

# PointMixer: MLP-Mixer for Point Cloud Understanding

Jaesung Choe<sup>1†</sup>Chunghyun Park<sup>2†</sup>Francois Rameau<sup>1</sup>Jaesik Park<sup>2</sup>In So Kweon<sup>1</sup>KAIST<sup>1</sup>POSTECH<sup>2</sup>

## Abstract

MLP-Mixer has newly appeared as a new challenger against the realm of CNNs and transformer. Despite its simplicity compared to transformer, the concept of channel-mixing MLPs and token-mixing MLPs achieves noticeable performance in visual recognition tasks. Unlike images, point clouds are inherently sparse, unordered and irregular, which limits the direct use of MLP-Mixer for point cloud understanding. In this paper, we propose **PointMixer**, a universal point set operator that facilitates information sharing among unstructured 3D points. By simply replacing token-mixing MLPs with a softmax function, PointMixer can "mix" features within/between point sets. By doing so, PointMixer can be broadly used in the network as inter-set mixing, intra-set mixing, and pyramid mixing. Extensive experiments show the competitive or superior performance of PointMixer in semantic segmentation, classification, and point reconstruction against transformer-based methods. [Code will be released soon.](#)

## 1. Introduction

3D scanning devices, such as LiDAR or RGB-D sensors, are widely used to capture a scene as 3D point clouds. Unlike images, point clouds are inherently sparse, unordered, and irregular. These properties make standard neural network architectures [16, 40] hardly applicable. To tackle these challenges, there have been numerous ad hoc solutions, such as sparse convolution networks [6, 10], graph neural networks [35, 39, 47, 48, 50, 53], point convolution networks [29, 42, 49, 52]. Despite their structural differences, these techniques have all been designed to extract meaningful feature representation from point clouds [13].

Among existing solutions, transformer [46] appears to be particularly beneficial to extract features from point clouds. Indeed, their self-attention layer that encompasses the dense relations between tokens is specifically relevant in the context of point clouds processing. The suitability of dot-product based self-attention for point cloud understanding

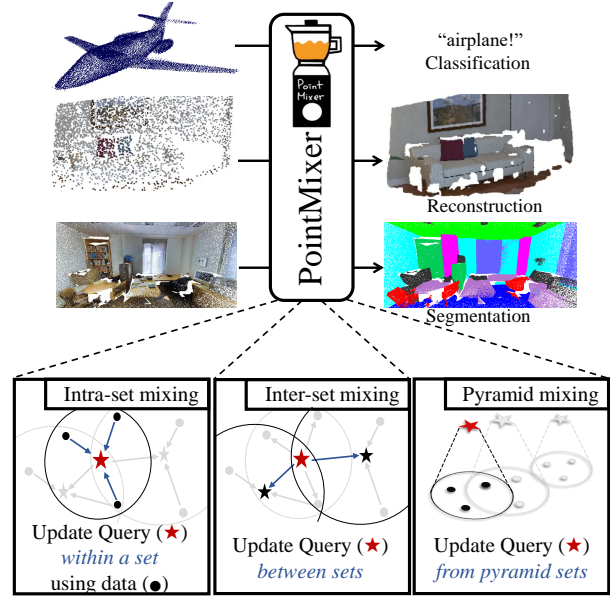


Figure 1. **PointMixer**. We introduce MLP-like architecture for various point cloud processing, which can consider various point-wise relations with larger receptive fields to "mix" information.

has recently been underlined by various studies [2, 12, 58] where significant improvements have been reported.

Beyond the well established realm of CNNs and transformer, MLP-Mixer [43] and Synthesizer [41] propose a new architecture that exclusively uses MLPs. Based on these pioneering studies [41, 43], concurrent works [5, 18, 23, 44] have been recently introduced. For instance, ResMLP [44] emphasizes that MLP-like architectures can take inputs of arbitrary length. Additionally, VisionPermutator [18] addresses the necessity of position-sensitive property in vanilla MLP-Mixer. We can also notice CycleMLP [5] which envisions the possibilities of MLPs for dense prediction tasks. Despite an increasing interest in MLP-like architectures and wonders about the necessity of self-attention, the application of MLP-like architectures for point cloud processing has not yet been explored.

In this paper, we propose **PointMixer**, the first MLP-like<sup>1</sup> architecture dedicated to point cloud understanding. To

<sup>†</sup> Both authors have contributed equally to this work.  
jaesung.choe@kaist.ac.kr / p0125ch@postech.ac.kr

this end, we construct a PointMixer network that consists of an encoder-decoder MLP architecture with PointMixer layers to mix point responses among an unstructured 3D point cloud. As illustrated in Fig. 1, PointMixer concentrates on sharing information (1) within points (intra-set mixing), (2) between point sets (inter-set mixing), or (3) points in different hierarchical sets. Intra-set mixing is close to the vanilla MLP-Mixer [43] layer. However, the idea of inter-set mixing and pyramid mixing comes from the simple retrospect: *Think about the points that treat me as their neighbor*. This is the clear difference compared to PointTransformer [58] that updates a query point only from its grouped neighbors. To validate the effectiveness of our layer design, we conduct extensive experiments on various point cloud tasks including dense prediction: semantic segmentation, point reconstruction, and object classification. These large experiment suggests that our method achieves compelling performances among recent transformer-based studies [12, 31, 31] in S3DIS, ShapeNetPart, ScanNet, ICL-NUIM, ModelNet40 datasets. Our contributions are summarized as below:

- First MLP-like<sup>1</sup> architecture for point cloud understanding, called PointMixer.
- Novel scheme to facilitate mixing of point features through intra-set, inter-set, and pyramid-mixing.
- Extensive experiments in various 3D point cloud tasks highlighting the efficacy of PointMixer compared to transformer-based studies.
- New analysis out an unsolved question: is dense token-wise interaction truly needed? [41], even for processing of unstructured point clouds?

## 2. Related work

In this section, we introduce previous studies about deep neural network for point clouds. Then, we detail the transformer and MLP-Mixer with their follow-ups studies.

