# FEAST-Mamba: FEAture and SpaTial Aware Mamba Network with Bidirectional Orthogonal Fusion for Cross-Modal Point Cloud Segmentation

**Chade Li[1,2], Pengju Zhang[1*], Bo Liu[1], Hao Wei[1], Yihong Wu[1,2*]**

[1]State Key Laboratory of Multimodal Artificial Intelligence,
Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.
lichade2021@ia.ac.cn, {pengju.zhang@ia, bo.liu@nlpr.ia, weihao2019@ia, yhwu@nlpr.ia}.ac.cn

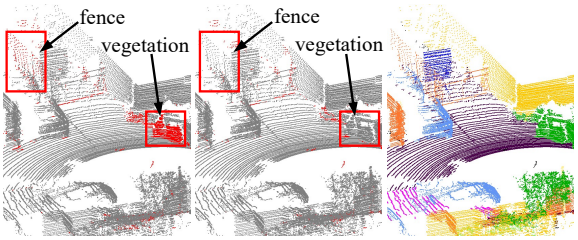
## Abstract

Point cloud segmentation has a wide range of applications in autonomous driving, augmented reality and virtual reality. Multi-modal fusion strategies have received increasing attention in point cloud segmentation recently. Despite the success, existing methods usually generate unnecessary information loss or redundancy. In this paper, we propose **FEAST-Mamba**, a novel **FEA**ture and **S**pa**T**ial aware **Mamba** network to tackle multi-modal point cloud segmentation. To exploit the complementarity between different modals, we propose a bidirectional orthogonal attention module, where features are first bidirectionally interacted with each other through cross-modal attention, and then orthogonal fusion is used to reduce feature redundancy. Furthermore, a reordering strategy is proposed for the Mamba architecture that takes into account both spatial and semantic information during cross-modal feature ordering. Experiments on indoor datasets, S3DIS and ScanNet, and outdoor datasets, nuScenes and SemanticKITTI, show that the proposed method achieves state-of-the-art performances.

## Introduction

Point clouds, a commonly used modality in computer vision, can provide reliable and accurate spatial information. Due to the lack of professional acquisition equipment and the difficulty of ground truth labeling, the number of labeled point cloud data is far less than that of other data such as images, and point cloud lacks texture information, so some point cloud segmentation methods fuse other modal data to enhance the input. PMF (Zhuang et al. 2021) projects the point cloud into a perspective view using residual-based fusion of multi-modal features. RPVNet (Xu et al. 2021) performs range image and voxel feature transfer to 3D points in each layer of the network. There are also methods that use large 2D semantic models such as Semantic-SAM (Li et al. 2023) to process the images and post-fuse the features or 2D labels with 3D data. However, these methods are incremental stacking in different modal data and lack effective constraints to reduce redundancy among fused data.

After obtaining the processed data like methods described above, some methods (Park et al. 2022; Robert, Raguet, and Landrieu 2023) employ the Transformer as the backbone for feature interaction between regions. Although their accuracy has been substantially improved, the Transformer, as a quadratic complexity network, has significant computational resources. In order to reduce computational resources and to better model long-range context, Mamba (Gu and Dao 2024) is proposed. As a linear processing network, Mamba only performs feature interactions between neighboring elements of the input, thus making it more sensitive to the input order. Consequently, if a disordered point cloud is used as input, the Mamba network may not be able to interact effectively with the input information, thus affecting the overall performance of the network. Most of the existing Mamba networks (Zhang et al. 2024b,a; Liang et al. 2024; Liu et al. 2024; Li et al. 2024; Wang et al. 2024a) simply consider the information at the 3D spatial coordinate level when sorting the unordered point clouds.




(a) Error by PTv3 (b) Error by Ours (c) Ground Truth

Figure 1: Visualization of the segmentation error maps of the proposed FEAST-Mamba and PTv3 (Wu et al. 2024a) in outdoor scenes.

To address the above problems, we propose FEAST-Mamba, a feature and spatial aware Mamba network for point cloud segmentation, where multi-modal data are fused in an orthogonal way. Specifically, according to the type of information characterized by each modality, we first divide the input into those containing spatial or texture information. We then project the augmented uni-modal data into the orthogonal space using independent projection layers and concatenate them in the order of voxels, range images, and points. Furthermore, a novel reordering strategy is proposed, which decides the order using the distribution of the data not only in 3D space but also in high-dimensional features.

*Corresponding authors

Experimental results on multiple indoor and outdoor datasets demonstrate that our method outperforms existing methods. Figure 1 shows the segmentation error maps of our method and the PTv3 (Wu et al. 2024a) in an outdoor scene, and it can be seen that we have better segmentation results in the region marked with red boxes. The main contributions of our work can be summarized as follows:

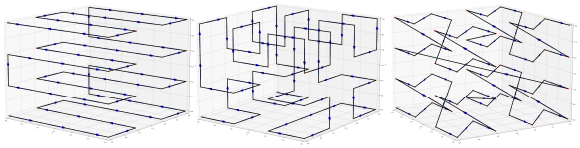
- We present an innovative cross-modal orthogonal fusion method. We augment the uni-modal features using bidirectional cross-modal attention, then project them into the same high-dimensional space via a projection layer, and apply the proposed loss term to make the two modal features orthogonal to each other.
- We propose an original reordering strategy that takes both feature similarity and coordinate proximity as the ordering criterion. For each element in the sorted sequence obtained by applying proposed reordering method, its adjacent elements are its neighbors in the original space or point cloud blocks with similar features.
- Exhaustive experiments on commonly used indoor and outdoor datasets show that our proposed method has excellent segmentation results, validating the effectiveness and robustness of our work.

