This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Point Clouds Meets Physics: Dynamic Acoustic Field Fitting Network for Point Cloud Understanding

Changshuo Wang
NTU, Singapore
wangchangshuo1@gmail.com

Shuting He
SUFE, China
shuting.he@sufe.edu.cn

Xiang Fang*
NTU, Singapore
xfang9508@gmail.com

Jiawei Han
BIT, China
hanjiawei@bit.edu.cn

Zhonghang Liu
SMU, Singapore
zhliu.2020@phdcs.smu.edu.sg

Xin Ning
Institute of Semiconductors, CAS
ningxin@semi.ac.cn

Weijun Li
Institute of Semiconductors, CAS
wjli@semi.ac.cn

Prayag Tiwari
Halmstad University, Sweden
prayag.tiwari@ieee.org

Abstract

*While existing pre-training-based methods have enhanced point cloud model performance, they have not fundamentally resolved the challenge of local structure representation in point clouds. The limited representational capacity of pure point cloud models continues to constrain the potential of cross-modal fusion methods and performance across various tasks. To address this challenge, we propose a **D**ynamic **A**coustic **F**ield Fitting **Net**work (**DAF-Net**), inspired by physical acoustic principles. Specifically, we represent local point clouds as acoustic fields and introduce a novel **A**coustic **F**ield **Conv**olution (**AF-Conv**), which treats local aggregation as an acoustic energy field modeling problem and captures fine-grained local shape awareness by dividing the local area into near field and far field. Furthermore, drawing inspiration from multi-frequency wave phenomena and dynamic convolution, we develop the **D**ynamic **A**coustic **F**ield **Conv**olution (**DAF-Conv**) based on AF-Conv. DAF-Conv dynamically generates multiple weights based on local geometric priors, effectively enhancing adaptability to diverse geometric features. Additionally, we design a **G**lobal **S**hape-**A**ware (**GSA**) layer incorporating EdgeConv and multi-head attention mechanisms, which combines with DAF-Conv to form the DAF Block. These blocks are then stacked to create a hierarchical DAFNet architecture. Extensive experiments demonstrate that DAFNet significantly outperforms existing methods across multiple tasks.*

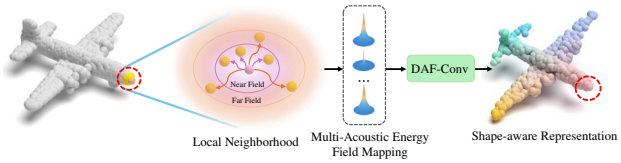


Figure 1. Illustration of local feature extraction using DAF-Conv. Given a local neighborhood of a point cloud (left), our method first models it as an acoustic field with near-field and far-field zones. Then, multiple acoustic energy field mappings are generated to capture various geometric patterns. Finally, DAF-Conv processes these mappings to produce shape-aware representations.

1. Introduction

Point cloud data, due to its direct representation of object surface information in three-dimensional space, has been widely applied in various scenarios including autonomous driving [6, 54, 72], robotic navigation [14, 45], and augmented reality [8, 43]. However, the sparsity, irregularity, and complexity of local structures in point cloud data make it challenging for existing models to fully capture geometric features, thus limiting their applications in 3D tasks.

Recently, several methods [33, 37, 38] have attempted to indirectly improve point cloud model performance through pre-training strategies or cross-modal contrastive learning. Although these approaches have enhanced point cloud processing tasks to some extent, they have not directly addressed the problem of local structure representation. Some researchers [19, 21, 25, 50] have also attempted to apply new architectures to point cloud understanding tasks. However, the local representational capacity of point cloud mod-

*denotes corresponding author.

els remains a bottleneck in improving overall model performance. Therefore, how to efficiently capture and represent the local structures of point clouds has become a crucial problem that urgently needs to be solved.

Physics, particularly the principles of acoustic field propagation [5], provides a new perspective for understanding local geometric structures. In acoustics, wave propagation is influenced by environmental geometric characteristics, and through constructing acoustic field models, we can effectively simulate wave behavior at different frequencies to obtain information in complex environments. Inspired by this, we propose a novel point cloud understanding method called **Dynamic Acoustic Field Fitting Network (DAFNet)**. This method borrows ideas from physical acoustic field propagation, representing local point clouds as acoustic fields and dynamically adjusting weights in different regions to capture local structural features more precisely.

Specifically, we propose a novel **Acoustic Field Convolution (AF-Conv)** that models local aggregation as an acoustic energy propagation problem. To more finely model the contributions of different neighboring points to the center point, AF-Conv divides the local region into near-field and far-field zones for separate geometric modeling. Furthermore, as shown in Fig. 1, based on AF-Conv, we propose **Dynamic Acoustic Field Convolution (DAF-Conv)**, which dynamically generates multiple weights based on geometric priors while simulating geometric feature variations under different "frequencies", thereby enhancing the model's adaptability to local geometric diversity. Finally, we combine a Global Shape-Aware (GSA) layer comprising EdgeConv and Multi-Head Attention (MHA) with DAF-Conv to build hierarchical DAF Blocks, forming the complete DAFNet. The experimental results show that DAFNet achieves superior performance over existing methods in multiple tasks, demonstrating its effectiveness in capturing complex geometric structures.

The main contributions of this work are as follows:

- We propose a physics-inspired end-to-end DAFNet that can learn rich contextual information and precise shape representations in irregular point clouds.
- We introduce a novel Acoustic Field Convolution (AF-Conv) that achieves fine-grained local shape awareness by modeling local point cloud regions as acoustic fields.
- We design Dynamic Acoustic Field Convolution (DAF-Conv) that adapts to geometric feature diversity by dynamically generating multiple frequency-aware weights from local geometry.
- The proposed DAFNet achieves state-of-the-art performance in multiple tasks, reaching 96.7% accuracy on the challenging ScanObjectNN with only 2.47M parameters.