**Point-based networks.** Without voxelization, point-based networks [26, 29, 34, 42, 50, 52, 58] directly process input point clouds. Since point cloud inherently has sparse, unordered, and irregular data structure; capturing proper local responses from points is desirable for the dense prediction tasks. To enforce such local estimations, PointNet++ [34] introduces a U-shaped network. To do so, this network adopts  $k$ -Neighbor Neighbor, Farthest Point Sampling, and tri-linear interpolation to build hierarchical point sets. PointTransformer [58] introduces self-attention into point cloud processing. Though PointTransformer [58] modifies basic layers using transformer, the rest of the network, such as an asymmetric and heuristic interpolation-based upsampling layer, remain analogous to PointNet++ [34].

<sup>1</sup>In [5], "MLP-like architecture" means variants of MLP-Mixer.

**Transformer and MLP-like architecture.** Transformer-based architecture has recently become a game changer in both natural language processing [8, 36, 41, 46] and computer vision [4, 9, 14, 27, 30, 38, 45, 58]. Due to high computational cost increasing quadratically with the number of tokens, several works [27, 38] conduct self-attention locally.

Recently, There exist trials to go beyond the hegemony of CNNs and transformer by introducing MLP-like architectures. Two pioneering papers, MLP-Mixer [43] and Synthesizer [41], introduce new architectural designs solely using MLPs. MLP-Mixer [43] presents a MLP-like network constituted of token-mixing MLPs and channel-mixing MLPs for visual recognition tasks. Synthesizer [41] investigates the true importance of dot-product based self-attention. Then, the paper proposes a simple but effective module that consists of MLPs and synthetic attention. In contrast to MLP-Mixer, Synthesizer conditions tokens independently (Dense Synthesizer) or design global attention weights that are not conditioned to tokens (Random Synthesizer). Also, these are clearly different from dot-product based self-attention that computes pairwise token interaction in transformer [46].

Based on these pioneering studies, concurrent papers [5, 11, 18, 23, 25, 28, 44, 56, 57] address new issues and potentials in MLP-like architectures. VisionPermutator [18] effectively preserves spatial dimensions of the input tokens by separately processing token representation along the different dimensions. CycleMLP [5] is the first paper that introduces a hierarchical MLP-like architecture for dense prediction tasks. For pixel-level inference, this paper proposes a cyclical style of pixel sampling strategy for spatial context aggregation. ResMLP [44] demonstrates the potential of MLPs for arbitrary length of input data which is relevant to point cloud processing [6, 10, 58].

Despite their success, these modern MLP-like approaches have not yet been applied to point clouds. In this paper, we propose PointMixer that demonstrates the applicability of MLP-like architectures for point cloud understanding, including dense prediction tasks. In contrast to PointTransformer, our network truly exploits the strength of MLPs to operate mixing within/beyond sets of points. By aggregating the point features in various dimensions, we successfully conduct various tasks.

## 3. Method

In this section, we describe the details of our PointMixer design. For the sake of clarity, we revisit the general formulation of MLP-Mixer in comparison with representative point-based approaches, such as PointNet, PointNet++ and Point Transformer (Sec. 3.1). Then, we underline the relevance of MLP-Mixer for point set operator (Sec. 3.2). Finally, we introduce our PointMixer layer (Sec. 3.3) that is adopted in our entire network (Sec. 3.4).

### 3.1. Preliminary

Let's assume that we have a point cloud  $\mathbf{P} \in \mathbb{R}^{N \times 3}$  and its corresponding point features  $\mathbf{X} \in \mathbb{R}^{N \times C}$  where  $C$  is the length of a channel dimension and  $N$  is the number of points. The objective is to learn a function  $f: \mathbf{X} \rightarrow \mathbf{Y}$  to produce the point features  $\mathbf{Y} \in \mathbb{R}^{N \times C'}$ . Instead of processing the entire point cloud directly, most approaches process the data locally based on the points' proximity. For this purpose,  $k$ -Nearest Neighbor ( $k$ NN) [20, 32, 59] has been widely used to get an index map of neighbor points, we denote this operation as:

$$\mathcal{M}_i = k\text{NN}(\mathbf{P}, \mathbf{p}_i), \quad (1)$$

where  $\mathcal{M}_i$  is an index map of the  $K$  closest points of a query point  $\mathbf{p}_i \in \mathbb{R}^{1 \times 3}$ . In general, given a query index  $i$ , the index map returns the index of the neighbor points ( $\mathcal{M}_i: i \rightarrow j$ ). Using this mapping relationship, we can assign each point  $\mathbf{p}_i$  in  $\mathbf{P}$  with its  $K$  closest point set  $\mathcal{P}_i = \{\mathbf{p}_k\}_{k=1}^K$ . This operation can be applied to all the points such that  $\mathcal{P} = \{\mathcal{P}_i\}_{i=1}^N$ . Similarly, we build  $\{\mathcal{X}_i\}_{i=1}^N$  for every 3D point where  $\mathcal{X}_i = \{\mathbf{x}_k\}_{k=1}^K$  is the neighboring set of 3D point features whose corresponding location is  $\mathbf{p}_i$ .

**PointNet** [33] is a pioneering study that introduces deep learning frameworks for the understanding of point clouds. Due to the inherent difference of data structure between images and point clouds, this paper designs an affine transformation layer as follows:

$$\mathbf{y} = \rho(\mathbf{W}\mathbf{x}), \quad (2)$$

where  $\mathbf{W}$  is an affine transform matrix,  $\mathbf{x}$  is a point feature vector,  $\mathbf{y}$  is the latent feature, and  $\rho$  is a non-linear activation function<sup>2</sup>. From the matrix  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^\top$ , the feature matrix is pooled along its column direction to obtain a permutation-invariant feature vector.

**PointNet++** [34] addresses the problem of PointNet that has difficulty in capturing local responses. To cope with this problem, this paper utilizes  $k$ NN and Farthest Point Sampling algorithm to build a pyramid encoder-decoder. Instead of dealing with the entire point cloud directly, PointNet++ aggregates set-level responses as:

$$\mathbf{y} = \rho(\mathbf{W}\mathbf{x}_k), \quad \text{s.t. } \forall k \in \mathcal{M}_i \quad (3)$$

where  $\mathcal{M}_i$  is an index map from  $k$ NN that indicates the indexes of  $K$  closest points to a query point  $\mathbf{p}_i$ . By adopting this grouping and sampling strategy, MLPs can capture local responses from unordered 3D points.

