SPATIAL-GEOMETRIC SELF-SUPERVISED LEARNING FOR LIDAR PANOPTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

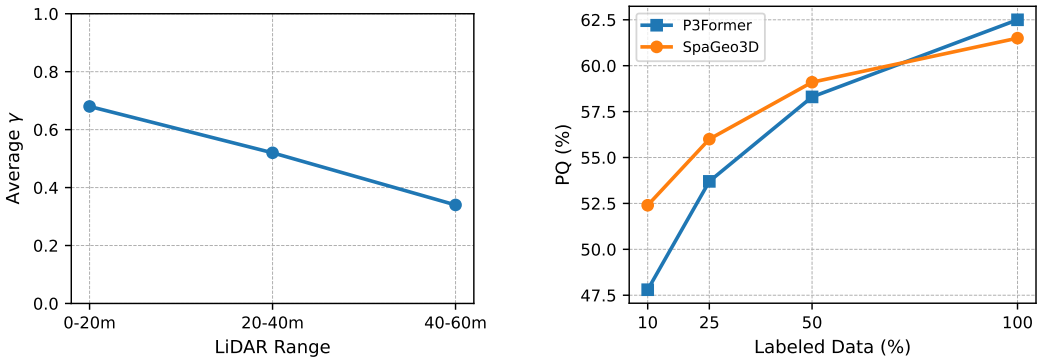
ABSTRACT

Recent advances in self-supervised learning (SSL) for LiDAR point clouds have significantly improved 3D scene understanding without requiring human annotations. However, existing approaches still face challenges in capturing fine-grained geometric structures from highly sparse outdoor LiDAR scenes, especially under limited annotation settings. Learning robust geometric representations is essential for LiDAR panoptic segmentation, which requires both accurate semantic recognition and precise instance localization. In this work, we introduce **SpaGeo3D**, a spatial-geometric self-supervised framework for LiDAR representation learning and panoptic segmentation. SpaGeo3D learns geometric-aware representations from unlabeled LiDAR scans through masked voxel reconstruction within visible local neighborhoods, enabling the model to capture spatial structures and surface geometry while avoiding unnecessary computation on large empty regions. To further exploit complementary spatial information, we propose a Multi-scale Polar Feature Fusion module that effectively integrates sparse 3D geometric features with polar-domain contextual representations. For instance-level prediction, an offset-density refinement mechanism jointly leverages geometric displacement and BEV density cues to improve instance center localization and object-level grouping. Extensive experiments on the SemanticKITTI and nuScenes datasets demonstrate that SpaGeo3D consistently outperforms existing approaches, achieving average improvements of +7.2% PQ and +12.2% mIoU in panoptic segmentation, while exhibiting strong generalization ability in large-scale outdoor LiDAR scenarios.

1 INTRODUCTION

In large-scale urban traffic scenarios, LiDAR point clouds are characterized by high sparsity, substantial variations in spatial scale, and pronounced non-uniformity in sampling density across near- and far-range regions. Meanwhile, the limited availability of annotated data poses a significant challenge to the development of effective 3D scene understanding models (Yang et al., 2024; Kong et al., 2023). Unlike regular 2D images, point clouds are inherently unstructured and exhibit strong spatial dependencies, making effective feature learning highly dependent on the ability to capture geometric structures and spatial distribution patterns. Consequently, learning robust 3D geometric representations from unlabeled point clouds under limited annotation has emerged as an important research challenge in 3D scene understanding.

In recent years, deep learning has driven substantial progress in 3D point cloud semantic and instance segmentation, with voxel-based, point-based, and hybrid representations providing different strategies for modeling geometric structures (Jiang et al., 2020; Zhu et al., 2021). More recently, LiDAR panoptic segmentation has increasingly explored label-efficient learning, multimodal representation, and unified 3D scene understanding (Pan et al., 2025; 2026). Despite these advances, panoptic segmentation continues to face a fundamental trade-off between representational capacity and computational efficiency. 3D convolution-based approaches can effectively capture local geometric structures, but often introduce considerable computational and memory overhead. In contrast, 2D projection-based representations, such as bird's-eye view (BEV), provide greater computational efficiency but may discard important 3D spatial information, particularly in sparse distant regions and around object boundaries. Moreover, both voxelization and projection introduce discretization



(a) Gamma analysis.

