Affective and Pleasurable Design, Vol. 164, 2025, 117–126
https://doi.org/10.54941/ahfe1006066

Construction of a PointNet-Based Autoencoder Using a 3D Scene Dataset for Feature Extraction From Indoor Space Point Clouds Excluding Interior Details

Takahiro Miki[1], Yusuke Osawa[1], and Keiichi Watanuki[1,2]

[1]Graduate School of Science and Engineering, Saitama University 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338–8570, Japan
[2]Advanced Institute of Innovative Technology, Saitama University 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338–8570, Japan

ABSTRACT

In this study, to automatically construct a virtual space with a high degree of freedom of expression that reflects the spatial shape of the real space and the arrangement of objects, we focused on the global shape of the indoor space without interior details as the first step and constructed a PointNet-based autoencoder to extract the features of the shape. To train the machine learning model, we used ScanNet++, which is a 3D indoor space dataset converted into point cloud data. Feature extraction was performed using two types of point cloud data: (1) point cloud data not used in the training of ScanNet++, and (2) point cloud data of an indoor space obtained through 3D scanning of a real environment. Feature extraction was evaluated by comparing the shapes of the input point cloud, restored output point cloud and distance error. As a result, both the ScanNet++ data and the indoor space data were output as rectangular shapes, and the general shapes of the walls and floors of the indoor space were generally consistent, indicating that spatial features were extracted. However, the interior furniture and other objects were removed. To investigate the applicability of the model, feature extraction was performed using 3D objects with elliptical shapes in an interior space. In future work, we will investigate the development of an autoencoder that performs feature extraction by focusing on the local shape around each point using a point-cloud convolution method, along with feature extraction following region classification within the interior space.

Keywords: 3D point clouds, Feature extraction, Pointnet, Autoencoder, Scannet++, Virtual space

INTRODUCTION

With recent advancements in extended reality (XR) technologies, such as virtual reality (VR) and mixed reality (MR), XR has gained popularity not only in the entertainment industry but also in the medical and welfare sectors. VR provides a high degree of expressive freedom in virtual space because the entire field of vision is transformed into a VR image when the user wears a head-mounted display (HMD). However, applying VR for practical use in arbitrary locations remains challenging because VR images do not reflect the actual surrounding environment. Moreover, the range of motion is limited (Ishizaka et al., 2018). In contrast, while MR imposes fewer restrictions on the range of operation, it is typically used for specific tasks, such as superimposing virtual objects or digital information in real space. Therefore, generating a VR space that accurately reflects the real-world environment in real time—particularly in scenes where the shape of the space changes—is essential. However, creating such a virtual space involves complex processes, making it difficult to implement in a widespread, general manner. This highlights the need for technology that can automatically construct a virtual space with a high degree of freedom, such as tilting or expanding the space, while still allowing the user to perceive the surrounding environment.

Generally, 3D objects represent virtual spaces, typically in the form of mesh data composed of points and surfaces. With the widespread use of 3D measurement devices, such as light detection and ranging (LiDAR), 3D point cloud processing technology has become increasingly important for generating mesh data. 3D point clouds, which describe shapes as sets of 3D points (x, y, and z), provide highly accurate information about the real space's geometry and the positional relationships of objects. Recently, numerous studies have focused on the automatic generation of 3D objects from 3D point clouds, with generative models based on adversarial generative networks (GANs), a deep learning technique. Many adversarial networks, such as l-GAN (Achlioptas et al., 2018), incorporate PointNet (Charles et al., 2017) as an encoder for feature extraction, a deep learning method that addresses the sequential inequality of point cloud data. Most evaluations of 3D point cloud generation methods rely on open-source datasets containing large CAD models across specific object categories, such as ModelNet (Wu et al., 2018). However, these datasets primarily consist of object categories like chairs and cars, and there is a lack of examples of deep learning or point cloud generation using entire spaces—such as indoor or outdoor environments—as datasets. Additionally, spatial objects, such as walls, floors, and furniture, have different characteristics compared to individual objects, making it challenging to apply features learned from the above datasets.

Therefore, to automatically construct a virtual space that reflects the spatial layout of the real environment and the arrangement of objects, this study focused on the global shape of indoor spaces as the first step. Specifically, it examined whether an autoencoder using PointNet could extract spatial features from real-world spaces. The machine learning model was trained on ScanNet++ (Yeshwanth et al., 2023), a 3D indoor space dataset converted into point cloud data. Feature extraction was performed using two types of point cloud data: one set of point cloud data not used in the training of ScanNet++, and another set obtained by scanning real indoor spaces in 3D. Spatial feature extraction was evaluated by comparing the shape of the input point cloud with the restored output point cloud, as well as by assessing the distance error.

