Uncertainty-Driven Gaussian Sphere Propagation for 3D Semantic Segmentation

Zhicheng Yan[1], Qingyong Li[1,2], Yixiao Song[1], and Wen Wang[1⋆]

[1] Laboratory of Big Data and Artificial Intelligence in Transportation, Beijing Jiaotong University, Beijing 100044, China
[2] Frontiers Science Center for Smart High-Speed Railway System, Beijing Jiaotong University, Beijing 100044, China
yanzhicheng@bjtu.edu.cn; liqy@bjtu.edu.cn; songyixiao@bjtu.edu.cn; wangwen@bjtu.edu.cn

Abstract. 3D point cloud semantic segmentation remains a fundamental challenge for autonomous driving and robotic perception. While Transformer-based architectures have achieved significant progress by capturing long-range dependencies, existing methods rely on deterministic point-wise predictions that fail to maintain semantic consistency across geometric voids such as occlusions and sparse boundaries. This can be attributed to the isotropic nature of conventional discrete feature aggregation, which lacks the directional awareness necessary to propagate reliable information across these discrete gaps. To address this, we propose Uncertainty-driven Gaussian Sphere Propagation (UGSP), a framework that transitions from discrete point processing to continuous geometric field reconstruction. By leveraging approximate Bayesian inference to identify reliable semantic anchors and high-uncertainty voids, a Spherical Harmonics (SH)-based aggregation mechanism is introduced that models the local scene as a collection of anisotropic Gaussian spheres. This approach enables the interpolation of semantic information along continuous spatial paths, allowing for direction-aware semantic propagation that effectively recovers structural integrity in high-uncertainty regions. Extensive experiments on indoor and outdoor benchmarks demonstrate the superiority of UGSP in semantically ambiguous and geometrically complex scenarios. The source code is publicly available at https://github.com/LENGYI1221/UGSP.

Keywords: 3D Point Cloud Segmentation · Uncertainty Modeling · Spherical Harmonics · Geometry-Aware Aggregation · Region-Level Refinement

1 Introduction

3D point clouds serve as a core representation for autonomous driving, robotic perception, and scene understanding by directly encoding 3D geometry. Recently, Transformer-based models, such as Point Transformer v3(PTv3) [47],

⋆ Corresponding author.

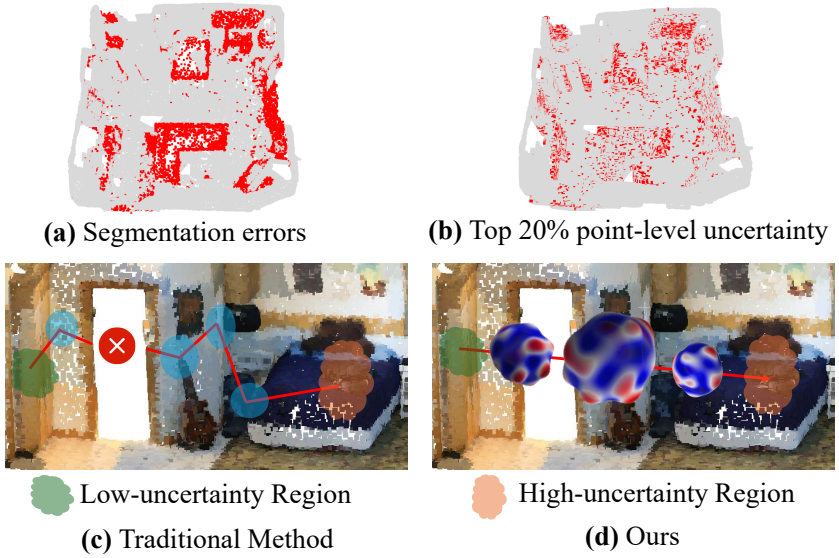

Fig. 1: Motivation of the proposed UGSP. (a) and (b) visualize the high spatial correlation between segmentation errors and predictive uncertainty, particularly in geometric voids. (c) Traditional methods fail to bridge discrete gaps due to the lack of directional awareness. (d) Our UGSP reformulates the process as a continuous field reconstruction, leveraging anisotropic Gaussian spheres to propagate reliable semantic information from anchors to high-uncertainty regions along continuous spatial paths.

have achieved state-of-the-art performance by leveraging vast receptive fields to capture long-range contextual dependencies. Despite these advancements, existing methods typically operate under a deterministic point-wise prediction paradigm, implicitly assuming that that spatial features across the entire domain possess uniform reliability during inference. However, this assumption frequently collapses in geometric voids characterized by complex occlusions, varying sampling densities, or structural boundaries, where the discrete nature of point clouds fails to provide continuous geometric support for stable inference [12, 15, 16, 22, 26–30, 33–35, 42, 45–48, 50, 54, 55].

We observe that segmentation errors are not randomly distributed but are highly correlated with predictive uncertainty, particularly in geometrically complex regions. As illustrated in Fig. 1, high uncertainty identifies geometric voids, such as object boundaries, sparse distant regions, or occluded surfaces. While low-uncertainty regions can serve as reliable semantic anchors, traditional discrete aggregation methods based on convolution or isotropic attention struggle to bridge these voids. This is because they rely on isotropic feature gathering, which lacks the geometric continuity and directional awareness necessary to selectively propagate valid information from certain anchors into ambiguous areas. To resolve this, we shift the perspective from discrete point processing to continuous geometric field reconstruction, proposing that semantic information should be propagated from low-uncertainty anchors to high-uncertainty regions via a direction-aware, structure-preserving mechanism.