(b) Label efficiency analysis.

Figure 1: Analysis of the proposed SpaGeo3D framework. (a) Effect of different gamma values. (b) Label efficiency comparison under limited annotation ratios.

into the representation process, which can inevitably lead to the loss of fine-grained geometric information and consequently limit segmentation performance.

In parallel, self-supervised learning (SSL) has emerged as an effective strategy for alleviating the reliance on large-scale annotated data by learning generalizable representations from unlabeled point clouds. Existing approaches typically learn geometric representations by reconstructing point cloud occupancy structures or geometric distributions. However, large-scale outdoor scenes contain a substantial number of empty voxels, making dense reconstruction computationally expensive and potentially allocating modeling capacity to uninformative regions. Meanwhile, efficient representations based on BEV or pillar structures provide favorable computational efficiency but may still lose fine-grained 3D geometric information. These limitations motivate a representation learning strategy that can exploit unlabeled LiDAR data while selectively focusing on informative geometric structures.

As illustrated in Fig. 1, **SpaGeo3D** exhibits robust performance under different hyperparameter settings and achieves favorable annotation efficiency. The gamma analysis demonstrates that the proposed framework maintains stable panoptic segmentation performance over a broad range of gamma values, highlighting the robustness of the feature interaction strategy. Furthermore, the label efficiency analysis shows that SpaGeo3D can achieve competitive performance with substantially fewer annotated samples, demonstrating the effectiveness of self-supervised geometric pretraining for large-scale LiDAR panoptic segmentation.

Motivated by these observations, we propose **SpaGeo3D**, a unified framework for large-scale LiDAR point cloud panoptic segmentation. The framework centers on self-supervised geometric representation learning, where unlabeled LiDAR scans are exploited to learn spatial and surface-level geometric priors before panoptic prediction. These learned representations are further enhanced through multi-scale polar-domain interaction, allowing complementary information from sparse 3D geometric and polar-view representations to be effectively integrated. We evaluate **SpaGeo3D** extensively on large-scale LiDAR scene datasets, demonstrating consistent improvements in panoptic segmentation performance. Our method achieves a **7.2% PQ gain** and a **12.2% mIoU gain**, highlighting the effectiveness of self-supervised geometric representation learning and multi-scale polar-domain interaction. To summarize, this work contains the following key contributions:

- **SpaGeo3D** is presented as an end-to-end framework for LiDAR panoptic segmentation, integrating self-supervised geometric pretraining with multi-scale polar-domain interaction to improve geometric representation and contextual feature modeling.
- A **visible-region masked pretraining scheme** is employed with a 3D convolutional encoder to learn geometric priors from unlabeled LiDAR scans. Local voxel occupancy is predicted within visible neighborhoods, enabling the encoder to capture spatial structures and surface characteristics while reducing redundant computation over empty regions.

- A **Multi-scale Polar Feature Fusion (MPFF)** module is introduced to establish complementary interactions between sparse 3D geometric features and polar-domain contextual representations. In addition, an offset–density refinement mechanism combines point-wise offset information with BEV density cues for instance-center localization and object clustering.

2 Related Work

With the rapid development of deep learning, LiDAR-based 3D scene understanding has achieved substantial progress in autonomous driving and robotics (Qi et al., 2017; Royen et al., 2024). Existing 3D instance and panoptic segmentation methods have evolved from top-down and bottom-up formulations toward hybrid, unified, and multimodal paradigms (Mohan et al., 2026). Nevertheless, balancing segmentation accuracy, geometric representation, and computational efficiency remains challenging for large-scale sparse LiDAR scenes.

