PointGP: Geometry-Primed Attention for Point Cloud Analysis

Yong Yang[1], **Jianming Huang**[1*], **Mengyuan Ge**[2], **Chunyang Huang**[1], **Bingbing Hu**[1] and **Junfeng Yao**[1,2,3*]
[1]School of Film, Xiamen University
[2]School of Informatics, Xiamen University
[3]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism
huangjm@xmu.edu.cn, yao0010@xmu.edu.cn

Abstract

Transformer-based architectures have demonstrated strong performance in 3D point cloud understanding, yet many existing methods generate attention weights mainly from semantic feature similarity. In deep networks, feature-centric attention may become less selective as point features are progressively smoothed, a behavior associated with feature homogenization and rank collapse, which can weaken the structural discrimination of local aggregation. We propose PointGP, a geometry-primed framework that uses rectified local geometric topology as the primary cue for attention generation. PointGP introduces a Semantic-Guided Manifold Rectifier to predict feature-conditioned local coordinate offsets, and a Dual-Stream Geometric Kernel to compute attention logits from both raw and rectified geometric cues. By reducing reliance on explicit query-key feature matching while implicitly incorporating semantic guidance through geometric rectification, PointGP provides an effective and efficient mechanism for local point aggregation. Experiments across five benchmarks covering classification, part segmentation, and indoor scene segmentation show that PointGP achieves competitive accuracy with strong parameter and computational efficiency compared with representative strong baselines.

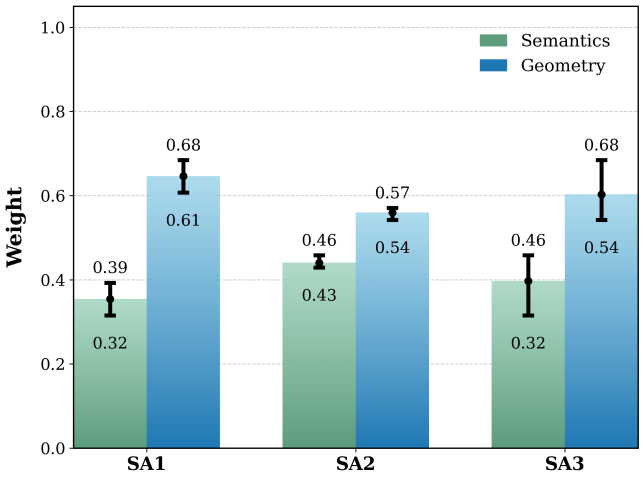

Figure 1: **Visualization of learned semantic and geometric weighting magnitudes.** We analyze the learned weighting magnitudes assigned to semantic features (green) and geometric cues (blue) across three Set Abstraction (SA) stages on ModelNet40 [Wu *et al.*, 2015]. Bar heights denote the mean weights, and error bars indicate the range across attention heads. The results suggest that geometric cues receive consistently higher weights across stages.

1 Introduction

The rapid evolution of 3D sensing technologies has established point clouds as a fundamental representation for digitizing the physical world. The increasing accessibility of high-quality 3D data has enabled a wide range of applications, including autonomous driving, robotic navigation, AR/VR, and large-scale reconstruction.

In response to this growing demand, extensive research has been devoted to developing learning architectures that can capture fine-grained geometry and semantic context in unstructured 3D spaces. Existing methods include convolutional neural networks (CNNs) [Thomas *et al.*, 2019], graph neural networks (GNNs) [Shi and Rajkumar, 2020], attention-based models [Zhao *et al.*, 2021], and recent Mamba-based frameworks [Bahri *et al.*, 2025] that leverage state-space representations for efficient sequence modeling.

Among these methods, Transformer-based architectures have shown strong effectiveness by modeling contextual dependencies through self-attention. However, most existing 3D Transformers still follow a feature-centric design paradigm, where attention weights are mainly derived from semantic feature similarity and geometric information is often introduced as auxiliary positional encoding. While this design is effective in many cases, directly transferring feature-similarity-based attention to point clouds may be suboptimal. Unlike text tokens or image patches, whose positions are defined on regular domains, point cloud coordinates are irregular, unordered, and closely tied to the underlying local shape structure. Therefore, local geometric relations provide informative structural cues that should play a more direct role in

*Corresponding authors.

attention generation.

As illustrated in Fig. 1, our empirical analysis shows that geometric cues tend to receive higher gating weights than semantic features across Set Abstraction stages, including deeper stages. This observation suggests that local geometry can provide useful structural cues in settings where feature representations may become progressively smoothed, motivating a geometry-primed perspective for point cloud attention. Based on this observation, we propose **PointGP**, a geometry-primed framework for point cloud analysis. Instead of directly using semantic feature similarity to determine attention weights, PointGP generates attention weights primarily from geometry-aware structural cues.

Our contributions are summarized as follows:

- We revisit feature-centric attention in 3D point cloud Transformers and provide empirical observations showing that local geometric cues can serve as stable structural anchors for 3D attention generation.
- We propose **PointGP**, a geometry-primed attention framework that reduces reliance on explicit query-key feature matching and generates attention weights from geometry-aware structural cues, enabled by feature-conditioned local coordinate rectification and attention logit computation from both raw and rectified geometric cues.
- Extensive experiments on ModelNet40 [Wu *et al.*, 2015], ScanObjectNN [Uy *et al.*, 2019], ShapeNet-Part [Chang *et al.*, 2015], S3DIS [Armeni *et al.*, 2016], and ScanNet v2 [Dai *et al.*, 2017] demonstrate that PointGP achieves competitive performance with strong parameter and computational efficiency compared with representative strong baselines.