# Related Work

## Deep Learning on 3D Understanding

The application of deep learning to 3D understanding tasks can be divided into two main categories, namely indirect processing and direct processing. For the indirect processing methods, the point cloud input is generally first projected into a 2D image under a specific viewpoint in a specific way (Lawin et al. 2017; Wu et al. 2018; Zhang et al. 2020), or be divided into voxels according to a specific space (Graham, Engelcke, and Maaten 2018; Choy, Gwak, and Savarese 2019), and to process the data under a new data format. However, indirect processing methods tend to result in some loss of information during data changes, which is not conducive to network processing. In contrast, the direct processing method performs feature processing directly on each input 3D point, avoiding unnecessary data loss.

The original direct processing methods use MLP to process the point cloud, such as PointNet (Charles et al. 2017), PointNet++ (Qi et al. 2017), and DeepGCN (Li et al. 2019). However, due to their limitations in effective receptive fields, it is also difficult for the network to better process the long sequence information just by applying the operation of stacking the receptive fields, such as in RandLA-Net (Hu et al. 2022). In recent years, inspired by the fact that Transformer networks, which are based on the attention mechanism, have a strong ability to model long-range context in the fields of NLP and 2D vision, some methods have also incorporated the idea of Transformer networks into point cloud processing tasks (Guo et al. 2021; Zhao et al. 2021; Park et al. 2022; Lai et al. 2022; Wu et al. 2022; Park et al. 2023; Robert, Raguet, and Landrieu 2023, 2024). Fortunately, these methods perform very well on 3D vision tasks.



(a) Serpentine-order (b) Hilbert-order (c) Z-order

Figure 2: Demonstration of reordering strategies used by existing methods (Zhang et al. 2024b,a; Liang et al. 2024; Liu et al. 2024; Li et al. 2024; Wang et al. 2024a) in 3D space.

However, Transformer networks have quadratic complexity. As a result, these methods incur significant computational overhead as the size of the input point cloud increases, thus limiting the performance of these methods.

## State Space Models

In order to alleviate the problem of excessive complexity and computational resource consumption by using Transformer network, the Mamba network (Gu and Dao 2024) was proposed. The core idea of this network is the State Space Model (SSM), which is able to model the network by connecting the inputs and outputs using potential states, and the network has strong long-range context modeling capability while the computational complexity is only linearity complexity. Several methods have applied Mamba's SSM to 2D and 3D understanding tasks, and have demonstrated its effectiveness in specific downstream tasks.

Since the point cloud is disordered compared to image and text, directly inputting the point cloud randomly into the SSM cannot show its strong long-range context modeling ability, in this regard, the existing methods always set up a reordering method in order to input the point cloud into the main network according to a specific order, and the commonly used reordering strategies are shown in Figure 2. Different reordering strategies determine the feature interaction and transfer paths in 3D space, so reordering strategies play a crucial role in the long-range context modeling capability of the network. However, for existing Mamba networks applied to 3D understanding tasks, the reordering strategies they use either only consider the distribution of the input in the 3D spatial information (Zhang et al. 2024b,a; Liang et al. 2024; Liu et al. 2024; Li et al. 2024; Wang et al. 2024a) or do not do a balanced treatment of all the samples in the input when only considering the distribution in the high-dimensional feature space (Wang et al. 2024b).

Therefore, we propose a reordering strategy with a dual perception of feature and spatial information.

## Cross-Modal Data Fusion

Due to the fact that point cloud is not easy to collect, the collection equipment is specialized, and the labeled data is difficult to obtain, there is very little ground truth data that can be used to train the network compared to data types such as 2D images and natural language. Moreover, when LiDAR point cloud is used as input alone, the network lacks access to information such as object texture, which limits the network's accuracy. In recent years, the methods of using fused

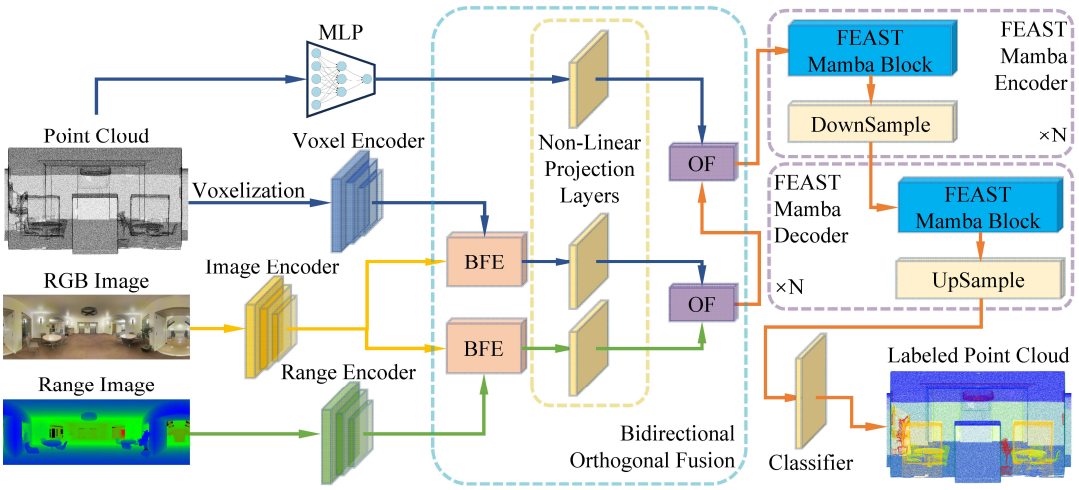



Figure 3: A general overview of the proposed network. Given the specific input data, the point cloud is firstly voxelized and then four features are obtained by applying MLP and different encoders respectively. For voxel features and range image features, we use RGB image features to enhance these two modal data representations through Bidirectional Feature Enhancement (BFE). The two enhanced modal features and the MLP features of 3D points are then sequentially applied to Orthogonal Fusion (OF) between cross-modal data features, separately. Next, the point cloud blocks with fused features are fed into our proposed FEAST Mamba block, which combines downsampling and upsampling for reordering and multi-scale feature processing.

cross-modal data as input have attracted much attention.

