# SAMBLE: Shape-Specific Point Cloud Sampling for an Optimal Trade-Off Between Local Detail and Global Uniformity

Chengzhi Wu[1] Yuxin Wan[1*] Hao Fu[1*] Julius Pfrommer[2]
Zeyun Zhong[1] Junwei Zheng[1†] Jiaming Zhang[1] Jürgen Beyerer[1,2]

[1]Karlsruhe Institute of Technology, Germany [2]Fraunhofer IOSB, Germany

{chengzhi.wu, zeyun.zhong, junwei.zheng, jiangming.zhang}@kit.edu,
{yuxin.wan, hao.fu}@student.kit.edu, {julius.pfrommer, juergen.beyerer}@iosb.fraunhofer.de

## Abstract

*Driven by the increasing demand for accurate and efficient representation of 3D data in various domains, point cloud sampling has emerged as a pivotal research topic in 3D computer vision. Recently, learning-to-sample methods have garnered growing interest from the community, particularly for their ability to be jointly trained with downstream tasks. However, previous learning-based sampling methods either lead to unrecognizable sampling patterns by generating a new point cloud or biased sampled results by focusing excessively on sharp edge details. Moreover, they all overlook the natural variations in point distribution across different shapes, applying a similar sampling strategy to all point clouds. In this paper, we propose a **S**parse **A**ttention **M**ap and **B**in-based **Le**arning method (termed SAMBLE) to learn shape-specific sampling strategies for point cloud shapes. SAMBLE effectively achieves an improved balance between sampling edge points for local details and preserving uniformity in the global shape, resulting in superior performance across multiple common point cloud downstream tasks, even in scenarios with few-point sampling.*

## 1. Introduction

Point cloud sampling is a less explored research area within the realm of this data representation. Traditional random sampling (RS) and farthest point sampling (FPS) remain the most commonly employed methods when sampling is required for point cloud learning. With the advancement of neural networks, several methods have emerged for point cloud sampling in a downstream task-oriented learning framework, including S-Net [7], SampleNet [14], MOPS-Net [34], etc. However, these methods first generate a new, smaller-sized point cloud as a proxy rather than directly sampling points from the original input, rendering the techniques akin to black-box neural network models with limited interpretability. Consequently, discerning geometric patterns in their qualitative results becomes challenging, as their outcomes closely resemble those obtained through random sampling. More recently, APES [46] pioneers the direction of using neural networks to learn point-wise sampling scores, with which it subsequently samples points whose scores are higher. However, with its score computation design and the Top-M sampling strategy, APES excessively focuses on local details of edge points, resulting in a deficiency in preserving good global uniformity of the input shapes. Consequently, the interpolation operation becomes impractical during the upsampling process, and the sampling quality of few-point sampling is notably subpar (see Fig. 1). In this paper, we introduce a novel point cloud sampling method that addresses the limitations of prior approaches, aiming to achieve a refined balance between capturing local details and preserving global uniformity.

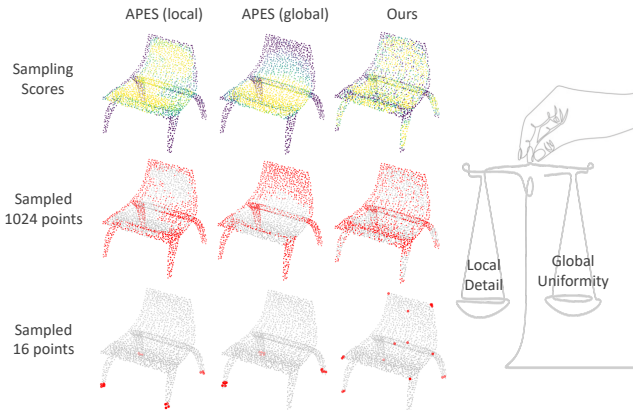


Figure 1. Our method achieves an improved trade-off between sampling local details and preserving global uniformity.

The concept originates from rethinking the mathematical characteristics of local details within point cloud shapes. Typically, these local details are represented by edge points

*Equal contribution. †Corresponding author.

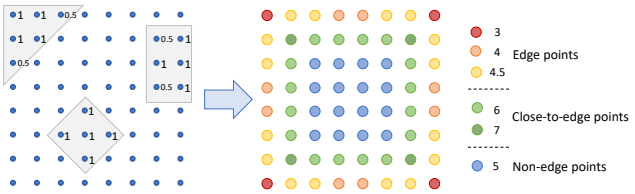


Figure 2. When selecting an equal number of neighbors for each point in the input point cloud, points at different positions are chosen as neighbors with varying frequencies (numbers on the right).

that define the shape's outline and sharpest features. Is there a point property that can easily distinguish between different categories, such as edge points and non-edge points? The answer is affirmative. In our investigation, we have uncovered an extremely fundamental yet crucial observation: if point $\mathbf{p}_i$ is one of the $k$-nearest neighbors of point $\mathbf{p}_j$, it does not necessarily imply that $\mathbf{p}_j$ is also among the $k$-nearest neighbors of $\mathbf{p}_i$. Consequently, it leads to the conclusion that the *frequency of each point being chosen as a neighbor* exhibits variation across a single point cloud.

We explore and demonstrate the importance of this point property with a simple example as illustrated in Fig. 2. Assume the input point cloud is a simple grid. When selecting 5 neighbors for each point, all three possible cases are given on the left (center point is self-contained as a neighbor). Note that in the triangular and rectangular cases, they each has a "quantum-entangled" twin point pair, in which two points share the possibility of being chosen as the neighbor. While an equal number of neighbors is selected for each point in the input point cloud, points at different positions are chosen as neighbors with varying frequencies, as listed on the right of Fig. 2. From it, we can observe that in addition to the edge point and non-edge point categories, there is also another noteworthy point category of close-to-edge points. Moreover, within each category, the points can be further grouped into more sub-categories. Overall, this point property effectively captures the local characteristics of a shape, especially for shape outline and sharp details. Building on it, we propose Sparse Attention Map (SAM) and introduce new methods for computing point-wise sampling scores to effectively balance the trade-off between local and global sampling. See more details in Sec. 3.2.

On the other hand, after the point-wise sampling scores are computed, previous methods employ a Top-M sampling strategy for all point cloud shapes, which exacerbates the issue of oversampling edge points. We argue that the naïve top-M sampling strategy may not be optimal across all point cloud shapes for downstream tasks. For example, sampling more non-edge points enhances global uniformity, while sampling more close-to-edge points "thickens" the edge, both of which can potentially improve the performance on downstream tasks [46]. To address this, we introduce a novel bin-based method to explore better sampling strategies shape-specifically by leveraging all point categories. This approach enables the sampling of points with smaller sampling scores, further optimizing the local-global trade-off. As a result, our method dynamically adjusts the sampling strategy for each shape, leading to a more tailored and interpretable sampling process for improved performance.