2 Related Work

2.1 Attention Mechanisms in Point Clouds

Transformers have been widely explored for 3D point cloud understanding due to their ability to model contextual dependencies in unordered point sets. PCT [Guo *et al.*, 2021] adapted global self-attention to point clouds with offset-attention, but its global design introduces quadratic complexity with respect to the number of points. To improve local feature aggregation, Point Transformer (PTv1) [Zhao *et al.*, 2021] introduced vector attention within local k-nearest neighborhoods. PTv2 [Wu *et al.*, 2022] further improved this design with Grouped Vector Attention (GVA) and partition-based pooling, while PTv3 [Wu *et al.*, 2024] scaled point Transformers using serialized representations based on space-filling curves.

2.2 Adaptive Geometry and Position Awareness

Geometric modeling is a central problem in point cloud analysis. KPConv [Thomas *et al.*, 2019] introduced kernel point convolution with deformable kernels to adapt receptive fields to local geometric structures. AdaptConv [Zhou *et al.*, 2021] dynamically generates graph convolution kernels from point-pair features, improving the flexibility of local aggregation. Position-aware operators, such as relative position encoding [Wu *et al.*, 2021] and point-specific positional encodings [Xiao *et al.*, 2025], further demonstrate the importance of spatial information in local feature learning. In point Transformers, PTv2 [Wu *et al.*, 2022] explicitly integrates relative position encodings into attention computation, strengthening spatial awareness in local neighborhoods. In addition, PointNet-style symmetric aggregation [Qi *et al.*, 2017a] has been widely used to summarize local or global point-set context, providing a simple yet effective way to encode neighborhood-level structural information.

2.3 Discussion and Positioning of PointGP.

PointGP is closely related to point Transformers and adaptive geometric modeling methods, but differs in the way attention weights are generated. Most point Transformer variants derive attention weights primarily from semantic feature interactions, such as dot products, feature subtraction, or MLP-based comparisons, while geometry is typically incorporated as positional encoding or an additive spatial term. In contrast, PointGP uses semantic features to rectify local geometric coordinates before attention computation, allowing geometry-aware structural cues to directly guide attention generation. In this way, PointGP shifts geometry from a supplementary positional term to a primary source for computing attention distributions, while still retaining semantic guidance through local geometric rectification.

3 Analysis: Geometry-Semantic Duality

In this section, we revisit feature-centric attention in point cloud Transformers and analyze why local geometry can provide complementary structural cues for attention generation. These analyses motivate a practical geometry-primed design, where semantic features modulate local geometric relations and attention weights are computed primarily from geometry-aware cues.

3.1 Motivation: Limitations of Feature-Centric Attention

Many point Transformer variants compute attention weights from semantic feature interactions, such as dot products, subtraction, or MLP-based feature comparisons. This design is effective when point features remain sufficiently discriminative. However, in deep networks, repeated local aggregation may progressively smooth feature representations. This phenomenon is closely related to feature homogenization and rank collapse [Dong *et al.*, 2021], and may make feature-similarity-based attention less selective.

Let $\mathbf{F}^{(l)} \in \mathbf{R}^{N\times C}$ denote the point feature matrix at layer l, where $\mathbf{f}_i^{(l)}$ is the feature of point i. When features within a local neighborhood become overly similar, i.e.,

$$\mathbf{f}_i^{(l)} \approx \mathbf{f}_j^{(l)}, \quad j \in \mathcal{N}(i), \tag{1}$$

the attention logits computed from feature similarity also become less distinguishable. For a standard dot-product attention formulation,

$$s_{ij} = \frac{(\mathbf{f}_i\mathbf{W}_Q)\,(\mathbf{f}_j\mathbf{W}_K)^T}{\sqrt{d_k}}, \tag{2}$$

similar neighboring features lead to logits with small relative variation. In the limiting case where $s_{ij} \approx S$ for all $j \in \mathcal{N}(i)$, the Softmax-normalized attention weights approach a uniform distribution:

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{k\in\mathcal{N}(i)} \exp(s_{ik})} \approx \frac{\exp(S)}{|\mathcal{N}(i)| \cdot \exp(S)} = \frac{1}{|\mathcal{N}(i)|}. \tag{3}$$

This simplified analysis indicates that feature-centric attention may lose local selectivity when semantic features are excessively smoothed, making the aggregation behave similarly to uniform local averaging. For clarity, we refer to this tendency as *attention collapse* in this paper.

In contrast, relative coordinates

$$\Delta\mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i \tag{4}$$

remain directly tied to the spatial arrangement of local neighborhoods. Although raw Euclidean geometry is not always sufficient to describe task-relevant semantic relations, it provides explicit structural cues that do not depend on feature similarity alone. This motivates a geometry-primed attention design, where local geometric relations serve as the primary input for attention-logit generation, while semantic features are used to guide geometric rectification.

3.2 Feature-Conditioned Local Coordinate Rectification

Raw Euclidean neighborhoods can be imperfect for point cloud analysis. For example, points that are close in Euclidean space may belong to different object parts or surfaces, while points on the same semantic structure may be separated by curvature, sparsity, occlusion, or sampling noise. Therefore, directly using raw coordinates may introduce a mismatch between Euclidean proximity and task-relevant local relations.