Inspired by the anisotropic modeling in 3D Gaussian Splatting, we propose Uncertainty-driven Gaussian Sphere Propagation (UGSP), a framework that models the local scene as a collection of anisotropic Gaussian spheres. Specifically, we first decompose the scene into low- and high-uncertainty regions via approximate Bayesian inference. To bridge these regions, we introduce a Spherical Harmonics (SH)-based aggregation mechanism that explicitly encodes the directional distribution of surrounding semantic information. Instead of restricting aggregation centers to existing points, we construct Gaussian aggregation spheres along continuous spatial paths connecting anchors to voids. This allows the aggregation process to sense the underlying geometric structure and selectively propagate information from reliable anchors into high-uncertainty voids along continuous spatial paths, as shown in Fig. 1(d). By transitioning from blind isotropic gathering to guided anisotropic propagation, our framework effectively recovers semantic consistency in regions where physical points are missing or unreliable.

Our main contributions are summarized as follows:

- We propose a unified uncertainty-driven framework for point cloud semantic segmentation, in which predictive uncertainty is explicitly modeled as a structural prior to regulate selective semantic propagation from reliable regions to ambiguous ones.
- We introduce a spherical harmonics-based feature aggregation module operating over anisotropic Gaussian spheres sampled along continuous spatial paths, encodes the anisotropic distribution of semantics, enabling coherent and direction-aware semantic propagation across sparse or occluded regions.
- Extensive experiments on standard point cloud benchmarks demonstrate the effectiveness of our approach, particularly in complex and semantically ambiguous regions.

2 Related Work

2.1 Semantic Segmentation of 3D Point Clouds

With the development of deep learning, point cloud semantic segmentation has evolved from handcrafted feature-based methods to end-to-end neural networks [12, 26, 30, 42]. Early point-based models such as PointNet and its variants directly process unordered point sets [33–35] and enhance geometric representation through local neighborhood modeling [15, 22, 46, 50, 54]. Subsequent works further improve feature learning using convolutional operators and geometric kernels [17, 19, 23, 43, 49, 53]. Graph-based and neighborhood aggregation methods also demonstrate strong capability for semantic segmentation [18,20,21,25,44]. In addition, voxel-based and sparse convolution methods enable efficient large-scale segmentation [7,10,11,31,38,39]. More recently, Transformer-based architectures leverage self-attention to capture long-range dependencies and achieve strong performance in complex scenes [16, 27–29, 45, 47, 48, 55].

While these methods significantly enhance contextual modeling in large-scale scenes, they primarily focus on deterministic mappings that overlook the inherent reliability of predictions. In contrast, we argue that modeling predictive uncertainty is crucial for capturing structural inconsistencies, especially in semantically ambiguous or sparsely sampled regions. By explicitly treating uncertainty as a structural prior, our method leverages reliable contextual information to resolve ambiguities, thereby achieving superior robustness and stability where conventional deterministic approaches often falter.

2.2 Segmentation with 3D Gaussian Splatting

Recently, 3D Gaussian Splatting (3DGS) [13] has emerged as a powerful representation for real-time high-fidelity rendering. Beyond novel view synthesis, assigning semantic or instance labels to 3D Gaussians has become a burgeoning research direction. A predominant line of research focuses on distilling knowledge from 2D foundation models into the 3D Gaussian space. For instance, *LangSplat* [36] and *Feature3DGS* [56] leverage CLIP [37] or DINO [6] features to enable open-vocabulary segmentation. Other works, such as GaussianGrouping [52], utilize 2D masks from SAM [14] as strong supervision to achieve instance-level decoupling. While these methods achieve impressive results by inheriting 2D priors, they often treat 3D Gaussians as mere carriers for external features, largely overlooking the intrinsic geometric structure and the inherent predictive uncertainty during the 3D reconstruction process.

While these methods rely on external 2D foundation priors, they predominantly treat 3D Gaussians as passive feature carriers, overlooking the intrinsic geometric structure and predictive uncertainty. In contrast, our approach eliminates this dependency on external models. By explicitly treating uncertainty as a structural prior, we leverage a 3D-native path aggregation mechanism to resolve semantic ambiguities. This self-contained strategy enables robust scene understanding without any external semantic supervision.

3 Method

3.1 Overview

Given an input point cloud $\mathcal{P} = \{p_i\}_{i=1}^{N}$, where $p_i \in \mathbb{R}^6$ denotes the 3D coordinates and color of the i-th point, our goal is to reconstruct a continuous geometric field that bridges the gap between discrete points in geometrically ambiguous regions. Unlike traditional deterministic paradigms that assume uniform spatial reliability, our framework explicitly models predictive uncertainty as a structural prior to guide anisotropic semantic propagation. Our overall framework consists of three core stages: Uncertainty Region Estimation, Path Construction and Spherical Harmonics (SH)-based Aggregation Mechanism. The general pipeline is illustrated in Fig. 2.