3D Instance and Panoptic Segmentation. Existing 3D instance segmentation methods can be broadly categorized into **top-down**, **bottom-up**, and **hybrid** approaches. Top-down methods first generate object proposals and then perform instance-level refinement (Zhang et al., 2020), making their performance dependent on proposal quality. Bottom-up methods directly learn point- or voxel-level relations and group features into instances (Xiang et al., 2020), avoiding proposal errors but remaining sensitive to sparse observations and large-object clustering. Hybrid approaches, such as SoftGroup and SoftGroup++ (Vu et al., 2022; 2023), combine efficient grouping and instance refinement, while PANet (Mei et al., 2023) exploits complementary multi-scale representations to improve feature interaction.

Panoptic segmentation further requires unified semantic and instance reasoning. Panoptic-PolarNet (Zhou et al., 2021), EvLPSNet (Sirohi et al., 2022), and CFNet (Li et al., 2023) employ efficient BEV or polar representations to improve large-scale LiDAR processing, while DS-Net (Hong et al., 2021) explicitly addresses non-uniform point density. Although such representations are computationally efficient, projection and discretization may weaken fine-grained 3D geometry, particularly for distant objects and boundaries. More recently, unified Transformer-based methods, including OneFormer3D (Kolodiazhnyi et al., 2023) and P3Former (Xiao et al., 2023), have improved joint semantic and instance modeling. Multimodal approaches such as IAL (Pan et al., 2025) and PanDA (Pan et al., 2026) further leverage image–LiDAR complementarity and domain adaptation. However, their increased modeling capacity often introduces additional computational complexity and dependence on rich multimodal information.

Self-supervised 3D Representation Learning. Self-supervised learning provides an effective way to exploit large-scale unlabeled LiDAR data and reduce reliance on dense annotations. MAELi (Krispel et al., 2023) introduced masked autoencoding tailored to sparse LiDAR observations, while BEV-MAE (Lin et al., 2024) incorporated BEV-guided masking and point-density information. Yan et al. (Yan et al., 2024) explored masked geometric feature prediction, and NOMAE (Abdelsamad et al., 2025a) further improved efficiency through multi-scale neighborhood occupancy reconstruction. These studies demonstrate the effectiveness of exploiting geometric priors from unlabeled LiDAR scans. However, most existing approaches focus on generic representation learning, detection, or semantic segmentation, leaving their integration with panoptic segmentation and multi-scale geometric–contextual interaction relatively underexplored.

Discussion and Motivation. Existing studies reveal a fundamental trade-off between geometric fidelity and computational efficiency. Sparse 3D representations preserve detailed local geometry but can incur considerable computation in large outdoor scenes, whereas polar or BEV representations efficiently capture global context at the cost of fine-grained geometric information. Meanwhile, self-supervised pretraining offers a promising means of learning geometric priors from unlabeled LiDAR scans, but has rarely been jointly explored with panoptic segmentation and multi-scale feature interaction. Motivated by these observations, we propose **SpaGeo3D**, which combines observation-aware self-supervised geometric pretraining with multi-scale polar-domain interaction to achieve efficient and geometry-aware LiDAR panoptic segmentation.

