Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 1006–1016
https://doi.org/10.54941/ahfe1006913

Semantic Segmentation-Guided 3D Shape Reconstruction of Indoor Scenes Using a PointNet-Based Autoencoder

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

This study aims to automatically construct virtual spaces that faithfully reflect the geometry and object arrangement in real-world environments. As a first step, we proposed a method for the three-dimensional (3D) shape reconstruction of indoor scenes using a PointNet-based autoencoder guided by semantic information. The proposed method first segmented a 3D point cloud into semantic classes and then applied a separately trained autoencoder to each class. To validate its effectiveness, we used the ScanNet++ indoor scene dataset and our own real-world data captured using a 3D scanner, performing qualitative visual comparisons and quantitative evaluations using metrics such as Chamfer distance (CD) and Earth mover's distance (EMD). The results demonstrated that the proposed method achieved high visual fidelity and low CD error (4.23×10^{-4}) on validation data similar to the training set. Although point scattering was observed in the unseen test data, the reconstruction fidelity still showed a clear improvement over prior work. Furthermore, we analyzed the counterintuitive observation that EMD showed an opposite trend to CD and showed that this was a statistical effect arising from the difference in the number of instances used for evaluation. A potential application of this method was also identified: by limiting the target classes, furniture could be intentionally excluded and only the skeletal structure of the space could be reconstructed. Future work will explore enhancing the local feature representation by adding normal information as an input feature and improving robustness through post-segmentation noise removal.

Keywords: 3D point clouds, Semantic segmentation, PointNet, Autoencoder, Virtual space

INTRODUCTION

In recent years, the advancement of extended reality (XR) technologies, including virtual reality (VR) and mixed reality (MR), has extended their applications beyond entertainment to fields such as medicine and welfare. VR offers a high degree of expressive freedom in virtual environments because head-mounted displays (HMDs) immerse the entire field of view of users. However, its use in arbitrary settings poses challenges because it isolates users from their physical surroundings and restricts their range of motion (Ishizaka et al., 2018). By contrast, MR imposes fewer movement constraints

but is generally limited to applications such as overlaying virtual objects onto the real world. These limitations underscore the need for technologies capable of generating VR environments that dynamically reflect the real world in real time. These capabilities are particularly relevant to applications involving spatial transformations and systems that support human activity and cognition, including indoor space design assistance and spatial awareness enhancement. Nonetheless, automating the construction of highly expressive virtual environments—capable of manipulations such as tilting or expanding space while preserving user awareness of physical surroundings—remains a difficult challenge because of the complexity of the underlying processes.

Three-dimensional objects, typically represented as mesh data comprising vertices and faces, are essential for constructing virtual environments. Recently, the automatic generation of 3D objects from point clouds has become an active area of research, with many generative models proposed based on deep learning methods such as generative adversarial networks (GANs). Numerous models, including the L-GAN (Achlioptas et al., 2018), employ PointNet (Qi et al., 2017) as an encoder for feature extraction, thereby leveraging its ability to preserve the permutation invariance of point cloud data. However, these models are typically evaluated using large-scale datasets of isolated CAD objects (Wu et al., 2015) that fail to capture the complexity of real-world indoor environments. Spatial objects such as walls, floors, and furniture exhibit characteristics distinct from those of individual CAD models; hence, a direct transfer of learned features is difficult.

Therefore, this study aims to achieve a high-fidelity 3D reconstruction of entire indoor scenes. We proposed a PointNet-based autoencoder guided by semantic segmentation. In this approach, a scene was first segmented into semantic classes and then an optimized autoencoder was applied to each class to reconstruct the individual object instances. The proposed method was trained and evaluated using the large-scale indoor dataset, ScanNet++ (Yeshwanth et al., 2023).

DESIGNING A SEMANTIC SEGMENTATION MODEL

Overview of the Framework

The framework proposed in this study is illustrated in Figure 1. In Step 1, semantic segmentation was performed on the indoor data to partition the scene into class-specific regions. In Step 2, an autoencoder was constructed for each corresponding class. Previous research that simultaneously trained on entire scenes faced the challenge of insufficiently capturing class-specific feature representations (Miki et al., 2025). Our method addressed this by training a dedicated autoencoder for each class and enabling the precise extraction and reconstruction of features for distinct regions such as walls, ceilings, floors, and tables. This approach aims to improve the segmentation accuracy and the expressive power of scene understanding.