To describe this mismatch, let the point cloud $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N} \subset \mathbf{R}^3$ be a discrete observation of an underlying surface or local structure. Practical point cloud encoders usually construct neighborhoods using the ambient Euclidean distance:

$$d_E(\mathbf{p}_i, \mathbf{p}_j) = \|\mathbf{p}_i - \mathbf{p}_j\|_2. \tag{5}$$

For curved, sparse, or noisy point clouds, this Euclidean neighborhood can deviate from the desired task-relevant neighborhood. We therefore introduce a learnable local coordinate rectification mechanism to adjust the geometric relation before attention generation.

Given a center point $\mathbf{p}_i$ and one of its neighbors $\mathbf{p}_j$, we predict a local coordinate offset from their feature difference:

$$\Delta\mathbf{o}_{ij} = \mathcal{H}_\theta\left(\mathbf{f}_j - \mathbf{f}_i\right), \tag{6}$$

where $\mathcal{H}_\theta$ is a learnable projection function. The rectified relative coordinate is then defined as

$$\widetilde{\Delta\mathbf{p}}_{ij} = \Delta\mathbf{p}_{ij} + \Delta\mathbf{o}_{ij}. \tag{7}$$

This formulation provides a task-driven local topology or metric adjustment: semantic feature differences guide the rectification of local geometric relations before attention generation. The resulting rectified coordinates remain in a compact geometric space suitable for efficient attention-logit computation.

Based on both raw and rectified geometric cues, PointGP computes geometry-aware attention logits:

$$s_{ij}^{\text{geo}} = \Psi_\theta\left(\Delta\mathbf{p}_{ij}, \widetilde{\Delta\mathbf{p}}_{ij}\right), \tag{8}$$

where Ψ_θ denotes the geometric kernel for attention-logit generation. In this way, semantic information affects attention indirectly through coordinate rectification, while the final attention distribution is derived from geometry-aware structural cues. This design reduces reliance on explicit query-key feature matching and provides an efficient alternative for local point aggregation.

3.3 Empirical Observation: Geometry Remains Informative

To empirically examine the roles of semantic and geometric cues, we design a probe attention module with two decoupled streams. The semantic stream computes logits from feature interactions, while the geometric stream computes logits from relative coordinates. Their contributions are controlled by learnable scalar gates w_{feat} and w_{geo}, which are normalized by a softmax constraint:

$$w_{\text{feat}} + w_{\text{geo}} = 1. \tag{9}$$

The combined probe logit is computed as

$$g_{ij} = w_{\text{geo}}\Psi_{\text{geo}}(\Delta\mathbf{p}_{ij}) + w_{\text{feat}}\Psi_{\text{sem}}(\mathbf{f}_i, \mathbf{f}_j). \tag{10}$$

The corresponding attention weight is then obtained by Softmax normalization:

$$\alpha_{ij} = \frac{\exp(g_{ij})}{\sum_{k\in\mathcal{N}(i)} \exp(g_{ik})}. \tag{11}$$

This decoupled design allows us to examine the relative contributions of semantic and geometric cues under the same local aggregation framework.

As shown in Fig. 1, in this probe setting, the learned gates suggest that geometric cues tend to receive higher weights than semantic features across Set Abstraction stages on ModelNet40 [Wu *et al.*, 2015]. This observation suggests that local geometry remains informative for attention generation, especially in settings where semantic representations may become progressively smoothed. Meanwhile, the non-negligible semantic gates suggest that semantic information remains complementary to geometry. These observations motivate the design of PointGP, where semantic features guide local coordinate rectification and attention weights are generated from both raw and rectified geometric cues.

4 Methodology

4.1 Overview

PointGP adopts a hierarchical encoder-decoder architecture for point cloud analysis. Given an input point cloud $\mathcal{P} \in \mathbf{R}^{N\times 3}$ and associated point features $\mathcal{F} \in \mathbf{R}^{N\times C}$, the basic building block is the **Local Aggregation** module. The core of this module is the proposed **Geometry-Primed Attention**

