation. In this experiment, we used seven methods, including Point Net, Point Net++, DG-CNN, KPConv, RandLA-Net, BAAF-Net, RandLA-Net+PnP-3D, as reference methods for comparison with this method. Point Net directly processes each point, uses multi-layer perceptrons (MLP) to extract features from point cloud data, and aggregates global features through global max pooling. Point Net++is an improved version of Point Net, which captures local and global features by introducing a hierarchical feature learning mechanism. It uses multiple PointNet modules to handle point cloud regions of different scales. DG-CNN uses dynamically constructed KNN maps for point cloud feature extraction. It captures local geometric structures by calculating neighbors in the feature space. KPConv is a convolutional based point cloud feature extraction method that uses learnable convolution kernels to process local neighborhood features of point clouds. RandLA-Net uses random sampling and local aggregation to efficiently process large-scale point cloud data. It captures features of different scales through a multi-layer attention mechanism. BAAF-Net uses a dual attention mechanism to aggregate local features, distinguishing categories with similar geometric structures but different appearances in point clouds. On the basis of RandLA-Net, the PnP-3D module is integrated to enhance feature representation by introducing more local context and global bilinear response. The above methods all adopt the default settings in the original paper, including network structure and training parameters.

The experimental results show that our method achieves an average intersection to union ratio (mIoU) of 63.56% and an overall accuracy (OA) of 82.63%, all higher than the other seven methods. Our method outperforms the benchmark network RandLA-Net by 20.56% and

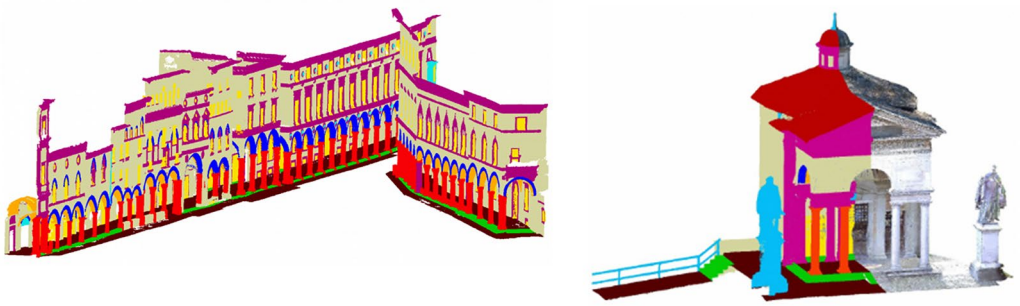


Fig. 6 Partial scene display of the ArCH dataset

Table 4 Detailed semantic segmentation results for the Ancient architecture dataset (numbers in bold indicate results higher than the corresponding baseline. In each column, the highest value is highlighted in red)

Method	mIou%	OA%	Tiebeam	Door	Floor	Roof	Step	Stylobate	Wall	Window	Column	Clutter
Point Net [21]	15.45	42.46	10.1	8.5	40.7	41.0	0.3	2.1	38.2	11.8	0.3	1.5
Point Net++ [22]	19.15	36.01	23.9	14.2	35.8	48.0	0.4	0.3	46.9	16.5	5.5	0.1
DG-CNN [6]	46.22	81.80	37.61	34.75	87.31	73.56	36.21	57.26	52.89	17.51	32.81	32.26
KpConv [19]	35.39	49.75	0.00	5.85	92.21	30.39	60.17	63.01	28.30	14.68	17.39	41.89
RandLA-Net [4]	43.00	70.59	1.30	15.48	86.35	48.95	46.44	46.71	79.17	51.58	47.87	6.24
BAAF-Net [39]	53.98	72.14	29.99	44.18	89.81	62.76	56.03	67.93	69.20	33.09	66.48	20.29
RnadLA-Net+PnP-3D [36]	60.20	80.57	20.52	19.08	94.33	78.71	82.79	71.02	86.40	43.03	53.93	52.13
Ours	**63.56**	**82.63**	23.12	33.85	95.01	79.02	84.52	70.44	85.90	48.80	59.92	55.02

12.04% in mIoU and OA evaluation metrics, respectively. Meanwhile, compared to the two improved methods BAAF-Net and RandLA-Net + PnP-3D on the benchmark network, our method is 9.58% and 10.49% higher than BAAF-Net. Compared with RandLA-Net + PnP-3D, our method has improved by 3.36% and 2.06%, respectively. Among them, the category of Fang has the highest segmentation accuracy on DG-CNN, with an IoU of 37.61%. BAAF-Net has the highest segmentation accuracy on doors, with an IoU of 44.18%. RandLA-Net + PnP-3D has an IoU of 71.02% on styleboard and 86% on walls. Our method has an IoU of 95.01%, 79.02%, 84.52%, 48.80%, 59.92%, and 55.02% for doors, roofs, steps, windows, columns, and others, respectively. The results have demonstrated the superiority of our method over the other seven methods in semantic segmentation of self built ancient architecture point clouds. The experimental results also show that the overall accuracy of the method using RandLA-Net as the benchmark is higher than the other four methods. We evaluated the performance of our method on the entire dataset using class accuracy (ACC,%) as an indicator, as shown in Fig. 7. As shown in Fig. 8, we present the visualization results of the Hall of Heavenly Gods (Area_2), which are the front