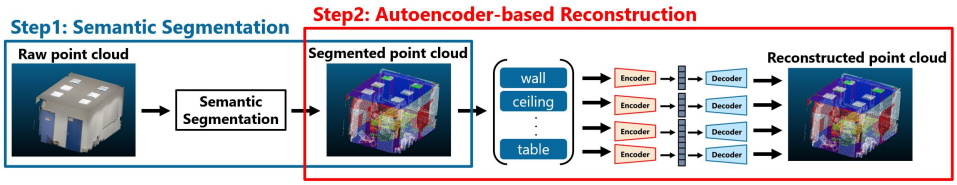

Figure 1: Semantic segmentation-guided 3D shape reconstruction framework.

Target Classes

We initially focused on four classes: wall, floor, ceiling, and table. These classes are foundational elements that define the overall structure of a scene and appear frequently. In particular, walls, floors, and ceilings are the primary components that form the boundaries of a space, making them indispensable for a geometric understanding of the entire scene. The table class was selected because it is a common object with a relatively simple shape, making it suitable for validating the effectiveness of an autoencoder. By limiting our scope to these high-frequency geometrically and semantically important classes, we could efficiently and robustly evaluate the efficacy of the proposed method.

Dataset

We used the 3D indoor scene dataset, ScanNet++, for training. ScanNet++ is a large-scale dataset that links high-quality 3D geometry with color information and is designed for various tasks including semantic segmentation (Figure 2). It currently consists of 1006 scenes with data from multiple sensors, including laser scanners, DSLR cameras, and iPhone LiDAR, along with semantic classes and instance labels. The dataset also provides an official data split for semantic segmentation, dividing it into 230 training, 50 validation, and 50 test scenes.

Figure 2: Example of a scene from the ScanNet++ dataset with semantic labels applied (adapted from Yeshwanth et al., 2023).

Training a Semantic Segmentation Model Using Point Transformer V3

We constructed a semantic segmentation model for 3D point clouds using Point Transformer V3 (Wu et al., 2024). Point Transformer V3 is a transformer-based model that efficiently extracts local and global features by applying self-attention to point cloud data. We trained the model for

100 epochs on four target classes: wall, ceiling, floor, and table. The model performance was evaluated using intersection over union (IoU) that measured the overlap between the predicted and ground-truth regions and accuracy.

The evaluation results for each class are listed in Table 1. These classes were generally classified with high accuracy. Floor and ceiling exhibited distinct geometric features, achieving high IoU scores of 0.9393 and 0.8706, respectively, and high accuracy scores of 0.9744 and 0.9297, respectively. The table class scored slightly lower, with an IoU of 0.7502 and accuracy of 0.8391, owing to its shape diversity. The wall class achieved an IoU of 0.8084 and accuracy of 0.9178. Based on these results, we constructed autoencoders for each class to perform feature extraction and reconstruction.

Table 1: Intersection over union (IoU) and accuracy of semantic segmentation for each target class.

Class	IoU	Accuracy
Wall	0.8084	0.9178
Ceiling	0.8706	0.9297
Floor	0.9393	0.9744
Table	0.7502	0.8391

3D SHAPE RECONSTRUCTION USING AN AUTOENCODER

Designing an Autoencoder

An autoencoder is a feature extraction algorithm that uses a neural network to compress an input point cloud into a low-dimensional latent space (encoding) and then reconstructs a point cloud similar to the input from that feature vector (decoding). This mechanism is used in applications such as image denoising, anomaly detection, clustering, and data generation.

In this study, we constructed a separate autoencoder for each class identified by semantic segmentation (wall, ceiling, floor, and table). This was performed to precisely extract the distinct geometric features of each class and improve the reconstruction accuracy. The input to the autoencoder was the coordinate information of the point cloud for each class, and the output was reconstructed in the same format. The Chamfer distance (CD) was used as a loss function to minimize the shape difference between the input and output point clouds. The encoder used the PointNet network structure that was designed to efficiently extract local and global geometric information from each point.