Our main contributions are summarized as follows:

- We propose a sparse attention map that directly integrates shape local and global information at the attention map level for point cloud sampling, introducing multiple methods for computing point-wise sampling scores.
- We present a novel approach for learning bin boundaries to partition points within individual shapes, enabling shape-specific sampling strategies by incorporating additional bin tokens during the attention computation.
- Our method successfully achieves an improved trade-off between capturing local details and preserving global uniformity for the sampling process, resulting in enhanced performance both qualitatively and quantitatively.

## 2. Related Work

**Point Cloud Sampling.** Point cloud sampling is a key process in 3D data handling for simplifying high-resolution dense point clouds. Over the past decades, non-learning-based methods [8, 10, 26] have predominantly been used for point cloud sampling. While Farthest Point Sampling (FPS) [8] is the most widely used one [18, 31, 33, 49, 60], Random Sampling (RS) has also been frequently adopted [10, 32, 61]. More recently, learning-based sampling methods have shown superior performance with task-oriented training. S-Net [7] represents a pioneering work of generating new point coordinates from global representations, while SampleNet [14] introduces a soft projection operation for better point approximation. Following S-Net, multiple learning-based methods have been proposed [21, 28, 40, 41]. MOPS-Net [34] learns a transformation matrix and multiplies it with the original point cloud to generate the sampled one. By employing the attention mechanism to learn point-wise sampling scores, APES [46] captures the edge points in the input point clouds with a strong focus.

**Deep Learning on Point Clouds.** In contrast to the voxelization-based methods [12, 17, 25] and multi-view-based methods [2, 3, 16, 37], point-based methods deal directly with point clouds. The pioneer studies of PointNet [30] and PointNet++ [31] tackle point clouds through point-wise Multi-Layer Perceptrons (MLPs) and max-pooling operations. Subsequently, other research shifts focus towards constructing more efficient building blocks for local feature extraction, such as convolution-based ones [1, 18, 20, 38, 49, 50, 62] and graph-based ones [4, 15, 22, 24, 36, 42, 54, 58]. More recently, while MLP-based methods like PointNeXt [33] and PointMetaBase [19] have rekin-

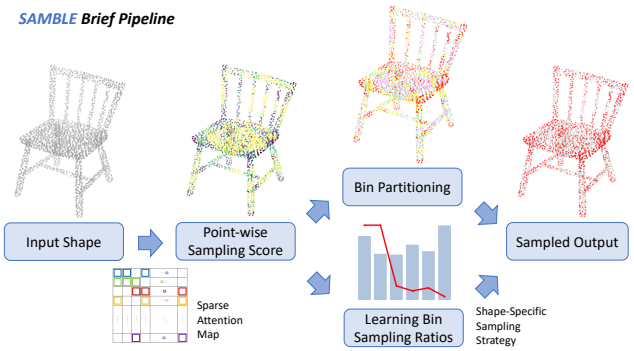


Figure 3. A brief pipeline of our proposed method SAMBLE. It learns shape-specific sampling strategies for point cloud shapes.

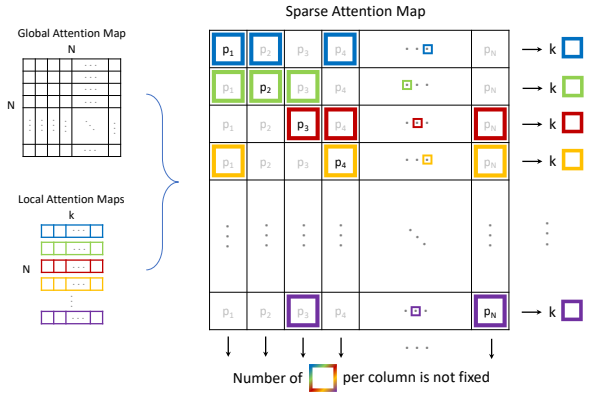


Figure 4. Sparse attention map. In each row, $k$ cells are selected based on the KNN neighbor indexes for each point. The values of other non-selected cells are all set to 0. Note that the number of cells selected within each column is variable.

dled people's interest, the application of attention mechanisms to point cloud analysis has also garnered widespread attention [9, 11, 39, 43, 45, 48, 57, 60]. For example, PT [51, 52, 60] series improve the model performance by introducing subtraction-based attention blocks, and [48] performs a large ablation study over attention module designs for point cloud processing. In addition, approaches that apply Transformers for point cloud self-supervised learning have also been proposed and explored [23, 29, 47, 57, 59].

## 3. Methodology

A brief pipeline of SAMBLE is illustrated in Fig. 3. It consists of three key steps: constructing a sparse attention map, computing point-wise sampling scores, and learning shape-specific sampling strategies through bin partitioning.

### 3.1. Sparse Attention Map

**Local and Global Attention Maps.** Both local and global attention maps are widely used in point cloud analysis. A global attention map is derived from the application of classical self-attention to point features of all points, while a local attention map concentrates on a point-centered area wherein cross-attention is specifically applied to the central point and its neighbors.

Denote $\mathcal{S}_i$ as the set of $k$-nearest neighbors of point $\mathbf{p}_i$, the local attention map for $\mathbf{p}_i$ is defined as

$$\mathbf{m}_i^l = \text{softmax}\left(Q(\mathbf{p}_i)K(\mathbf{p}_{ij} - \mathbf{p}_i)_{j\in\mathcal{S}_i}^\top / \sqrt{d}\right), \quad (1)$$

where $Q$ and $K$ stand for the linear layers applied on the query and key input, and the square root of the feature dimension count $\sqrt{d}$ serves as a scaling factor [39].

For the global attention map which is equivalent to taking all points as neighbors for each point, it is defined as

$$\mathbf{M}^g = \text{softmax}\left(Q(\mathbf{p}_i)K(\mathbf{p}_j)_{i,j\in\mathcal{S}}^\top / \sqrt{d}\right), \quad (2)$$

where $\mathcal{S}$ denotes the set of all input points.