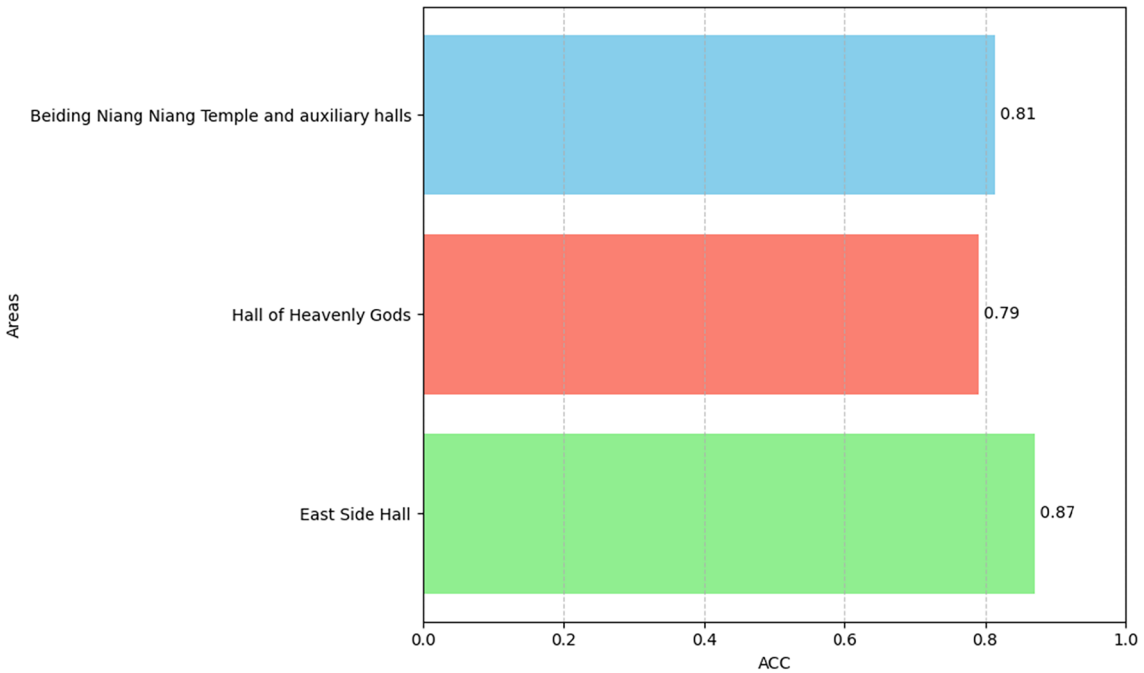


Fig. 7 Accuracy of testing for each area of the ancient architecture dataset (ACC,%)

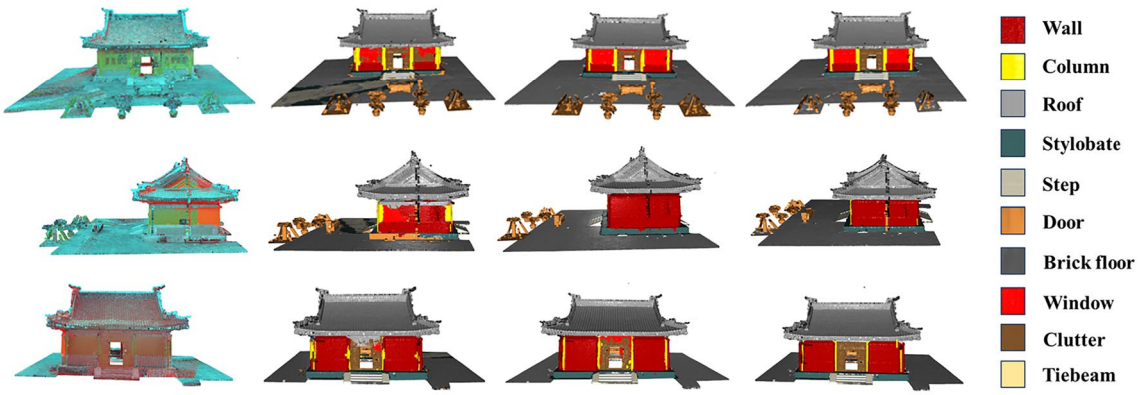
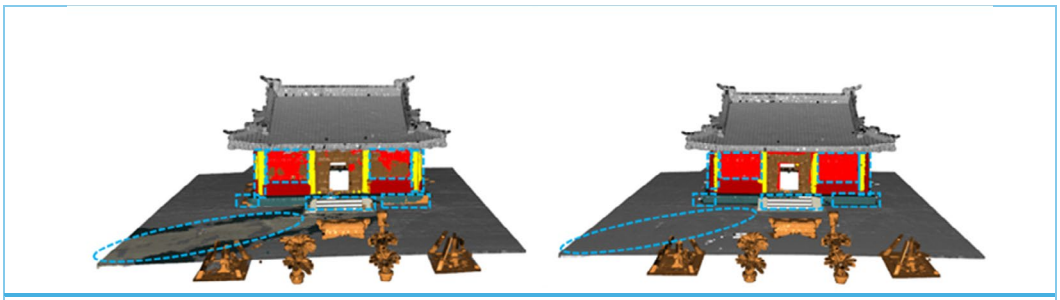
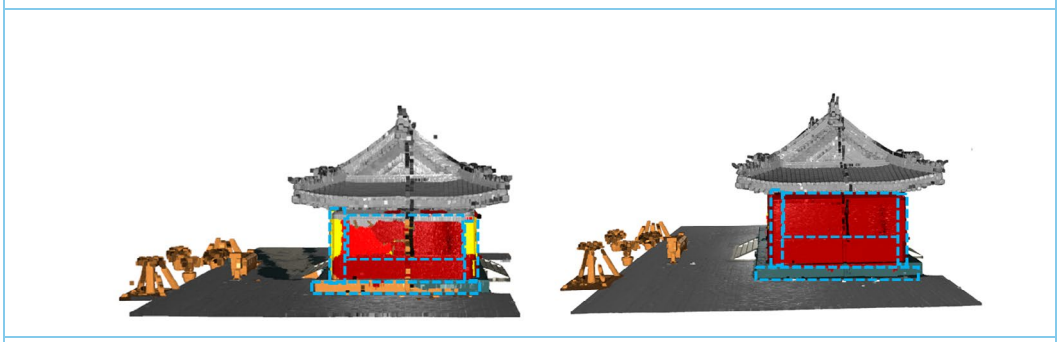