PointNet

PointNet 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 3, even when 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 that is difficult to manage using traditional deep learning methods.

PointNet addresses this challenge by introducing symmetric functions that ensure that the output remains invariant to the order of the input data. The PointNet architecture combines a shared multilayer 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 $(p_1,\ p_2,\ \cdots,\ p_i,\ \cdots, p_n)$ is input, the output is $(f(p_1),\ f(p_2),\ \cdots,\ f(p_i),\ \cdots,\ f(p_n))$. Max pooling is then used to aggregate features from all points in the point cloud, and this pooling operation is applied channel-by-channel. Using the maximum value as the pooling function, the result remains unchanged irrespective of the input point order, ensuring that the output is independent of the point order. As described, the combination of the shared MLP and max pooling generates the same output irrespective of the point order, enabling the construction of a symmetric function via a neural network. Figure 4 illustrates the network structure.

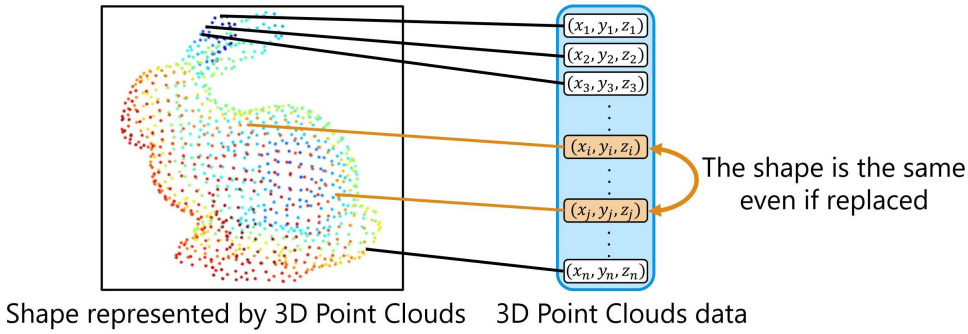

Figure 3: Unordered 3D point clouds.

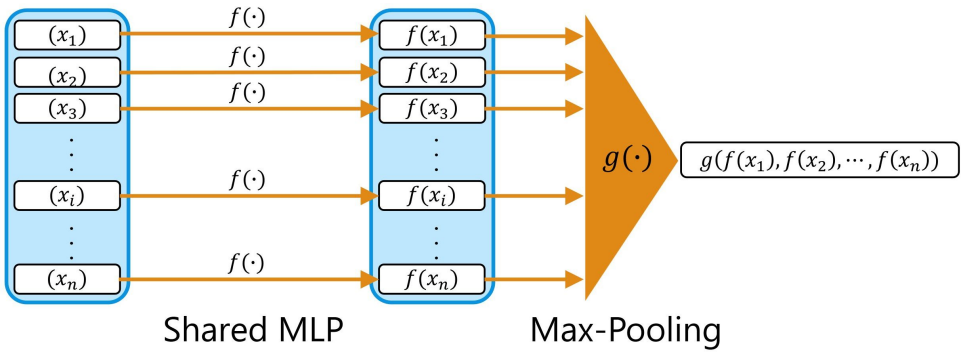

Figure 4: Symmetric function of PointNet.

Machine Learning Model Structure

The structure and hyperparameters of our machine learning model were based on a PointNet-based autoencoder (Achlioptas et al., 2018). Figure 5 illustrates the model architecture. The input point cloud size was set to 4096

points. The encoder consisted of three 1D convolutional layers, each followed by a ReLU activation function. This convolutional operation used shared weights across all the input points. Next, max pooling aggregated the global features of the point cloud to obtain a feature vector. The decoder comprised three fully connected layers, with the ReLU activation function applied to all but the output layer. The model was trained to minimize the CD between the input and output point clouds. A successful reconstruction, where the input and output shapes matched, indicated that the feature extraction was performed properly. For training, we used the CD as the loss function, the Adam optimizer, a batch size of 16, and a learning rate of 0.001 for 50 epochs.