**Sparse Attention Map.** Instead of using local or global attention maps solely, we propose sparse attention map, which combines the knowledge from both local and global information, to compute point-wise sampling scores. The idea is illustrated in Fig. 4. After obtaining the global attention map with Eq. 2, $k$-NN is employed locally to find $k$ neighbors for each point. In this case, $k$ cells are being selected in each row. However, as discussed before, please notice that each point is chosen as a neighbor with varying frequencies. This means while for each row $k$ cells are selected, for each column, the number of selected cells varies. The selected cells are then "carved" out to form the sparse attention map, with the values of other non-selected cells being set to 0. More vividly, consider the global attention map as a grid stone slab of size $N \times N$. For carve-based SAM, values of all cells are pre-computed and hidden in the slab grid cells, and only the selected cells are carved out.

### 3.2. Computing Point-wise Sampling Score

**Indexing Mode.** When sampling points, the points are indexed based on the computed point-wise sampling scores. We call the method of computing point-wise sampling scores from the full/sparse attention map as Indexing Mode. With the original full attention map, following APES, there are two possible indexing modes: (i) row standard deviation; and (ii) column sum. For a global attention map $\mathbf{M}^g$ of size $N \times N$, denote $m_{ij}$ as the value of $i$th row and $j$th column in $\mathbf{M}^g$. These two indexing modes are formulated as modes (i) and (ii) in Tab. 1. To avoid possible confusion, we use notation $\mathbf{p}_o$ to denote a point only in this subsection.

With the proposed sparse attention map, there are many other possible indexing modes. As discussed in Sec. 1, to achieve an improved trade-off between sampling edge points and preserving global uniformity, the frequency of each point being chosen as a neighbor, i.e., *the number of selected cells in each column* is the key. The following indexing modes are designed and explored for comparison:

| Indexing Mode | Attention Map | Formula | Remark |
|---|---|---|---|
| (i) Row standard deviation | Full | $a_{\mathbf{p}_o} = f_{\text{std}}(\{m_{oj} \mid j = 1, 2, \ldots, N\})$ | $f_{\text{std}}$: Computes standard deviation for a set of values |
| (ii) Column sum | Full | $a_{\mathbf{p}_o} = \sum_{i=1}^{N} m_{io}$ | |
| (iii) Row standard deviation | Sparse | $a_{\mathbf{p}_o} = f_{\text{std}}(\{m^s_{oj} \mid j \in S_o\})$ | $S_o$: Set of indices of selected cells in $o$th row |
| (iv) Row sum | Sparse | $a_{\mathbf{p}_o} = \sum_{j=1}^{N} m^s_{oj}$ | Non-selected cells are all of 0s |
| (v) Column sum | Sparse | $a_{\mathbf{p}_o} = \sum_{i=1}^{N} m^s_{io}$ | |
| (vi) Column average | Sparse | $a_{\mathbf{p}_o} = \sum_{i=1}^{N} m^s_{io}/n_o$ | $n_o$: Number of selected cells in $o$th column |
| (vii) Column square-divided | Sparse | $a_{\mathbf{p}_o} = \sum_{i=1}^{N} m^s_{io}/n_o^2$ | $n_o$: Number of selected cells in $o$th column |

Table 1. Proposed different indexing modes for computing point-wise sampling scores.

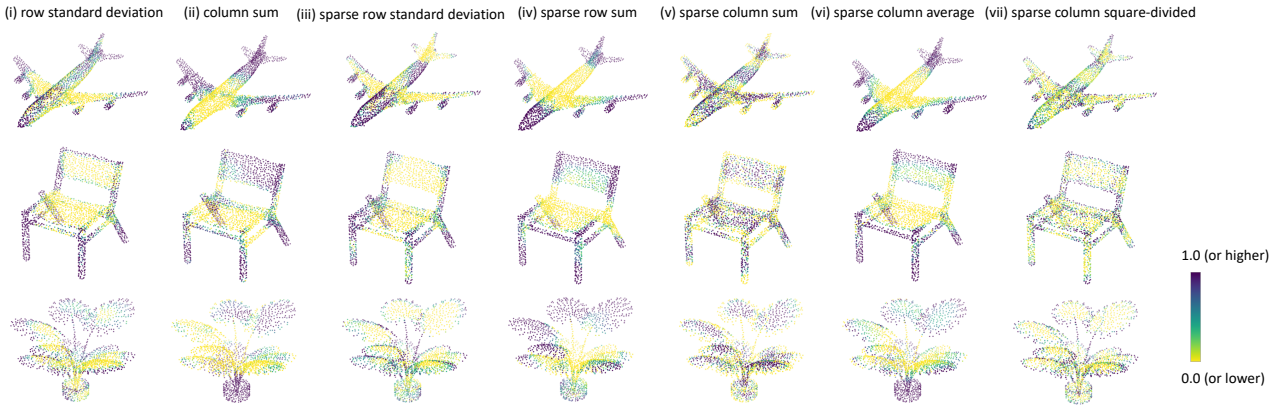


Figure 5. Point sampling score heatmaps under different indexing modes. Scores are normalized to $\mathcal{N}(0.5, 1)$ for better visualization.

(iii) sparse row standard deviation; (iv) sparse row sum; (v) sparse column sum; (vi) sparse column average; and (vii) sparse column square-divided. Again, for a sparse attention map $\mathbf{M}^s$ of size $N \times N$, denote $m^s_{ij}$ as the value of $i$th row and $j$th column in $\mathbf{M}^s$. For point $\mathbf{p}_o$, we denote the set of indexes of the selected $k$ cells (indexes of $k$-nearest neighbors) in $o$th row as $S_o$, and denote the number of selected cells in $o$th column as $n_o$. Details and respective formulas of the proposed indexing modes are listed in Tab. 1.

**Heatmap.** To analyze the behavior of each indexing mode, we train a separate model for each mode, ensuring that all other settings remain consistent. The sampling score distributions are visualized as heatmaps in Fig. 5, enhancing the interpretability of our method. From these heatmaps, we can see that both row-standard-deviation-based modes (modes i and iii) concentrate heavily on edge points. However, because they consistently prioritize thin or detailed regions, some areas may be overlooked. In contrast, modes ii and iv show less emphasis on edge points and instead distribute focus across a broader range of points, with a tendency toward other non-edge regions in a biased manner.

More interestingly, the comparison of modes v, vi, and vii, which utilize column-wise information from SAM, reveals distinct sampling preferences and strategies across different point categories. Mode v prioritizes non-edge points, mode vi emphasizes the global shape, and mode vii focuses slightly more on edge points. This is because edge points typically have a smaller number of $n_o$. Despite these differences and unique characteristics, all three modes capture the overall shape more uniformly compared to the former four. In our case, we aim to sample edge points without over-emphasizing them. For instance, when sampling detailed areas like chair legs, we want to capture some edge points without selecting them all, while also ensuring that non-edge points are sampled to preserve better global uniformity. Given this balance, we chose mode vii as the primary indexing mode for most of the experiments in the following sections. The detailed ablation study over different indexing modes is presented in Sec. 4.4.