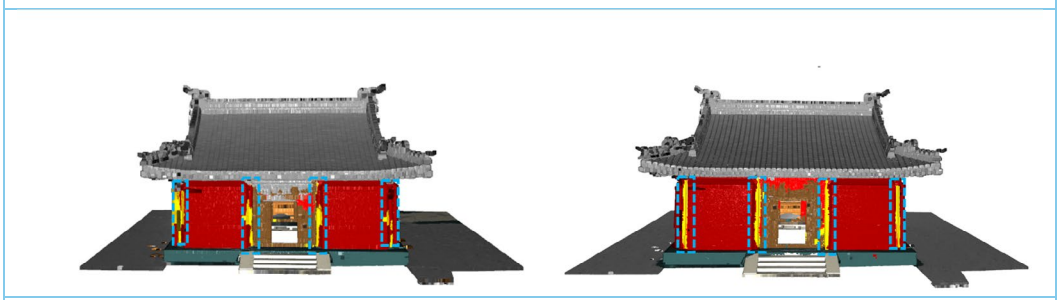
Fig. 8 Display of segmentation results for the Hall of Heavenly Gods (from left to right are: RGB color input point cloud, Randla Net prediction results, this method prediction results, and Ground Truth)



(a) Front view detail of semantic segmentation effect in Area_2 (left: RandLA Net, right: Our method)



(b) Side view details of semantic segmentation results for Area_2 (left: RandLA Net, right: Our method)



(c) Rear view detail of semantic segmentation effect in Area_2 (left: RandLA Net, right: Our method)

Fig. 9 Details of semantic segmentation effect in Area_2 (left: RandLA Net, right: Our method)

view, side view, and rear view of the the Hall of Heavenly Gods from top to bottom. Figure 9 shows a comparison of the segmentation details of the front, side, and back views of the Hall of Heavenly Gods (Area_2) (with blue dashed lines indicating the comparison of details). Our method shows that it is more accurate than the benchmark method RandLA-Net in segmenting similar geometric structures and boundaries, especially in difficult to distinguish categories such as doors, windows, columns, and walls.

This is because for large-scale scenes such as ancient architecture, the amount of point cloud data is usually large. RandLA-Net's efficient random downsampling strategy can process this data more quickly. At the same time, to overcome the problem of accidentally discarding key features during the random downsampling process, a local feature aggregation module is introduced to gradually increase the acceptance domain of each 3D point, effectively preserving geometric details. The high computational complexity of the other four methods results in slower processing speed, larger memory usage, and affects overall accuracy. This has been demonstrated in reference [4] to demonstrate the superiority of this benchmark method in handling large-scale complex scenes. However, for ancient architecture, which has more complex detailed structures and high

geometric similarity between different categories, our method can adapt to complex geometric shapes and fine decorations, and capture complex overall layouts and structures to cope with more challenging complex scenes of ancient architecture. Specifically:

(1) Firstly, the enhanced dual attention pooling module first extracts topological features, learning edge, corner, and curvature information of categories such as tiebeam, doors and windows, and roofs, in order to gain a deeper understanding of the structure and form of buildings. The subsequent extraction of appearance features involves information such as color, texture, and material, such as the color of walls and the texture of doors and windows.
(2) Secondly, the global contextual semantic feature module captures the global information of the entire point cloud scene, helping the model understand how various architectural elements are interrelated in space. This is particularly important for distinguishing architectural elements that are similar in location but of different types, such as decorative columns next to windows. This module achieves accurate segmentation of components such as doors, windows, columns, and roofs by analyzing the overall shape and structure of the building, and can clarify their boundaries with the surrounding environment. In addition, it also helps to identify and segment various types of ancient architectural elements, such as accurately identifying the shape of roofs and the height and diameter of columns by analyzing the overall shape and scale of the building.