2. Related Work

2.1. CNN-based Methods

Early works [30, 42, 56, 63] in point cloud processing primarily involved projecting them onto 2D images for 2D CNN processing or converting them into regular 3D grid structures for 3D convolution operations. However, these transformation processes inevitably led to spatial information loss and increased computational costs. To address this issue, researchers began applying deep learning techniques directly in 3D space. PointNet [35] pioneered this direction. However, this method overlooked local structural information. Subsequently, PointNet++ [36] enhanced the local perception capability of point cloud models by grouping point clouds into different local neighborhoods. Based on PointNet++ [36], existing works [23, 49, 51, 52, 55, 66] have developed methods based on point-wise perceptrons, convolution operations, and attention mechanisms. Nevertheless, these methods still struggle to effectively model local structural shape information.

2.2. Transformer-based Methods

Transformers have demonstrated remarkable success in various tasks [9, 10, 13, 53] due to their inherent ability to capture global features. Leveraging this strength, researchers have progressively developed a range of Transformer-based point cloud models. PCT [17] enhances input embeddings through furthest point sampling and nearest neighbor search while using self-attention mechanisms to capture global features. Point Transformer [71] employs vector attention to extract aggregated local features. Subsequently, updated versions [57, 58] of Point Transformer explored different architectures to improve performance and efficiency. Inspired by Swin Transformer [27], works such as SST [12] and SWFormer [46] explored window attention mechanisms to improve efficiency. Although these methods excel at capturing global features, their performance hasn't shown significant advancement over CNN-based approaches.

2.3. State Space Model-based Methods

State Space Models (SSMs) [16] are mathematical models for describing dynamic systems. Recently, SSMs have emerged as a reliable alternative architecture to CNNs and Transformers. In computer vision, Vision Mamba [76] successfully compresses visual representations through bidirectional state space models, while VMamba [26] innovatively introduces cross-scanning modules, enabling global receptive fields in 2D images through 1D selective scanning. Researchers have explored SSM applications in point cloud processing. PCM [69] first introduced the Mamba architecture to point cloud processing through octree serialization strategies. Mamba3D [19] enhanced local geometric feature extraction through LNP blocks and designed bidi-

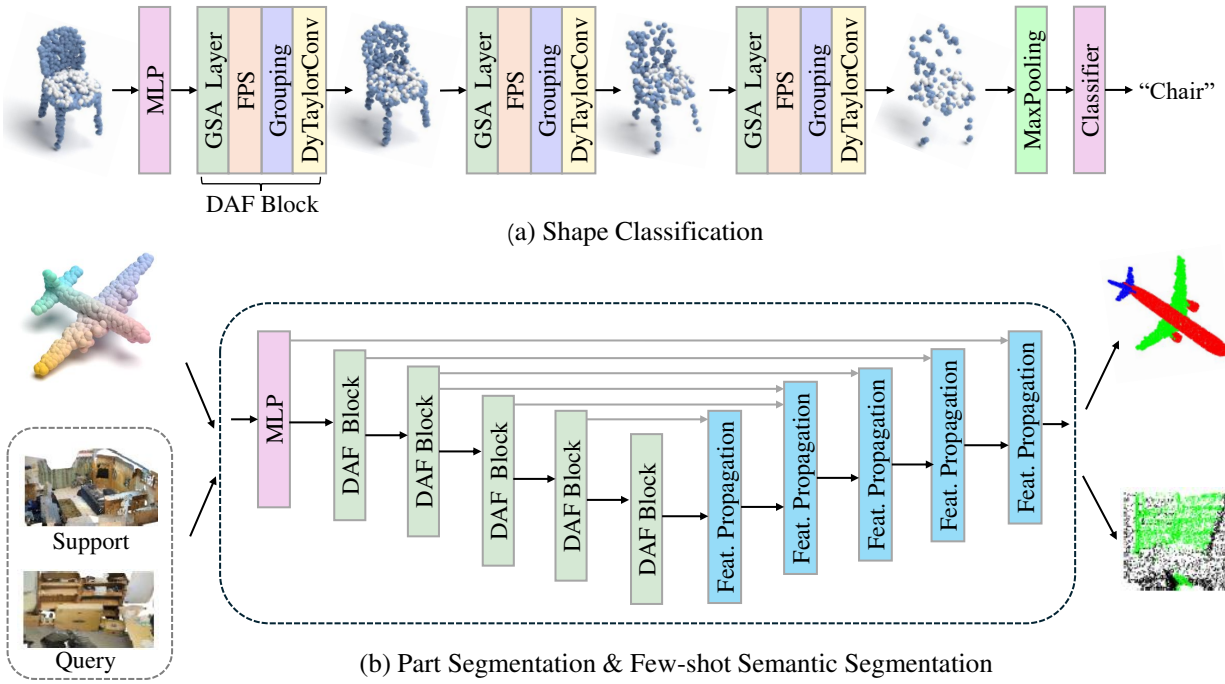


Figure 2. Architecture of DAFNet. (a) For shape classification, the network consists of three sequential DAF Blocks followed by max pooling and classifier layers. Each DAF Block contains a GSA layer and DAF-Conv for feature extraction. (b) For part segmentation and few-shot semantic segmentation tasks, the network adopts an encoder-decoder structure with skip connections, where the encoder shares the same architecture as the classification network, and the decoder uses feature propagation to restore point cloud resolution. The few-shot version additionally incorporates support and query mechanisms for processing different point cloud inputs.

rectional SSM structures. PointMamba [25], another significant work, customized SSM architecture for point cloud analysis tasks, substantially reducing computational costs.

2.4. Pre-training-based Methods

Point cloud pre-training methods primarily comprise generative approaches [4, 33, 67] and cross-modal contrastive learning methods. In generative approaches, Point-MAE [33] extends the MAE concept by pre-training through random masking and reconstruction of point cloud segments PCP-MAE [70] learns shape priors by predicting missing regional point cloud structures. PointDif [74] utilizes diffusion models for pre-training, learning hierarchical geometric priors through point cloud recovery at different noise levels. In cross-modal contrastive learning, ReCon [37] combines generative modeling and contrastive learning paradigms, addressing representation overfitting in contrastive learning and data padding issues in generative methods. ULIP [61] achieves cross-modal knowledge transfer through a language-image-point cloud pre-training framework, while ACT [11] employs cross-modal autoencoders as teacher models to acquire knowledge from other modalities. Although these pre-training methods have advanced point cloud processing performance, they haven't fundamentally resolved the point cloud representation challenge.