DATASET CREATION FOR SPATIAL FEATURE EXTRACTION

In this study, the ScanNet++ indoor spatial 3D dataset was used for training. ScanNet++ is a large-scale dataset that integrates the 3D geometry of a room with color information, ensuring high-quality and high-accuracy geometric data. The dataset contains over 280,000 images and 3.7 million RGBD frames across 460 scenes, captured using a high-end laser scanner with sub-millimeter resolution. For this study, mesh data from 323 scenes were used to train a machine-learning model. The maximum room dimensions in the dataset are 22 m (width, x), 22.5 m (depth, y), and 6.3 m (height, z).

Conversion of 3D Objects to Point Cloud Data

To convert 3D objects (mesh data) into point cloud data, the Poisson disk sampling method (Yuksel, 2015) was initially applied. This method controls the minimum distance between sampled points, ensuring that it does not fall below a specified value. As a result, all points in the sampled cloud were separated by a uniform distance, enabling sampling with high spatial uniformity (Figure 1).

Conditions for Dataset Creation

The conditions for dataset creation were as follows: the number of input points was 10,000; the number of data points was 323; the dimensions of the latent variable were 128; and the number of layers was three.

During training, normalization was applied within the range of 0–1 based on the maximum and minimum lengths of the interior space for each dataset. The dataset was divided into 90% training data for weight optimization and 10% test data to evaluate generalization performance.

CREATING A POINTNET-BASED AUTOENCODER

Creating an Autoencoder

An autoencoder is a neural network algorithm designed to compress (encode) an input point cloud int a lower dimension, extract feature vectors in the latent space, and extract features from the feature vectors to restore (decode) an output point cloud similar to the input. This mechanism is commonly used in applications such as image denoising, anomaly detection, clustering, and data generation. For feature extraction using a 3D point cloud, the PointNet network structure was employed as an encoder to extract feature vectors.

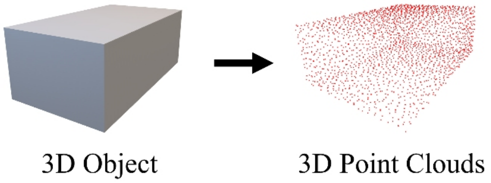

Figure 1: Conversion from a 3D object to 3D point clouds.

PointNet

Point Net is a deep learning method for point clouds that accepts point-cloud data as direct input. The 3D point clouds lack an inherent order or grid structure for any of their data elements. As shown in Figure 2, even if two points in the 3D point cloud are swapped, the overall shape of the cloud remains unchanged. This type of data is referred to as out-of-order data, which is difficult to handle with traditional deep learning methods. PointNet addresses this challenge by introducing symmetric functions that ensure the output remains invariant to the order of the input data. The PointNet architecture combines a shared Multi-Layer Perceptron (MLP) and max pooling. In shared MLP, the same MLP is applied to each point along the channel direction. Let $f(p,\ \theta)$ (where p is a 3D point and θ is a weight parameter of MLP) be a shared MLP. For example, when a 3D point cloud $(p_1,\ p_2,\ \cdots,\ p_i,\ \cdots,\ p_j,\ \cdots p_n)$ is input, the output is $(f(p_1), f(p_2), \cdots, f(p_i), \cdots, f\left(p_j\right), \cdots, f(p_n))$. Max pooling is then used to aggregate features from all points in the point cloud, with this pooling operation applied channel-by-channel. By using the maximum value as the pooling function, the result remains unchanged regardless of the input point order, ensuring that the output is independent of point order. As described, the combination of shared MLP and max pooling generates the same output regardless of point order, enabling the construction of a symmetric function via a neural network. The network structure is illustrated in Figure 3.

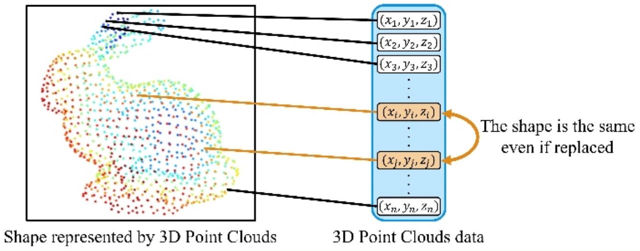

Figure 2: Unordered 3D point clouds.