**PointTransformer** [58] adopts vector subtraction attention as a similarity measurement to capture local responses. Given a query point feature  $\mathbf{x}_i$  and its neighbor feature

set  $\mathcal{X}_i$ , self attention densely relates token interaction as:

$$\mathbf{y}_i = \sum_{k \in \mathcal{M}_i} \text{softmax}\left(\psi(\mathbf{W}_1\mathbf{x}_i - \mathbf{W}_2\mathbf{x}_k + \delta)\right) (\mathbf{W}_3\mathbf{x}_k + \delta), \quad (4)$$

where  $\mathbf{W}$  indicates linear transformation matrix,  $\mathbf{y}_i$  is the output feature for the  $i$ -th point.  $\psi$  denotes an MLPs to calculate vector similarities, and  $\delta$  is a positional encoding vector to embed local structures of 3D points in the operator.

### 3.2. MLP-Mixer as Point Set Operator

MLP-Mixer [43] has achieved remarkable success by only using MLPs for image classification. However, when dealing with sparse and unordered points, the direct application of the MLP-Mixer network is restricted. Let us revisit MLP-Mixer to ease the understanding of our PointMixer.

The MLP-Mixer layer<sup>3</sup> consists of token-mixing MLPs and channel-mixing MLPs. MLP-Mixer takes  $K$  tokens having  $C$ -dimensional features, denoted as  $\mathbf{X} \in \mathbb{R}^{K \times C}$ , where tokens are features from image patches. It begins with token-mixing MLPs that transposes the spatial axis and channel axis to mix spatial information.

$$\mathbf{X}' = \mathbf{X} + (\mathbf{W}_2 \rho(\mathbf{W}_1 (\text{LayerNorm}(\mathbf{X}))^\top))^\top, \quad (5)$$

Then, it continues with channel-mixing MLPs so that input tokens are mixed in spatial and channel dimensions.

$$\mathbf{Y} = \mathbf{X}' + \mathbf{W}_4 \rho(\mathbf{W}_3 \text{LayerNorm}(\mathbf{X}')), \quad (6)$$

where  $\mathbf{W}$  is the weight matrix of a linear function and  $\rho$  is GeLU [17]. Mixing spatial and channel information are permutation equivariant operators, and the global average pooling is a symmetric function. Therefore, MLP-Mixer network can be used for set operators where the outputs are invariant to the permutation of input tokens [22, 43]. This property is also satisfied in prior works [33, 58] and is essential for point features extraction.

However, since MLP-Mixer network is designed for the image classification, the mixed features pass *global average pooling* without positional embeddings, which regards equal importance for each token. Moreover, as a point set operator, vanilla MLP-Mixer layer only computes *intra-set* relations as PointTransformer layer [58] does. From this analysis, we observe room for improvement, especially in the context of 3D point cloud analysis.

### 3.3. PointMixer

We introduce an approach to embed geometric relations between points features into the MLP-Mixer's framework. As illustrated in Fig. 2, the PointMixer layer takes a point set  $\mathcal{P}_i$  and its associated point features set  $\mathcal{X}_i$  as inputs in

<sup>2</sup>We omit the transformation module for simplicity.

<sup>3</sup>We set the relation of terminologies as layer  $\subset$  block  $\subset$  network.

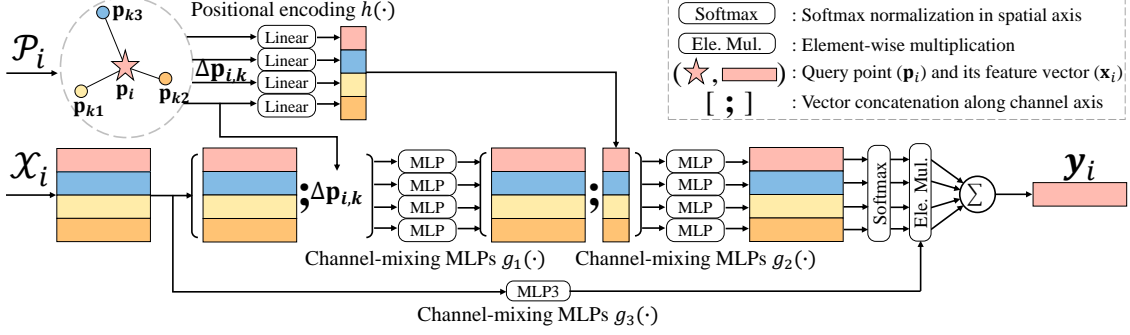


Figure 2. **PointMixer layer.** Given a set of points  $\mathcal{P}_i$  and their corresponding features  $\mathcal{X}_i$ , PointMixer produces a new query feature vector  $\mathbf{y}_i$ . In contrast to MLP-Mixer [43], we utilize relative position vector  $\Delta \mathbf{p}_{i,k}$  for processing unstructured 3D points.

order to compute a feature vector  $\mathbf{y}_i$ . For a point  $\mathbf{p}_i$ , we can aggregate the features of its neighboring points as follows:

$$s_k = g_2\left([g_1([\mathbf{x}_k; \Delta \mathbf{p}_{i,k}]); h(\Delta \mathbf{p}_{i,k})]\right) \text{ s.t. } \forall k \in \mathcal{M}_i, \quad (7)$$

where  $g(\cdot)$  is the channel-mixing MLPs,  $h(\cdot)$  is a linear function, and  $\Delta \mathbf{p}_{i,k} := \mathbf{p}_k - \mathbf{p}_i$  is the relative position vector.  $\mathbf{x}_k$  is a  $k$ -th element of the feature vector set  $\mathcal{X}_i$  referenced to a query point  $\mathbf{p}_i$ .  $[\cdot]$  represents vector concatenation along the channel axis. As a result of this operation, we obtain a score vector  $\mathbf{s} = [s_1, \dots, s_k, \dots, s_K]$ . Finally, the features of the  $K$  adjacent points are gathered to produce a new feature vector  $\mathbf{y}_i$  as:

$$\mathbf{y}_i = \sum_{k \in \mathcal{M}_i} \text{softmax}(s_k) \odot g_3(\mathbf{x}_k) \text{ s.t. } \forall k \in \mathcal{M}_i, \quad (8)$$

where  $\text{softmax}(\cdot)$  is the softmax function that normalizes the spatial dimension, and  $\odot$  indicates an element-wise multiplication. Note that this operation is different from the global average pooling proposed in MLP-Mixer. As shown in Fig. 2, PointMixer layer is defined as below:

$$\mathbf{y}_i = \text{PointMixer}(\mathcal{X}_i, \mathcal{P}_i). \quad (9)$$

As a set operator, PointMixer layer has different characteristics compared to the vanilla MLP-Mixer layer. First, PointMixer layer sees relative positional information ( $\Delta \mathbf{p}$ ) to encode the local structure of a point set. Second, the vanilla MLP-Mixer layer does not have a softmax function. Last, PointMixer layer does not have token-mixing MLPs. Let us explain the reasons behind these differences.