(1) Uncertainty Region Estimation. The process begins by passing the point cloud through a feature extraction backbone to obtain high-dimensional point

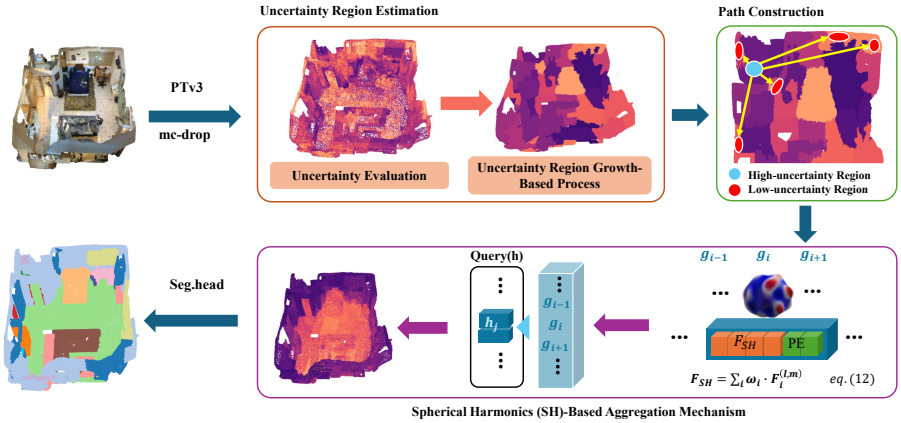

Fig. 2: Overview of the proposed UGSP framework. The pipeline begins with uncertainty region estimation, where point-wise uncertainty is identified via Monte Carlo Dropout. Subsequently, an uncertainty-based region growth process partitions the scene into high- and low-uncertainty regions. During path construction, geometric bridges are established to connect these regions, facilitating structured context transfer. Along each path, spherical harmonic (SH)-based aggregation mechanism then samples anisotropic Gaussian spheres g_i, employing SH encoding to capture direction-aware features F_{SH}. Then, a cross-attention refinement module integrates these features to refine original point features for the final segmentation head.

representations and initial semantic predictions. To quantify the reliability of these local inferences, we employ approximate Bayesian inference via Monte Carlo Dropout during the inference phase. Subsequently, an Uncertainty-Aware Region Growing algorithm utilizes these estimates to partition the scene into reliable low-uncertainty anchors and ambiguous high-uncertainty voids, which typically correspond to complex boundaries or occluded surfaces.

(2) Path construction. To bridge the semantic gaps identified in the previous stage, we establish connectivity between unreliable voids and their surrounding stable anchors. This stage facilitates structured information flow by constructing continuous spatial paths that transcend the limitations of discrete point sampling. By establishing these bridges, we provide a geometric scaffold for propagating valid semantic context into regions lacking physical point support.

(3) Spherical Harmonics (SH)-based aggregation mechanism. Along the established paths, a sequence of Aggregation Centers is sampled to perform anisotropic Gaussian Sphere Aggregation. We model the local scene as a collection of Gaussian-weighted spherical neighborhoods, where Spherical Harmonics (SH) are utilized to encode the directional distribution of semantic features. This enables the model to sense the underlying structural orientations and selectively propagate information along the continuous paths. Finally, a Cross-Attention refinement module integrates these path-derived contextual features with the original point features to output the final refined semantic predictions.

3.2 Uncertainty Region Estimation

Uncertainty Evaluation We denote by f_i the point-wise feature extracted by the backbone network:

$$f_i = \text{Backbone}(p_i). \tag{1}$$

To estimate predictive uncertainty, we enable Dropout layers during inference and perform K stochastic forward passes. Let $\hat{y}_i^{(k)}$ denote the predicted class probability for point p_i at the k-th pass. The predictive mean is computed as:

$$\bar{y}_i = \frac{1}{K} \sum_{k=1}^{K} \hat{y}_i^{(k)}. \tag{2}$$

The point-wise uncertainty u_i is measured by the standard deviation across Monte Carlo samples:

$$u_i = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\hat{y}_i^{(k)} - \bar{y}_i)^2}. \tag{3}$$

Intuitively, a higher u_i indicates that the model is more uncertainty about the prediction at point p_i, and such points are more likely to lie near semantic boundaries or in sparse and noisy regions.

Uncertainty Region Growth-based Process To explicitly delineate regions with varying predictive uncertainty, we propose a hierarchical region partitioning algorithm. The algorithm consists of three main stages: (1) high-uncertainty region generation, (2) low-uncertainty region generation, and (3) final assignment of remaining points.

Voxelization and Uncertainty Computation Let the voxel grid size be g. The voxelized set of voxels is $\mathcal{V} = \{v_i\}_{i=1}^{M}$. For each voxel v, its representative uncertainty value u_v is computed as the average uncertainty of all points within that voxel:

$$u_v = \frac{1}{|\mathcal{N}_v|} \sum_{p_i \in v} u_i, \tag{4}$$

where $\mathcal{N}_v$ denotes the number of points falling into voxel v.

High-Uncertainty Region Generation To capture the structural patterns of semantic ambiguity, the scene is partitioned into distinct high-uncertainty regions. Specifically, seed voxels are selected via Farthest Point Sampling (FPS) from the set of voxels containing the top 40% highest uncertainty values, ensuring spatial diversity.

Starting from these seeds, we perform an agglomerative merging process to form N_H coherent regions. The affinity between two adjacent regions R_a and R_b is governed by a distance-uncertainty joint cost function:

$$\mathcal{C}(R_a, R_b) = \min_{v_i \in R_a, v_j \in R_b} \{d(v_i, v_j)\} \cdot (1 + \alpha \cdot |\bar{u}_{R_a} - \bar{u}_{R_b}|), \tag{5}$$

where $d(\cdot)$ denotes the Euclidean distance between voxel centroids and $\bar{u}$ represents the mean uncertainty within a region. The hyperparameter α balances geometric proximity and uncertainty homogeneity. We iteratively merge pairs with the minimum cost until the target granularity N_H is achieved, resulting in the final set of regions $\mathcal{R} = \{R_1, \ldots, R_{N_H}\}$. Detailed implementation of the merging priority queue is provided in the supplementary material.