### 3.3. Sampling with Bins

After point-wise sampling scores are computed with SAM, points are sampled according to certain rules. The simplest approach is top-M sampling, where points with the highest scores are sampled. In our case, as we aim to enhance the local-global trade-off and leverage all point categories during the sampling process, we suggest employing a bin-based sampling strategy to allow for the possible sampling of certain close-to-edge points or even non-edge points.

**Bin Partitioning.** The process begins with processing the distribution of normalized point-wise sampling scores $a_{\mathbf{p}_i}$ across the shapes within the current batch. Let $n_b$ repre-

sent the number of bins used for partitioning, from which $n_b - 1$ bin boundary values need to be derived from the distribution. During each training step, a vector $\boldsymbol{\nu}_c = (\nu_1, \nu_2, \cdots, \nu_{n_b-1})$ is computed based on the point score distribution, ensuring an equitable division of points across all shapes in the current batch. Note that while $\boldsymbol{\nu}_c$ facilitates an even division at the batch level, the points within each individual shape are not necessarily evenly partitioned according to the batch-based bin boundary values.

During the training, for the first iteration, we directly use the boundary values derived from the first batch of data as the dynamic boundary values. Subsequently, since the second iteration, boundaries are updated adaptively in a momentum-based manner:

$$\boldsymbol{\nu}_t = \gamma \boldsymbol{\nu}_{t-1} + (1-\gamma)\boldsymbol{\nu}_c\,, \tag{3}$$

where $\boldsymbol{\nu}_{t-1}$ stands for the bin partitioning boundaries used in the last iteration, and $\boldsymbol{\nu}_t$ is the updated dynamic boundaries used for the current iteration. $\gamma \in (0, 1)$ is the momentum update factor. With updated boundary values $\boldsymbol{\nu}_t$, points in each shape are divided into $n_b$ subsets of $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{n_b}\}$ based on their sampling scores.

The core idea presented here is that, instead of using pre-engineered bin boundary values, we employ adaptive learning for bin boundaries, and they are gradually learned from the entire training dataset. These values are intended to evenly partition the distribution of point-wise sampling scores across all shapes and points in the training dataset. Consequently, for each individual shape, the acquired boundary values can effectively partition its points into bins with a shape-specific strategy, capturing the unique characteristics of the shape while maintaining a degree of proximity to other shapes within the dataset.

**Tokens for Learning Bin Weights.** With points already being partitioned into bins for each shape, the next step is to learn a shape-specific sampling strategy, i.e., to learn shape-specific sampling weights for each bin. Inspired by ViT [6], VilT [13], and Mask3D [35] — which leverage additional tokens during the computation of attention maps to extract and convey information across the entire feature map or specific groups of points or pixels — we introduce additional tokens specifically for learning bin sampling weights. In our case, attention maps are computed shape-specific during the downsampling process, facilitating the learning of bin sampling weights also in a shape-specific manner.

Using the former proposed bin partitioning method, points in each shape are partitioned into $n_b$ subsets of $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{n_b}\}$. The sampling weight $\omega_j$ for bin $\mathcal{B}_j (j = 1, 2, \ldots, n_b)$ is established based on the distinctive features of each shape. Fig. 6 gives the network structure of our proposed downsampling layer and illustrates the idea of using additional tokens. $n_b$ bin tokens are introduced during the attention computation, where each token corresponds to a

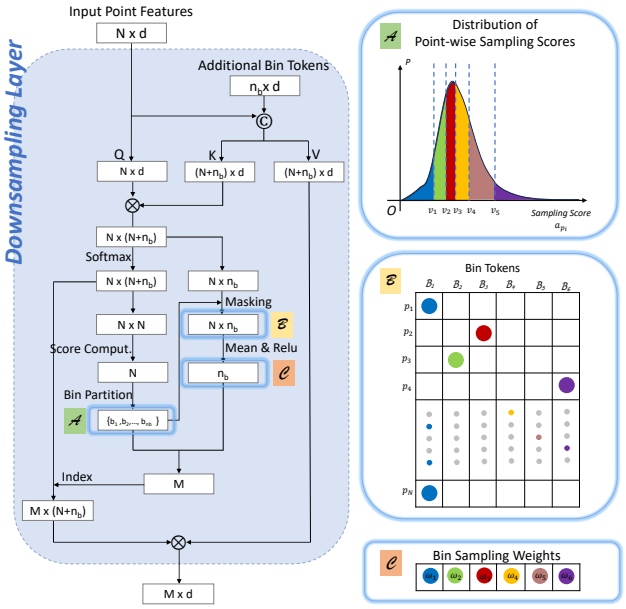


Figure 6. Network structure of our proposed downsampling layer. Block $\mathcal{A}$: Points in each shape are partitioned into $n_b$ bins. Block $\mathcal{B}$: Masking the split-out point-to-token sub-attention map. Block $\mathcal{C}$: Learned bin sampling weights.

specific bin. As shown in Fig. 6, the bin tokens are initially concatenated with the input point-wise features for $Key$ and $Value$. Subsequently, the combined features are subjected to a cross-attention mechanism with the original point-wise features as $Query$. The attention map is split into two parts of a point-to-point sub-attention map and a point-to-token sub-attention map. For the point-to-point attention map, the methods proposed in Sec. 3.1 and Sec. 3.2 are applied to it to obtain point-wise sampling scores. Note that in this case, the row-wise sum is not exactly equal to 1 but still very close to 1 since $n_b$ is of a very small quantity compared to $N$. With computed point scores, dynamic boundary values $\mathbf{v}_t$ are obtained for bin partitioning. Using the information regarding the allocation of points to respective bins, a mask operation is performed on the point-to-token sub-attention map as illustrated in Block B of Fig. 6. The sampling weights $\omega_j$ are then subsequently acquired with

$$\omega_j = \text{ReLU}\left(\frac{1}{\beta_j}\sum_{\mathbf{p}_i \in \mathcal{B}_j} m_{\mathbf{p}_i, \mathcal{B}_j}\right)\,, \tag{4}$$

where $\beta_j$ stands for the number of points in bin $\mathcal{B}_j$, and $m_{\mathbf{p}_i, \mathcal{B}_j}$ represents the element in the energy matrix corresponding to point $\mathbf{p}_i$ in row and $\mathcal{B}_j$ in column.