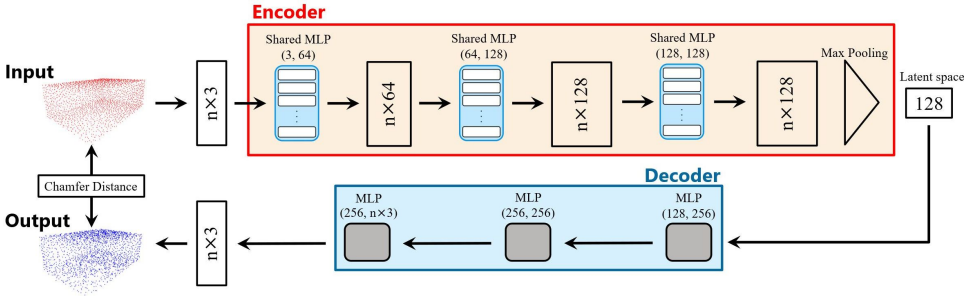

Figure 5: Structure of the autoencoder machine learning model.

EVALUATION OF 3D SHAPE RECONSTRUCTION

Evaluation Methods

To validate the effectiveness of the proposed method, we evaluated the 3D shape reconstruction performed using a combined semantic segmentation model and autoencoder. We used three types of data for evaluation: (1) validation data from the ScanNet++ dataset, (2) test data from the ScanNet++ dataset, and (3) real-world indoor point cloud data acquired using 3D scanning. This allowed us to confirm not only the performance on an existing dataset but also the generalization capability to real-world environments. The real-world data were acquired using an iPad Pro (2nd generation, Apple Inc.), equipped with a direct time-of-flight (dToF) LiDAR, and the Scaniverse application. The dToF method measures the distance to an object by detecting the flight time of light that allows the scanning of large areas in a short amount of time. The acquired data were first meshed using Scaniverse and then sampled into a point cloud for evaluation. The workflow from scanning to sampling is illustrated in Figure 6.

For evaluation, we performed a qualitative assessment by visually comparing the input and output point clouds. For a quantitative assessment, we measured the distance error between the input and output point clouds using two metrics: CD and Earth mover's distance (EMD). These are defined by the following equations (Achlioptas et al., 2018).

$$CD(X_1, X_2) = \sum_{x \in X_1} \min_{\xi \in X_2} \|x - \xi\|_2^2 + \sum_{x \in X_2} \min_{\xi \in X_1} \|x - \xi\|_2^2 \quad (1)$$

$$EMD(X_1, X_2) = \min_{\phi: X_1 \to X_2} \sum_{x \in X_1} \|x - \phi(x)\|_2 \quad (2)$$

Here, X_1 and X_2 represent the point clouds that are assumed to have the same number of points for the EMD calculation. Term ϕ represents a bijection from X_1 to X_2, and $\| \cdot \|_2$ denotes the $\mathbb{R}^3$ Euclidean distance. The CD is suitable for evaluating the overall shape reproducibility—that is, the extent to which the reconstructed shape matches the original—because it calculates the sum of distances from each point in one point cloud to the nearest point in the other point cloud. In contrast, the EMD excels at rigorously evaluating the similarity of point distributions and density by considering the optimal correspondence between the clouds. In this study, we performed a multifaceted accuracy verification using the CD to evaluate global shape reproducibility and complementarily using the EMD to assess correspondence at the point distribution level.

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

Evaluation Results

We performed qualitative and quantitative evaluations to validate the effectiveness of the proposed method.

For a qualitative evaluation, visual comparisons of the reconstruction results are shown in Figures 7–9. For the validation data (Figure 7), fine details were accurately reproduced, and an extremely high reconstruction accuracy was confirmed. However, for the test (Figure 8) and real-world data (Figure 9), point scattering was observed, and the reconstruction accuracy was limited. However, compared with our prior work that did not use semantic segmentation (Miki et al., 2025), the basic shape of the objects was preserved, demonstrating the improvement of the proposed method.