**No token mixing.** While a given pixel in an image systematically admits 8 adjacent pixels, in a point cloud, each point can have an arbitrary number of neighbors. In this context, MLPs are particularly desirable since they can handle variable input lengths [44]. However, in MLP-Mixer layer, the combined use of channel-mixing MLPs and token-mixing MLPs prevents the process of an arbitrary number of points in a set. Let us briefly explain the

reason. As stated in Eq. 2, MLPs require the pre-defined dimensionality in channels for the affine transformation of the incoming data. In token-mixing MLPs (Eq. 5), we switch the channel axis and spatial axis, which fix the number of input tokens (*e.g.*, points). Thus, when we use both types of MLPs in a layer, we can only take a pre-defined number of points with fixed channel dimensionality.

Inspired by Synthesizer [41], we alleviate this problem by replacing the token-mixing MLPs with a softmax function. In our point of view, the softmax function weakly binds token-wise information in a non-parametric manner (*i.e.*, normalization). By doing so, PointMixer is able to calculate various cardinality of point sets. Thus, as in Fig. 3, PointMixer becomes a universal point operator for mixing various types of points: inter-set mixing, intra-set mixing, and pyramid-mixing.

**Intra-set mixing.** Given a point set  $\mathcal{P}_i$  and its corresponding feature set  $\mathcal{X}_i$ , intra-set mixing aims to compute point-wise interaction within each set. This is close to the popular set operators, transformer [46] and MLP-Mixer [43]. Based on the index map  $\mathcal{M}_i$  from  $k$ NN, the PointMixer layer updates a query feature  $\mathbf{x}_i$  using its neighbor feature set  $\mathcal{X}_i$ , as in Eq. 7 and Eq. 8. While this layer well mixes the point-wise information, the receptive field is bounded within a set.

**Inter-set mixing.** Inter-set mixing is a new concept of spreading point features between different sets. Using the index mapping  $\mathcal{M}_i$  from  $k$ NN, we can trace back to find another set  $\mathcal{X}_j$  that includes a query point  $\mathbf{p}_i$  as their neighbors (*i.e.*,  $\mathcal{M}_i^{-1} : j \rightarrow i$ ). This process can be viewed as the inverse version of clustering. For example in Fig. 3, we apply  $k$ NN on each query point where  $K=3$ . As a result (Fig. 3-(a)), red point ( $\star$ ) is included in three sets ( $\mathcal{X}_{\star}, \mathcal{X}_{\star}, \mathcal{X}_{\star}$ ), but green point ( $\star$ ) is clustered to two sets ( $\mathcal{X}_{\star}, \mathcal{X}_{\star}$ ). When we apply the inverse mapping  $\mathcal{M}^{-1}$  (Fig. 3-(b)), we collect the red point from three sets and the green points from two sets<sup>4</sup>. In shorts, inverse map-

<sup>4</sup>Note that we have a chance to collect variable number of points after we conduct an inverse mapping  $\mathcal{M}^{-1}$ .



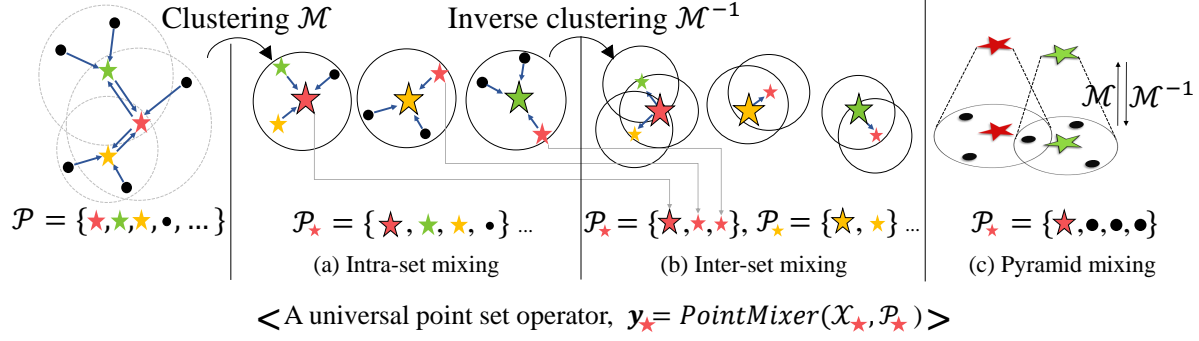


Figure 3. **Applicability of PointMixer: a universal point set operation for inter-set mixing, intra-set mixing, and pyramid mixing.**

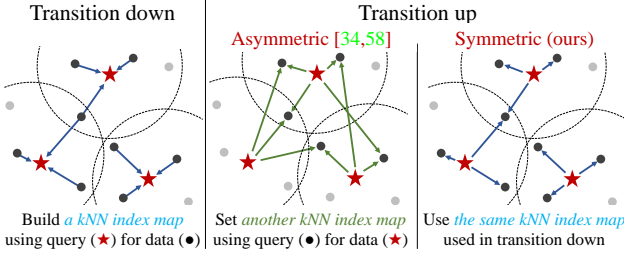


Figure 4. **Comparison of two transition down schemes: symmetric (ours) and asymmetric [34, 58].**

ping  $\mathcal{M}^{-1}: j \rightarrow i$  finds the set index  $j$  that includes a target point  $\mathbf{p}_i$ . The purpose of inter-set mixing is to aggregate the point responses that are originated from the same point  $\mathbf{p}_i$  but grouped in various sets  $\mathcal{X}_j$ .