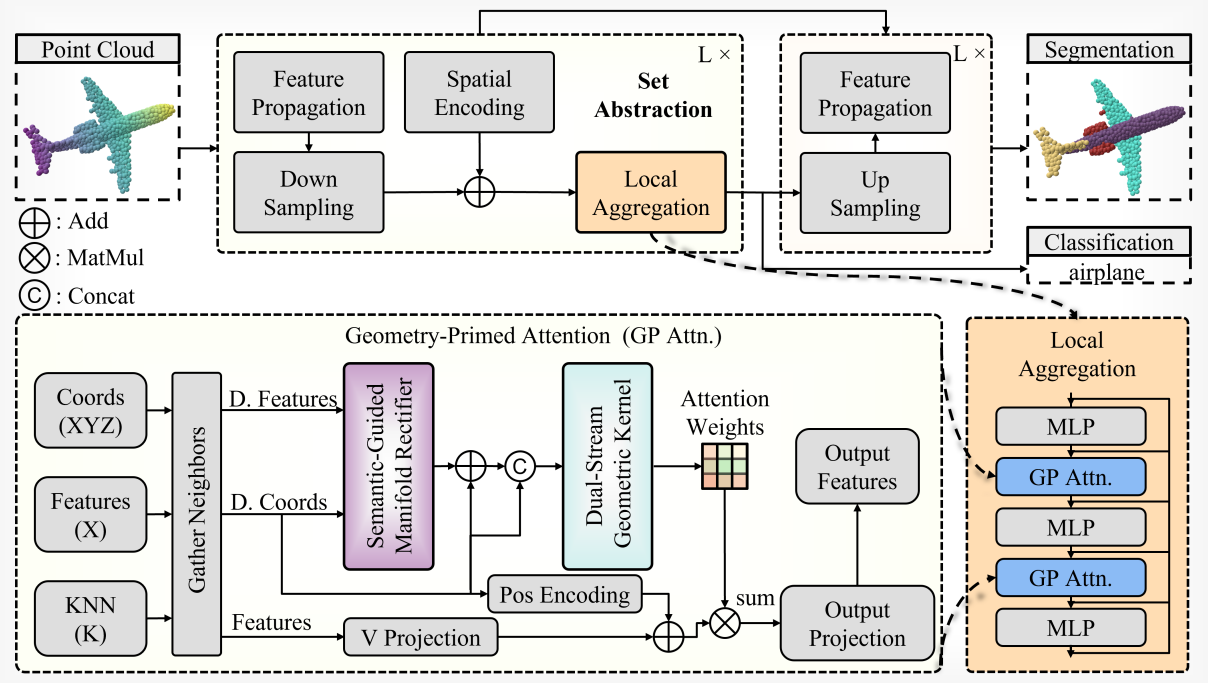

Figure 2: **Overview of the PointGP architecture.** PointGP adopts a hierarchical encoder architecture, with decoder-style feature propagation used for segmentation tasks. The core component is the Geometry-Primed Attention (GPA) module (bottom-left), which consists of the Semantic-Guided Manifold Rectifier and the Dual-Stream Geometric Kernel. The Local Aggregation block (bottom-right) stacks GPA and Feed-Forward Network layers with residual connections.

(GPA) mechanism, which generates attention weights primarily from geometry-aware structural relations. Semantic features do not directly form query-key attention logits; instead, they influence attention generation through local coordinate rectification, as illustrated in Fig. 2.

4.2 Semantic-Guided Manifold Rectifier

Standard k-NN grouping is usually constructed in the raw Euclidean space, which may not always align well with task-relevant local relations under semantic inconsistency, curvature, sparsity, occlusion, or sampling irregularity. To alleviate this mismatch, we introduce the *Semantic-Guided Manifold Rectifier* (SGMR), a learnable local topology/metric rectifier that predicts feature-conditioned coordinate offsets before attention-logit generation.

Let $\mathcal{N}(i)$ denote the k-nearest neighbors of center point $\mathbf{p}_i$ in the current abstraction layer. For each neighbor $\mathbf{p}_j$, SGMR first computes the local feature difference:

$$\delta \mathbf{f}_{ij} = \mathbf{f}_j - \mathbf{f}_i, \tag{12}$$

which provides semantic-aware guidance for local geometric rectification. A lightweight MLP $\Phi(\cdot)$ then maps this feature difference to a 3D coordinate offset:

$$\Delta \mathbf{o}_{ij} = \Phi(\delta \mathbf{f}_{ij}; \Theta_\Phi), \tag{13}$$

where Θ_Φ denotes learnable parameters.

Given the raw relative coordinate

$$\Delta \mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i, \tag{14}$$

the rectified relative coordinate is obtained as:

$$\Delta \tilde{\mathbf{p}}_{ij} = \Delta \mathbf{p}_{ij} + \Delta \mathbf{o}_{ij}. \tag{15}$$

We initialize the last projection layer of $\Phi(\cdot)$ with zeros, so SGMR starts from the raw Euclidean geometry and gradually learns task-adaptive coordinate rectification during training. The rectified coordinate $\Delta \tilde{\mathbf{p}}_{ij}$ then serves as a feature-conditioned geometric cue for subsequent attention-logit generation, helping the attention kernel better differentiate task-relevant neighbors from spatially close but semantically inconsistent points without explicit query-key feature matching.

4.3 Dual-Stream Geometric Kernel

To generate geometry-aware attention logits, we introduce the *Dual-Stream Geometric Kernel* (DSGK). DSGK first concatenates the raw and rectified relative coordinates into a composite geometric descriptor:

$$\mathbf{g}_{ij} = [\Delta \mathbf{p}_{ij} \oplus \Delta \tilde{\mathbf{p}}_{ij}] \in \mathbf{R}^6, \tag{16}$$

where $\oplus$ denotes channel-wise concatenation. The raw coordinate preserves the original local spatial organization, while the rectified coordinate provides task-adaptive geometric refinement.

DSGK processes $\mathbf{g}_{ij}$ with two parallel streams. The local stream models fine-grained edge-level geometry using a shared MLP ψ_{loc}:

$$\mathbf{h}_{ij}^{\text{loc}} = \psi_{\text{loc}}(\mathbf{g}_{ij}). \tag{17}$$

The contextual stream summarizes neighborhood-level geometric structure by applying a shared MLP ψ_{ctx} followed by symmetric max pooling:

$$\mathbf{h}_i^{\text{ctx}} = \max_{j \in \mathcal{N}(i)} \left(\psi_{\text{ctx}}(\mathbf{g}_{ij}) \right). \tag{18}$$

The context vector $\mathbf{h}_i^{\text{ctx}}$ is broadcast back to each neighbor and fused with the local feature to produce vector attention logits:

$$\mathcal{A}_{ij} = \psi_{\text{fuse}} \left([\mathbf{h}_{ij}^{\text{loc}} \oplus \mathbf{h}_i^{\text{ctx}}] \right), \tag{19}$$

where ψ_{fuse} is a fusion MLP and $\mathcal{A}_{ij} \in \mathbf{R}^{C_v}$ matches the channel dimension of the value feature.