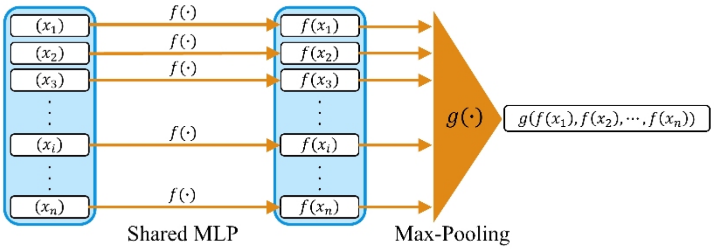

Figure 3: Symmetric function of PointNet.

Machine Learning Model Structure

The structure of the machine learning model and its hyperparameters were set based on the PointNet-based autoencoder proposed by Achlioptas et al. (2018). Figure 4 presents the structure of the machine learning model, with a three-layer example. The encoder consists of three 1D convolutional layers with batch normalization, followed by the application of the ReLU activation function after each layer. Convolution was performed on the coordinates and features of each input point using a shared weight across all points. Max pooling was then applied to aggregate the global features of the point cloud, producing a feature vector. The decoder consisted of three fully connected layers, excluding the output layer, with the ReLU function applied after each layer. The model was trained to minimize the Chamfer Distance between the input and output point clouds. When the shapes of the input and output point clouds align, appropriate feature extraction is achieved. The loss function for training was the Chamfer Distance, the optimization method was Adam, the batch size was 32, and the learning rate was set to 0.0005, with training conducted over 500 epochs.

EVALUATION OF SPATIAL FEATURE EXTRACTION

Evaluation Methods

Two types of data were used for the evaluation: ScanNet++ point cloud data not used in training the machine learning model and indoor space point cloud data obtained by 3D scanning of the real space. The indoor space point cloud data from ScanNet++ are shown in Figure 5. The real-space indoor point cloud data were acquired using an iPad Pro 2nd generation (Apple Inc.), equipped with a direct flight (dToF) LiDAR sensor and a 3D scanning application (Scaniverse). The dToF method measures distance by detecting the time difference between the light emitted from the source and the light reflected from the object, until it reaches the sensor. The acquired point cloud data were exported as a 3D object in Scaniverse, then sampled to 10,000 points (the required number of input points) using the Poisson Disk Sampling method. Figure 6 illustrates the process from data acquisition to sampling. The interior space is approximately 5.2 m wide, 2.8 m high, and 8 m deep.

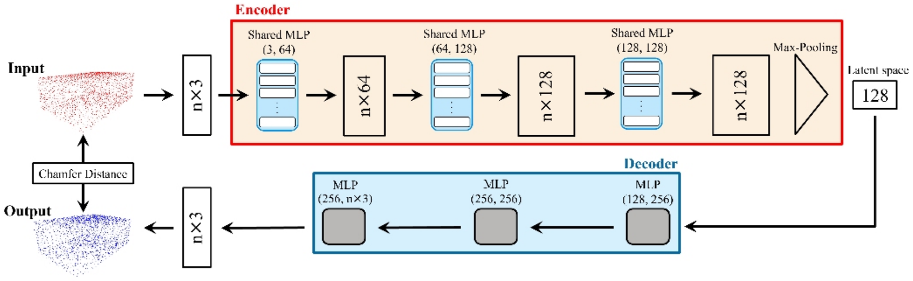

Figure 4: Structure of the autoencoder machine learning model.

The point cloud data were normalized to the range of 0–1 based on the maximum and minimum dimensions of the rectangular point cloud data in the dataset and then input into a trained autoencoder for restoration. The normalization parameters were later used to denormalize the restored output point cloud.

To quantitatively evaluate the precision of the restoration, the input and restored output point clouds were visualized, their shapes compared, and the distance error calculated. The input point cloud of the autoencoder was the source point cloud S, and the output point cloud after restoration was the target point cloud T. Mapping was performed using the kd-tree method, which is the nearest-neighbor search method. Next, the distance between the points was calculated using the mean squared error (MSE) [m^2], as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} \|p_{S_i} - q_{T_i}\|_2^2[m^2] \tag{1}$$

Here p_{Si} denotes the i-th coordinate vector of the source point group, q_{Ti} denotes the i-th coordinate vector of the target point group, and n denotes the overall number of points. Figure 7 shows the sequence of the evaluation methods.

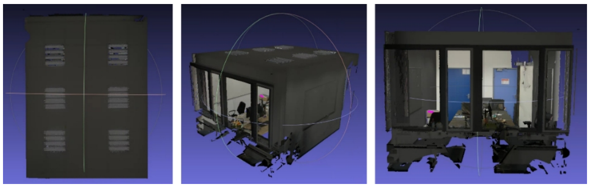

Figure 5: ScanNet++ indoor spatial point cloud data for evaluation.