**Symmetric pyramid mixing.** PointMixer is generally applicable for the transition of two point clouds. For instance, let’s prepare a point set  $\mathcal{P}_s = \{\mathbf{p}_i\}_{i=1}^{N'}$  by sampling points from original set  $\mathcal{P}_o$  (i.e.,  $\mathcal{P}_s \subset \mathcal{P}_o$ ). Using a point  $\mathbf{p}_i \in \mathcal{P}_s$ , we calculate its neighbor among  $\mathcal{P}_o$  and obtain index mapping  $\mathcal{M}_i$ . By putting  $\mathcal{M}_i$  in Eq. 7 and Eq. 8, we can readily pass the feature from  $\mathcal{P}_o$  to  $\mathcal{P}_s$ . This is the case for the point cloud *downsampling*.

For dense prediction tasks, such as semantic segmentation or point reconstruction, it is often required to build U-shaped networks (U-Nets) to utilize both low- and high-level features, and the U-Nets involve point downsampling and upsampling. Interestingly, we notice that conventional of U-Nets in both PointNet++ [34] and PointTransformer [58] is not symmetric in terms of downsampling and upsampling. It is because the spatial grouping is performed asymmetrically. For instance, the conventional approach builds an index map  $\mathcal{S}$  for the downsampling:

$$\mathcal{M}_i = kNN(\mathcal{P}_o, \mathbf{p}_i) \text{ s.t. } \forall i \in \mathcal{S} \quad (10)$$

After that, conventional approaches build another  $kNN$  map  $\mathcal{O}$  for the upsampling.

$$\mathcal{M}_j = kNN(\mathcal{P}_s, \mathbf{p}_j) \text{ s.t. } \forall j \in \mathcal{O} \quad (11)$$

However, this is not symmetric because *nearest neighbor is not a symmetric function*: even if  $\mathbf{p}_i$ ’s nearest neighbor is  $\mathbf{p}_j$ ,  $\mathbf{p}_j$ ’s nearest neighbor may not  $\mathbf{p}_i$ . Fig. 4 shows an illustration of asymmetric downsampling and upsampling.

Instead, we can use the inverse mapping of  $\mathcal{M}_i^{-1}: j \rightarrow i$  (Fig. 3-(b)) instead of making a new index mapping. See Fig. 4 for the symmetric upsampling. In this manner, the proposed PointMixer layer keeps using the same  $k$ -Nearest Neighbor mapping for the upsampling. The benefit of this approach is that it does not introduce additional  $k$ -Nearest Neighbor search. As illustrated in Fig. 5-(b) of the transition down block, we use Farthest point sampling [34] to produce  $\mathcal{P}_s$  from  $\mathcal{P}_o$ . Then, we utilize  $kNN$  to cluster  $\mathcal{P}_s$  from  $\mathcal{P}_o$ . The resulting point locations are used to calculate the relative distance  $\Delta \mathbf{p}$ . In the transition up block, we keep using the index map  $\mathcal{M}$  calculated in the transition down block. To this end, we apply PointMixer layer (Eq. 9) in transition up/down while maintaining the symmetric relation between the sampled point cloud  $\mathcal{P}_s$  and the original points  $\mathcal{P}_o$ .

### 3.4. Network Architecture

In this section, we describe the details of our MLP-like encoder-decoder architecture as in Fig. 5. Our network is composed of several MLP blocks, such as the transition down blocks, transition up blocks, and Mixer blocks. For a fair comparison, our network mainly follows the proposed hyper-parameters in PointTransformer [58] for network composition. Overall, our network takes a deep pyramid-style network that progressively downsamples points to obtain global features. For dense prediction tasks, such as semantic segmentation, we include upsampling blocks for per-point estimation. Finally, our header block is designed for task-specific solutions. For classification, we take fully-connected layers, dropout layers, and global average pooling. For semantic segmentation and point reconstruction, the header block consists of MLPs without pooling layers. Please refer to our supplementary material for more details of our architecture.

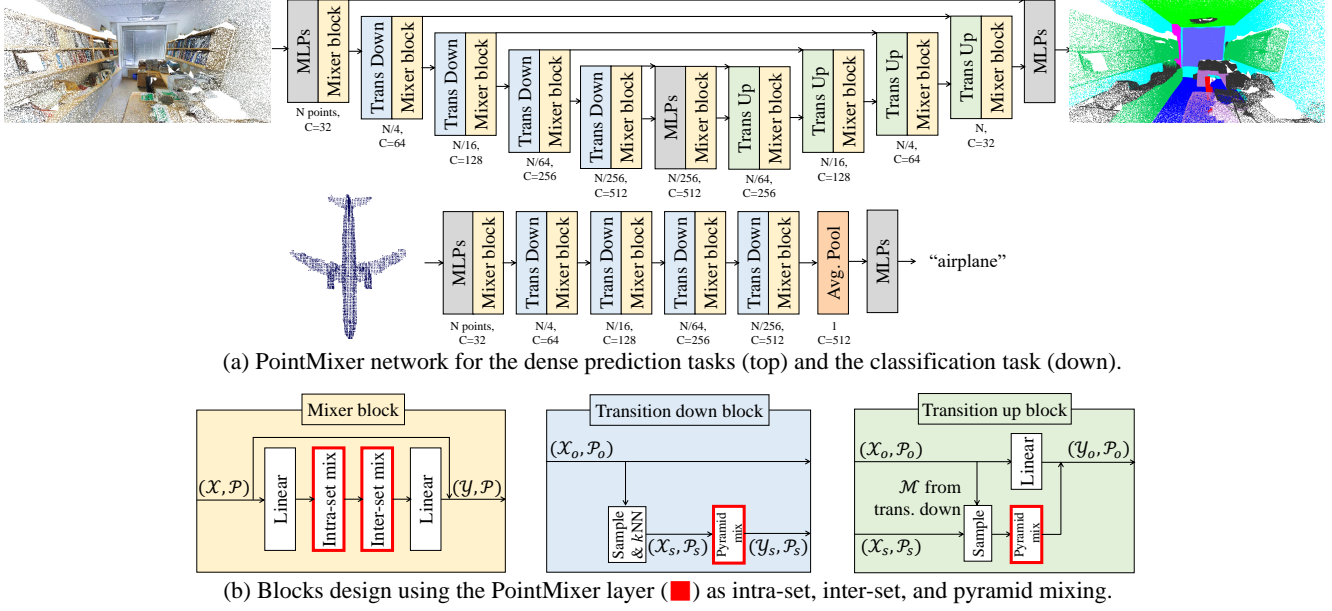


Figure 5. Overall architecture of PointMixer network.