To better validate the effectiveness of our method, we have listed several sets of results for object categories with similar geometric structures that are difficult to distinguish in Table 5. In terms of geometric structure, doors and windows are represented as planes perpendicular to the ground, which are mainly distinguished by color and texture. The proposed feature aggregation strategy fully utilizes the geometric and appearance information in the points. Columns and walls are difficult to distinguish due to their similar textures and close geometric positions, resulting in poor segmentation results. Our global contextual semantic feature module effectively improves this issue by more accurately grasping spatial location and overall structure.

Evaluation on ArCH

In this experiment, we evaluated the ArCH [10] dataset, which consists of 17 annotated point clouds and an additional 10 unannotated point clouds. The ArCH dataset contains numerous scenes that are part of the UNESCO World Heritage List (WHL), showcasing multiple historical periods and architectural styles. In this benchmark dataset, 15 scenarios were used for training, while 2 scenarios were used for testing. Due to some scenarios not covering all nine categories, five scenarios were selected for analysis in this experiment: 5-SMV_chapel_1, 6-SMV_chapel_2to4, 7-SMV_chapel_24, 15-OTT_church, and

Table 5 Results of geometric structure similarity types between the proposed DSC-Net, RandLA-Net, and RandLA-Net + PnP-3D on the Ancient architecture dataset (evaluation metric is mIoU, %)

Similar Classes	RandLA-Net	BAAF-Net	RandLA-Net + PnP-3D	Ours
Wall/Door	79.17/15.48	69.20/44.18	86.40/19.08	85.90/33.85
Window/Door	51.58/15.48	33.09/44.18	43.03/19.08	48.80/33.85
Column/Wall	47.87/79.17	66.48/69.20	53.93/86.40	59.92/85.90

Table 6 Key Features of the Partial ArCH Dataset

Name	Number of points	Scene	Data acquisition	Number of classes(excluded Clutter)
5_SMV_chapel_1	3,783,412	Outdoor	TLS + UAV	9/9
6_SMV_chapel_2to4	6,326,871	Indoor/outdoor	TLS + UAV	9/9
7_SMV_chapel_24	3,571,064	Outdoor	TLS + UAV	9/9
15_OTT_church	13,264,040	Indoor/outdoor	TLS	9/9
A_SMG_portico	16,165,924	Outdoor	TLS + UAV	9/9

A-SMG_portico. We use five fold (K = 5) cross validation to evaluate the final results, selecting one fold as the test set and the other two folds as the training set (where each fold corresponds to an area). Table 6 presents detailed information on the selected data, including the number of point clouds, experimental scenarios, data acquisition methods, and categories of the dataset. We use overall accuracy (OA) and mean intersection to union ratio (mIoU) as evaluation metrics. Given the limited evaluation of this dataset, we conducted comparative experiments using six methods: PointNet PointNet++, DG-CNN、RandLA-Net, BAAF-Net, RandLA-Net + PnP-3D. Table 7 provides a detailed list of quantitative segmentation results for each method. Our method outperforms RandLA Net by 0.5% in mIoU and 1.04% in OA compared to baseline based methods; Compared with BAAF Net, it is 0.17% and 0.15% higher, respectively. Compared with RandLA Net + PnP3D, it is 1.7% and 1.06% higher, respectively. The segmentation performance in the 6_SMV_chapel_2to4 scene is shown in Fig. 10, and our method is also higher than the other six methods, demonstrating the highest segmentation results.

Evaluation on S3DIS

In order to make the effectiveness of our method more convincing, we conducted experiments on the recognized public dataset S3DIS [11] and demonstrated semantic segmentation results from sixfold cross validation. As evaluation indicators, we used mean union intersection (mIoU) and overall accuracy (OA), and the detailed comparison results of these indicators are shown in Table 8 and Fig. 11. Compared to the baseline network, our method demonstrates stronger competitiveness in all evaluation metrics. Compared with RandLA Net using the same random downsampling strategy, the sixfold cross validation results showed that our mIoU increased by 1.03% and OA increased by 0.4%. Our IoU in ceiling, beam, table, board, and clutter reached 93.8%, 63.9%, 71.4%, 67.1%, and 60.9%, respectively, showing significant advantages. These results clearly demonstrate the superiority of our method on the benchmark dataset S3DIS. In Fig. 12, we visualize the raw and predicted results of three classic scenarios from S3DIS. The comparison with the baseline method RandLA Net shows that our method can accurately distinguish similar geometric categories, demonstrating its robustness in various scenarios.