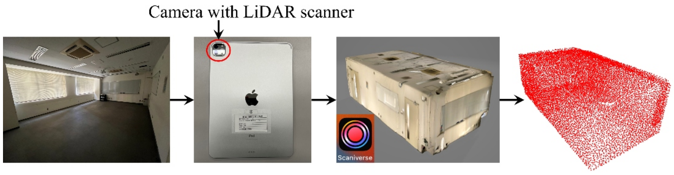

Figure 6: Flow from acquisition to sampling of indoor spatial point cloud data.

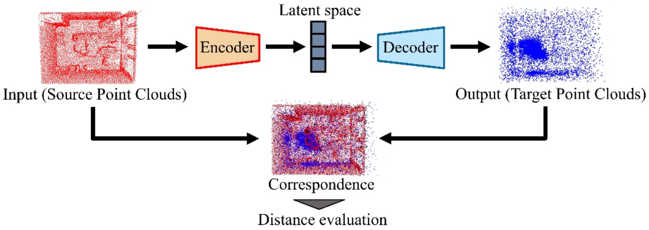

Figure 7: Flowchart of evaluation methods.

Evaluation Results

The results of visualizing the shapes of the input and output point clouds and the distance error (MSE), are presented. Figure 8 shows the results for the ScanNet++ data. These results indicate that the general shapes of the interior space, such as walls and floors, were consistent, and spatial features were successfully extracted; however, interior furniture and other objects were omitted. Figure 9 shows the results for the indoor space data. The results show that the general shapes of the walls and floors were consistent in some areas, but scattered in others. A comparison of the MSE between the ScanNet++ and indoor space data showed that the error for the indoor space data was larger than for the ScanNet++ data.

DISCUSSION

Visualization and MSE Results

Both the ScanNet++ and indoor space data were output in a rectangular shape, with the general shapes of the walls and floors being consistent. This indicates that spatial features could be successfully extracted; however, interior furniture and other objects were omitted. This suggests that the model can extract the global shape of an interior space by focusing on the walls and floors, while disregarding the finer details of the interior through training with realistic point cloud data.

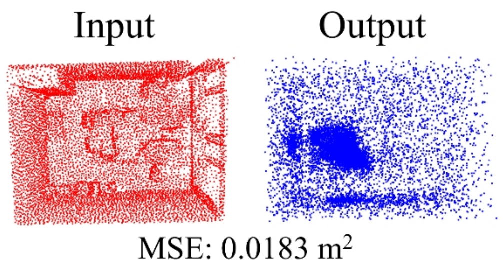

Figure 8: Evaluation results of ScanNet++ data.

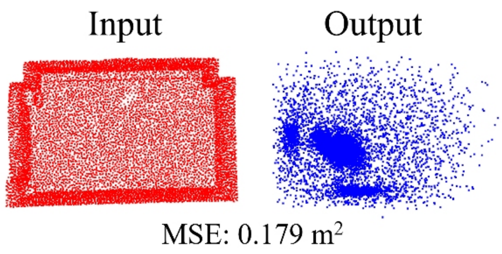

Figure 9: Evaluation results of indoor space data.

Two possible reasons for this outcome are that first, the training data were sampled using the Poisson Disk Sampling method, which maintains a constant distance between points across the entire room. This sampling approach may have led to the exclusion of smaller objects like furniture. Second, the walls and floors share similar shapes across different rooms, making them easier to learn compared to the more varied interior furniture Therefore, feature extraction primarily focused on the global shape, omitting the interior details of the space.

To reproduce the entire room, including both walls, floors, and furniture, future work could involve classifying and separating these components before performing feature extraction. Specifically, after classifying walls, floors, and furniture (as shown in Figure 10), the number of points for the furniture should be increased relative to the walls and floors. Using an autoencoder with PointNet to extract features for each object separately could improve the representation of the entire room.

The PointNet used in this model generates a single vector representing the entire point cloud's features. To ensure the features are independent of point order, the entire point cloud's features were aggregated through global pooling. However, this approach only processes the data as a whole or as individual points, making it difficult to capture local features effectively. Neural networks in image processing have successfully captured local features through convolution, suggesting that similar techniques could enhance 3D shape extraction. By focusing on local shapes around each point, it may be possible to improve feature extraction beyond the capabilities of pure PointNet. Thus, we are considering the use of point-cloud convolutional methods, such as Edge Conv (Wang et al., 2019), for future training.