## 4. Experiments

In this section, we evaluate the efficacy and versatility of the proposed PointMixer for various point cloud understanding tasks, including: semantic segmentation, point reconstruction, and object classification.

### 4.1. Semantic Segmentation

To validate the effectiveness of PointMixer for semantic segmentation, we propose an evaluation on the large-scale point cloud dataset S3DIS [3] consisting of 271 room reconstructions. Each 3D point of this dataset is assigned to one label among 13 semantic categories. For the sake of fairness, we meticulously follow the widely used evaluation protocol proposed by PointTransformer [58]. For training, we set the batch size as 2 and use the SGD optimizer with momentum and weight decay set to 0.9 and 0.0001 respectively<sup>5</sup>. For evaluation, we use the class-wise mean Intersection of Union (mIoU), class-wise mean accuracy (mAcc), and overall point-wise accuracy (OA).

As shown in Table 1, PointMixer achieves the state-of-the-art performance in S3DIS [3] Area 5. Even in class-wise accuracy, PointMixer network outperforms PointTransformer [58], except for a few classes. While various studies [12, 31, 58] underline the necessity of dense point-wise interaction through dot-product attention, PointMixer successfully outperforms these approaches, even without dense token-wise computation. Similar to Synthesizer [41], our work questions the necessity for dot-product attention.

<sup>5</sup>The training detail are described in the supplementary material.

### 4.2. Point Cloud Reconstruction

To highlight the versatility of our approach, we propose a large-scale assessment for the newly introduced task of point cloud reconstruction [2] where an inaccurate point cloud is jointly denoised, densified, and completed. This experiment is also particularly interesting to evaluate the generalization capabilities of networks since it allows us to test methods in unmet environments. In details, this paper [2] uses synthetic/object dataset (ShapeNetPart [55]) for training, then evaluates the reconstruction accuracy in the unmet scenes, such as real-world/room-scale 3D scans (ScanNet [7]) and synthetic/room-scale point cloud (ICL-NUIM [15]). Specifically, we train on the synthetic objects of ShapeNetPart [55] and evaluate the reconstruction accuracy on unmet scenes from ScanNet [7] (real reconstruction) and ICL-NUIM [15] (synthetic data).

Note that in [2], the point reconstruction is performed in two stages: 1) point upsampling via a sparse hourglass network, 2) denoising and refinement via transformer network. For this evaluation, we propose to replace the second stage of this pipeline with various architectures (i.e. PointMixer network and previous studies [33, 34, 58]) to compare their performances. Under the same data augmentation and data pre-processing as in [2, 24], we train PointMixer and previous studies [33, 34, 58]. For evaluation, we utilize the Chamfer distance (i.e., Chamf.) to measure the discrepancy between the predictions and the ground truth point clouds. Additionally, we use the accuracy (i.e., Acc.), completeness (i.e., Comp.), and F1 score to calculate the accuracy of the reconstruction in occupancy aspects [1, 19, 21, 54].

Method	OA	mAcc	mIoU	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
PointNet [33]	-	49.0	41.1	88.8	97.3	69.8	0.1	3.9	46.3	1.8	59.0	52.6	5.9	40.3	26.4	33.2
PCCN [49]	-	67.0	58.3	92.3	96.2	75.9	<b>0.3</b>	6.0	69.5	63.5	66.9	65.6	47.3	68.9	59.1	46.2
MinkowskiNet [6]	-	71.7	65.4	91.8	98.7	86.2	0.0	34.1	48.9	62.4	81.6	89.8	47.2	74.9	74.4	58.6
KPConv [42]	-	72.8	67.1	92.8	97.3	82.4	0.0	23.9	58.0	69.0	81.5	91.0	75.4	75.3	66.7	58.9
PointTrans [58]	<b>90.8</b>	76.5	70.4	94.0	<b>98.5</b>	<b>86.3</b>	0.0	38.0	<b>63.4</b>	74.3	89.1	<b>82.4</b>	<b>74.3</b>	<b>80.2</b>	76.0	59.3
PointMixer (ours)	<b>90.8</b>	<b>77.4</b>	<b>71.3</b>	<b>94.2</b>	98.2	86.0	0.0	<b>43.8</b>	62.1	<b>78.5</b>	<b>90.6</b>	82.2	73.9	79.8	<b>78.5</b>	<b>59.4</b>

Table 1. Quantitative results on semantics segmentation in S3DIS dataset.

Method	ShapeNet-Part				ScanNet				ICL-NUIM			
	Chamf.(↓)	Acc.(↑)	Comp.(↑)	F1(↑)	Chamf.(↓)	Acc.(↑)	Comp.(↑)	F1(↑)	Chamf.(↓)	Acc.(↑)	Comp.(↑)	F1(↑)
PointNet [33]	1.33	63.2	38.8	48.2	3.05	37.51	27.79	32.58	2.98	46.89	33.23	38.11
PointNet++ [33]	1.25	65.1	39.0	50.1	2.97	38.25	29.54	33.41	2.88	48.77	35.77	39.85
PointRecon [2]	1.19	<b>81.02</b>	40.41	53.4	2.86	40.42	30.24	34.08	2.78	54.11	38.09	43.55
PointTrans [58]	1.12	75.9	40.9	52.7	2.79	41.10	32.14	35.56	2.57	51.06	36.38	41.64
PointMixer (ours)	<b>1.10</b>	77.12	<b>41.5</b>	<b>53.5</b>	<b>2.74</b>	<b>42.12</b>	<b>33.53</b>	<b>37.81</b>	<b>2.43</b>	<b>56.51</b>	<b>38.16</b>	<b>44.74</b>

Table 2. Point cloud reconstruction results. Note that ↑ means higher the better, and vice versa for ↓.