3. Methods

This section first reviews point convolution and dynamic convolution, followed by introducing acoustic field convolution. Then, we present our proposed dynamic acoustic field convolution. Finally, we describe DAFNet (as shown in Fig. 2).

3.1. Background

3.1.1. Point Convolution

Local feature aggregation in point clouds typically follows a "sampling-grouping-aggregation" procedure. The aggregation operation is crucial as it directly affects local feature representation by modeling local point clouds. We denote center points as $\{(p_i, f_i)\}_{i=1}^{M}$, where M represents the number of points, $p_i \in \mathbb{R}^{1\times 3}$ and $f_i \in \mathbb{R}^{1\times C}$ represent the coordinates and features of the i-th point respectively, and C denotes the number of feature channels. The most commonly used method for local feature aggregation based on coordinate space is to construct point convolution, which can be formalized as:

$$f_i' = \mathcal{A}(\{\mathcal{M}(p_i, p_j) \cdot \mathcal{T}(f_i, f_j) | p_j \in \mathcal{N}(p_i)\}), \quad (1)$$

where $\mathcal{A}$ represents the aggregation function, typically max-pooling. $\mathcal{M}$ and $\mathcal{T}$ are mapping functions, usually imple-

mented through Multi-Layer Perceptrons (MLPs). $\mathcal{N}(p_i)$ represents the local neighborhood of p_i, and p_j denotes the neighboring points of p_i.

An alternative aggregation approach operates in feature space, commonly known as Dynamic Graph Convolution (DGC) [55]. It aggregates local features in feature space, meaning that points close in feature space might be distant in coordinate space. DGC can be formalized as:

$$f_i' = \mathcal{A}(\{\mathcal{T}(f_i, f_j) | f_j \in \mathcal{N}(f_i)\}). \tag{2}$$

3.1.2. Dynamic Convolution

Dynamic convolution [62] aims to enhance network modeling capability by dynamically generating multiple convolution kernels based on input data, adapting to different data structures. Taking PAConv's [60] dynamic convolution as an example, dynamic convolution in point clouds can be expressed as:

$$g_i = \mathcal{A}(\{w(p_j) \cdot f_j | p_j \in \mathcal{N}(p_i)\}), \tag{3}$$

where $w(p_j)$ denotes the weight corresponding to feature f_j of point p_j. $w(p_j)$ typically consists of multiple weights $w_h(p_i)$ and their corresponding attention coefficients $\alpha_h(p_j)$ generated in a data-driven manner:

$$w(p_j) = \sum_{h=1}^{H} \alpha_h(p_j) \cdot w_h(p_j), \tag{4}$$

where $\sum_{h=1}^{H} \alpha_h(p_j) = 1$, H represents the number of $w_h(p_j)$.

Therefore, formula 3 can be expressed as:

$$g_i = \mathcal{A}(\{\sum_{h=1}^{H} \alpha_h(p_j) \cdot w_h(p_j) \cdot f_j | p_j \in \mathcal{N}(p_i)\}). \tag{5}$$

3.2. Acoustic Field Convolution

In physical acoustics [22, 31], the sound pressure field $p(x)$ generated by a sound source at spatial point x can be represented by Green's function:

$$p(x) = \frac{e^{-jk\|x-y\|}}{4\pi\|x-y\|}, \tag{6}$$

where y is the source position, k is the wave number, and $\|x-y\|$ represents the distance between observation point x and source y. This formula reveals the fundamental physical laws of sound wave propagation: sound pressure intensity decays exponentially with distance while being affected by spherical spreading attenuation.

Notably, based on the distance between the source and observation point, the acoustic field can be divided into near-field and far-field regions. In the near-field region, the relationship between sound pressure and distance is more complex, exhibiting strong local correlations; in the far-field region, the sound pressure demonstrates simple spherical wave attenuation characteristics.

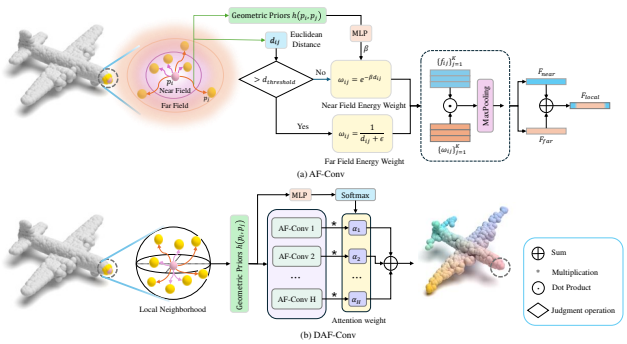


Figure 3. Illustration of AF-Conv and DAF-Conv. (a) AF-Conv divides local neighborhood into near-field and far-field regions and calculates energy weights based on geometric priors. (b) DAF-Conv extends AF-Conv with multiple parallel acoustic field convolutions to achieve dynamic feature aggregation.

Inspired by this theory, we view local feature aggregation as acoustic field propagation, where the center point p_i can be viewed as the sound source and neighboring points p_j as observation points. Meanwhile, we refine the local neighborhood into near-field and far-field regions to extract fine-grained information, constructing the Acoustic Field Convolution (AF-Conv). As shown in Fig. 3(a), it can be formalized as:

$$g_i = \mathcal{A}(\{W_{AF}(p_i, p_j) \cdot f_j | p_j \in \mathcal{N}(p_i)\}), \tag{7}$$

where the weight function $W_{AF}(p_i, p_j)$ directly draws inspiration from acoustic field propagation characteristics and designs different weight calculation methods for near-field and far-field regions:

$$W_{AF}(p_i, p_j) = \begin{cases} e^{-\beta d_{ij}}, & \text{if } d_{ij} < d_{threshold} \\ \frac{1}{d_{ij}+\epsilon}, & \text{otherwise} \end{cases}, \tag{8}$$

where β represents learnable coefficient. d_{ij} denotes the Euclidean distance between p_j and p_i. In $W_{AF}(p_i, p_j)$, for the near-field region ($d_{ij} < d_{threshold}$), we use the exponential decay term $e^{-\beta d_{ij}}$ to capture subtle changes in local geometric structures, corresponding to the complex attenuation characteristics of the acoustic near-field region; for the far-field region ($d_{ij} \geq d_{threshold}$), we adopt a simple reciprocal relationship $\frac{1}{d_{ij}+\epsilon}$ to better perceive broader contextual information.