Through this design, DSGK combines pairwise geometric details with compact neighborhood-level structural context for attention-logit generation. Although the contextual stream adopts PointNet-style symmetric pooling, it operates on geometric descriptors rather than directly aggregating point features.

4.4 Geometry-Primed Attention Aggregation

Given the attention logits generated by DSGK, GPA aggregates local features in a geometry-primed manner. Input features are first projected into value vectors:

$$\mathbf{v}_j = \mathbf{f}_j \mathbf{W}_v, \tag{20}$$

where $\mathbf{v}_j \in \mathbf{R}^{C_v}$. To preserve explicit spatial information in the aggregated content, we add value-side positional encoding from the raw relative coordinate:

$$\tilde{\mathbf{v}}_{ij} = \mathbf{v}_j + \zeta(\Delta \mathbf{p}_{ij}), \tag{21}$$

where $\zeta(\cdot)$ maps $\Delta \mathbf{p}_{ij}$ to the value-channel dimension. Raw geometry is used for value encoding to retain physical spatial information, while rectified geometry guides attention-weight generation.

The vector attention weights are obtained by applying softmax normalization over the neighbor dimension:

$$\boldsymbol{\alpha}_{ij} = \frac{\exp(\mathcal{A}_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(\mathcal{A}_{ik})}, \tag{22}$$

where normalization is performed over $j \in \mathcal{N}(i)$ for each channel. In the multi-head implementation, value channels are split into multiple heads and normalized independently.

The output feature of point i is computed as:

$$\mathbf{y}_i = \text{BN} \left(\sum_{j \in \mathcal{N}(i)} \boldsymbol{\alpha}_{ij} \odot \tilde{\mathbf{v}}_{ij} \right) + \mathbf{f}_i, \tag{23}$$

where $\odot$ denotes element-wise multiplication.

In GPA, attention weights are not directly computed from query-key feature similarity. Instead, semantic features influence attention through SGMR, while DSGK derives logits from raw and rectified geometric cues, making geometry-aware structural relations the primary cues for local attention generation.

4.5 Computational Characteristics

Compared with feature-centric attention mechanisms based on high-dimensional query-key interactions, GPA generates attention logits from compact low-dimensional geometric descriptors. DSGK operates on the 6-D descriptor $\mathbf{g}_{ij}$ through shared MLPs, while SGMR adds only lightweight offset prediction from local feature differences. Thus, the main attention-logit generation cost comes from efficient geometric MLP operations rather than full pairwise feature-channel comparisons. This enables PointGP to preserve adaptive local aggregation with favorable parameter and computational efficiency.

5 Experiments

5.1 Experimental Setup

We evaluate **PointGP** on five widely used point cloud benchmarks covering three representative tasks. For shape classification, we use ModelNet40 [Wu *et al.*, 2015] and ScanObjectNN [Uy *et al.*, 2019]; for part segmentation, we use ShapeNetPart [Chang *et al.*, 2015]; for indoor scene semantic segmentation, we use S3DIS [Armeni *et al.*, 2016] and ScanNet v2 [Dai *et al.*, 2017]. These datasets cover clean synthetic objects, noisy real-world objects, fine-grained part-level segmentation, and large-scale indoor scene understanding, providing a comprehensive evaluation of the proposed geometry-primed attention design. The network configurations for different datasets are summarized in Table 1. The number of parameters is measured over all learnable weights, and FLOPs are measured with a batch size of 1 under the corresponding input resolution for each benchmark.

5.2 Shape Classification

ModelNet40. ModelNet40 [Wu *et al.*, 2015] is a standard benchmark for 3D shape classification, containing 12,311 CAD models from 40 categories, with 9,843 samples for training and 2,468 samples for testing. As shown in Table 2, PointGP achieves 94.1% OA and 91.6% mAcc with 4.9M parameters. Compared with PTv2 [Wu *et al.*, 2022], PointGP obtains comparable accuracy while using substantially fewer parameters. These results suggest that geometry-primed attention provides an effective and parameter-efficient design for clean synthetic shape classification.

ScanObjectNN. ScanObjectNN [Uy *et al.*, 2019] is a challenging real-world benchmark containing noisy and incomplete point clouds with background clutter and occlusion. We evaluate PointGP on the hardest PB_T50_RS split, which better reflects real-world object recognition scenarios than clean synthetic datasets. As shown in Table 3, PointGP achieves 90.3% OA and 89.6% mAcc with 4.9M parameters.

5.3 Part Segmentation

ShapeNetPart. ShapeNetPart [Chang *et al.*, 2015] is a large-scale benchmark for fine-grained 3D part segmentation. It contains 16,881 shapes from 16 categories with 50 part labels. The task requires models to capture local geometric details and part boundaries, making it suitable for evaluating the effectiveness of local attention mechanisms. As

Dataset	Channels	Depth	K	**Epochs**	**Batch**	**LR**	**Warmup**	Params (M)	FLOPs (G)
ModelNet40	96-192-384	2-2-2	20	600	32	2e-3	60	4.9	0.8
ScanObjectNN	96-192-384	2-2-2	24	400	32	3e-3	40	4.9	0.8
ShapeNetPart	96-192-288-384	2-2-2-2	20	250	32	2e-3	20	6.5	2.0
S3DIS	64-128-256-512	2-2-4-2	24	100	8	6e-3	10	5.4	29.0
ScanNet v2	64-96-160-288-512	1-2-2-4-2	24	100	8	6e-3	10	6.9	169.3

Table 1: **Network architecture and training specifications across five benchmarks.** "Channels" denotes feature dimensions; "Depth" is the number of stacked blocks; "K" is the neighborhood size; "LR" is the initial learning rate; "Warmup" is the number of warmup epochs.