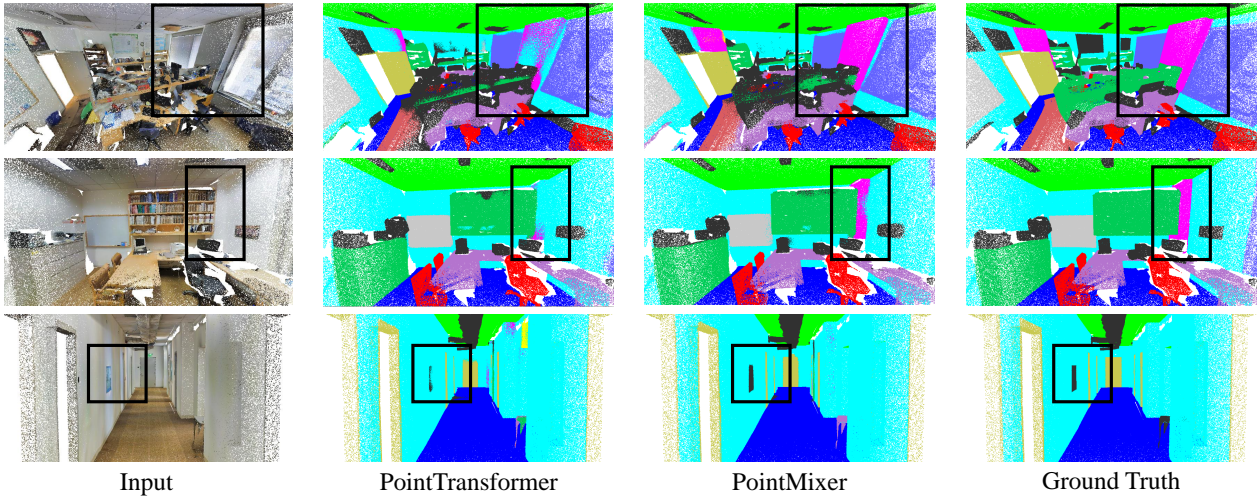


Figure 6. Qualitative results in semantic segmentation.



Figure 7. Qualitative results in point reconstruction.

Method	input	mAcc	OA
PointNet [33]	point	86.2	89.2
PointNet++ [34]	point	-	91.9
DGCNN [50]	point	90.2	92.2
KPConv [42]	point	-	92.9
PointTrans [58]	point	<b>90.6</b>	<b>93.7</b>
PointMixer (ours)	point	90.3	92.7

Table 3. Shape classification results in ModelNet40 dataset.

Although our network is solely trained in a synthetic/object dataset, as in Table 2, our network can generalize towards unmet scenes including real-world 3D scans (ScanNet) and synthetic/room-scale point clouds [15]. Moreover, our method compares favorably to previous transformer-based studies [2, 58] as well as conventional MLP-based strategies [33, 34]. In particular, the performance gap between ours and previous studies become larger in the ScanNet and ICL-NUIM dataset.



### 4.3. Shape Classification

The ModelNet40 [51] dataset has 12K CAD models with 40 object categories. We follow the official train/val/test splits to train and evaluate ours and the previous studies. For fair comparison, we follow the data augmentation and pre-processing as proposed in PointNet++ [34], which is also adopted in the recent studies [29, 42, 58]. For evaluation, we adopt the mean accuracy within each category (mAcc), and the overall accuracy over all classes (OA). The results presented in Table 3 show that our approach remains competitive against recently proposed techniques. Classification is the only presented task where PointMixer demonstrates lower results than Point Transformer. It might suggest that certain global description tasks might not fully require the larger receptive field offered by PointMixer.

### 4.4. Ablation Study

We conduct an ablation study about PointMixer. First, we proceed case study of our PointMixer (Table 4). Second, we analyze the influence of positional encoding on PointMixer (Table 5). Last, we compare PointMixer that is free from token-mixing MLPs (Table 6).

**A universal point operator.** PointMixer can be applicable to act as inter-set mixing, intra-set mixing, and pyramid mixing. Each mixing operator consists of simple MLPs with the softmax function but covers a different set of points. As in Table 4, it turns out that each type of mixing consistently improves the accuracy of both semantic segmentation and point cloud reconstruction. It implies that the various ways of sharing points’ responses are beneficial for point cloud understanding.

**Positional encodings.** MLP-Mixer [37] does not explicitly emphasize the necessity of positional encodings [9, 46] for image classification tasks. However, to handle unstructured 3D points, relative position information is the starting key to understanding 3D points. As in Table 5, there is a large performance gap when we intentionally omit positional encodings in our PointMixer layer. In particular, positional encoding is most precious in intra-set mixing.

**Unnecessary token-mixing MLPs.** We validate our claims about the role of a softmax function, *i.e.*, weakly binding tokens. While maintaining the rest of the condition as we design for the final PointMixer layer, we intentionally erase a softmax function in all PointMixer layers and include token-mixing MLPs, as proposed in the vanilla MLP-Mixers. Table 6 demonstrates that even without explicit use of token-mixing MLPs, our PointMixer successfully achieves similar accuracy in semantic segmentation and point reconstruction.

Preserve (✓)			Semantic seg		Point recon	
PointMixer			mIoU	mAcc	Chamf.(↓)	F1(↑)
Inter-set	Intra-set	Pyramid-mix				
✓			0.6918	0.7582	1.125	53.48
✓	✓		0.6910	0.7565	1.120	53.50
✓		✓	0.6926	0.7655	1.111	53.52
✓	✓	✓	<b>0.7136</b>	<b>0.7740</b>	<b>1.109</b>	<b>53.65</b>

Table 4. Ablation study about mixing operations.