Ablation experiments

Our method has been validated through testing on the ancient architecture dataset, ArCH dataset, and S3DIS dataset. In order to gain a deeper understanding of the mechanism of the network, we conducted two sets of ablation experiments on the ancient architecture dataset and evaluated the ablation results using standard threefold cross validation. Meanwhile, considering the widespread use of S3DIS as a common dataset in 3D point cloud semantic segmentation research, conducting ablation experiments on a standardized S3DIS set can help demonstrate the generality and robustness of our

Table 7 Quantitative Segmentation Results of Different Methods on the ArCH Dataset (numbers in bold indicate results higher than the corresponding baseline. In each column, the highest value is highlighted in red)

Methods	mIou%	OA%	Arch	Column	Moldings	Floor	Window	Wall	Stair	Vault	Roof
Point Net [21]	34.16	63.21	0.1	0.0	11.4	74.3	2.6	60.1	8.0	65.1	85.8
Point Net++ [22]	40.91	72.13	3.8	67.2	12.1	82.5	7.5	74.6	30.3	44.2	46.0
DG-CNN [6]	32.47	74.64	6.36	4.56	8.43	73.30	3.13	69.17	21.40	30.74	75.18
RandLA-Net [4]	55.34	81.36	19.08	72.81	19.45	84.20	26.44	79.41	69.27	61.87	65.58
BAAF-Net [39]	55.67	82.25	18.33	76.69	18.73	84.59	14.70	77.56	71.34	69.11	69.97
RandLA-Net + PnP-3D [36]	54.14	81.34	7.64	70.44	11.64	86.26	32.74	77.57	66.17	65.70	69.07
Ours	**55.84**	**82.40**	18.73	72.25	17.25	91.05	19.36	76.51	62.64	69.26	75.58

Fig. 10 Front view of 6_SMV_chapel_2to4 (left: original point cloud, center: predicted result, right: real point cloud)

Table 8 Quantitative Segmentation Results of Different Methods on the S3DIS Dataset (numbers in bold indicate results higher than the corresponding baseline. In each column, the highest value is highlighted in red)

Method	mIou (%)	OA (%)	Ceiling	Floor	Wall	Beam	Column	Window	Door	Table	Chair	Sofa	Book-case	Board	Clutter
Point Net [21]	47.6	78.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
Point Net + + [22]	54.5	81.0	89.7	91.0	70.1	49.7	35.6	49.5	59.0	61.9	48.9	45.0	40.3	28.3	39.4
SPG [37]	62.1	86.4	89.9	95.1	76.4	62.8	47.1	55.3	68.4	73.5	69.2	63.2	45.9	8.7	52.9
DG-CNN [6]	56.1	84.1	90.8	91.6	72.1	51.3	36.2	50.1	62.6	63.2	52.3	47.1	42.3	29.4	40.4
KP-Conv [19]	70.6	88.3	93.6	92.4	83.1	63.6	54.3	66.1	76.6	57.8	64.0	69.3	74.9	61.3	60.3
ShellNet [38]	66.8	87.1	90.2	93.6	79.9	60.4	44.1	64.9	52.9	71.3	84.7	53.8	64.6	48.9	59.4
RandLA-Net [4]	70.0	88.0	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
Ours	**71.03**	**88.4**	93.8	93.7	80.9	63.9	47.4	64.6	70.2	71.4	78.9	66.3	64.4	67.1	60.9

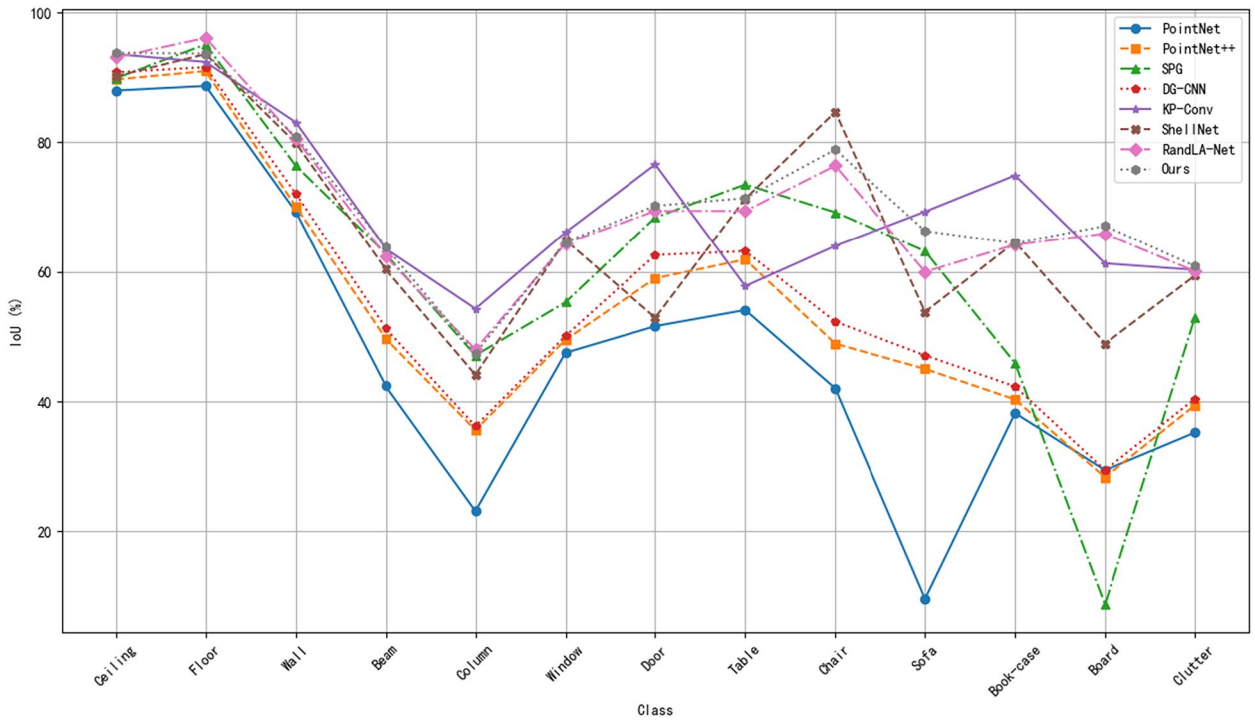


Fig. 11 Comparison of semantic segmentation effects of different categories using different methods (mIou,%)