3.3. Dynamic Acoustic Field Convolution

In sound wave propagation [31], waves of different frequencies exhibit distinct propagation characteristics. Inspired by

this multi-frequency phenomenon and dynamic convolution [62], as shown in Fig. 3(b), we propose Dynamic Acoustic Field Convolution (DAF-Conv), which extends AF-Conv by dynamically generating multiple sets of weights to simulate local geometric feature variations under different "frequencies". It can be formalized as:

$$F_h(p_i, p_j) = W_h(p_i, p_j) \cdot f_j, \tag{9}$$

$$W_h(p_i, p_j) = \begin{cases} \alpha_h e^{-\beta_h d_{ij}}, & \text{if } d_{ij} < d_{threshold} \\ \frac{\alpha_h}{d_{ij}+\epsilon}, & \text{otherwise,} \end{cases} \tag{10}$$

$$\{\alpha_h, \beta_h\}_{h=1}^{H} = \text{MLP}(h_j), \tag{11}$$

where $d_{threshold}$ is a scalar that defines the boundary between near-field and far-field regions. The weight coefficient α_h and β_h are generated from local geometric prior $h_j = [p_i, p_j, p_j - p_i, \|p_j - p_i\|]$ through MLP to simulate propagation characteristics of different frequency components.

The local aggregated features are then separated into near-field and far-field components based on the distance threshold:

$$F_h^{near}(p_i) = \mathcal{A}(\{F_h(p_i, p_j) \mid j \in \mathcal{N}_{near}(i)\}), \tag{12}$$

$$F_h^{far}(p_i) = \mathcal{A}(\{F_h(p_i, p_j) \mid j \in \mathcal{N}_{far}(i)\}), \tag{13}$$

where $\mathcal{N}_{near}(i) = \{j \mid p_j \in \mathcal{N}(p_i), d_{ij} < d_{threshold}\}$ represents the set of near-field neighbors of the center point p_i, and $\mathcal{N}_{far}(i) = \{j \mid p_j \in \mathcal{N}(p_i), d_{ij} \geq d_{threshold}\}$ represents the set of far-field neighbors.

The component for each frequency h can be obtained by:

$$F_h(p_i) = F_h^{near}(p_i) + F_h^{far}(p_i). \tag{14}$$

To achieve adaptive fusion of features from different frequency components, we introduce an attention mechanism based on geometric prior h_j. The attention weights are generated as fellow:

$$\gamma_h = \text{softmax}(\text{MaxPooling}(\text{MLP}(h_j))). \tag{15}$$

The final local aggregated feature is obtained by weighted summation of features from all frequency components:

$$g_i = \sum_{h=1}^{H} \gamma_h \cdot F_h(p_i). \tag{16}$$

DAF-Conv achieve adaptive modeling of local geometric features through separately processing near-field and far-field features for each frequency component, independent max-pooling aggregation for near-field and far-field features, and adaptive fusion of multiple frequency components through learned attention weights.

Compared to the single weight function in AF-Conv, DAF-Conv achieves more sophisticated feature extraction capabilities through its multi-frequency design and separate processing of near-field and far-field features.

3.4. Dynamic Acoustic Field Convolution Neural Network

To ensure each point has rich feature representation before constructing local neighborhoods, we build a Global Shape-Aware (GSA) layer before applying DAF-Conv. The GSA layer first aggregates each point through EdgeConv [55] without reducing point cloud resolution, then applies Multi-Head Attention (MHA) for global context modeling on the enhanced points. This layer effectively enhances DAF-Conv's perception ability for local structure modeling.

As shown in Fig. 2, based on the GSA layer and DAF-Conv, we construct the DAF Block. By stacking these blocks, we design a novel pretraining-free network, DAFNet, for point cloud analysis. We implement models for three major tasks:

Point Cloud Classification. For point cloud classification, we employ three DAF Blocks, followed by a classifier consisting of max-pooling and fully connected layers. The output feature dimensions of the three blocks are 64, 128, 256 respectively. Notably, with only 2.47M parameters, it achieves the best performance on ScanObjectNN [48].

Part Segmentation and Few-shot Point Cloud Semantic Segmentation. For part segmentation, the encoder structure remains identical to the classification task. In the decoder stage, we employ feature propagation to restore point cloud resolution. Between the encoder and decoder, we apply U-Net-like skip connections to fully utilize contextual information. In few-shot segmentation, while maintaining the same encoder-decoder structure as part segmentation, we incorporate the Query module from SegPN[77] to reduce domain gaps between query and support features.

4. Experiments

4.1. 3D Shape Classification

Real-world Object Classification on ScanObjectNN. We evaluate our method on ScanObjectNN [48], a challenging real-world point cloud dataset containing 2,902 objects across 15 categories. Following previous works, we conduct experiments on its most challenging variant (PB_T50_RS), where objects are presented with background and occlusion.

As shown in Table 1, DAFNet achieves state-of-the-art performance on ScanObjectNN [48], reaching 96.7% overall accuracy (OA) and 96.3% mean class accuracy (mAcc) with voting mechanism. These results significantly surpass both existing supervised methods (e.g., GPSFormer [52]: 95.4% OA) and pre-training approaches (e.g., ReCon++

Table 1. Classification results on the ScanObjectNN [48] dataset. "-" denotes unknown.

Method	Year	mAcc (%)	OA (%)
Supervised Methods			
PointNet [35]	2017	63.4	68.2
PointNet++ [36]	2018	69.8	73.7
DGCNN [55]	2019	73.6	78.1
SimpleView [15]	2021	-	80.5
MVTN [18]	2021	83.1	85.5
PointMLP [28]	2022	84.4	85.7
RepSurf-U [40]	2022	83.1	86.0
PointNeXt [39]	2022	85.8	87.7
Point-PN [68]	2023	-	87.1
DeLA [2]	2023	89.3	90.4
GPSFormer [52]	2024	93.8	95.4
Mamba3D [19]	2024	-	92.6
Pre-training Methods			
Point-BERT [65]	2021	-	83.1
Point-MAE [33]	2022	-	85.2
ULIP-2 + PointNeXt [61]	2023	91.2	91.5
PointGPT [3]	2023	-	93.6
ReCon [41]	2023	-	91.3
ReCon++ [38]	2024	-	95.25
PointMamba [25]	2024	-	89.3
PCP-MAE [70]	2024	-	90.4
Ours			
DAFNet (w/o voting)	-	96.1	96.4
DAFNet (w/ voting)	-	**96.3**	**96.7**

Table 2. Classification results on ModelNet40 dataset. "-" denotes unknown.