**In-Bin Point Sampling.** For each shape, by considering the number of points contained within bins $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{n_b})$ alongside the determined bin sampling weights $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_{n_b})$, the specific numbers of points to be sampled from each bin $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_{n_b})$

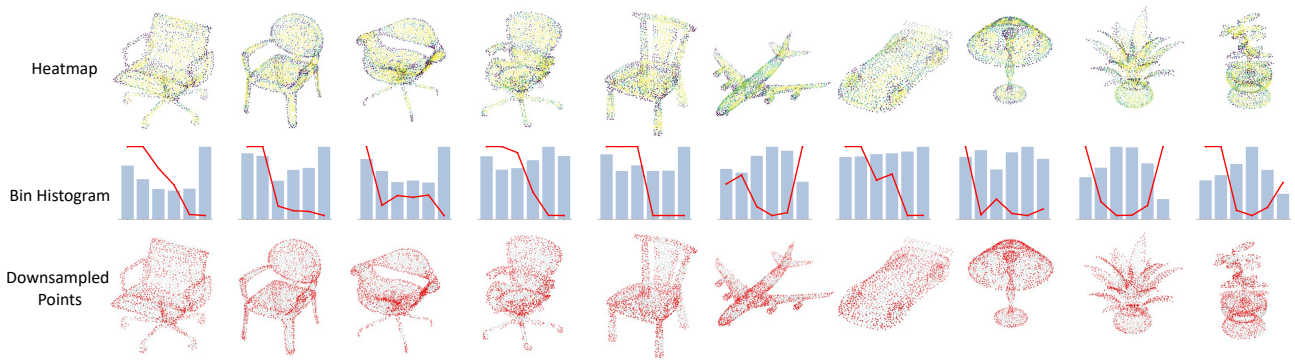


Figure 7. Qualitative results of our proposed SAMBLE. Apart from the sampled results, sampling score heatmaps and bin histograms along with bin sampling ratios are also given. All shapes are from the test set. Zoom in for optimal visual clarity.

need to be determined. Direct multiplication of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ does not yield a sum that aligns with the total number of down-sampled points $M$ required by the network structure. To address this discrepancy, a scaling method is applied to first scale bin sampling weights $\omega_j$. Furthermore, to prevent $\kappa_j$ from surpassing the available point number $\beta_j$ in any bin, any excess points are proportionately redistributed to other bins that have not been fully sampled. The detailed algorithm is presented in the supplementary materials.

Finally, within bin $\mathcal{B}_j$, $\kappa_j$ points are selected through random sampling with priors. The sampling probability $\rho_{\mathbf{p}_i}$ is obtained via softmax operation over the normalized point sampling score $a_{\mathbf{p}_i}$ with a temperature parameter $\tau$:

$$\rho_{\mathbf{p}_i} = \frac{e^{a_{\mathbf{p}_i}/\tau}}{\sum_{\mathbf{p}_i \in \mathcal{B}_j} e^{a_{\mathbf{p}_i}/\tau}}. \tag{5}$$

## 4. Experiments

Like most related works, such as S-NET [7], SampleNet [14], and APES [46], SAMBLE is specifically designed for point cloud shapes. To ensure a fair comparison, we conduct experiments using the same base network architecture as APES [46] on standard point cloud shape datasets, including ModelNet40 and ShapeNet-Part. It is important to note that point cloud sampling is not a standalone task; its effectiveness must be validated through downstream tasks.

### 4.1. Classification

**Experiment Setting.** ModelNet40 classification benchmark [53] includes 12,311 CAD models across 40 categories. For a fair comparison, we use the official train-test split, with 9,843 for training and 2,468 for testing. Points are uniformly sampled from the mesh surface and normalized to the unit sphere. Only 3D coordinates are used as input, with random scaling, rotation, and shifting applied for data augmentation. We use $n_b = 6$ bins for point partitioning. The momentum update factor $\gamma = 0.99$ for updating bin boundary values. The temperature parameter $\tau = 0.1$.

| Method | Cls. | Seg. | |
|---|---|---|---|
| | OA (%) | Cat. mIoU (%) | Ins. mIoU (%) |
| PointNet [30] | 89.2 | 80.4 | 83.7 |
| PointNet++ [31] | 91.9 | 81.9 | 85.1 |
| SpiderCNN [55] | 92.4 | 82.4 | 85.3 |
| DGCNN [42] | 92.9 | 82.3 | 85.2 |
| PointConv [49] | 92.5 | 82.8 | 85.7 |
| PT[1] [9] | 92.8 | - | 85.9 |
| PT[2] [60] | 93.7 | 83.7 | 86.6 |
| PCT [11] | 93.2 | - | 86.4 |
| PRA-Net [5] | 93.7 | 83.7 | 86.3 |
| CurveNet [27] | 93.8 | - | 86.6 |
| DeltaConv [44] | 93.8 | - | 86.6 |
| PointNeXt [33] | 93.2 | 84.4 | **86.7** |
| PointMetaBase [19] | - | 84.3 | **86.7** |
| APES (local) [46] | 93.5 | 83.1 | 85.6 |
| APES (global) [46] | 93.8 | 83.7 | 85.8 |
| SAMBLE | **94.2** | **84.5** | **86.7** |

Table 2. Classification and segmentation results on the ModelNet40 and ShapeNet-Part benchmarks. In comparison with other SOTA methods that also only use raw point cloud data as input.

**Qualitative and Quantitative Results.** Qualitative results of SAMBLE are presented in Fig. 7, including sampling score heatmaps, learned bin partitioning strategy with bin sampling ratios, and final sampled results. From it, we can observe that SAMBLE effectively samples sufficient edge points to capture the shape structure. Moreover, it maintains better global uniformity by avoiding excessive focus on edge points, particularly in those thin or sharp regions like chair legs. Logged shape bin histograms confirm the learning of shape-specific sampling strategies. More visualization results are provided in the supplementary materials, highlighting an intriguing pattern where shapes of the same category exhibit similar histogram distributions and sampling strategies. Overall, SAMBLE successfully achieves an improved trade-off between sampling edge points and preserving shape global uniformity. Quantitative results are provided in Tab. 2. Our method performs better than previous approaches and achieves state-of-the-art performance.

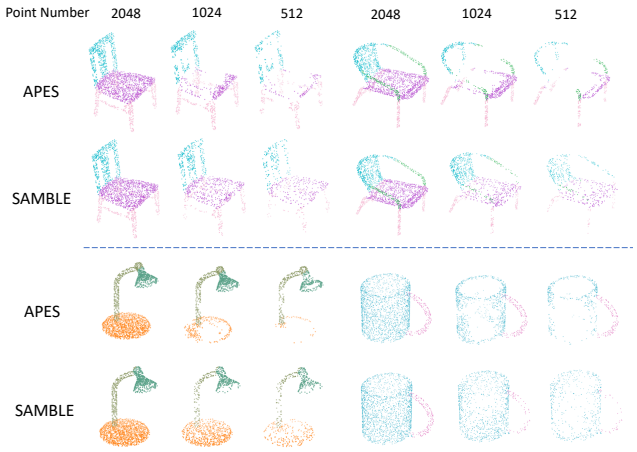


Figure 8. Segmentation results of our proposed SAMBLE in comparison with APES. All shapes are from the test set.