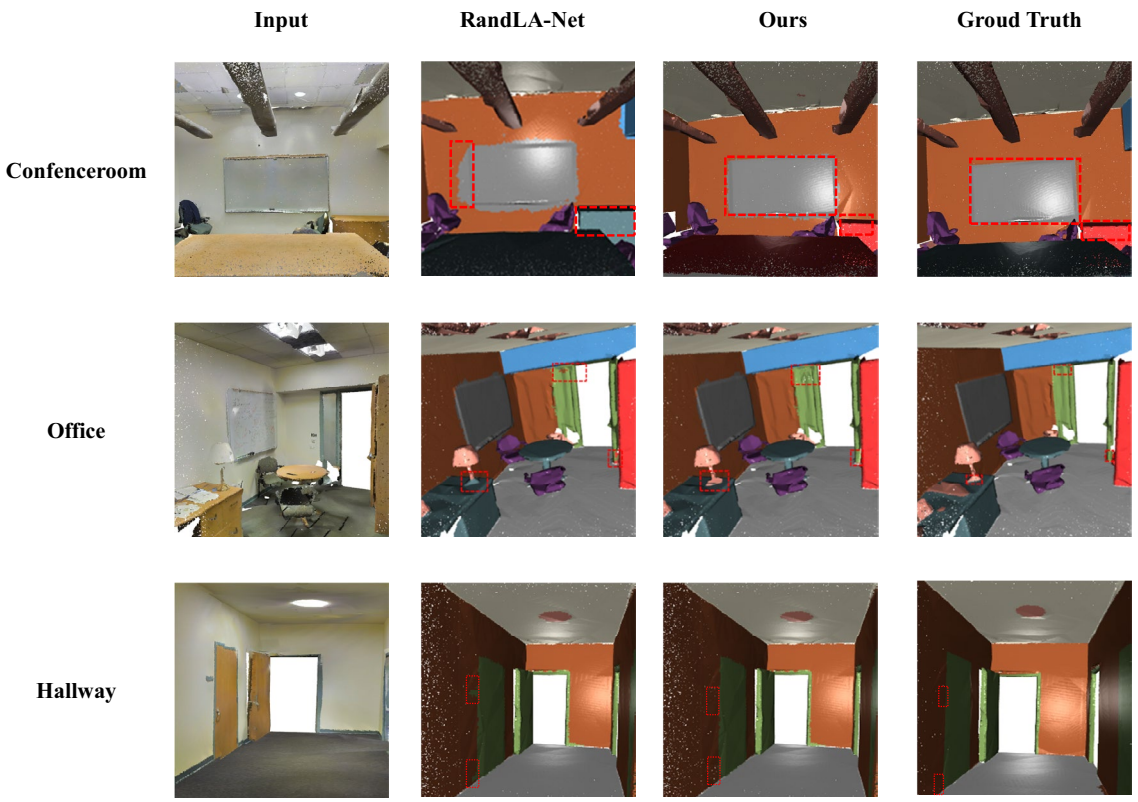


Fig. 12 Visualization Examples of S3DIS Dataset in Three Typical Indoor Scenarios

Table 9 Results of ablation experiments on self built ancient architecture datasets

Methods	mIou (%)
Remove coordinate position encoding	60.25
Remove all weights	57.32
Using semantic weights alone	60.88
Using topology weights alone	61.79
Remove all fusion weights from ReLU	62.37
Remove multi-level fusion features	62.15
Remove global contextual semantic features	59.28
Ours	63.56

Table 10 Ablation experiments on the S3DIS dataset Area_5

Methods	mIou (%)
Remove coordinate position encoding	62.88
Remove all weights	59.81
Using semantic weights alone	61.54
Using topology weights alone	62.90
Remove all fusion weights from ReLU	64.62
Remove multi-level fusion features	62.52
Remove global contextual semantic features	60.46
Ours	65.55

method. We also conducted two sets of ablation experiments on Area_5 of the S3DIS dataset.

We evaluated the effectiveness of various modules of DSC Net under different configurations. Specifically, we designed four control experiments: removing coordinate encoding operations, removing all weights, removing fused features, and removing global contextual semantic feature modules. To further explore the effectiveness of different components of enhanced dual attention, we employed three different methods to evaluate the impact of attention forms: using topological weights alone, using semantic weights alone, and not applying ReLU activation before weight fusion. As shown in Tables 9 and 10, our enhanced dual attention pool module and global context feature module significantly improve the accuracy of point cloud segmentation. In the enhanced dual attention module, compared to the method without coordinate encoding, our encoding strategy increased mIoU by 3.31% and 2.67% on both datasets, respectively. Our encoding module achieves this by calculating the distance between the centroid and adjacent points, as well as their offsets in the x, y, and z directions, which provides information that the original coordinates do not have and is crucial for local geometric perception. When evaluating the importance of different weights in our dual attention module, we found that the lack of weights hinders effective learning and aggregation of local adjacent points, where topological weights typically have a more significant impact on feature information learning than semantic weights. In addition, if ReLU activation is not applied before feature fusion, the weights of these two types will interfere with each other, resulting in a decrease in mIoU. Our experiments have shown that by integrating enhanced feature information, this module can adapt to complex ancient architectural data structures and perform well in other data scenarios, thereby improving feature differentiation ability. Learning global contextual features can enable networks to extend their understanding of complex objects from local to global, enriching feature representation. The experimental results have demonstrated the effectiveness of the module and significantly improved its performance.