Methods	Year	mAcc (%)	OA (%)
Supervised Methods			
PointNet [35]	2017	86.0	89.2
PointNet++[36]	2017	-	90.7
DGCNN[55]	2019	90.2	92.9
Point Transformer [71]	2021	90.6	93.7
PAConv [60]	2021	-	93.9
CurveNet [32]	2021	-	94.2
PointMLP [28]	2022	91.4	94.5
RepSurf-U [40]	2022	-	94.7
GPSFormer [52]	2024	92.2	94.2
Mamba3D [19]	2024	-	94.1
Pre-training Methods			
Point-BERT [65]	2021	-	93.8
Point-MAE [33]	2022	-	94.0
ULIP + PointMLP [61]	2023	92.4	94.7
PointGPT [3]	2023	-	94.9
ReCon [41]	2023	-	94.5
ReCon++ [38]	2024	-	94.8
PointMamba [25]	2024	-	93.6
PCP-MAE [70]	2024	-	94.0
Ours			
DAFNet (w/o voting)	-	91.9	93.8
DAFNet (w/ voting)	-	92.2	94.1

[38]: 95.25% OA). The superior performance demonstrates the effectiveness of our acoustic field-inspired architecture in handling real-world point cloud data.

Synthetic Object Classification on ModelNet40. We further evaluate DAFNet on ModelNet40 [59], a widely-used synthetic dataset containing 12,311 CAD models from 40 categories, with 9,843 for training and 2,468 for testing.

As presented in Table 2, DAFNet achieves competitive results on ModelNet40 [59] with 94.1% overall accuracy (OA) and 92.2% mean class accuracy (mAcc). While these results are slightly lower than some recent methods like PointGPT [3] (94.9% OA), the performance gap is minimal (less than 1%). This is mainly because ModelNet40 [3], as a synthetic dataset, has been well-addressed by existing methods. In contrast, DAFNet shows more significant advantages on real-world datasets, as demonstrated by our results on ScanObjectNN [48].

4.2. Part Segmentation

To evaluate our method's capability in fine-grained shape understanding, we conduct experiments on ShapeNet-Part [64], which contains 16,881 shapes from 16 categories with 50 part annotations. Following convention, we use 13,807 samples for training and 2,874 for testing.

As shown in Table 3, DAFNet achieves competitive performance with 85.2% class mIoU and 86.8% instance mIoU, comparable to state-of-the-art methods. These results demonstrate that our acoustic field-inspired approach effectively captures both global shape information and local geometric details necessary for part segmentation tasks.

4.3. Few-shot Point Cloud Semantic Segmentation

4.3.1. Results analysis on the S3DIS dataset

We evaluate our model on S3DIS [1], which contains 3D point clouds from 272 rooms with 13 semantic categories. Following [73], we divide scenes into 1m × 1m blocks and sample 2048 points per block, resulting in 7547

Table 3. Part segmentation results (%) on ShapeNetPart. We report both class mIoU and instance mIoU.

Methods	Year	class mIoU	instance mIoU
Supervised Methods			
PointNet [35]	2017	80.4	83.7
PointNet++[36]	2017	81.9	85.1
DGCNN [55]	2019	82.3	85.1
KPConv [47]	2019	85.1	86.4
Point Transformer [71]	2021	83.7	86.6
CurveNet [32]	2021	-	86.8
PointNeXt [39]	2022	85.2	87.1
SPoTr [34]	2023	85.4	87.2
GPSFormer [52]	2024	85.4	86.8
Pre-training Methods			
PointGPT[3]	2023	84.1	86.2
ReCon [41]	2023	84.8	86.4
PointMamba[25]	2024	84.4	86.2
PCP-MAE [70]	2024	84.9	86.1
Ours			
DAFNet	-	85.2	86.8

Table 4. **Few-shot Results (%) on S3DIS.** S_i denotes the split i is used for testing. *Avg* is their average mIoU. The best results are shown in **bold**.

Method	2-way						3-way					
	1-shot			5-shot			1-shot			5-shot		
	S_0	S_1	*Avg*	S_0	S_1	*Avg*	S_0	S_1	*Avg*	S_0	S_1	*Avg*
DGCNN [55]	36.34	38.79	37.57	56.49	56.99	56.74	30.05	32.19	31.12	46.88	47.57	47.23
ProtoNet [44]	48.39	49.98	49.19	57.34	63.22	60.28	40.81	45.07	42.94	49.05	53.42	51.24
MPTI [73]	52.27	51.48	51.88	58.93	60.56	59.75	44.27	46.92	45.60	51.74	48.57	50.16
AttMPTI [73]	53.77	55.94	54.86	61.67	67.02	64.35	45.18	49.27	47.23	54.92	56.79	55.86
BFG [29]	55.60	55.98	55.79	63.71	66.62	65.17	46.18	48.36	47.27	55.05	57.80	56.43
2CBR [75]	55.89	61.99	58.94	63.55	67.51	65.53	46.51	53.91	50.21	55.51	58.07	56.79
PAP3D [20]	59.45	66.08	62.76	65.40	70.30	67.85	48.99	56.57	52.78	61.27	60.81	61.04
Seg-PN [77]	64.84	67.98	66.41	67.63	71.48	69.56	59.11	60.42	59.77	59.48	64.72	62.10
DAFNet	**68.13**	**70.27**	**69.20**	**70.51**	**73.15**	**71.83**	**61.33**	**65.55**	**63.44**	**65.25**	**68.67**	**66.96**
Improvement	+3.29	+2.29	+2.79	+2.88	+1.67	+2.27	+2.22	+5.13	+3.67	+5.77	+3.95	+4.86

blocks. As shown in Table 4, DAFNet achieves state-of-the-art performance across all few-shot settings. In 2-way experiments, our method achieves 69.20% mIoU in 1-shot and 71.83% mIoU in 5-shot settings, surpassing the previous best method Seg-PN [77] by 2.79% and 2.27% respectively. For the more challenging 3-way scenario, DAFNet maintains its superiority with 63.44% mIoU in 1-shot and 66.96% mIoU in 5-shot settings, outperforming Seg-PN [77] by 3.67% and 4.86%. These improvements across different settings demonstrate DAFNet's strong feature extraction capability and generalization ability.