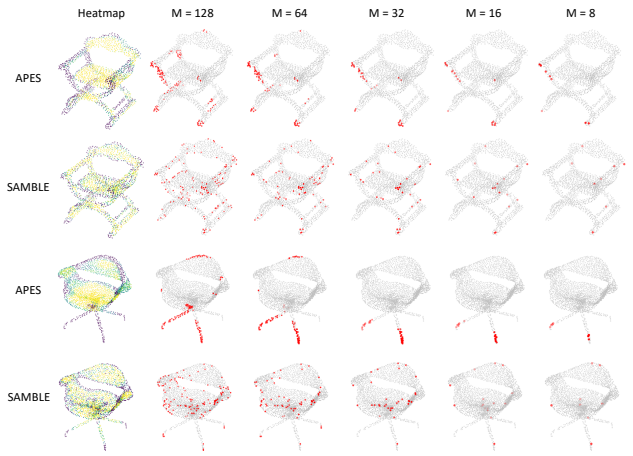


Figure 9. Sampled results of few-point sampling in comparison with APES. Zoom in for optimal clarity.

| Method | PointNeXt | | | SAMBLE | | |
|---|---|---|---|---|---|---|
| Point Number | 2048 | 1024 | 512 | 2048 | 1024 | 512 |
| Cat. mIoU (%) | 84.40 | 83.79 | 82.77 | 84.51 | 84.84 | **85.04** |
| Ins. mIoU (%) | 86.70 | 86.18 | 85.18 | 86.67 | 86.93 | **87.12** |

Table 3. Additional segmentation performances evaluated on the intermediate downsampled sub-point clouds.

## 4.2. Segmentation

**Experiment setting.** The ShapeNet-Part dataset [56] is used for 3D object part segmentation. It includes 16,880 3D models across 16 categories, with 14,006 models for training and 2,874 for testing. Each category contains 2–6 parts, totaling 50 distinct parts. We use the sampled point sets produced in [30] for a fair comparison with prior work. For evaluation metrics, we report both category mIoU and instance mIoU. We use $n_b = 4$ bins for point partitioning. The momentum update factor $\gamma = 0.99$ for updating bin boundary values. The temperature parameter $\tau = 0.1$.

**Qualitative and Quantitative Results.** Qualitative results are presented in Fig. 8, where we observe that, compared to APES, which tends to focus excessively on edge points, SAMBLE achieves a significantly improved balance between sampling edge points and preserving shape global uniformity. For example, SAMBLE demonstrates a more balanced use of non-edge points, as seen in the chair seat and lamp base, reflecting a thoughtful sampling strategy that accounts for different point categories and provides a more comprehensive representation of the overall shape. The quantitative results in Tab. 2 further highlight that SAMBLE achieves state-of-the-art performance.

For the part segmentation benchmark, we further report the performance on the intermediate downsampled sub-point clouds in Tab. 3. Additionally, results from PointNeXt [33] are also presented, which is a prominent point cloud learning method that employs FPS for downsampling. It is evident that FPS-based methods exhibit poorer performance when evaluated on intermediate downsampled sub-point clouds. In contrast, a notable observation is that SAMBLE achieves superior performance on intermediate downsampled points. This indicates that the learned sampled points contribute more to the overall performance, while the upsampling layer cannot fully reconstruct the features of the discarded points. Although SAMBLE enables the interpolation-based upsampling and it outperforms the upsampling layer used in APES (see details in Sec. 4.4), there remains potential for further improvement by designing a more meticulously crafted upsampling layer. However, this is another topic and beyond the scope of this paper.

## 4.3. Few-Point Sampling

**Experiment setting.** We further compare our sampling method to previous approaches, including RS, FPS, and more recent learning-based methods such as S-Net, SampleNet, LightTN, APES, and others. The evaluation follows the same framework as in [7, 41, 46]. First, the point cloud is downsampled to a limited number of points, and the resulting subset is then fed into a task network for evaluation. For this comparison, we use the ModelNet40 classification task with the vanilla PointNet network. All sampling methods are evaluated across multiple sampling sizes of $M$.

**Qualitative and Quantitative Results.** Quantitative results are presented in Tab. 4. Note that APES [46] uses FPS to pre-process the input into $2M$ points while we do not. For a fair comparison, additional results of APES without the pre-processing step are also tested and reported. Nonetheless, even without pre-processing, SAMBLE achieves state-of-the-art results in the few-point sampling task as the number of sampled points decreases to extremely smaller ones.

| $M$ | Voxel | RS | FPS [8] | S-NET [7] | PST-NET [40] | SampleNet [14] | MOPS-Net [34] | DA-Net [21] | LighTN [41] | APES [46] (w/ pre-pro.) | APES [46] (w/o pre-pro.) | SAMBLE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 512 | 73.82 | 87.52 | 88.34 | 87.80 | 87.94 | 88.16 | 86.67 | 89.01 | 89.91 | **90.81** | 89.81 | 90.58 |
| 256 | 73.50 | 77.09 | 83.64 | 82.38 | 83.15 | 84.27 | 86.63 | 86.24 | 88.21 | **90.40** | 86.78 | 90.18 |
| 128 | 68.15 | 56.44 | 70.34 | 77.53 | 80.11 | 80.75 | 86.06 | 85.67 | 86.26 | 89.77 | 84.87 | **90.02** |
| 64 | 58.31 | 31.69 | 46.42 | 70.45 | 76.06 | 79.86 | 85.25 | 85.55 | 86.51 | 89.57 | 79.23 | **89.81** |
| 32 | 20.02 | 16.35 | 26.58 | 60.70 | 63.92 | 77.31 | 84.28 | 85.11 | 86.18 | 88.56 | 75.63 | **89.45** |

Table 4. Comparison with other sampling methods. Evaluated on the ModelNet40 classification benchmark with multiple sampling sizes. For APES, we additionally report its performance when pre-processing is not performed for a fair comparison.

| Indexing Mode | | i | ii | iii | iv | v | vi | vii |
|---|---|---|---|---|---|---|---|---|
| Cls. | OA (%) | 93.92 | 93.78 | 93.63 | 93.66 | 93.40 | **94.11** | 94.08 |
| Seg. | Cat. mIoU (%) | 83.98 | 83.85 | 83.62 | 83.51 | 83.47 | 84.12 | **84.22** |
| | Ins. mIoU (%) | 86.16 | 85.99 | 85.74 | 85.60 | 85.49 | 86.38 | **86.46** |

Table 5. Classification and segmentation performance with different indexing modes.

| Number of Bins | | 1 | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|---|---|
| Cls. | OA (%) | 93.95 | 93.91 | 93.98 | **94.18** | 94.02 | 93.80 | 93.84 |
| Seg. | Cat. mIoU (%) | 84.22 | 84.14 | **84.51** | 84.40 | 84.19 | 83.98 | 84.36 |
| | Ins. mIoU (%) | 86.46 | 86.28 | **86.67** | 86.61 | 86.48 | 86.23 | 86.43 |

Table 6. Classification and segmentation performance with different number of bins.