Low-Uncertainty Region Generation and Remaining Points The generation of Low-Uncertainty Regions follows the same pipeline as High-Uncertainty Regions but uses the bottom 40% of voxels. The remaining intermediate-uncertainty points are assigned via KNN to their nearest high- or low-uncertainty region, ensuring complete partitioning of the point cloud.

3.3 Path Construction

The next critical step is to establish structured pathways that propagate reliable semantic context from low-uncertainty anchors (R_l) into high-uncertainty voids (R_h). We clarify two complementary directions to avoid confusion: the *geometric path construction* proceeds *outward*, building a bridge from each high-uncertainty region (R_h) to its nearest low-uncertainty anchors (R_l); the *information flow* then runs *inward*, propagating reliable features from the anchors back toward the voids. Thus the path defines only the spatial scaffold, while feature propagation always transfers reliable context from R_l to R_h.

Region Center Identification: For any segmented region $R \in \{R_h, R_l\}$, its representative center, denoted as C_R, is calculated as the geometric centroid of all points belonging to that region in the input point cloud $\mathcal{P}$:

$$C_R = \frac{1}{|R|} \sum_{p_i \in R} p_i. \tag{6}$$

This centroid C_R serves as the anchor point representing the overall spatial location and geometric characteristics of the entire region.

Path Generation and Center Sampling. We establish connectivity by constructing directed pathways between the representative centers of high-uncertainty regions and their spatially adjacent low-uncertainty counterparts. For each C_{R_h}, we identify its k-nearest low-uncertainty anchors $\{C^j_{R_l}\}_{j=1}^m$ based on Euclidean distance, where m is a hyperparameter governing the density of the contextual

graph. Along each path, we sample a sequence of N_s uniformly spaced intermediate points to act as Aggregation Centers.

To ensure these centers reside in geometrically meaningful areas, we impose a density-aware adaptive radius constraint. Specifically, for each sampled center p_s, we initially define a spherical neighborhood $\mathcal{N}(p_s, r)$ with a base radius r. If the occupancy within this neighborhood $|\{v \in \mathcal{V} \mid v \in \mathcal{N}(p_s, r)\}|$ is below the minimum threshold $T_{\min}$, we iteratively expand the radius r until the occupancy requirement is satisfied. This mechanism ensures that each Gaussian sphere captures a sufficiently rich local context even in sparse regions, thereby bridging the semantic gaps between certain and uncertain regions without breaking the connectivity of the established pathways.

3.4 Spherical Harmonics (SH)-based Aggregation Mechanism

Spherical Aggregation To simultaneously model directional information and local feature distributions within spherical neighborhoods, we introduce a Gaussian-weighted spherical feature aggregation mechanism. This mechanism encodes the direction of points within the spherical neighborhood using spherical harmonic functions, and combines it with a distance-dependent Gaussian weight to achieve fine-grained modeling of the local geometry. Given an aggregation center p_s, let the set of points within its spherical neighborhood be $\{\mathbf{x}_i\}$. We first compute the direction vector for each point relative to the aggregation center:

$$\mathbf{d}_i = \mathbf{x}_i - p_s, \tag{7}$$

where $\mathbf{d}_i \in \mathbb{R}^3$ is the direction vector. To eliminate scale influence and retain only directional information, we normalize the direction vector and map it to spherical coordinates:

$$\hat{\mathbf{d}}_i = \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|} \quad \longrightarrow \quad (\theta_i, \phi_i), \tag{8}$$

where θ_i and ϕ_i represent the polar and azimuthal angles, respectively.

Next, we compute the expansion of this direction in spherical harmonics up to order L:

$$y_i = Y(\theta_i, \phi_i) \in \mathbb{R}^{(L+1)^2}, \tag{9}$$

where $Y(\theta, \phi)$ represents the spherical harmonic basis functions that capture the directional information of the points within the spherical neighborhood.

To inject directional information into the feature representation, we couple the point feature with the corresponding spherical harmonic basis function for each order (l, m) as follows:

$$\mathbf{F}_i^{(l,m)} = \mathbf{f}_i \cdot Y_l^m(\theta_i, \phi_i), \tag{10}$$

where $l = 0, \ldots, L$, and $m = -l, \ldots, l$. This operation elevates the original point features into the spherical harmonic frequency domain, enhancing the model's sensitivity to directional geometry.

We further introduce a distance-dependent Gaussian weight to perform a weighted aggregation of the points. Specifically, the weight for point $\mathbf{x}_i$ is defined as:

$$w_i = \psi(r_i), \quad r_i = \|\mathbf{x}_i - p_s\|, \tag{11}$$

where $\psi(r_i)$ is a neural network that encodes the radial distance r_i from the aggregation center p_s and produces the corresponding weight.

Finally, the aggregated feature at the aggregation center p_s for the (l, m)-th order spherical harmonic is computed as:

$$F_{SH} = \sum_i w_i \cdot \mathbf{F}_i^{(l,m)} = \sum_i w_i \cdot \mathbf{f}_i \cdot Y_l^m(\theta_i, \phi_i). \tag{12}$$

This weighted aggregation operation effectively combines local point features with directional information and applies distance-dependent weighting, ensuring that points closer to the aggregation center contribute more significantly to the final feature representation. The result is a set of refined geometric features that are sensitive to both the local geometry and the relative directions of the points within the neighborhood.