4.3.2. Results analysis on the ScanNet dataset

We further evaluate DAFNet on ScanNet [7], which contains 1,513 real-world scanned scenes with 20 semantic categories. Following [73], we process the scenes into 36,350 blocks with 2,048 points each. As shown in Table 5, DAFNet achieves consistent improvements across all experimental settings. For 2-way scenarios, our method obtains 69.37% mIoU in 1-shot and 70.76% mIoU in 5-shot settings, surpassing Seg-PN [77] by 5.63% and 2.69% respectively. In the more challenging 3-way settings, DAFNet achieves 66.42% mIoU in 1-shot and 70.46% mIoU in 5-shot settings, outperforming Seg-PN [77] by 2.85% and 4.86%. These substantial improvements across different settings demonstrate DAFNet's strong generalization ability and robustness in handling complex real-world scenes with limited training samples.

Table 5. Few-shot Results (%) on ScanNet. S_i denotes the split i is used for testing. *Avg* is their average mIoU. The best results are shown in **bold**.

Method	2-way						3-way					
	1-shot			5-shot			1-shot			5-shot		
	S_0	S_1	*Avg*	S_0	S_1	*Avg*	S_0	S_1	*Avg*	S_0	S_1	*Avg*
DGCNN [55]	31.55	28.94	30.25	42.71	37.24	39.98	23.99	19.10	21.55	34.93	28.10	31.52
ProtoNet [44]	33.92	30.95	32.44	45.34	42.01	43.68	28.47	26.13	27.30	37.36	34.98	36.17
MPTI [73]	39.27	36.14	37.71	46.90	43.59	45.25	29.96	27.26	28.61	38.14	34.36	36.25
AttMPTI [73]	42.55	40.83	41.69	54.00	50.32	52.16	35.23	30.72	32.98	46.74	40.80	43.77
BFG [29]	42.15	40.52	41.34	51.23	49.39	50.31	34.12	31.98	33.05	46.25	41.38	43.82
2CBR [75]	50.73	47.66	49.20	52.35	47.14	49.75	47.00	46.36	46.68	45.06	39.47	42.27
PAP3D [20]	57.08	55.94	56.51	64.55	59.64	62.10	55.27	55.60	55.44	59.02	53.16	56.09
Seg-PN [77]	63.15	64.32	63.74	67.08	69.05	68.07	61.80	65.34	63.57	62.94	68.26	65.60
DAFNet	**68.79**	**69.95**	**69.37**	**70.91**	**70.60**	**70.76**	**66.14**	**66.70**	**66.42**	**68.97**	**71.95**	**70.46**
Improvement	+5.64	+5.63	+5.63	+3.83	+1.55	+2.69	+4.34	+1.36	+2.85	+6.03	+3.69	+4.86

Table 6. Effect of different components of DAFNet.

GSA Layer		DAF-Conv	OA (%)
EdgeConv	MHA		
Individual Components			
✓	✗	✗	88.6
✗	✓	✗	89.4
✗	✗	✓	91.7
Component Combinations			
✓	✓	✗	91.3
✓	✗	✓	94.8
✗	✓	✓	93.5
✓	✓	✓	**96.7**

4.4. Ablation Study

All experiments are conducted on the ScanObjectNN [48] dataset unless explicitly stated otherwise.

4.4.1. Ablation Study of Each Component

We conduct ablation experiments to analyze the effectiveness of each component in DAFNet, as shown in Table 6. Among individual components, DAF-Conv shows the strongest performance with 91.7% accuracy, notably higher than single GSA layer components (EdgeConv: 88.6%, MHA: 89.4%). When combining components, the fusion of DAF-Conv with EdgeConv achieves 94.8%, while DAF-

Table 7. Effect of different local geometric priors h_j.

Local Geometric Prior h_j	OA(%)
$[p_j]$	93.6
$[p_j, p_i]$	94.7
$[p_j, p_i p_j - p_i]$	96.2
$[p_j, p_i, p_j - p_i, \|\|p_j - p_i\|\|]$	**96.7**

Conv with MHA reaches 93.5%, both showing substantial improvements over single component performance. The full model integrating all components achieves the best performance of 96.7%, demonstrating the strong complementary effects between DAF-Conv's local geometric modeling and the GSA layer's global context capturing.

4.4.2. Ablation Study of Local Geometric Prior

As shown in Table 7, we conduct ablation studies on the composition of local geometric prior h_j. Starting with only center point coordinates $[p_j]$ yields 93.6% accuracy. Incorporating neighbor point information $[p_j, p_i]$ brings a 1.1% improvement. Adding relative position $[p_j, p_i, p_j - p_i]$ further boosts performance by 1.5%. The complete geometric prior with Euclidean distance $||p_j - p_i||$ achieves the best result of 96.7%, validating the importance of comprehensive geometric information in local structure modeling.

4.4.3. Analysis of Model Hyperparameters

We conduct comprehensive ablation studies on three key hyperparameters of DAFNet, as shown in Fig. 4.

Frequency Component Number H: Fig. 4(a) shows that model performance improves as H increases from 1 to 6, achieving peak accuracy at $H = 6$ (96.7%). Further increasing H leads to performance degradation, dropping to 95.1% at $H = 10$. This indicates that while multiple frequency components enhance feature diversity, excessive components may introduce redundant information.

Distance Threshold $\mathbf{d_{threshold}}$: As shown in Fig. 4(b), we examine the impact of $d_{threshold}$ which divides the local neighborhood into near-field and far-field regions. The optimal performance (96.7%) is achieved at $d_{threshold} = 0.6$. Lower thresholds limit near-field information, while higher values may include irrelevant points in the near-field region, both resulting in suboptimal performance.