Existing fusion methods can be mainly categorized into early, late, and intermediate fusion methods. For early fusion, existing methods generally operate by overlaying cross-modal data at the element level (addition, multiplication, attention, etc.) or by simple concatenation. The former approach intuitively fuses multi-modal data into one modality, which may lead to an imbalance in the network's focus on different modal representations, while the latter approach may have feature redundancy between modal data containing similar information. Late fusion methods simply fuse the predictions of different branches of the network, which can easily affect the final prediction results due to noise in individual modal data. For intermediate fusion, such as RPVNet (Xu et al. 2021), the fusion operation is often required to be carried out at each layer of the network. In summary, the existing cross-modal data fusion methods lack the comprehensive consideration of complementarity and redundancy between different modal data.

Hence, we propose an orthogonal fusion method that combines bidirectional cross-modal attention mechanisms.

# Methodology

## Framework Overview

The framework of our proposed FEAture-SpaTial dual-aware Mamba network with bidirectional orthogonal fusion of cross-modal data for point cloud segmentation is shown in Figure 3. Firstly, we input the data containing spatial information (point clouds and range images) and texture information (RGB images). Then we design the corresponding encoders to extract the features respectively, where the point cloud used in points by MLP and voxelized point cloud by voxel encoder. After that, we fused the features of cross-modal data through two steps, that is, bidirectional data augmentation and orthogonal fusion. For bidirectional data enhancement, we apply cross-modal attention for bidirectional feature complementation among modal features characterizing different information; for orthogonal fusion, we constrain the projection layer by the proposed loss so that the enhanced features are orthogonal to each other in the high-dimensional space, and then concatenate the orthogonal features to reduce the redundancy between the corresponding elements of different modalities in the fused data.

After completing the above cross-modal feature fusion process, we apply the proposed FEAST Mamba block for the subsequent processing of the features in both forward and backward directions, as shown in Figure 4. In the FEAST Mamba block, we design a reordering strategy that incorporates both 3D spatial distribution and high-dimensional feature distribution into the ordering criterion.

## Multi-Modal Data Fusion

**Bidirectional Cross-Modal Feature Enhancement** Using the known LiDAR and camera internal and external parameters in the dataset, the correspondence between voxel blocks in the point cloud, pixel regions in the RGB image and range image can be determined. After obtaining the above correspondences, we apply cross-modal attention computation for bidirectional data enhancement. Assuming that there are $n_{rgb}$ regions in an RGB image, $n_{range}$ regions in a range image, and $n_{voxel}$ voxels in a point cloud, there are two steps for voxel and RGB image data enhancement:

1) To use the spatial information of the voxels in point cloud to enhance the RGB image data, first we use the features of a central voxel and its $n_{nbr}$ neighboring voxels $x_{voxels} \in \mathbb{R}^{(n_{nbr}+1)\times n_d}$ as Key and Value, and use the fea-

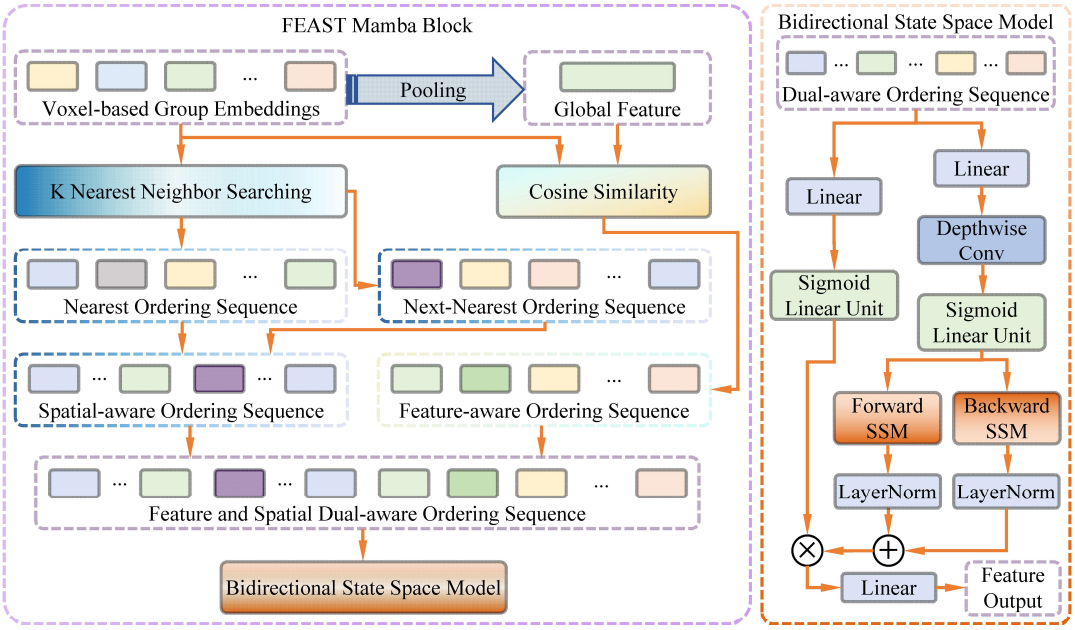



Figure 4: Framework of the proposed FEAST (Feature and Spatial-aware) Mamba block.