Path Aggregation After performing feature aggregation within a single Gaussian sphere, we extend the modeling to multiple Gaussian spheres distributed along the region connection path. Let the Gaussian spheres be sequentially constructed along the path , where each F_{SH} corresponds to an aggregation center p_s, and its feature representation is obtained through the direction-aware aggregation process described earlier.

To explicitly model the relative spatial relationships between Gaussian spheres, we introduce trajectory direction encoding. For the i-th Gaussian sphere, its displacement relative to the path origin C_{R_h} is defined as:

$$t_i = (p_s - C_{R_h}) \cdot \hat{\mathbf{d}}_i, \tag{13}$$

This projection maps the three-dimensional displacement to a one-dimensional scalar, capturing the relative order and distance of the spheres along the path.

Next, we apply a multi-layer perceptron (MLP) to embed the displacement vector t_i, yielding the trajectory encoding vector:

$$PE = \mathrm{MLP}(t_i). \tag{14}$$

The MLP applies nonlinear transformations to t_i, producing a higher-dimensional representation that not only retains the spatial distance but also models the relative relationships between spheres along the path.

Let F_{SH} represent the feature of the i-th Gaussian sphere after aggregation. The final output g_i for the i-th Gaussian sphere is obtained by concatenating the trajectory encoding PE with the in-sphere feature F_{SH}:

$$g_i = [F_{SH}; PE]. \tag{15}$$

To capture long-range dependencies and path-level context, we apply a Transformer module to encode the sequence:

$$h = \text{Transformer}(g_0, ..., g_{N_s}), \tag{16}$$

where h represents the result of feature aggregation along the path. By jointly using the Transformer, we enable effective feature propagation along the path, with global context helping to refine features in regions of high uncertainty.

High-Uncertainty Feature Update After obtaining the path-level contextual modeling results, we first project the point features in the low-uncertainty regions and the sequence of Gaussian sphere features into a common feature space. This is achieved using learnable linear transformations to obtain the Query, Key, and Value representations. Specifically, the Gaussian sphere (path) features serve as Queries, while the low-uncertainty region features (f_{R_l}) serve as Keys and Values—a standard cross-attention design that imports reliable context into uncertain regions rather than an inversion of it. Information exchange then occurs via the Cross-Attention mechanism:

$$\tilde{f} = \text{Transformer}(Q(h), K(f_{R_l}), V(f_{R_l})), \tag{17}$$

The Cross-Attention mechanism allows the model to selectively attend to relevant contextual information from the Gaussian spheres, effectively mitigating local semantic ambiguities.

We further clarify how this region-level refinement maps back to points. Rather than being directly projected, the single region-level feature $\tilde{f}$ from the cross-attention above is concatenated with each point's backbone feature f_i and fed to the shared segmentation head for per-point classification. Thus $\tilde{f}$ injects reliable context uniformly along the path while f_i preserves per-point detail, avoiding region-boundary artifacts.

Through the above Gaussian sphere sequence modeling and cross-region feature propagation mechanism, the model is able to effectively correct predictions in high-uncertainty regions by leveraging contextual information from low-uncertainty regions, all while maintaining local geometric consistency. This mechanism improves the model's robustness when handling challenging regions with ambiguous geometry and semantics.

4 Experiments

4.1 Datasets and Evaluation Protocol

Our method is validated across five diverse 3D semantic segmentation benchmarks: S3DIS [2] and ScanNet V2 [9] for indoor scenes, and nuScenes [5], Waymo Open Dataset [40], and SemanticKITTI [3] for large-scale outdoor environments. These datasets provide a comprehensive testbed covering various sensor modalities (RGB-D and LiDAR) and scene complexities. We follow standard evaluation

Table 1: Indoor 3D semantic segmentation results (mIoU).

Method	ScanNet V2	S3DIS(Area 5)
3D-only Methods		
KPConv [43]	69.2	67.1
PointCNN [22]	46.9	57.3
ST [16]	74.3	72.0
PointNeXt [35]	71.5	70.5
PointMetaBase [24]	72.8	72.0
MinkUNet [4]	72.2	65.4
Swin3D [51]	76.4	72.5
Uncertainty-aware Methods		
SalsaNet [1]	52.3	58.7
SalsaNext [8]	55.8	61.5
PTv3 [47]	76.8	73.3
ours	**77.9**	**75.6**

Table 2: Outdoor 3D semantic segmentation results (mIoU).

Method	nuScenes	Sem.KITTI	Waymo
3D-only Methods			
SPVNAS [41]	77.4	64.7	68.1
OA-CNNs [32]	78.9	70.6	71.2
PointNeXt [35]	76.4	68.4	70.2
MinkUNet [4]	73.6	61.1	68.5
Swin3D [51]	78.9	72.1	73.5
Uncertainty-aware Methods			
SalsaNext [8]	68.2	59.5	65.6
PTv3 [47]	80.2	72.3	71.3
ours	**82.2**	**73.8**	**72.1**

protocols for each: reporting results on Area-5 for S3DIS, and utilizing the official validation splits for ScanNetV2 and the three outdoor datasets. The mean Intersection over Union (mIoU) serves as the primary metric, with strict adherence to official ignore-label settings to ensure fair comparison with state-of-the-art methods.