Neighborhood Size K: Fig. 4(c) demonstrates the effect of neighborhood size K in EdgeConv. Performance reaches its peak (96.7%) at $K = 20$. Smaller K values provide insufficient local context, while larger values may incorporate irrelevant points, both leading to decreased accuracy.

4.5. Model Complexity

We compare the model complexity of DAFNet with existing methods on ScanObjectNN [48], as shown in Table 8. While achieving state-of-the-art performance (96.7% OA), DAFNet maintains moderate computational requirements with only 2.47M parameters and 1.1G FLOPs. This is significantly more efficient than models like MVTN [18] (11.2M parameters, 43.7G FLOPs) and PointMLP [28] (12.6M parameters, 31.4G FLOPs), demonstrating that our acoustic field-inspired design achieves superior accuracy without sacrificing computational efficiency.

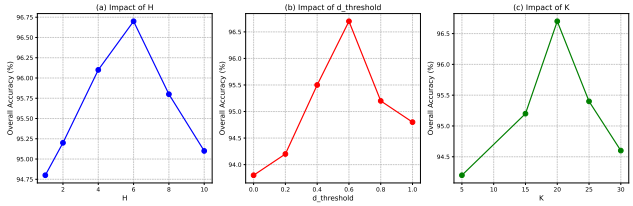


Figure 4. Ablation studies on key hyperparameters. (a) Impact of frequency component number H in DAF-Conv. (b) Effect of distance threshold $d_{threshold}$ for near-field and far-field separation. (c) Influence of neighborhood size K in EdgeConv layer.

Table 8. Model complexity comparison on ScanObjectNN.

Methods	OA(%)	Params.(M)	FLOPs(G)
PointNet [35]	68.0	3.5	0.5
PointNet++[36]	77.9	1.5	1.7
DGCNN[55]	78.1	1.8	2.4
PointCNN [24]	78.5	0.6	-
MVTN [18]	82.8	11.2	43.7
PointMLP [28]	85.4	12.6	31.4
PointNeXt [39]	87.4	1.4	3.6
DAFNet (Ours)	**96.7**	2.47	1.1

5. Conclusion

In this paper, we presented DAFNet, a physics-inspired architecture for point cloud understanding that addresses local structure representation challenges. Inspired by acoustic field principles, we introduced AF-Conv modeling local feature aggregation through acoustic propagation, and extended it to DAF-Conv which dynamically adapts to diverse geometric features through multi-frequency simulation. By integrating these with GSA, our approach captures both fine-grained local structures and global context. Experiments demonstrate DAFNet achieves state-of-the-art performance across multiple tasks. Future work will extend acoustic field modeling to other domains, exploring applications in additional 3D tasks.

Acknowledgment

This project is sponsored by Shanghai Pujiang Programme 24PJD030. The computational work for this article is partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[1] 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 *CVPR*, pages 1534–1543, 2016. 6

[2] Binjie Chen, Yunzhou Xia, Yu Zang, Cheng Wang, and Jonathan Li. Decoupled local aggregation for point cloud learning. *arXiv preprint arXiv:2308.16532*, 2023. 6

[3] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *arXiv preprint arXiv:2305.11487*, 2023. 6, 7

[4] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *NeurIPS*, 36, 2024. 3

[5] Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *CVPR*, pages 21886–21896, 2024. 2

[6] Pranav Singh Chib and Pravendra Singh. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles*, 2023. 1

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 7

[8] Jeevan S Devagiri, Sidike Paheding, Quamar Niyaz, Xiaoli Yang, and Samantha Smith. Augmented reality and artificial intelligence in industry: Trends, tools, and future challenges. *Expert Systems with Applications*, 207:118002, 2022. 1

[9] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, pages 2694–2703, 2023. 2

[10] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *ICCV*, pages 20224–20234, 2023. 2

[11] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022. 3

[12] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *CVPR*, pages 8458–8468, 2022. 2

[13] Xiang Fang, Wanlong Fang, Changshuo Wang, Daizong Liu, Keke Tang, Jianfeng Dong, Pan Zhou, and Beibei Li. Multi-pair temporal sentence grounding via multi-thread knowledge transfer network. *arXiv preprint arXiv:2412.15678*, 2024. 2

[14] Ruchi Goel and Pooja Gupta. Robotics and industry 4.0. *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*, pages 157–169, 2020. 1

[15] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *ICML*, pages 3809–3820. PMLR, 2021. 6

[16] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2

[17] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021. 2

[18] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *ICCV*, pages 1–11, 2021. 6, 8

[19] Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. In *ACM MM*, pages 4995–5004, 2024. 1, 2, 6

[20] Shuting He, Xudong Jiang, Wei Jiang, and Henghui Ding. Prototype adaption and projection for few-and zero-shot 3d point cloud semantic segmentation. *TIP*, 32:3199–3211, 2023. 7

[21] Limin Jiang, Changshuo Wang, Xin Ning, and Zaiyang Yu. Lttpoint: a mlp-based point cloud classification method with local topology transformation module. In *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*, pages 783–789. IEEE, 2023. 1

[22] Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders. *Fundamentals of acoustics*. John wiley & sons, 2000. 4

[23] Artem Komarichev, Zichun Zhong, and Jing Hua. A-cnn: Annularly convolutional neural networks on point clouds. In *CVPR*, pages 7421–7430, 2019. 2

[24] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *NeurIPS*, 31, 2018. 8

[25] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024. 1, 3, 6, 7

[26] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *NeurIPS*, 2024. 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 *ICCV*, pages 10012–10022, 2021. 2

[28] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022. 6, 8

[29] Yongqiang Mao, Zonghao Guo, LU Xiaonan, Zhiqiang Yuan, and Haowen Guo. Bidirectional feature globalization for few-shot semantic segmentation of 3d point cloud scenes. In *2022 International Conference on 3D Vision (3DV)*, pages 505–514. IEEE, 2022. 7

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

[31] Philip McCord Morse and K Uno Ingard. *Theoretical acoustics*. Princeton university press, 1986. 4

[32] AAM Muzahid, Wanggen Wan, Ferdous Sohel, Lianyao Wu, and Li Hou. Curvenet: Curvature-based multitask learning deep networks for 3d object recognition. *IEEE/CAA Journal of Automatica Sinica*, 8(6):1177–1187, 2020. 6, 7