Qualitative results are presented in Fig. 9. For few-point sampling, APES relies on FPS to pre-sample the input into $2M$ points due to its limitations . In contrast, our method preserves better global uniformity, allowing direct few-point sampling from the input while still achieving satisfactory sampled results, as demonstrated in Fig. 9. When sampling very few points, APES tends to concentrate on the sharpest regions, whereas our SAMBLE method preserves better global uniformity throughout the point cloud shape.

### 4.4. Ablation Study

In this subsection, our emphasis is directed toward the novel designs introduced within this paper, excluding common topics such as network width. More ablation studies and further design justifications are provided in the supplementary materials to enhance our method's interpretability.

**Different Indexing Modes.** Apart from the visualized heatmaps given in Fig. 5, we also report their respective experimental results in Tab. 5. The tests are performed using top-M as the sampling strategy. From it, we can observe that indexing modes vi and vii achieve best performances.

**Number of Bins.** As a key parameter in SAMBLE, an ablation study is performed over the number of bins $n_b$. The results are presented in Tab. 6. Remarkably, increasing the number of bins does not yield improved performance. This phenomenon is likely attributable to the subdivision of shapes into an excessive number of point categories, leading to the gradual diminishment of score disparities across the bins. In our case, $n_b = 6$ and 4 yield the best performance for the classification and segmentation tasks respectively, and we use it for the corresponding experiments.

| Upsample | Interpolation | | | Cross-Attention |
|---|---|---|---|---|
| | $K_{up} = 3$ | $K_{up} = 8$ | $K_{up} = 16$ | |
| APES (local) | 82.89 / 85.40 | 82.95 / 85.44 | 82.96 / 85.42 | 83.11 / 85.58 |
| APES (global) | 83.16 / 85.53 | 83.19 / 85.59 | 83.17 / 85.55 | 83.67 / 85.81 |
| SAMBLE | **84.51 / 86.67** | 84.35 / 86.48 | 84.31 / 86.43 | 84.36 / 86.44 |

Table 7. Segmentation results with different upsampling layers on ShapeNet-Part. The number before "/" is the category mIoU, and the number after is the instance mIoU.

**Upsampling layer.** An important aspect to highlight is the upsampling layer. Most point cloud network models employ neighbor-based interpolation [31, 33, 60] for upsampling, as FPS is used during the downsampling process. However, APES introduces a cross-attention layer for upsampling to address its limitations of overemphasizing edge points, which renders traditional neighbor-based interpolation impractical. In contrast, our method achieves an improved balance between sampling edge points and preserving global uniformity, allowing the use of interpolation operations during upsampling. An ablation study for evaluating various upsampling layers and interpolation with different $K_{up}$ values is conducted, and the results are presented in Table 7. The results show a performance drop for APES when interpolation is used in place of cross-attention, while SAMBLE demonstrates superior performance with it.

## 5. Conclusion

In this paper, a novel point cloud sampling method SAMBLE is proposed to learn shape-specific sampling strategies for point cloud shapes. Based on a sparse attention map that integrates both local and global information, multiple indexing modes are designed and explored. By partitioning the points in each shape into bins and learning respective sampling ratios for each bin, shape-specific sampling strategies are acquired for individual point cloud shapes. SAMBLE achieves an optimal trade-off between sampling local details and preserving global uniformity, resulting in improved performance on downstream tasks. For future directions, advancements in upsampling layers could further improve the model's performance. Additionally, adapting the proposed method for point cloud scenes is another promising area to explore.

## References


[1] Pyunghwan Ahn, Juyoung Yang, Eojindl Yi, Chanho Lee, and Junmo Kim. Projection-based point convolution for efficient point cloud segmentation. *IEEE Access*, 10:15348–15358, 2022. 2

[2] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Asian Conference on Computer Vision*, pages 180–196. Springer, 2016. 2

[3] Alexandre Boulch, Bertrand Le Saux, and Nicolas Audebert. Unstructured point cloud semantic labeling using deep segmentation networks. *3DOR@ Eurographics*, 3, 2017. 2

[4] Can Chen, Luca Zanotti Fragonara, and Antonios Tsourdos. GAPointNet: Graph attention based point neural network for exploiting local feature of point cloud. *Neurocomputing*, 438:122–132, 2021. 2

[5] Silin Cheng, Xiwu Chen, Xinwei He, Zhe Liu, and Xiang Bai. PRA-Net: Point relation-aware network for 3d point cloud analysis. *IEEE Transactions on Image Processing*, 30: 4436–4448, 2021. 6

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[7] Oren Dovrat, Itai Lang, and Shai Avidan. Learning to sample. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2760–2769, 2019. 1, 2, 6, 7, 8

[8] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. 2, 8

[9] Nico Engel, Vasileios Belagiannis, and Klaus Dietmayer. Point transformer. *IEEE Access*, 9:134826–134840, 2021. 3, 6

[10] Fabian Groh, Patrick Wieschollek, and Hendrik PA Lensch. Flex-convolution: Million-scale point-cloud learning beyond grid-worlds. In *Asian Conference on Computer Vision (ACCV)*, pages 105–122. Springer, 2018. 2

[11] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. PCT: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021. 3, 6

[12] Mingyang Jiang, Yiran Wu, Tianqi Zhao, Zelin Zhao, and Cewu Lu. PointSift: A sift-like network module for 3d point cloud semantic segmentation. *arXiv preprint arXiv:1807.00652*, 2018. 2

[13] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 5

[14] Itai Lang, Asaf Manor, and Shai Avidan. SampleNet: Differentiable point cloud sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7578–7588, 2020. 1, 2, 6, 8