4.2 Main Result

We compare our method with the state-of-the-art Point Transformer v3 (PTv3) baseline across both indoor and outdoor benchmarks. Tables 1 and 2 report quantitative results on indoor and outdoor datasets, respectively.Overall, our method consistently achieves superior performance across all datasets. The substantial gains observed in both scenarios underscore the efficacy of modeling uncertainty as a structural prior to guide semantic refinement.

On ScanNet V2 and S3DIS, UGSP outperforms the baseline by 1.1% and 2.3% mIoU, respectively. The improvement is particularly pronounced in scenes with complex layouts and numerous semantic categories. This confirms our hypothesis that Geometric Voids, such as object boundaries and occluded surfaces,

Setting	Uncertainty region	SH Encode	Attention	mIoU
Baseline (PTv3)	-	-	-	73.3
+ RAND-KNN	✗	✓	✓	72.8
+ UND-KNN	✗	✓	✓	74.1
+ Ours	✓	✓	✓	75.6
w/o SH	✓	✗(max)	✓	72.8
w/o SH	✓	✗(average)	✓	73.1
w/o Attention	✓	✓	✗(w/o path-dir)	73.3
w/o Attention	✓	✓	✗(concate instead)	73.5

Table 3: Ablation study on key components of the proposed method on S3DIS Area5.

which typically exhibit high predictive uncertainty, can be effectively recovered by propagating information from reliable anchors via our SH-based mechanism.

For large-scale outdoor datasets like nuScenes, SemanticKITTI and Waymo, our method also sets new performance benchmarks. Specifically, on SemanticKITTI, while the overall performance of all existing methods is inherently constrained by the sparsity of single-view LiDAR scans, UGSP still yields a consistent improvement of 1.5% over the baseline. This demonstrates that even in sparse environments with restricted fields of view, our path-based aggregation effectively senses underlying structural orientations to maintain semantic consistency.

Beyond quantitative results, we provide extensive qualitative visualizations and a robustness stress test in the Supplementary Material. In the latter, we intentionally remove 20% of points in localized regions to simulate incomplete observations. These experiments demonstrate that our uncertainty-driven path aggregation effectively recovers semantic information in fragmented areas and refines boundaries in complex scenes, further substantiating the model's robustness and generalization ability.

4.3 Ablations

Effect of Uncertainty-driven Region Construction To verify the necessity of explicit uncertainty modeling, we compare our approach with two variants: RAND-KNN (random seed initialization without uncertainty) and UND-KNN (partial uncertainty modeling for initialization but not for growth). As shown in Table 3, RAND-KNN yields the lowest performance (72.8% mIoU) due to its neglect of uncertainty. While UND-KNN improves the result to 74.1% by identifying core regions, it still struggles with imprecise boundaries. In contrast, our full uncertainty-driven strategy achieves the best performance (75.6% mIoU), confirming that integrating uncertainty throughout the region-growing process effectively refines structural boundaries and ensures feature coherence across complex regions.

Role of Spherical Harmonic Encoding This experiment assesses the importance of directional geometry modeling by replacing spherical harmonic (SH) encoding with simpler pooling operations, such as max pooling and average pooling, within Gaussian neighborhoods.

We hypothesize that removing SH encoding results in a performance drop, as SH encoding stores specific spatial information for different directions. Unlike pooling methods, which aggregate information across all directions indiscriminately, SH encoding preserves the fine-grained spatial details essential for accurate 3D geometry modeling. Each aggregation center in SH encoding can capture multiple categories of points with specific orientations, leading to more structured and coherent region representations.

In contrast, pooling methods like max pooling and average pooling tend to mix points from different directions, losing the critical directional information. While attention mechanisms later in the model can capture some distance-related features, they still fail to fully recover the fine-grained directional information that SH encoding preserves. As shown in Table 3, the model's performance drops when SH encoding is replaced with max pooling, resulting in a mIoU of 72.8%. Replacing it with average pooling improves performance slightly to 73.1%, but it still does not match the performance achieved with SH encoding, reinforcing the crucial role SH encoding plays in preserving directionally-specific spatial information.

Path-based Context Propagation We further analyze the role of the Spherical Harmonics (SH) encoding and the path-direction (path-dir) modeling in Table 3. We observe that SH brings consistent improvement only when path-dir information is explicitly preserved. Once path-dir modeling is removed, the performance gain from SH becomes marginal. This phenomenon is expected, as SH is inherently designed to model angular variations and directional signals. Without explicit path-direction cues, the encoded harmonic coefficients cannot be fully aligned with the underlying geometric structure. Therefore, SH and path-dir modeling are complementary: SH provides a structured angular basis representation, while path-dir supplies the necessary geometric context that enables SH to be effectively exploited.

We further study the effect of the Transformer-based attention module. Although the sequence length in our setting is relatively short, the underlying representation is structurally complex. The SH encoding introduces strong inter-channel dependencies due to the coupled nature of spherical harmonic bases. The coefficients are not independent; instead, they jointly encode directional information with intrinsic harmonic correlations. Consequently, simple sequential modeling operations (e.g., naive ordering or pooling-based aggregation) are insufficient to capture such high-order cross-channel interactions. The attention mechanism mitigates this issue by enabling global, content-adaptive interactions among SH coefficients, thereby improving the aggregation of structured angular information and leading to consistent performance gains.

4.4 Parameter Sensitivity & Calculation Efficiency

To further evaluate the robustness of the proposed framework, we conduct comprehensive sensitivity studies on three key hyperparameters: the spherical harmonic order L, the number of Gaussian spheres sampled along each propagation

Table 4: Ablation studies on S3DIS Area5. We investigate the performance impact of (a) Spherical Harmonic order L, (b) the number of Gaussian spheres along the path, and (c) the number of Monte Carlo samples K. All experiments are conducted on the PTv3 baseline.