[33] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, pages 604–621. Springer, 2022. 1, 3, 6

[34] Jinyoung Park, Sanghyeok Lee, Sihyeon Kim, Yunyang Xiong, and Hyunwoo J Kim. Self-positioning point-based transformer for point cloud understanding. In *CVPR*, pages 21814–21823, 2023. 7

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 2, 6, 7, 8

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NIPS*, 30, 2017. 2, 6, 7, 8

[37] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *ICML*, pages 28223–28243. PMLR, 2023. 1, 3

[38] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *ECCV*, pages 214–238. Springer, 2025. 1, 6

[39] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *NIPS*, 35:23192–23204, 2022. 6, 7, 8

[40] Haoxi Ran, Jun Liu, and Chengjie Wang. Surface representation for point clouds. In *CVPR*, pages 18942–18952, 2022. 6

[41] Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David Choffnes. Recon: Revealing and controlling pii leaks in mobile network traffic. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pages 361–374, 2016. 6, 7

[42] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, pages 3577–3586, 2017. 2

[43] Mickael Sereno, Xiyao Wang, Lonni Besançon, Michael J Mcguffin, and Tobias Isenberg. Collaborative work in augmented reality: A survey. *TVCG*, 28(6):2530–2549, 2020. 1

[44] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *NIPS*, 30, 2017. 7

[45] Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 3:54–70, 2023. 1

[46] Pei Sun, Mingxing Tan, Weiyue Wang, Chenxi Liu, Fei Xia, Zhaoqi Leng, and Dragomir Anguelov. Swformer: Sparse window transformer for 3d object detection in point clouds. In *ECCV*, pages 426–442. Springer, 2022. 2

[47] 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 *ICCV*, pages 6411–6420, 2019. 7

[48] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, pages 1588–1597, 2019. 5, 6, 7, 8

[49] Changshuo Wang, Xin Ning, Linjun Sun, Liping Zhang, Weijun Li, and Xiao Bai. Learning discriminative features by covering local geometric space for point cloud analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15, 2022. 2

[50] C Wang, H Wang, X Ning, S Tian, and W Li. 3d point cloud classification method based on dynamic coverage of local area. *J. Softw*, 34(4):1962–1976, 2022. 1

[51] Changshuo Wang, Xin Ning, Weijun Li, Xiao Bai, and Xingyu Gao. 3d person re-identification based on global semantic guidance and local feature aggregation. *TCSVT*, 34 (6):4698–4712, 2023. 2

[52] Changshuo Wang, Meiqing Wu, Siew-Kei Lam, Xin Ning, Shangshu Yu, Ruiping Wang, Weijun Li, and Thambipillai Srikanthan. Gpsformer: A global perception and local structure fitting-based transformer for point cloud understanding. *arXiv preprint arXiv:2407.13519*, 2024. 2, 5, 6, 7

[53] Changshuo Wang, Rongsheng Cao, and Ruiping Wang. Learning discriminative topological structure information representation for 2d shape and social network classification via persistent homology. *Knowledge-Based Systems*, page 113125, 2025. 2

[54] Ruiping Wang, Siew-Kei Lam, Meiqing Wu, Zhijian Hu, Changshuo Wang, and Jing Wang. Destination intention estimation-based convolutional encoder-decoder for pedestrian trajectory multimodality forecast. *Measurement*, 239: 115470, 2025. 1

[55] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 2, 4, 5, 6, 7, 8

[56] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *CVPR*, pages 1850–1859, 2020. 2

[57] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 2

[58] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *CVPR*, pages 4840–4851, 2024. 2

[59] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 6

[60] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *CVPR*, pages 3173–3182, 2021. 4, 6

[61] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *CVPR*, pages 1179–1189, 2023. 3, 6

[62] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *NIPS*, 32, 2019. 4, 5

[63] Ze Yang and Liwei Wang. Learning relationships for multi-view 3d object recognition. In *ICCV*, pages 7505–7514, 2019. 2

[64] 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

[65] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, pages 19313–19322, 2022. 6

[66] Huang Zhang, Changshuo Wang, Long Yu, Shengwei Tian, Xin Ning, and Joel Rodrigues. Pointgt: A method for point-cloud classification and segmentation based on local geometric transformation. *IEEE Transactions on Multimedia*, 2024. 2

[67] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *NeurIPS*, 35:27061–27074, 2022. 3

[68] Renrui Zhang, Liuhui Wang, Ziyu Guo, Yali Wang, Peng Gao, Hongsheng Li, and Jianbo Shi. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. *arXiv preprint arXiv:2303.08134*, 2023. 6

[69] Tao Zhang, Xiangtai Li, Haobo Yuan, Shunping Ji, and Shuicheng Yan. Point could mamba: Point cloud learning via state space model. *arXiv preprint arXiv:2403.00762*, 2024. 2

[70] Xiangdong Zhang, Shaofeng Zhang, and Junchi Yan. Pcp-mae: Learning to predict centers for point masked autoencoders. *Advances in Neural Information Processing Systems*, 37:80303–80327, 2024. 3, 6, 7

[71] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pages 16259–16268, 2021. 2, 6, 7

[72] Jingyuan Zhao, Wenyi Zhao, Bo Deng, Zhenghong Wang, Feng Zhang, Wenxiang Zheng, Wanke Cao, Jinrui Nan, Yubo Lian, and Andrew F Burke. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications*, page 122836, 2023. 1

[73] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *CVPR*, pages 8873–8882, 2021. 6, 7

[74] Xiao Zheng, Xiaoshui Huang, Guofeng Mei, Yuenan Hou, Zhaoyang Lyu, Bo Dai, Wanli Ouyang, and Yongshun Gong. Point cloud pre-training with diffusion models. In *CVPR*, pages 22935–22945, 2024. 3

[75] Guanyu Zhu, Yong Zhou, Rui Yao, and Hancheng Zhu. Cross-class bias rectification for point cloud few-shot segmentation. *TMM*, 25:9175–9188, 2023. 7

[76] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 2

[77] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Jiaming Liu, Han Xiao, Chaoyou Fu, Hao Dong, and Peng Gao. No time to train: Empowering non-parametric networks for few-shot 3d scene segmentation. In *CVPR*, pages 3838–3847, 2024. 5, 7