tures of the RGB image region corresponding to the central voxel $x_{rgb} \in \mathbb{R}^{1\times n_d}$ as Query for the cross-modal attention computation. The complete calculation procedure is shown in Equations 1 to 5:

$$Q = Linear_{q}^{\mathbb{R}^{1\times n_d}\to\mathbb{R}^{(n_{nbr}+1)\times n_d}}(x_{rgb}), \quad (1)$$

$$K, V = Linear_{k,v}^{\mathbb{R}^{(n_{nbr}+1)\times n_d}\to\mathbb{R}^{(n_{nbr}+1)\times n_d}}(x_{voxels}), \quad (2)$$

$$attn_i^{rgb} = softmax(Q \cdot K_i), \quad (3)$$

$$out^{rgb} = \sum_{i=1}^{n_{nbr}+1} (attn_i^{rgb} \times V_i), \quad (4)$$

$$x_{rgb}^{enh} = Linear^{\mathbb{R}^{(n_{nbr}+1)\times n_d}\to\mathbb{R}^{1\times n_d}}(out^{rgb}), \quad (5)$$

where $Linear$ represents the linear layer operation and its superscript represents the input and output dimensions of this linear layer, $n_d$ is the dimension of features, $attn_i^{rgb}$ is the attention value obtained by dot product computation of a region in the RGB image with the corresponding vectors of the $i$th corresponding voxel region, and $out^{rgb}$ is the aggregated feature obtained after the whole attention computation process. Finally, the features are projected back to the dimension consistent with the input features through an additional linear layer to obtain the final enhanced RGB region's feature $x_{rgb}^{enh}$.

2) Then the features of each voxel $x_{voxel} \in \mathbb{R}^{1\times n_d}$ are used as the Query, the enhanced features of the pixel region corresponding to the voxel in the RGB image and its 8 nearest neighboring pixel regions $x_{rgbs}^{enh} \in \mathbb{R}^{9\times n_d}$ are used as the Key and Value. Thus, the augmented color texture information is used to augment the features of the voxel again, similar to the calculation process described above.

The bidirectional enhancement between range images and RGB images is similar to the process described above.

**Cross-Modal Orthogonal Fusion** After obtaining the enhanced feature, we propose an orthogonal fusion method to reduce the redundancy among different modals.

Specifically, we first project the enhanced voxel features $x_{voxel}^{enh}$ and the enhanced range image features $x_{range}^{enh}$ into a high-dimensional feature space $x_{voxel}^{high}, x_{range}^{high} \in \mathbb{R}^D$ with the same number of dimensions through independent nonlinear projection layers. Then we concatenate the high-dimensional features $x_{con}^{vr} \in \mathbb{R}^{2D}$ from the two modal data to obtain the fused features, as shown in Equation 6:

$$x_{con}^{vr} = concat(x_{voxel}^{high}, x_{range}^{high}). \quad (6)$$

To increase the independence of features from different data sources, we set up $\mathcal{L}_{reg}^{vr}$ with two hyper-parameters $\lambda_1$ and $\lambda_2$ to constrain the above nonlinear projection layers $Linear_{proj}^{\mathbb{R}^{n_d}\to\mathbb{R}^D}$, as shown in Equation 7:

$$\mathcal{L}_{reg}^{vr} = \lambda_1 |x_{voxel}^{high} \cdot x_{range}^{high}| + \lambda_2 (x_{voxel}^{high} \cdot x_{range}^{high})^2. \quad (7)$$

The above constraint $\mathcal{L}_{reg}^{vr}$ encourages the projected two modal high-dimensional features to be orthogonal to each other through the independent projection layers, which in terms of helping the network to capture more different and complementary information from various modalities of data.

After obtaining the orthogonal fusion features of the two modalities through the above process, we set up an additional nonlinear projection layer with the same constraint (as

| Method | OA | mAcc | mIoU | ceiling | floor | wall | beam | column | window | door | table | chair | sofa | bookcase | board | clutter |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PointNet | - | 49.0 | 41.1 | 88.8 | 97.3 | 69.8 | 0.1 | 3.9 | 46.3 | 10.8 | 59.0 | 52.6 | 5.9 | 40.3 | 26.4 | 33.2 |
| KPConv | - | 72.8 | 67.1 | 92.8 | 97.3 | 82.4 | 0.0 | 23.9 | 58.0 | 69.0 | 91.0 | 81.5 | 75.3 | 75.4 | 66.7 | 58.9 |
| PCT | - | 67.7 | 61.3 | 92.5 | 98.4 | 80.6 | 0.0 | 19.4 | **61.6** | 48.0 | 85.2 | 76.6 | 67.7 | 46.2 | 67.9 | 52.3 |
| ST | 91.5 | 78.1 | 72.0 | 96.2 | **98.7** | 85.6 | 0.0 | 46.1 | 60.0 | **76.8** | 92.6 | 84.5 | 77.8 | 75.2 | 78.1 | 64.0 |
| SPT | 89.5 | 77.3 | 68.9 | 92.6 | 97.7 | 83.5 | **0.2** | 42.0 | 60.6 | 67.1 | 88.8 | 81.0 | 73.2 | 86.0 | 63.1 | 60.0 |
| PTv2 | 91.6 | 78.0 | 72.6 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| PTv3 | - | - | 73.4 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| PTv3+PPT | 92.0 | 80.1 | **74.7** | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Ours | **92.5** | **82.1** | **74.7** | **96.7** | 98.6 | **88.5** | 0.0 | **55.3** | 61.5 | 76.6 | **93.2** | **85.7** | **81.9** | **87.1** | **81.2** | **64.5** |

Table 1: Indoor semantic segmentation results of S3DIS (Armeni et al. 2016) on Area 5.