The results of the quantitative evaluation are presented in Table 2 and Figure 10. The CD for the validation data (4.23×10^{-4}) was approximately five times better than that for the test data (21.1×10^{-4}), a result that correlated with the visual evaluation. In contrast, the EMD showed a

counterintuitive result, with the test data (0.191) scoring better than the validation data (0.316). This trend was similar for the wall, floor, and table classes, with only the ceiling class exhibiting a significantly lower EMD value (Figure 10). The discrepancies in these trends between the metrics and classes are analyzed in detail in the next section.

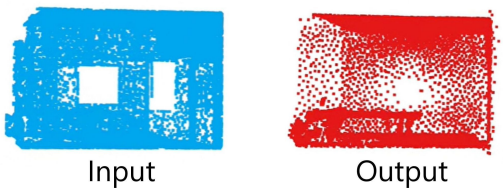

Figure 7: Reconstruction result for the validation data.

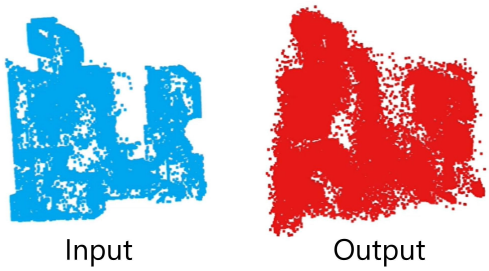

Figure 8: Reconstruction result for the test data.

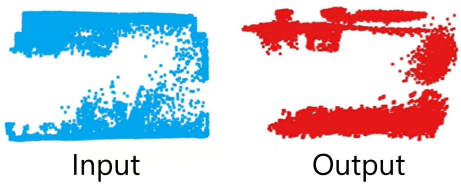

Figure 9: Reconstruction result for the real-world scanned data.

Table 2: Quantitative comparison of the 3D shape reconstruction.

	Validation	Test
Number of Instances	12	50
CD ($\times 10^{-4}$)	4.23	21.1
EMD	0.316	0.191

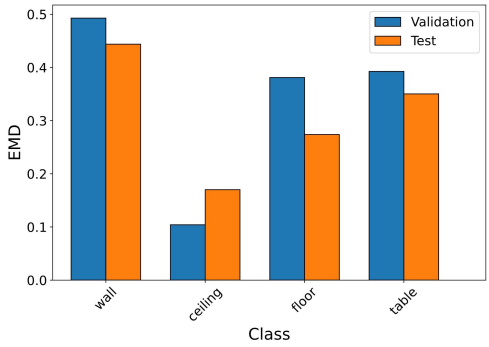

Figure 10: Earth mover's distance (EMD) evaluation results per class.

DISCUSSION

This section discusses the effectiveness, challenges, and future prospects of the proposed method based on the evaluation results presented in the previous section.

Discussion of Qualitative Results

The visual comparison results suggested that the proposed method had significant potential for structural reconstruction of indoor scenes. In particular, for the validation data from ScanNet++, large-scale structures that defined the space, such as walls, floors, and ceilings, were reconstructed with a high fidelity that they were nearly indistinguishable from the input point cloud. This indicated that the strategy of training an autoencoder for each semantic class effectively extracted and represented the global geometric features inherent to each class.

However, the results for the test and real-world data revealed several challenges in terms of reconstruction accuracy. In the test and real-world data, noise-like points were scattered around the objects, indicating that the model did not fully generalize to shape variations that were not included in the training data. The PointNet-based autoencoder aggregated the features of an entire point cloud into a single global feature vector; hence, fully preserving the local and fine-grained geometric information for objects with diverse shapes within the same class was difficult. This could cause information loss during decoding, contributing to point scattering. Furthermore, a phenomenon was observed in which the parts that existed as walls in the input data were missing from the output. As presented in Table 1, the semantic segmentation accuracy was high, suggesting that the introduction of post-processing steps such as noise removal and smoothing after segmentation could be effective in preventing these defects. Additionally, a tendency was observed for curved surfaces and complex irregularities in the input point cloud to be reconstructed as flat surfaces. This was likely because the current model learnt only from the coordinate information and did not consider normal information (that is, surface orientation and curvature).