Method	OA	mAcc	Params
PointNet++ [Qi *et al.*, 2017b]	91.9	-	1.5
DGCNN [Wang *et al.*, 2019]	92.9	90.2	1.8
PCT [Guo *et al.*, 2021]	93.2	-	2.9
PTv1 [Zhao *et al.*, 2021]	93.8	90.6	4.9
PointNeXt [Qian *et al.*, 2022]	94.0	91.1	1.4
PointMLP [Ma *et al.*, 2022]	94.1	91.3	12.6
PTv2 [Wu *et al.*, 2022]	**94.2**	**91.6**	11.4
DualMLP [Paul *et al.*, 2024]	93.7	-	6.5
SI-Mamba [Bahri *et al.*, 2025]	92.7	90.5	12.3
LFE-PM [Zhou *et al.*, 2025]	93.0	-	9.0
PointGP (Ours)	94.1	**91.6**	4.9

Table 2: Classification results on ModelNet40. **OA**: Overall Accuracy (%). **mAcc**: Mean per-class Accuracy (%). **Params**: number of parameters (M).

Method	OA	mAcc	Params
PointNet++ [Qi *et al.*, 2017b]	86.2	84.4	1.5
DGCNN [Wang *et al.*, 2019]	86.1	84.3	1.8
PTv2 [Wu *et al.*, 2022]	88.6	-	11.4
PointMAE [Pang *et al.*, 2022]	85.2	-	22.1
PointMLP [Ma *et al.*, 2022]	85.4	83.9	12.6
PointNeXt [Qian *et al.*, 2022]	88.2	86.8	1.4
Point2vec [Zeid *et al.*, 2023]	87.5	86.0	-
KPConvX [Thomas *et al.*, 2024]	89.3	88.1	14.3
SI-Mamba [Bahri *et al.*, 2025]	87.3	-	12.3
LFE-PM [Zhou *et al.*, 2025]	88.5	-	9.0
PointGP (Ours)	**90.3**	**89.6**	4.9

Table 3: Classification results on ScanObjectNN PB_T50_RS. **OA**: Overall Accuracy (%). **mAcc**: Mean per-class Accuracy (%). **Params**: number of parameters (M).

shown in Table 4, PointGP achieves 87.3% instance-averaged mIoU and 85.3% class-averaged mIoU with 6.5M parameters. PointGP outperforms several representative point-based backbones while maintaining a moderate model size.

5.4 Semantic Segmentation

S3DIS Area 5. The S3DIS dataset [Armeni *et al.*, 2016] contains 271 rooms across six areas with 13 semantic categories. We follow the standard Area 5 evaluation protocol, training on Areas 1–4 and 6 and testing on Area 5. This benchmark evaluates scene-level semantic understanding under complex indoor layouts and diverse object scales. As shown in Table 5, PointGP achieves 91.7% OA and 73.2% mIoU.

ScanNet v2. ScanNet v2 [Dai *et al.*, 2017] is a large-scale indoor RGB-D dataset containing 1,513 reconstructed 3D scenes annotated with 20 semantic categories. We follow the standard evaluation protocol, using 1,201 scenes for training, 312 scenes for validation, and 100 hidden scenes for testing. Compared with object-level benchmarks, ScanNet v2 contains larger scenes, more diverse spatial layouts, and stronger variations in point density. As shown in Table 6, PointGP achieves 75.8% mIoU on the validation set and 75.4% mIoU on the hidden test set.

5.5 Ablation Studies

We conduct ablation studies on ModelNet40 to analyze the effectiveness and efficiency of the proposed design, as it provides a standard and efficient benchmark for comparing architectural variants under consistent training conditions. All variants follow the same network configuration and training protocol, and Overall Accuracy (OA) is reported as the primary metric.

Effect of Geometry-Primed Attention. To compare GPA with feature-centric attention mechanisms, we replace GPA with several representative attention modules, including standard scalar attention [Vaswani *et al.*, 2017], PCT-style offset attention [Guo *et al.*, 2021], PTv1-style vector attention [Zhao *et al.*, 2021], and PTv2-style grouped vector attention [Wu *et al.*, 2022]. As shown in Table 7, GPA achieves the best accuracy with fewer parameters and lower computational cost. Compared with grouped vector attention, GPA improves OA from 93.5% to 94.1%, while reducing parameters from 6.4M to 4.9M and FLOPs from 5.3G to 0.8G. This comparison suggests that deriving attention logits from geometry-aware cues provides an efficient alternative to direct feature-similarity-based attention.

Efficiency Analysis. Table 7 also shows that GPA substantially reduces computational cost. Compared with standard scalar attention, GPA reduces FLOPs from 2.9G to 0.8G. The efficiency gain mainly comes from avoiding explicit high-dimensional query-key feature matching for attention-logit generation. Instead, GPA computes attention logits from low-dimensional geometric descriptors, while semantic information is introduced through feature-conditioned coordinate rectification. This design improves the accuracy-efficiency trade-off of local attention.