[15] Itai Lang, Dvir Ginzburg, Shai Avidan, and Dan Raviv. DPC: Unsupervised deep point correspondence via cross and self construction. In *2021 International Conference on 3D Vision (3DV)*, pages 1442–1451. IEEE, 2021. 2

[16] Felix Järemo Lawin, Martin Danelljan, Patrik Tosteberg, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Deep projective 3d semantic segmentation. In *International Conference on Computer Analysis of Images and Patterns*, pages 95–107. Springer, 2017. 2

[17] Truc Le and Ye Duan. PointGrid: A deep network for 3d shape understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9204–9214, 2018. 2

[18] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. PointCNN: Convolution on x-transformed points. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 2

[19] Haojia Lin, Xiawu Zheng, Lijiang Li, Fei Chao, Shanshan Wang, Yan Wang, Yonghong Tian, and Rongrong Ji. Meta architecture for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17682–17691, 2023. 2, 6

[20] Yiqun Lin, Zizheng Yan, Haibin Huang, Dong Du, Ligang Liu, Shuguang Cui, and Xiaoguang Han. FPConv: Learning local flattening for point convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4293–4302, 2020. 2

[21] Yanan Lin, Yan Huang, Shihao Zhou, Mengxi Jiang, Tianlong Wang, and Yunqi Lei. DA-Net: Density-adaptive downsampling network for point cloud classification via end-to-end learning. In *2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 13–18. IEEE, 2021. 2, 8

[22] Zhi-Hao Lin, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1809, 2020. 2

[23] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision (ECCV)*, pages 657–675. Springer, 2022. 3

[24] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8895–8904, 2019. 2

[25] Daniel Maturana and Sebastian Scherer. VoxNet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015. 2

[26] Carsten Moenning and Neil A Dodgson. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory, 2003. 2

[27] AAM Muzahid, Wanggen Wan, Ferdous Sohel, Lianyao Wu, and Li Hou. CurveNet: Curvature-based multitask learning

deep networks for 3d object recognition. *IEEE/CAA Journal of Automatica Sinica*, 8(6):1177–1187, 2020. 6

[28] Ehsan Nezhadarya, Ehsan Taghavi, Ryan Razani, Bingbing Liu, and Jun Luo. Adaptive hierarchical down-sampling for point cloud classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12956–12964, 2020. 2

[29] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 604–621. Springer, 2022. 3

[30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 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)*, pages 652–660, 2017. 2, 6, 7

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 2, 6, 8

[32] Haozhe Qi, Chen Feng, Zhiguo Cao, Feng Zhao, and Yang Xiao. P2B: Point-to-box network for 3d object tracking in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6329–6338, 2020. 2

[33] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:23192–23204, 2022. 2, 6, 7, 8

[34] Yu Qian, Junhui Hou, Yiming Zeng, Qijian Zhang, Sam Tak Wu Kwong, and Ying He. MOPS-Net: A matrix optimization-driven network fortask-oriented 3d point cloud downsampling. *ArXiv*, abs/2005.00383, 2020. 1, 2, 8

[35] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 5

[36] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3693–3702, 2017. 2

[37] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3887–3896, 2018. 2

[38] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. KPConv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019. 2

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[40] Xu Wang, Yi Jin, Yigang Cen, Congyan Lang, and Yidong Li. PST-Net: Point cloud sampling via point-based transformer. In *11th International Conference on Image and Graphics (ICIG)*, pages 57–69. Springer, 2021. 2, 8

[41] Xu Wang, Yi Jin, Yigang Cen, Tao Wang, Bowen Tang, and Yidong Li. LighTN: Light-weight transformer network for performance-overhead tradeoff in point cloud downsampling. *IEEE Transactions on Multimedia*, 2023. 2, 7, 8

[42] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019. 2, 6

[43] Cheng Wen, Baosheng Yu, and Dacheng Tao. Learnable skeleton-aware 3d point cloud sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17671–17681, 2023. 3

[44] Ruben Wiersma, Ahmad Nasikun, Elmar Eisemann, and Klaus Hildebrandt. DeltaConv: anisotropic operators for geometric deep learning on point clouds. *ACM Transactions on Graphics (TOG)*, 41(4):1–10, 2022. 6

[45] Chengzhi Wu, Xuelei Bi, Julius Pfrommer, Alexander Cebulla, Simon Mangold, and Jürgen Beyerer. Sim2real transfer learning for point cloud segmentation: An industrial application case on autonomous disassembly. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4531–4540, 2023. 3

[46] Chengzhi Wu, Junwei Zheng, Julius Pfrommer, and Jürgen Beyerer. Attention-based point cloud edge sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5333–5343, 2023. 1, 2, 6, 7, 8

[47] Chengzhi Wu, Qianliang Huang, Kun Jin, Julius Pfrommer, and Jürgen Beyerer. A cross branch fusion-based contrastive learning framework for point cloud self-supervised learning. In *2024 International Conference on 3D Vision (3DV)*, pages 528–538. IEEE, 2024. 3

[48] Chengzhi Wu, Kaige Wang, Zeyun Zhong, Hao Fu, Junwei Zheng, Jiaming Zhang, Julius Pfrommer, and Jürgen Beyerer. Rethinking attention module design for point cloud analysis. In *International Conference on Pattern Recognition (ICPR)*, 2024. 3

[49] Wenxuan Wu, Zhongang Qi, and Li Fuxin. PointConv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9621–9630, 2019. 2, 6

[50] Wenxuan Wu, Li Fuxin, and Qi Shan. PointConvFormer: Revenge of the point-based convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21802–21813, 2023. 2

[51] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:33330–33342, 2022. 3

[52] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang

Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, 2024. 3

[53] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 6

[54] Qiangeng Xu, Xudong Sun, Cho-Ying Wu, Panqu Wang, and Ulrich Neumann. Grid-GCN for fast and scalable point cloud learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5661–5670, 2020. 2

[55] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. SpiderCNN: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018. 6

[56] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 7

[57] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19313–19322, 2022. 3

[58] Kuangen Zhang, Ming Hao, Jing Wang, Xinxing Chen, Yuquan Leng, Clarence W de Silva, and Chenglong Fu. Linked dynamic graph cnn: Learning through point cloud by linking hierarchical features. In *2021 27th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pages 7–12. IEEE, 2021. 2

[59] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-M2AE: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27061–27074, 2022. 3

[60] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 2, 3, 6, 8

[61] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, 2018. 2

[62] Wei Zhu, Yue Ying, Jin Zhang, Xiuli Wang, and Yayu Zheng. Point cloud registration network based on convolution fusion and attention mechanism. *Neural Processing Letters*, pages 1–21, 2023. 2