The results of this study suggested new potential applications. By limiting the target classes of this method to walls, floors, and ceilings, indoor furniture

such as tables and chairs could be intentionally excluded, and only the skeletal structure of the space could be reconstructed. This could be a highly effective approach for applications requiring only pure spatial structures, such as creating digital twins of indoor spaces for architectural design or assessing structures prior to renovation.

Discussion of Quantitative Results

The quantitative evaluation results corroborated the insights gained from the qualitative assessment with objective numerical data while offering deeper insights into the characteristics of the evaluation metrics. As presented in Table 2, the CD results strongly correlated with the visual evaluation, and the fact that the validation data were approximately five times better than the test data clearly quantified the generalization challenge of the model. In contrast, the EMD showed a counterintuitive trend in which the test data scored better than the validation data. This was thought to be a statistical effect arising from differences in the number and scale of instances in the datasets. Specifically, because the validation data were evaluated for an average of 12 instances, a high EMD score from only a few difficult-to-reconstruct instances significantly increased the overall average. By contrast, the test data were evaluated on a larger set of 50 instances, making the overall average less sensitive to the impact of a few instances with large errors. This was considered the reason for the reversal phenomenon observed in the average values.

Furthermore, as shown in Figure 10, a more detailed analysis was possible by examining the EMD for each class. The ceiling class had a significantly lower EMD than the other classes, indicating an extremely high reconstruction accuracy. This was likely because ceilings generally had fewer occlusions and were often simple flat surfaces; hence, the model could easily learn their features. In contrast, the wall and floor classes had relatively high EMD scores. This was presumed to be due to the higher complexity and diversity of their shapes, as walls include features such as doors and windows and floors were often partially hidden by furniture. This per-class EMD analysis quantitatively confirmed that the reconstruction accuracy of the proposed method was highly dependent on the simplicity and diversity of the shape of the target object.

CONCLUSION

In this study, aiming for the automatic construction of virtual spaces that reflected real-world geometry, we proposed a 3D shape reconstruction method for indoor scenes using a PointNet-based autoencoder guided by semantic segmentation. By segmenting a scene into semantic classes and applying an optimized autoencoder to each class, we aimed for high-fidelity reconstruction.

The evaluation results demonstrated that the proposed method achieved visually accurate reconstructions and an extremely low CD error (4.23×10^{-4}) on validation data similar to the training set. This result suggested that the proposed semantic-guided approach was effective,

particularly for reproducing large-scale structures. However, challenges in the generalization and local shape representation, such as increased CD values and point scattering, were identified for unseen test and real-world scan data. This was attributed to the structure of the PointNet-based autoencoder that aggregated the features of the entire point cloud into a single global feature vector. Furthermore, by applying the characteristics of the method, the study demonstrated that by limiting the target classes to walls, floors, and ceilings, furniture could be intentionally excluded and only the skeletal structure of the space could be extracted. Additionally, this study analyzed the phenomenon in which evaluation metrics CD and EMD exhibited opposite trends and showed that this was a statistical effect arising from the difference in the number of instances used for evaluation, rather than a difference in the reconstruction quality.

In future work, we aim to improve the representation of local shapes, such as curved surfaces, by adding the normal information of each point as an input feature. We will also improve the robustness to real-world data by applying noise removal and smoothing to the point cloud after segmentation.

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.

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

Miki, T., Osawa, Y., and Watanuki, K. (2025) "Construction of a PointNet-Based Autoencoder Using a 3D Scene Dataset for Feature Extraction From Indoor Space Point Clouds Excluding Interior Details", proceedings of the AHFE International Conference, Volume 164, pp. 117–126.

Qi, C. R., Su, H., Mo, K., 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.

Wu, X., Jiang, L., Wang, P. S., Liu, Z., Liu, X., Qiao, Y., Quyang, W., He, T., and Zhao, H. (2024) "Point Transformer v3: Simpler Faster Stronger", proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4840–4851.

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", proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12–22.