shown in Equation 8) on the point-level features obtained after MLP processing $x_{mlp} \in \mathbb{R}^{n_d}$. Then we project the point features into the high-dimensional space $x_{mlp}^{high} \in \mathbb{R}^{2D}$ with the same dimension as that of the voxel-range fusion features, and concatenate the features again to obtain the final multi-modal orthogonal fusion feature $x_{con}^{vrp}$, as shown in Equations 9 to 10:

$$\mathcal{L}_{reg}^{vrp} = \lambda_1 |x_{con}^{vr} \cdot x_{mlp}^{high}| + \lambda_2 (x_{con}^{vr} \cdot x_{mlp}^{high})^2, \tag{8}$$

$$x_{mlp}^{high} = Linear_{proj}^{\mathbb{R}^{n_d} \to \mathbb{R}^{2D}}(x_{mlp}), \tag{9}$$

$$x_{con}^{vrp} = concat(x_{con}^{vr}, x_{mlp}^{high}). \tag{10}$$

## Feature and Spatial Aware Reordering

Since the SSM used in Mamba is a linear processing model, it is better at working with structured data, and directly working with unstructured data does not demonstrate its powerful long-range modeling capability. For disordered data such as point clouds, an additional reordering strategy is required to reorder the point cloud blocks into an ordered one-dimensional sequence.

In order to solve the above issue, we design a feature and spatial aware reordering strategy. In terms of the block's feature, we use the features $x_{con}^{vrp}$ as the $f_{block}$. For the spatial coordinates, we perform an average pooling operation on all $N_p$ 3D point coordinates $p_i$ within the voxel to obtain $p_{block}$, as shown in Equation 11:

$$p_{block} = \frac{\sum_{p_i \in \mathcal{P}_{block}} p_i}{N_p}. \tag{11}$$

In the spatial-aware ordering strategy, we randomly select a point cloud block as the first element of the ordering sequence, and then we obtain the nearest neighbor point cloud block in the 3D space by the K-Nearest-Neighbor (KNN) algorithm for the spatial coordinates defined in Equation 11, and use it as the second element. Next, a weighted average of the identified elements in the sorted sequence is computed, where the weight of each element is the number of points contained in the corresponding block, thereby obtaining the spatial coordinates of the identified elements of the sorted sequence as a whole. The spatial coordinates of the sorted elements are then used to calculate the nearest neighbors among the remaining undefined point cloud blocks by the KNN algorithm to obtain the next element of the sorting sequence, and so on. In order to enlarge the effective receptive fields, a similar sorting process is performed, except that the criterion for determining the subsequent elements is changed from nearest neighbors to next-nearest neighbors.

In the feature-aware ordering strategy, we calculate the similarity of the high-dimensional features after multi-modal orthogonal fusion within all the point cloud blocks. Specifically, the pooling operation is performed on high-dimensional features of all input point cloud blocks to obtain the global features, and the cosine similarity between each point cloud block feature and the global feature is calculated. Then all point cloud blocks are sorted in ascending order according to the value of the corresponding similarity.

In summary, we obtain three 1D sequences from the above operations and merge them to obtain a sequence that is three times the length of the number of point cloud blocks. We then input the orthogonal cross-modal fusion features of all the point cloud blocks into the forward and backward SSM in the order of the sequence just obtained for subsequent feature processing.

## Objective Function

The overall objective function for the network training process is divided into two main parts: one of which is the classification loss consisting of the cross-entropy loss and the Lovasz-softmax loss (Berman, Triki, and Blaschko 2018) for the overall network training process as shown in Equation 12, and the other part is the orthogonal loss for only the high-dimensional projection layer of the uni-modal features, as shown in Equation 13:

$$\mathcal{L}_{class} = \alpha \mathcal{L}_{WCE} + \beta \mathcal{L}_{lovasz}, \tag{12}$$

$$\mathcal{L}_{ortho} = \gamma \mathcal{L}_{reg}^{vr} + \delta \mathcal{L}_{reg}^{vrp}. \tag{13}$$

# Experiments

## Experimental Settings

**Datasets** Following the practice of new point cloud segmentation models in recent years (Wang et al. 2024a; Wu et al. 2024a; Peng et al. 2024), we conduct experiments on five popular benchmarks, including the indoor scene datasets S3DIS (Armeni et al. 2016), ScanNet (Dai et al. 2017), as well as the outdoor scene datasets SemanticKITTI (Behley et al. 2019) and nuScenes (Caesar et al. 2020).

| Method | mIoU |
|---|---|
| OA-CNNs | 76.1 |
| KPConvX-L | 76.3 |
| OneFormer3D | 76.6 |
| SPM | 76.8 |
| ODIN | **77.8** |
| PPT+SparseUNet | 76.4 |
| PonderV2+SparseUNet | 77.0 |
| Swin3D-L | <u>77.5</u> |
| Ours | **77.8** |

Table 2: Indoor semantic segmentation results of ScanNet (Dai et al. 2017) on the validation set.

| Method | SemanticKITTI mIoU | nuScenes mIoU |
|---|---|---|
| SphereFormer | 67.8 | 79.5 |
| WaffleIron | 68.0 | 79.1 |
| 2DPASS | 69.3 | - |
| FRNet | - | 79.0 |
| PTv2 | 70.3 | 80.2 |
| OA-CNNs | 70.6 | 78.9 |
| PPT+SparseUNet | 71.4 | 78.6 |
| PTv3+PPT | <u>72.3</u> | <u>80.4</u> |
| Ours | **73.2** | **80.8** |

Table 3: Outdoor semantic segmentation results of SemanticKITTI (Behley et al. 2019) on the validation set and nuScenes (Caesar et al. 2020).

**Metrics** Following (Wang et al. 2024a; Wu et al. 2024a; Peng et al. 2024), we use the Overall Accuracy (OA), Mean Accuracy (mAcc), and mean Intersection-over-Union (mIoU) as evaluation metrics. The best indicators are highlighted in **bold** and the next best indicators are <u>underlined</u>.