(a) SH Order		**(b) # Spheres**		**(c) MC Samples K**			
L	mIoU (%)	Num.	mIoU (%)	K	mIoU (%)	K	mIoU (%)
2	74.2	3	74.2	1 (None)	73.3	12	75.2
3	**75.6**	5	**75.6**	4	74.2	16	**75.6**
4	75.6	7	75.5	8	74.7	20	75.6

path N_s, and the Monte Carlo (MC) sampling number K for uncertainty estimation. The results are summarized in Table 4. Due to space constraints, additional sensitivity analyses regarding the k-nearest high-uncertainty neighbors and the minimum threshold T_{min} within each Gaussian sphere are provided in the Supplementary Material, further validating the stability of our uncertainty-driven approach across diverse settings.

Effect of Spherical Harmonic Order. As shown in Table 4, increasing the SH order from $L = 2$ to $L = 3$ significantly improves performance (74.23 $\rightarrow$ 75.62 mIoU), indicating that higher-order harmonics better capture anisotropic and direction-sensitive geometric structures. However, further increasing to $L = 4$ yields only marginal improvement (75.64 mIoU), suggesting that moderate harmonic order is sufficient to model dominant directional distributions, while higher frequencies contribute limited additional information. Therefore, we adopt $L = 3$ as the default configuration.

Sensitivity to the Number of Gaussian Spheres. Table 4 analyzes the impact of the number of aggregation spheres along each propagation path. Using only 3 spheres results in insufficient contextual modeling (74.2 mIoU). Increasing to 5 spheres substantially enhances performance (75.62 mIoU), demonstrating the benefit of richer intermediate context aggregation. When further increasing to 7 spheres, performance slightly decreases (75.53 mIoU), suggesting that excessive sampling introduces redundant context and may dilute stable directional cues. Hence, 5 spheres provide the best balance between contextual coverage and stability.

Impact of Monte Carlo Sampling Number. Table 4 presents the effect of different MC sampling numbers. Without uncertainty modeling ($K = 1$), the model achieves 73.3 mIoU, equivalent to the PTv3 baseline. As K increases, segmentation performance steadily improves and saturates around $K = 16$. Beyond this point, further increasing sampling (e.g., $K = 20$) does not yield additional gain, indicating that uncertainty estimation becomes stable once sufficient stochastic samples are collected.

Computational Efficiency. Table 5 reports efficiency analysis on an H100 GPU on S3DIS Area 5, together with three simpler alternatives. For our UGSP, the 16 MC passes are executed in parallel via batch inference on a single GPU rather than sequentially. Among the compared alternatives, a matched-parameter MLP head (+MLP Refine) reaches only 74.4%, showing the gain is not from ex-

Table 5: Comparison with simpler refinement alternatives on S3DIS Area 5. Latency, throughput (FPS), and GPU memory are measured on a single H100 GPU.

Method	mIoU (%)	Lat. (ms)	FPS	Mem. (GB)
PTv3 (backbone)	73.3	105	9.5	5.8
+ MLP Refine	74.4	121	8.3	11.4
+ TTA (16×)	73.7	2,194	0.5	8.8
+ RAND-KNN	72.8	183	5.5	8.2
+ UGSP (Ours)	**75.6**	**183**	**5.5**	**8.2**

tra parameters. Test-time augmentation (TTA, 16×), which averages predictions over 16 randomly augmented copies of the input, is far costlier (2,194 ms, 0.5 FPS) yet attains only 73.7%, confirming that naive multi-pass ensembling cannot match uncertainty-guided routing; and RAND-KNN falls below the PTv3 baseline (72.8%) at the same cost, isolating the causal role of the uncertainty signal. Params/FLOPs grow from 46.17M/102.3G to 58.72M/135.46G, an overhead that remains favorable given the +2.3% mIoU gain since refinement operates only on uncertainty-guided regions.

5 Conclusion

In this paper, we propose an uncertainty-driven Gaussian sphere propagation framework for 3D semantic segmentation, treating predictive uncertainty as a structural prior to identify semantic ambiguities. By integrating a spherical harmonics-based aggregation mechanism, our method enables robust information propagation along continuous spatial paths, effectively bridging geometric gaps in sparsely sampled areas. Through trajectory-aware modeling and cross-attention refinement, we selectively inject reliable contextual features into high-uncertainty regions, significantly enhancing discriminative power. Extensive evaluations on five indoor and outdoor benchmarks demonstrate that our approach consistently outperforms state-of-the-art models like PTv3, particularly in complex and ambiguous scenarios.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62276019, 62306028, 62501043, in part by the Fundamental Research Funds for the Central Universities under Grant 2022JBQY007, 2026XKRC005, in part by the Beijing Natural Science Foundation under Grant L231019.