Conclusion

In this study, we propose a network based on Enhanced Dual Attention Pooling and Global Context Feature Aggregation, named DSC-Net, aimed at accurately analyzing and understanding 3D point cloud data obtained from complex, large-scale scenes. By guiding local feature fusion at both the dimensional and point levels, the network enhances its ability to recognize objects with similar geometric structures. The DSC module can be easily embedded into various network architectures for point cloud segmentation; we have embedded it into an encoder-decoder architecture, resulting in the DSC-Net architecture featured in this work. We tested it on the Ancient architecture dataset, ArCH, and S3DIS, where the proposed DSC-Net not only outperforms the most advanced point cloud segmentation methods based on rapid random sampling (such as RandLA-Net) in terms of accuracy but also shows excellent performance in handling diverse architectural environments. Particularly in the ancient architecture dataset, the model not only achieves high precision in spatial resolution but also effectively identifies and classifies key structural elements in cultural heritage, such as doors, windows, roofs, and decorative details. With our designed point cloud semantic segmentation network, we provide strong technical support for cultural heritage preservation, further advancing the scientific and precise approach to the conservation and restoration of ancient architectures. Additionally, the widespread application of this model also helps to strengthen the systematic study and management of cultural heritage, offering new perspectives and methods for protecting precious historical monuments worldwide.

Acknowledgements
We thank F. Matrone and all contributors for the ArCH dataset and appreciate the use of the Stanford 3D Indoor Spaces Dataset (S3DIS) in this study. This dataset was provided by the Vision and Geometry Group at Stanford University and played a crucial role in the completion of this study. We also thank the authors of the relevant papers on this dataset, whose work has provided a foundation and inspiration for our research. The relevant methods and dataset details have been cited in our references.

Author contributions
Jianghong Zhao conceived the presented idea and put forward experimental suggestions. Rui Liu conducted and refined the analysis process and wrote the manuscript. Huaxin Nan performed the data processing, as well as providing some of the network comparison experiment results. All authors approved the final manuscript.

Funding
This work was supported by the Project of National Key R&D Program Project (2018YFC0807806); Open Fund Project of State Key Laboratory of Geographic Information Engineering (SKLGIE2019-Z-3-1); Open Fund of State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering of Wuhan University (19E01); Open Research Fund Project of the Key Laboratory of Digital Mapping and Land Information Application of the Ministry of Natural Resources (ZRZYBWD202102); Software Science Research Project of the Ministry of Housing and Urban Rural Development (R20200287); Beijing Social Science Foundation Decision Consulting Major Project (21JCA004); National Natural Science Foundation of China(42171416).

Availability of data and materials
All data generated or analyzed during this study are included in this published article.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 13 May 2024 Accepted: 10 July 2024
Published online: 02 August 2024