**Implementation Details** Experiments are implemented on a server equipped with four Titan RTX (24G $\times$ 4 GPU memory). Consistent with PTv3 (Wu et al. 2024a), we use the AdamW optimizer with a cosine scheduler during training and set 5% of the training process as a warm-up.

## Experiments on Indoor Scene Datasets

We summarize the performance on indoor scene of FEAST-Mamba and SOTA methods (Charles et al. 2017; Thomas et al. 2019; Guo et al. 2021; Lai et al. 2022; Robert, Raguet, and Landrieu 2023; Wu et al. 2022, 2024a,b; Peng et al. 2024; Thomas et al. 2024; Kolodiazhnyi et al. 2024; Wang et al. 2024a; Jain et al. 2024; Zhu et al. 2024; Yang et al. 2023) in Table 1 and Table 2.

In the experiments in S3DIS Area-5 (Table 1), FEAST-Mamba achieves a superior result, outperforming the method (Wu et al. 2024a) that using multiple datasets as training data in terms of overall accuracy metrics. In terms of category mIoU, our method achieves the best results in nine categories, and has significant accuracy gains in object categories such as chairs, sofas, and blackboards, which are relatively small in overall scene. This suggests that our proposed feature-aware reordering strategy enables SSM to efficiently model between similar small objects at long distances.

| Method | Input | Para.(M) | FLOPs(G) | Speed(FPS) | mIoU |
|---|---|---|---|---|---|
| PTv2 | Point | 12.8 | 39.5 | 5.2 | 70.3 |
| PTv3+PPT | Point, Normal | 46.3 | 52.3 | 15.9 | <u>72.3</u> |
| Ours | Point | 11.3 | 36.8 | 19.2 | 72.0 |
| Ours | Point, RGB, Range | 36.2 | 47.2 | 15.2 | **73.2** |

Table 4: Runtime evaluation and modality comparison of SemanticKITTI (Behley et al. 2019) on the validation set.

Experiments on the ScanNet dataset (Table 2) show that FEAST-Mamba reaches the first place equal to the ODIN (Jain et al. 2024) method on this dataset. The superiority of our method can be seen in the fact that it achieves the same level of accuracy as ODIN, which employs Swin-B, the base model of Swin Transformer (Liu et al. 2021), as its backbone, while FEAST-Mamba uses an SSM with a smaller number of parameters and only linear complexity as the backbone. Furthermore, although PPT (Wu et al. 2024b) uses data from multiple datasets for training, FEAST-Mamba still has an accuracy improvement of 1.4%. Compared to other Mamba networks SPM (Wang et al. 2024a), our method has an accuracy advantage of 1.0%, which shows that our proposed reordering strategy can better apply SSM to point cloud segmentation.

## Experiments on Outdoor Scene Datasets

We showcase the results on outdoor scene of SOTA methods (Lai et al. 2023; Puy, Boulch, and Marlet 2023; Yan et al. 2022; Wu et al. 2022; Peng et al. 2024; Wu et al. 2024b,a; Xu et al. 2024) and ours in Table 3 and Table 4.

Experiments on the SemanticKITTI validation set (Table 3) show that FEAST-Mamba has an accuracy improvement of at least 0.9% in mIoU compared to existing methods. Experiments on the nuScenes datasets (Table 3) also show that FEAST-Mamba outperforms the methods including training on other datasets for mIoU. The results on parameters, FLOPs, speed and mIoU from Table 4 indicate that our approach maintains a lower model complexity while also ensuring a fast running speed. In addition, our method demonstrate competitiveness in terms of accuracy.

Moreover, we also conduct qualitative results of our method on the SemanticKITTI validation set, as shown in Figure 5. It can be seen that our method has better segmentation results in the regions marked with red boxes, showing that our method has stronger recognition ability for the categories with fewer samples (person, sign, etc.), and for the similar ground categories (sidewalk, terrain, etc.). This is because our method fuses complementary representational information between different modal data during data fusion in a better way, which enables the network to discriminate different ground categories. In addition, since our reordering strategy can establish associations between similar objects over long distances, there are also enough features for network learning for categories with fewer samples.

## Ablation Experiments

Table 5 summarizes the impact on each part of our proposed reordering strategy and of the two-part data fusion

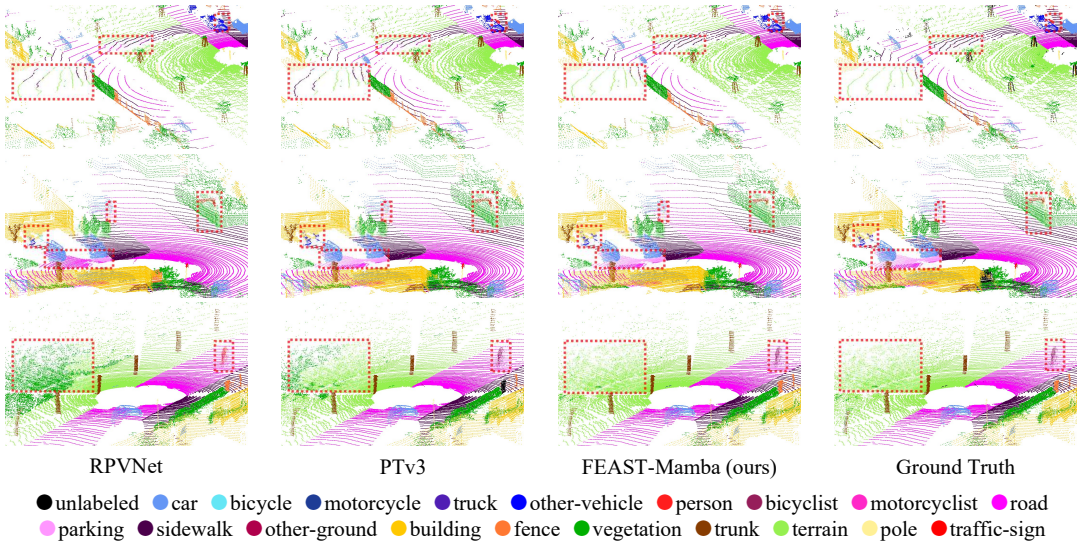



Figure 5: Qualitative results of outdoor semantic segmentation experiment on SemanticKITTI (Behley et al. 2019).