Feature Extraction for Different Room Shapes

Both ScanNet++ and indoor space data featured rooms with corners and rectangular shapes. However, real-world rooms come in various shapes, such as elliptical and triangular. To understand the limitations of the model, we created a 3D object of a room with an elliptical shape using Blender, a 3DCG software bank, as shown in Figure 11. The point cloud data were converted and feature extraction was performed using the PointNet-based autoencoder.

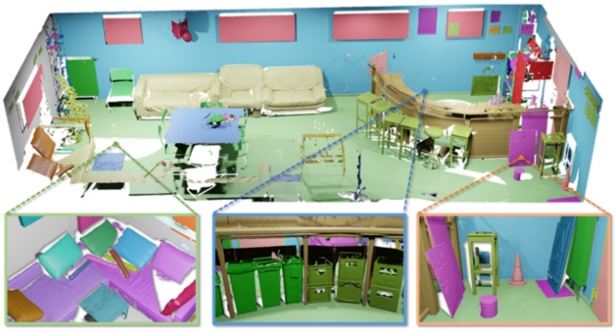

Figure 10: Example of region classification (adapted from Yeshwanth et al., 2023).

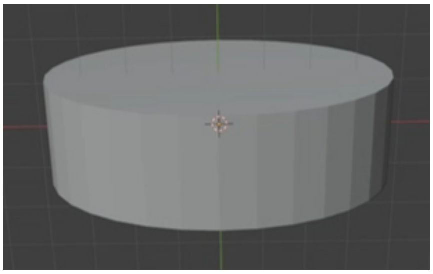

Figure 11: Creation of an elliptical-shaped interior space using Blender.

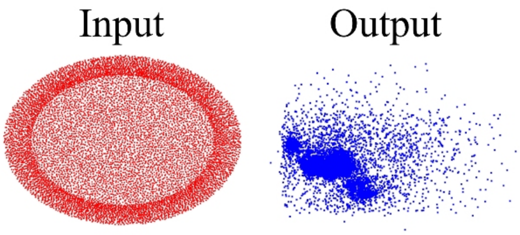

Figure 12: Results of the evaluation of interior space data with an elliptical shape.

The results, shown in Figure 12, indicate that the lower right portion of the output point cloud accurately extracted features, but other regions were scattered, with lower feature extraction accuracy compared to the ScanNet++ data. Therefore, the model developed in this study is more suitable for analyzing rectangular room shapes, but not for more irregular room geometries.

CONCLUSION

This study aimed to automatically construct a virtual space that reflects the spatial shape of a real environment, focusing on the global shape of indoor spaces without interior details. We constructed a PointNet-based autoencoder for feature extraction, training the model on ScanNet++, a 3D indoor space dataset converted into point cloud data. Feature extraction was performed using two types of point cloud data: one set that was not used for training ScanNet++, and another obtained by 3D scanning a real indoor space. The feature extraction was evaluated by comparing the input and restored output point clouds, along with calculating the distance error. The results showed that both the ScanNet++ and indoor space data were output as rectangular shapes, with the general shapes of the walls and floors consistent, suggesting the extraction of spatial features. However, interior furniture and other objects were omitted. This suggests that the model can successfully extract the global shape of interior spaces through training with real-world point cloud data while excluding interior details.

Future work will focus on creating an autoencoder that extracts features by focusing on the local shape around each point using point-cloud convolution methods. Additionally, we plan to perform feature extraction after classifying interior elements, such as furniture and walls, within the indoor space.

REFERENCES

Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L. (2018) “Learning Representations and Generative Models for 3D Point Clouds”, proceedings of the 35th International Conference on Machine Learning, pp. 40–49.

Charles R. Qi, Hao Su, Kaichun Mo, and Guibas, L. J. (2017) “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”, proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660.

Ishizaka, N., Osawa, Y., Watanuki, K., Kaede, K., and Muramatsu, K. (2018) “Measurement of Braking and Driving Forces during Walking on Virtual Slope”, proceedings of the Design and Systems Conference, p. 1206.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2019) “Dynamic graph CNN for learning on point clouds”, ACM Transactions on Graphics (tog), Vol. 38, No. 5, pp. 1–12.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. (2015) “3d shapenets: A deep representation for volumetric shapes”, proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920.

Yeshwanth, C., Liu, Y. C., Nießner, M., and Dai, A. (2023) “Scannet++: A high-fidelity dataset of 3d indoor scenes”, In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12–22.

Yuksel, C. (2015) “Sample Elimination for Generating Poisson Disk Sample Sets” Computer Graphics Forum, Volume 34, No. 2. pp. 25–32.