References

1. Zhang R, Zhou X, Zhao J, Min C. A semantic segmentation algorithm of ancient building's point cloud data. Geomat Inform Sci Wuhan Univ. 2020;45(5):753–9.
2. Liu X, Ren T, Sun X, Xu C, Zhou M. A review of three dimensional digitalization methods for Ancient Chinese Architecture. J Shanxi Univ (Nat Sci Ed). 2023;3:592–603.
3. Zhao J, Hua X, Yang J, Yin L, Liu Z, Wang X. A review of point cloud segmentation of architectural cultural heritage. ISPRS Ann Photogramm Remote Sens Sp Inform Sci. 2023;10:247–54.
4. Hu Q, Yang B, Xie L, Rosa S, Guo Y, Wang Z, Markham A. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2020; 11108–11117.
5. Chen S, Niu S, Lan T, Liu B. Large-scale 3D point cloud representations via graph inception networks with applications to autonomous driving. 2019. arXiv preprint arXiv:1906.11359.
6. Wang L, Huang Y, Hou Y, Zhang S, Shan J. Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019; 10296–10305.
7. Feng M, Zhang L, Lin X, Gilani SZ, Mian A. Point attention network for semantic segmentation of 3D point clouds. Pattern Recognit. 2020;107:107446. https://doi.org/10.1016/j.patcog.2020.107446
8. Zhang H, Dana K, Shi J, Zhang Z, Wang X, Tyagi A, Agrawal A. Context encoding for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2018; 7151–7160.
9. Fan S, Dong Q, Zhu F, Lv Y, Ye P, Wang F. SCF-Net: learning spatial contextual features for large-scale point cloud segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2021; 14504–14513.
10. Matrone F, Lingua A, Pierdicca R, Malinverni E, Paolanti M, Grilli E, Landes T. A benchmark for large-scale heritage point cloud semantic segmentation. Int Arch Photogramm Remote Sens Spat Inf Sci. 2020;43:1419–26.
11. Armeni I, Sener O, Zamir A, Jiang H, Brilakis I, Fischer M, Savarese S. 3D semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2016; 1534–1543.
12. Lawin F, Danelljan M, Tosteberg P, Bhat G, Khan F, Felsberg M. Deep projective 3D semantic segmentation. In: Proceedings of the Computer Analysis of Images and Patterns (CAIP). 2017; 95–107.
13. Tatarchenko M, Park J, Koltun V, Zhou Q. Tangent convolutions for dense prediction in 3D. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2018; 3887–3896.
14. Graham B, Engelcke M, Maaten L. 3D Semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2018; 9224–9232.
15. Choy C, Gwak J, Savarese S. 4D Spatio-Temporal ConvNets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2019; 3075–3084.
16. Riegler G, Ulusoy A, Geiger A. OctNet: learning deep 3D representations at high resolutions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2017; 3577–3586.
17. Hua B, Tran M, Yeung S. Pointwise convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2018; 984–993.
18. Wang S, Suo S, Ma W, Pokrovsky A, Urtasun R. Deep parametric continuous convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2018; 2589–2597.
19. Thomas H, Qi CR, Deschaud JE, Marcotegui B, Goulette F, Guibas L. KPConv: flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2019; 411–6420.
20. Engelmann F, Kontogianni T, Leibe B. Dilated point convolutions: on the receptive field size of point convolutions on 3d point clouds. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA).2020; 9463–9469.
21. Charles RQ, Su H, Kaichun M, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2017; 652–660.
22. Charles RQ, Yi L, Su H, Guibas LJ. PointNet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st international conference on neural information processing systems. 2017; 5105–5114.
23. Haznedar B, Bayraktar R, Ozturk A, Arayici Y. Implementing PointNet for point cloud segmentation in the heritage context. Heritage Sci. 2023;11(1):2.
24. Malinverni ES, Pierdicca R, Paolanti M, Martini M, Morbidoni C, Matrone F, Lingua A. Deep learning for semantic segmentation of 3D point cloud. Int Arch Photogramm Remote Sens Spat Inf Sci. 2019;42:735–42.
25. Huang Q, Wang W, Neumann U. Recurrent slice networks for 3D segmentation of point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2018; 2626–2635.
26. Liu F, Li S, Zhang L, Zhou C, Ye R, Wang Y, Lu L. 3DCNN-DQN-RNN: A deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds. In: Proceedings of the IEEE international conference on computer vision(ICCV). 2017; 5678–5687.
27. Morbidoni C, Pierdicca R, Paolanti M, Quattrini R, Mammoli R. Learning from synthetic point cloud data for historical buildings semantic segmentation. J Comput Cult Heritage (JOCCH). 2020;13(4):1–16.
28. Ji Y, Dong Y, Hou M, Qi Y, Li A. An extraction method for roof point cloud of ancient building using deep learning framework. Int Arch Photogramm Remote Sens Spat Inf Sci. 2021;46:321–7.

29. Pierdicca R, Paolanti M, Matrone F, Martini M, Morbidoni C, Malinverni ES, Lingua AM. Point cloud semantic segmentation using a deep learning framework for cultural heritage. Remote Sens. 2020;12(6):1005.
30. Matrone F, Grilli E, Martini M, Paolanti M, Pierdicca R, Remondino F. Comparing machine and deep learning methods for large 3D heritage semantic segmentation. ISPRS Int J Geo Inf. 2020;9(9):535.
31. Yang J, Zhang Q, Ni B, Li L, Liu J, Zhou M, Tian Q. Modeling point clouds with self-attention and gumbel subset sampling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2019; 3323–3332.
32. Chen L, Li X, Fan D, Cheng M, Wang K, Lu S. LSANet: Feature learning on point sets by local spatial aware layer. 2019. arXiv preprint arXiv:1905.05442.
33. Wang F, Yang Y, Wu Z, Zhou J, Zhang W. Real-time semantic segmentation of point clouds based on an attention mechanism and a sparse tensor. Appl Sci. 2023;13(5):3256.
34. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In Proceedings of the Medical image computing and computer-assisted intervention(MICCAI). 2015; 234–241.
35. Zhao H, Jia J, Koltun V. Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR). 2020; 10076–10085.
36. Qiu S, Anwar S, Barnes N. Pnp-3d: a plug-and-play for 3d point clouds. IEEE Trans Pattern Anal Mach Intell. 2021;45(1):1312–9.
37. Landrieu L, Simonovsky M. Large-scale point cloud semantic segmentation with superpoint graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2018; 4558–4567.
38. Zhang Z, Hua BS, Yeung SK. ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics. In: Proceedings of the IEEE/CVF international conference on computer vision. (ICCV). 2019; 1607–1616.
39. Qiu S, Anwar S, Barnes N. Semantic segmentation for real point cloud scenes via bilateral augmentation and adaptive fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR). 2021; 1757–1767.

Publisher's Note