| Method | mIoU | Δ |
|---|---|---|
| w/o Feature and Spatial aware reordering | 71.2 | +0.0 |
| w/o Feature-aware reordering | 72.3 | +1.1 |
| w/o Spatial-aware reordering | 72.6 | +1.4 |
| w/o Next-Nearest reordering | 73.0 | +1.8 |
| w/o BFE with orthogonal loss | 72.4 | +1.2 |
| w/o Bidirectional Enhancement | 72.7 | +1.5 |
| w/ all | 73.2 | +2.0 |

Table 5: Ablation study of different proposed modules on SemanticKITTI (Behley et al. 2019) validation set.

| Reordering Strategy | mIoU | Δ |
|---|---|---|
| Random input | 71.2 | +0.0 |
| Serpentine curve | 71.9 | +0.7 |
| Hilbert curve | 72.3 | +1.1 |
| Z-order curve | 72.1 | +0.9 |
| Proposed FEAST dual-aware reordering | 73.2 | +2.0 |

Table 6: Comparisons of different reordering strategies on SemanticKITTI (Behley et al. 2019) validation set.

approach. It can be seen that the network accuracy is improved by 1.4%, 1.1%, and 2.0% when the feature-aware, spatial-aware, and both proposed reordering strategies are applied individually and simultaneously, respectively. This is because our proposed feature-aware reordering strategy enables similar objects that are far away in 3D space to be placed on the neighboring elements in the sorting sequence, thus constructing effective long-range contextual associations. Our proposed spatial-aware reordering strategy effectively improves the effective receptive fields of the local point cloud blocks. As for the two components of the spatial-aware reordering strategy, we also designed experiments to verify that the nearest and the next-nearest reordering strategies bring 0.4% and 0.2% improvement to the network, respectively. The comparison of the last three rows of data also shows that the cross-modal data fusion using the BFE and Bidirectional Enhancement module proposed can improve the network accuracy by 0.8% and 0.5%, which further demonstrates that our proposed fusion method can effectively exploit the complementarities between different modal data so that the augmented features have more complete representational information.

### Effect of Different Ordering Strategies

Table 6 shows the results of the analysis experiment for the proposed reordering strategy and other reordering methods. Compared to other reordering methods, our proposed reordering strategy can improve the accuracy by 0.9% to 1.3%. This is because our feature-spatial dual-aware reordering strategy makes better use of the spatial nearest-neighbor information and feature distributions of specific inputs than other methods that apply a fixed sort order for different inputs. And our strategy can adaptively construct correlations between similar objects over long distances as well as spatial nearest-neighbors for different input scenarios.

## Conclusions

We propose the FEAST-Mamba, a multi-modal point cloud segmentation network. In order to take fully exploit the complementarities between different modal and to reduce the redundancy between the data, we introduce an orthogonal data fusion method with bidirectional feature enhancement. Furthermore, our reordering strategy addresses the shortcomings of existing Mamba networks in modeling long-range context, and can efficiently construct associations between spatially proximate neighbors and blocks with similar features. Satisfactory segmentation results are obtained in experiments on both indoor and outdoor datasets, demonstrating the effectiveness and robustness of our method.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62403459, the Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems under Grant No. MAIS2024214 and Beijing Natural Science Foundation under Grant No. L241012. We thank the anonymous Program Committees and Program Chairs so much for their helpful comments and suggestions.

## References

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1534–1543.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4413–4421.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Charles, R. Q.; Su, H.; Kaichun, M.; and Guibas, L. J. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 77–85.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3070–3079.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2432–2443.

Graham, B.; Engelcke, M.; and Maaten, L. v. d. 2018. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9224–9232.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R. R.; and Hu, S.-M. 2021. PCT: Point cloud transformer. *Computational Visual Media*, 7(2): 187–199.

Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; and Markham, A. 2022. Learning Semantic Segmentation of Large-Scale Point Clouds With Random Sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8338–8354.

Jain, A.; Katara, P.; Gkanatsios, N.; Harley, A. W.; Sarch, G.; Aggarwal, K.; Chaudhary, V.; and Fragkiadaki, K. 2024. ODIN: A Single Model for 2D and 3D Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3564–3574.

Kolodiazhnyi, M.; Vorontsova, A.; Konushin, A.; and Rukhovich, D. 2024. OneFormer3D: One Transformer for Unified Point Cloud Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20943–20953.

Lai, X.; Chen, Y.; Lu, F.; Liu, J.; and Jia, J. 2023. Spherical Transformer for LiDAR-Based 3D Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17545–17555.

Lai, X.; Liu, J.; Jiang, L.; Wang, L.; Zhao, H.; Liu, S.; Qi, X.; and Jia, J. 2022. Stratified Transformer for 3D Point Cloud Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8490–8499.

Lawin, F. J.; Danelljan, M.; Tosteberg, P.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2017. Deep Projective 3D Semantic Segmentation. In Felsberg, M.; Heyden, A.; and Krüger, N., eds., *Computer Analysis of Images and Patterns*, 95–107. Cham: Springer International Publishing. ISBN 978-3-319-64689-3.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; and Gao, J. 2023. Semantic-SAM: Segment and Recognize Anything at Any Granularity. arXiv:2307.04767.