Preserve (✓)			Semantic seg		Point recon	
Positional encodings in			mIoU	mAcc	Chamf.(↓)	F1(↑)
Inter-set	Intra-set	Pyramid-mix				
			0.6471	0.7178	1.130	52.98
		✓	0.6593	0.7252	1.130	52.95
	✓		0.6398	0.7047	1.130	53.04
	✓	✓	0.6612	0.7292	1.123	53.59
✓			0.6959	0.7627	1.116	53.59
✓		✓	0.7020	0.7677	1.134	52.45
✓	✓		0.6993	0.7654	1.127	53.50
✓	✓	✓	<b>0.7136</b>	<b>0.7740</b>	<b>1.109</b>	<b>53.65</b>

Table 5. Ablation study about positional encoding.

Preserve (✓)			Semantic seg		Point recon	
Mix of what?			mIoU	mAcc	Chamf.(↓)	F1(↑)
Channel	Token	Softmax				
✓			0.5831	0.6647	1.233	48.55
✓	✓		0.7105	0.7733	1.119	53.27
✓		✓	<b>0.7136</b>	<b>0.7740</b>	<b>1.109</b>	<b>53.65</b>

Table 6. Ablation study about token-mixing MLPs.

## 5. Limitation

PointMixer network shows the best performance in dense prediction tasks, but not in object classification task. Since MLP-Mixer is originally designed for image classification, we do not expect MLPs are the reason. Since we did not provide the reasonable explanation, we will further investigate this problem for the upgraded PointMixer design.

## 6. Conclusion

We propose the first MLP-like architecture for point cloud understanding. Our network is equipped with a universal point set operator, *PointMixer*. Regardless of the cardinality of a point set, PointMixer layer can be used to - mix - point responses in various ways, such as inter-set mixing, intra-set mixing, and pyramid-mixing. Extensive experiments demonstrate the competitive or superior performance of PointMixer network compared to transformer-based studies, including dense prediction tasks. The results imply that MLPs with geometric encodings are still promising compared to the transformer-based networks with dense token-wise interaction or the vanilla MLP-Mixer network using token/channel-wise mixing, respectively.



## References

- [1] Henrik Aanaes, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjarholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. [6](#)
- [2] Anonymous. Deep point cloud reconstruction. In *Submitted to The Tenth International Conference on Learning Representations*, 2022. under review. [1](#), [6](#), [7](#)
- [3] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016. [6](#)
- [4] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *International Conference on Learning Representations*, 2020. [2](#)
- [5] Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*, 2021. [1](#), [2](#)
- [6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. [1](#), [2](#), [7](#)
- [7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. [6](#)
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. [2](#)
- [9] 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. [2](#), [8](#)
- [10] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. [1](#), [2](#)
- [11] Jianyuan Guo, Yehui Tang, Kai Han, Xinghao Chen, Han Wu, Chao Xu, Chang Xu, and Yunhe Wang. Hire-mlp: Vision mlp via hierarchical rearrangement. *arXiv preprint arXiv:2108.13341*, 2021. [2](#)
- [12] 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(2):187–199, 2021. [1](#), [2](#), [6](#)
- [13] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. [1](#)
- [14] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021. [2](#)
- [15] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *IEEE Int. Conf. on Robotics and Automation.*, 2014. [6](#), [7](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [1](#)
- [17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. [3](#)
- [18] Qibin Hou, Zihang Jiang, Li Yuan, Ming-Ming Cheng, Shuicheng Yan, and Jiashi Feng. Vision permutator: A permutable mlp-like architecture for visual recognition. *arXiv preprint arXiv:2106.12368*, 2021. [1](#), [2](#)
- [19] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanaes. Large scale multi-view stereopsis evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. [6](#)
- [20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 2019. [3](#)
- [21] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 2017. [6](#)
- [22] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019. [3](#)
- [23] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. Convmlp: Hierarchical convolutional mlps for vision. *arXiv preprint arXiv:2109.04454*, 2021. [1](#), [2](#)
- [24] Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. Point cloud upsampling via disentangled refinement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. [6](#)
- [25] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. As-mlp: An axial shifted mlp architecture for vision. *arXiv preprint arXiv:2107.08391*, 2021. [2](#)
- [26] 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*, pages 1800–1809, 2020. [2](#)
- [27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. [2](#)
- [28] Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. Sparse-mlp: A fully-mlp architecture with conditional computation. *arXiv preprint arXiv:2109.02008*, 2021. [2](#)

- [29] Jiageng Mao, Xiaogang Wang, and Hongsheng Li. Interpolated convolutional networks for 3d point cloud understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1578–1587, 2019. 1, 2, 8
- [30] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3164–3173, 2021. 2
- [31] Kirill Mazur and Victor Lempitsky. Cloud transformers: A universal approach to point cloud processing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10715–10724, October 2021. 2, 6
- [32] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009. 3
- [33] 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 conference on computer vision and pattern recognition*, pages 652–660, 2017. 3, 6, 7
- [34] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 2, 3, 5, 6, 7, 8
- [35] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5199–5208, 2017. 1
- [36] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2
- [37] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Int. Conf. on Machine Learning.*, 2019. 8
- [38] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 68–80, 2019. 2
- [39] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3693–3702, 2017. 1
- [40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1
- [41] Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention in transformer models. *CoRR*, 2020. 1, 2, 4, 6
- [42] 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*, pages 6411–6420, 2019. 1, 2, 7, 8
- [43] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 1, 2, 3, 4
- [44] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021. 1, 2, 4
- [45] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2
- [46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2, 4, 8
- [47] Chu Wang, Babak Samari, and Kaleem Siddiqi. Local spectral graph convolution for point set feature learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–66, 2018. 1
- [48] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10296–10305, 2019. 1
- [49] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continuous convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2589–2597, 2018. 1, 7
- [50] 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. 1, 2, 7
- [51] 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 conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 8
- [52] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021. 1, 2
- [53] 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*, pages 5661–5670, 2020. 1
- [54] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Eur. Conf. Comput. Vis.*, 2018. 6

- [55] 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. 6
- [56] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li. S<sup>2</sup>-mlp: Spatial-shift mlp architecture for vision. *arXiv preprint arXiv:2106.07477*, 2021. 2
- [57] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li. S<sup>2</sup>-mlpv2: Improved spatial-shift mlp architecture for vision. *arXiv preprint arXiv:2108.01072*, 2021. 2
- [58] 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*, pages 16259–16268, 2021. 1, 2, 3, 5, 6, 7, 8
- [59] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 3