Method	Ins.	Cls.	Params
PointNet++ [Qi *et al.*, 2017b]	85.1	81.9	1.5
DGCNN [Wang *et al.*, 2019]	85.2	82.3	1.8
RSCNN [Liu *et al.*, 2019]	86.2	84.0	-
PTv1 [Zhao *et al.*, 2021]	86.6	83.7	7.8
PTv2 [Wu *et al.*, 2022]	86.6	-	11.4
PointMLP [Ma *et al.*, 2022]	86.9	85.2	12.6
PointNeXt [Qian *et al.*, 2022]	87.0	85.2	22.5
OTMae3D [Wang *et al.*, 2025]	86.8	85.1	-
AdaCrossNet [Putra *et al.*, 2025]	-	85.1	-
LFE-PM [Zhou *et al.*, 2025]	85.9	84.0	9.0
PointGP (Ours)	**87.3**	**85.3**	6.5

Table 4: Part segmentation results on ShapeNetPart. **Ins.**: instance-averaged mIoU (%). **Cls.**: class-averaged mIoU (%). **Params**: number of parameters (M).

Method	OA	mAcc	mIoU
PTv1 [Zhao *et al.*, 2021]	90.8	76.5	70.4
FPT [Park *et al.*, 2022]	-	77.6	71.0
PointNeXt [Qian *et al.*, 2022]	91.0	77.2	71.1
PointMixer [Choe *et al.*, 2022]	-	77.4	71.4
StratifiedFormer [Lai *et al.*, 2022]	91.5	**78.1**	72.0
PTv2 [Wu *et al.*, 2022]	91.6	78.0	72.0
DU-Net [Xiu *et al.*, 2023]	-	-	72.2
PointMamba [Liang *et al.*, 2024]	-	-	70.3
PointGP (Ours)	**91.7**	77.5	**73.2**

Table 5: Semantic segmentation results on S3DIS Area 5. **OA**: Overall Accuracy (%); **mAcc**: Mean Class Accuracy (%); **mIoU**: Mean Intersection-over-Union (%).

Component-wise Effectiveness. To evaluate the individual contributions of SGMR and DSGK, we construct a baseline that uses only raw relative coordinates and a single-stream MLP for attention-logit generation. As shown in Table 8, adding SGMR improves OA from 92.8% to 93.5%, showing that feature-conditioned coordinate rectification provides useful geometric cues. Adding DSGK improves OA to 93.6%, indicating the benefit of combining edge-level geometric details with neighborhood-level context. Combining both components achieves the best result of 94.1%, suggesting that SGMR and DSGK are complementary.

6 Conclusion and Limitations

In this paper, we presented PointGP, a geometry-primed framework for 3D point cloud analysis. Different from conventional feature-centric attention mechanisms that mainly rely on semantic feature similarity, PointGP derives attention weights from geometry-aware local relations. Specifically, the proposed Semantic-Guided Manifold Rectifier predicts feature-conditioned local coordinate offsets, while the Dual-Stream Geometric Kernel computes attention logits from both raw and rectified geometric cues. In this formulation, semantic information guides local geometric rectification, and attention generation is driven by geometry-aware structural relations. Extensive experiments on classification and segmentation benchmarks demonstrate that PointGP achieves competitive performance with strong parameter efficiency.

Method	Val	Test	Params
PointNet++ [Qi *et al.*, 2017b]	53.5	33.9	1.5
O-CNN [Wang *et al.*, 2017]	74.0	72.8	-
PointConv [Wu *et al.*, 2019]	61.0	55.6	2.5
KPConv [Thomas *et al.*, 2019]	69.2	68.0	14.0
MinkowskiNet [Choy *et al.*, 2019]	72.2	73.4	-
BPNet [Hu *et al.*, 2021]	73.9	74.9	-
SFormer [Lai *et al.*, 2022]	74.3	73.7	-
PTv2 [Wu *et al.*, 2022]	75.4	75.2	11.3
PointGP (Ours)	**75.8**	**75.4**	6.9

Table 6: Semantic segmentation results on ScanNet v2 validation and test sets. **Val** and **Test** denote mIoU (%). **Params**: number of parameters (M).

Attention Module	OA Range	Params	FLOPs
Standard Scalar Attn.	91.4–91.8	5.6	2.9
Offset Attn.	92.1–92.4	6.0	2.9
Vector Attn.	93.1–93.3	6.8	6.4
Grouped Vector Attn.	93.0–93.5	6.4	5.3
GPA (Ours)	**93.5–94.1**	**4.9**	**0.8**

Table 7: Ablation study on attention mechanisms. **OA Range**: observed Overall Accuracy range over repeated runs (%). **Params**: Parameters (M). **FLOPs**: Floating point operations (G).

Limitations. Despite its effectiveness, PointGP still has several limitations. First, the proposed design relies on local neighborhoods constructed by k-NN. When point clouds are extremely sparse, highly noisy, or dominated by outliers, the neighborhood structure may become unreliable, which can affect the quality of coordinate rectification and attention generation. Second, PointGP mainly emphasizes geometric relations. For scenarios where semantic differences depend heavily on non-geometric attributes such as color, texture, or intensity, geometric cues alone may not always be sufficient to distinguish ambiguous regions. Future work will explore multi-modal extensions that combine geometry with richer appearance or sensor cues for more robust point cloud understanding.

Variant	SGMR	DSGK	OA Range	$\Delta_{\max}$
Base.	×	×	92.1–92.8	—
Base. + SGMR	√	×	92.7–93.5	+0.7
Base. + DSGK	×	√	92.7–93.6	+0.8
PointGP	√	√	**93.1–94.1**	**+1.3**

Table 8: Component-wise ablation on ModelNet40. SGMR: Semantic-Guided Manifold Rectifier; DSGK: Dual-Stream Geometric Kernel. **OA Range**: observed Overall Accuracy range over repeated runs (%); $\Delta_{\max}$: gain of the best observed OA relative to the baseline best observed OA (%).