Li, G.; Müller, M.; Thabet, A.; and Ghanem, B. 2019. DeepGCNs: Can GCNs Go As Deep As CNNs? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9266–9275.

Li, Z.; Ai, Y.; Lu, J.; Wang, C.; Deng, J.; Chang, H.; Liang, Y.; Yang, W.; Zhang, S.; and Zhang, T. 2024. Mamba24/8D: Enhancing Global Interaction in Point Clouds via State Space Model. arXiv:2406.17442.

Liang, D.; Zhou, X.; Xu, W.; Zhu, X.; Zou, Z.; Ye, X.; Tan, X.; and Bai, X. 2024. PointMamba: A Simple State Space Model for Point Cloud Analysis. arXiv:2402.10739.

Liu, J.; Yu, R.; Wang, Y.; Zheng, Y.; Deng, T.; Ye, W.; and Wang, H. 2024. Point Mamba: A Novel Point Cloud Backbone Based on State Space Model with Octree-Based Ordering Strategy. arXiv:2403.06467.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Park, C.; Jeong, Y.; Cho, M.; and Park, J. 2022. Fast Point Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16928–16937.

Park, J.; Lee, S.; Kim, S.; Xiong, Y.; and Kim, H. J. 2023. Self-Positioning Point-Based Transformer for Point Cloud Understanding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21814–21823.

Peng, B.; Wu, X.; Jiang, L.; Chen, Y.; Zhao, H.; Tian, Z.; and Jia, J. 2024. OA-CNNs: Omni-Adaptive Sparse CNNs for 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21305–21315.

Puy, G.; Boulch, A.; and Marlet, R. 2023. Using a Waffle Iron for Automotive Point Cloud Semantic Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3356–3366.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Robert, D.; Raguet, H.; and Landrieu, L. 2023. Efficient 3D Semantic Segmentation with Superpoint Transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17149–17158.

Robert, D.; Raguet, H.; and Landrieu, L. 2024. Scalable 3D Panoptic Segmentation As Superpoint Graph Clustering. In *2024 International Conference on 3D Vision (3DV)*, 179–189.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. KPConv: Flexible and Deformable Convolution for Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Thomas, H.; Tsai, Y.-H. H.; Barfoot, T. D.; and Zhang, J. 2024. KPConvX: Modernizing Kernel Point Convolution with Kernel Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5525–5535.

Wang, T.; Wen, W.; Zhai, J.; Xu, K.; and Luo, H. 2024a. Serialized Point Mamba: A Serialized Point Cloud Mamba Segmentation Model. arXiv:2407.12319.

Wang, Z.; Chen, Z.; Wu, Y.; Zhao, Z.; Zhou, L.; and Xu, D. 2024b. PoinTramba: A Hybrid Transformer-Mamba Framework for Point Cloud Analysis. arXiv:2405.15463.

Wu, B.; Wan, A.; Yue, X.; and Keutzer, K. 2018. SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 1887–1893.

Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; and Zhao, H. 2024a. Point Transformer V3: Simpler Faster Stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4840–4851.

Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; and Zhao, H. 2022. Point Transformer V2: Grouped Vector Attention and Partition-based Pooling. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 33330–33342. Curran Associates, Inc.

Wu, X.; Tian, Z.; Wen, X.; Peng, B.; Liu, X.; Yu, K.; and Zhao, H. 2024b. Towards Large-scale 3D Representation Learning with Multi-dataset Point Prompt Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19551–19562.

Xu, J.; Zhang, R.; Dou, J.; Zhu, Y.; Sun, J.; and Pu, S. 2021. RPVNet: A Deep and Efficient Range-Point-Voxel Fusion Network for LiDAR Point Cloud Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 16004–16013.

Xu, X.; Kong, L.; Shuai, H.; and Liu, Q. 2024. FRNet: Frustum-Range Networks for Scalable LiDAR Segmentation. arXiv:2312.04484.

Yan, X.; Gao, J.; Zheng, C.; Zheng, C.; Zhang, R.; Cui, S.; and Li, Z. 2022. 2DPASS: 2D Priors Assisted Semantic Segmentation on LiDAR Point Clouds. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 677–695. Cham: Springer Nature Switzerland. ISBN 978-3-031-19815-1.

Yang, Y.-Q.; Guo, Y.-X.; Xiong, J.-Y.; Liu, Y.; Pan, H.; Wang, P.-S.; Tong, X.; and Guo, B. 2023. Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding. arXiv:2304.06906.

Zhang, G.; Fan, L.; He, C.; Lei, Z.; Zhang, Z.; and Zhang, L. 2024a. Voxel Mamba: Group-Free State Space Models for Point Cloud based 3D Object Detection. arXiv:2406.10700.

Zhang, T.; Li, X.; Yuan, H.; Ji, S.; and Yan, S. 2024b. Point Cloud Mamba: Point Cloud Learning via State Space Model. arXiv:2403.00762.

Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; and Foroosh, H. 2020. PolarNet: An Improved Grid Representation for Online LiDAR Point Clouds Semantic Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9598–9607.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P.; and Koltun, V. 2021. Point Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 16239–16248.

Zhu, H.; Yang, H.; Wu, X.; Huang, D.; Zhang, S.; He, X.; Zhao, H.; Shen, C.; Qiao, Y.; He, T.; and Ouyang, W. 2024. PonderV2: Pave the Way for 3D Foundation Model with A Universal Pre-training Paradigm. arXiv:2310.08586.

Zhuang, Z.; Li, R.; Jia, K.; Wang, Q.; Li, Y.; and Tan, M. 2021. Perception-Aware Multi-Sensor Fusion for 3D LiDAR Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16280–16290.