References

1. Aksoy, E.E., Baci, S., Cavdar, S.: Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving. In: 2020 IEEE intelligent vehicles symposium (IV). pp. 926–932. IEEE (2020)

2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1534–1543 (2016)
3. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9297–9307 (2019)
4. Cabrera, J.J., Santo, A., Gil, A., Viegas, C., Payá, L.: Minkunext: Point cloud-based large-scale place recognition using 3d sparse convolutions. Array p. 100569 (2025)
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3075–3084 (2019)
8. Cortinhal, T., Tzelepis, G., Erdal Aksoy, E.: Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In: International Symposium on Visual Computing. pp. 207–222. Springer (2020)
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5828–5839 (2017)
10. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Nießner, M.: Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4578–4587 (2018)
11. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 9224–9232 (2018)
12. Huang, J., You, S.: Point cloud labeling using 3d convolutional neural network. In: Proc. Int. Conf. Pattern Recognit. pp. 2670–2675 (2016)
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
15. Komarichev, A., Zhong, Z., Hua, J.: A-cnn: Annularly convolutional neural networks on point clouds. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 7421–7430 (2019)
16. Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J.: Stratified transformer for 3d point cloud segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 8500–8509 (2022)
17. Lan, S., Yu, R., Yu, G., Davis, L.S.: Modeling local geometric structure of 3d point clouds using geo-cnn. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 998–1008 (2019)
18. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4558–4567 (2018)

19. Lei, H., Akhtar, N., Mian, A.: Seggcn: Efficient 3d point cloud segmentation with fuzzy spherical kernel. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11611–11620 (2020)
20. Li, D., Shi, G., Wu, Y., Yang, Y., Zhao, M.: Multi-scale neighborhood feature extraction and aggregation for point cloud segmentation. IEEE Trans. Circuits Syst. Video Technol. **31**(6), 2175–2191 (2020)
21. Li, G., Muller, M., Thabet, A., Ghanem, B.: Deepgcns: Can gcns go as deep as cnns? In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 9267–9276 (2019)
22. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. Adv. Neural Inf. Process. Syst. **31** (2018)
23. Li, Y., Li, X., Zhang, Z., Shuang, F., Lin, Q., Jiang, J.: Densekpnet: Dense kernel point convolutional neural networks for point cloud semantic segmentation. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2022)
24. Lin, H., Zheng, X., Li, L., Chao, F., Wang, S., Wang, Y., Tian, Y., Ji, R.: Meta architecture for point cloud analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17682–17691 (2023)
25. Liu, C., Zeng, D., Akbar, A., Wu, H., Jia, S., Xu, Z., Yue, H.: Context-aware network for semantic segmentation toward large-scale point clouds in urban environments. IEEE Trans. Geosci. Remote Sens. **60**, 1–15 (2022)
26. Liu, F., Li, S., Zhang, L., Zhou, C., Ye, R., Wang, Y., Lu, J.: 3dcnn-dqn-rnn: A deep reinforcement learning framework for semantic parsing of large-scale 3d point clouds. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 5678–5687 (2017)
27. Liu, S., Chi, J., Wu, C., Xu, F., Yu, X.: Sgt-net: A transformer-based stratified graph convolutional network for 3d point cloud semantic segmentation. Comput. Mater. Contin. **79**(3) (2024)
28. Liu, Z., Yang, X., Tang, H., Yang, S., Han, S.: Flatformer: Flattened window attention for efficient point cloud transformer. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1200–1211 (2023)
29. Luo, S., Sun, Z., Wang, Y., Sun, Y., Zhu, C.: Ldcnet: Long-distance context modeling for large-scale 3d point cloud scene semantic segmentation. In: Proc. ACM Multimedia. pp. 1321–1330 (2024)
30. Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. pp. 922–928 (2015)
31. Meng, H.Y., Gao, L., Lai, Y.K., Manocha, D.: Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 8500–8508 (2019)
32. Peng, B., Wu, X., Jiang, L., Chen, Y., Zhao, H., Tian, Z., Jia, J.: OA-CNNs: Omni-adaptive sparse cnns for 3d semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 21305–21315 (2024)
33. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 652–660 (2017)
34. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. **30** (2017)
35. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Adv. Neural Inf. Process. Syst. **35**, 23192–23204 (2022)
36. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)

37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
38. Riegler, G., Osman Ulusoy, A., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3577–3586 (2017)
39. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 1746–1754 (2017)
40. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
41. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: European conference on computer vision. pp. 685–702. Springer (2020)
42. Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S.: Segcloud: Semantic segmentation of 3d point clouds. In: Proc. 3DV. pp. 537–547 (2017)
43. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 6411–6420 (2019)
44. Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J.: Graph attention convolution for point cloud semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 10296–10305 (2019)
45. Wang, P.S.: Octformer: Octree-based transformers for 3d point clouds. ACM Trans. Graph. **42**(4), 1–11 (2023)
46. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM TOG **38**(5), 1–12 (2019)
47. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4840–4851 (2024)
48. Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H.: Point transformer v2: Grouped vector attention and partition-based pooling. Adv. Neural Inf. Process. Syst. **35**, 33330–33342 (2022)
49. Xu, M., Ding, R., Zhao, H., Qi, X.: Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (2021)
50. Yan, X., Zheng, C., Li, Z., Wang, S., Cui, S.: Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5589–5598 (2020)
51. Yang, Y.Q., Guo, Y.X., Xiong, J.Y., Liu, Y., Pan, H., Wang, P.S., Tong, X., Guo, B.: Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. Computational Visual Media **11**(1), 83–101 (2025)
52. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: European conference on computer vision. pp. 162–179. Springer (2024)
53. Zhang, Z., Hua, B.S., Yeung, S.K.: Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 1607–1616 (2019)

54. Zhao, H., Jiang, L., Fu, C.W., Jia, J.: Pointweb: Enhancing local neighborhood features for point cloud processing. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 5565–5573 (2019)
55. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 16259–16268 (2021)
56. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21676–21685 (2024)