Acknowledgments

The paper is supported by the National Natural Science Foundation of China (No. 62072388), Fujian Provincial Science and Technology Major Project (No. 2024HZ022003), Jiangxi Provincial Natural Science Foundation Key Project (No. 20244BAB28039), Xiamen Public Technology Service Platform (No. 3502Z20231043), and Fujian Sunshine Charity Public Welfare Foundation.

References

[Armeni *et al.*, 2016] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1534–1543, 2016.

[Bahri *et al.*, 2025] Ali Bahri, Moslem Yazdanpanah, Mehrdad Noori, Sahar Dastani, Milad Cheraghalikhani, Gustavo Adolfo Vargas Hakim, David Osowiechi, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. Spectral informed mamba for robust point cloud processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11799–11809, June 2025.

[Chang *et al.*, 2015] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[Choe *et al.*, 2022] Jaesung Choe, Chunghyun Park, Francois Rameau, Jaesik Park, and In So Kweon. PointMixer: MLP-mixer for point cloud understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 620–640, 2022.

[Choy *et al.*, 2019] 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 (CVPR)*, pages 3075–3084, 2019.

[Dai *et al.*, 2017] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017.

[Dong *et al.*, 2021] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2793–2803, 2021.

[Guo *et al.*, 2021] 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.

[Hu *et al.*, 2021] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14373–14382, 2021.

[Lai *et al.*, 2022] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3D point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8500–8509, 2022.

[Liang *et al.*, 2024] 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. In *Advances in Neural Information Processing Systems*, volume 37, pages 32653–32677, 2024.

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

[Ma *et al.*, 2022] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[Pang *et al.*, 2022] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 604–621, 2022.

[Park *et al.*, 2022] Chunghyun Park, Yoonwoo Jeong, Minsu Cho, and Jaesik Park. Fast point transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16949–16958, June 2022.

[Paul *et al.*, 2024] Sneha Paul, Zachary Patterson, and Nizar Bouguila. DualMLP: A two-stream fusion model for 3D point cloud classification. *The Visual Computer*, 40(8):5435–5449, 2024.

[Putra *et al.*, 2025] Oddy Virgantara Putra, Kohichi Ogata, Eko Mulyanto Yuniarno, and Mauridhi Hery Purnomo. Adacrossnet: Adaptive dynamic loss weighting for cross-modal contrastive point cloud learning. *International Journal of Intelligent Engineering and Systems*, 18(1):134–146, 2025.

[Qi *et al.*, 2017a] 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 (CVPR)*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical fea-

ture learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5105–5114, 2017.

[Qian *et al.*, 2022] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 23192–23204, 2022.

[Shi and Rajkumar, 2020] Weijing Shi and Ragunathan Rajkumar. Point-GNN: Graph neural network for 3D object detection in a point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1711–1719, 2020.

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

[Thomas *et al.*, 2024] Hugues Thomas, Yao-Hung Hubert Tsai, Timothy D. Barfoot, and Jian Zhang. KPConvX: Modernizing kernel point convolution with kernel attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5535, 2024.

[Uy *et al.*, 2019] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1588–1597, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, 2017.

[Wang *et al.*, 2017] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-CNN: Octree-based convolutional neural networks for 3D shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

[Wang *et al.*, 2019] 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.

[Wang *et al.*, 2025] Chuxin Wang, Yixin Zha, Jianfeng He, Wenfei Yang, and Tianzhu Zhang. Rethinking masked representation learning for 3d point cloud understanding. *IEEE Transactions on Image Processing*, 34:247–262, 2025.

[Wu *et al.*, 2015] 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 (CVPR)*, pages 1912–1920, 2015.

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

[Wu *et al.*, 2021] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10033–10041, 2021.

[Wu *et al.*, 2022] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer V2: Grouped vector attention and partition-based pooling. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 33330–33342, 2022.

[Wu *et al.*, 2024] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, 2024.

[Xiao *et al.*, 2025] Yu Xiao, Hui Wu, Yisheng Chen, Chongcheng Chen, Ruihai Dong, and Ding Lin. Hybrid offset position encoding for large-scale point cloud semantic segmentation. *Remote Sensing*, 17(2):256, 2025.

[Xiu *et al.*, 2023] Haoyi Xiu, Xin Liu, Weimin Wang, Kyoung-Sook Kim, Takayuki Shinohara, Qiong Chang, and Masashi Matsuoka. Diffusion unit: Interpretable edge enhancement and suppression learning for 3d point cloud segmentation. *Neurocomputing*, 559:126780, 2023.

[Zeid *et al.*, 2023] Karim Abou Zeid, Jonas Schult, Alexander Hermans, and Bastian Leibe. Point2Vec for self-supervised representation learning on point clouds. In *Proceedings of the DAGM German Conference on Pattern Recognition*, pages 131–146, 2023.

[Zhao *et al.*, 2021] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021.

[Zhou *et al.*, 2021] Haoran Zhou, Yidan Feng, Mingsheng Fang, Mingqiang Wei, Jing Qin, and Tong Lu. Adaptive graph convolution for point cloud analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4965–4974, 2021.

[Zhou *et al.*, 2025] Bowen Zhou, Lixin Zhan, and Jie Jiang. Lfe-pointmamba: Point cloud learning via local feature enhancement and state space model. *Knowledge-Based Systems*, 329:114350, 2025.