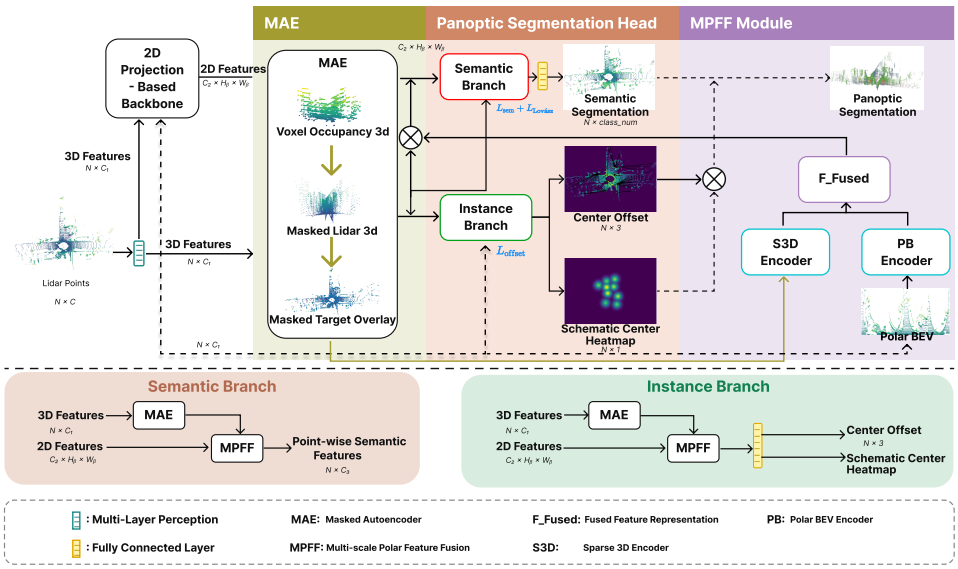


Figure 2: Overall architecture of **SpaGeo3D**. The framework consists of observation-aware self-supervised geometric pretraining, **Multi-scale Polar Feature Fusion (MPFF)**, and offset–density refinement for instance modeling. Sparse 3D geometric features and polar-domain contextual features are aligned and interacted for unified LiDAR panoptic segmentation.

3 METHOD

3.1 PRELIMINARY

Our framework builds upon self-supervised masked voxel modeling and multi-scale polar representation learning for large-scale LiDAR panoptic segmentation. As shown in Fig. 2, given a LiDAR point cloud $\mathcal{P} = \{(x_i, f_i)\}_{i=1}^{N}$, where x_i and f_i denote the coordinate and feature of point i, respectively, the point cloud is first voxelized and randomly divided into visible and masked voxels. The visible voxels are processed by the proposed self-supervised encoder to learn multi-scale geometric representations from unlabeled LiDAR scans.

The learned geometric features are further enhanced through multi-scale interaction between sparse 3D and polar-domain representations, where complementary local geometric and contextual information is effectively integrated. The fused features are then fed into the panoptic segmentation head, where geometric offsets and BEV density cues are jointly exploited for instance localization and grouping.

3.2 VISIBLE MASKED MULTI-SCALE SELF-SUPERVISED ENCODER

The proposed Visible Masked Multi-scale Self-Supervised Encoder learns geometric representations from unlabeled LiDAR point clouds through local occupancy reconstruction. Given an input point cloud, we first voxelize the scene and randomly divide the resulting voxels into visible voxels V_m and masked voxels V_n. For each visible voxel $v \in V_m$, its representation is constructed by combining the aggregated point feature and spatial positional encoding:

$$x_v = \phi(v) + e(v), \tag{1}$$

where $\phi(\cdot)$ denotes the aggregated geometric feature and $e(\cdot)$ denotes the positional encoding. The visible voxel features are processed by a Transformer encoder to capture long-range geometric dependencies.

To learn hierarchical geometric representations, the visible voxels are progressively downsampled into multiple scales. Features at different resolutions are subsequently fused through a top-down propagation strategy, enabling the encoder to jointly preserve fine-grained local structures and large-scale spatial context.

Different from conventional masked autoencoders that reconstruct the entire voxelized scene, SpaGeo3D restricts reconstruction to the local neighborhoods of visible voxels. Specifically, for each $v \in V_m$, we define

$$\mathcal{N}(v) = \{u \mid \|u - v\|_\infty \leq 4\}, \tag{2}$$

corresponding to a local voxel block of size $9 \times 9 \times 9$. Occupancy prediction is performed only for voxels within these neighborhoods, while isolated masked voxels outside the reconstruction regions are excluded from supervision. A sparse convolutional decoder predicts occupancy probabilities at each scale, and the binary cross-entropy loss is used for optimization. By focusing reconstruction on visible local neighborhoods, the proposed strategy avoids redundant computation over massive empty regions and encourages the encoder to learn spatial continuity and surface-level geometric structures. The resulting multi-scale representation is then transferred to the downstream panoptic segmentation network.

3.3 Multi-scale Polar Feature Fusion

To achieve effective interaction between 3D geometric features and 2D semantic features under different receptive fields, we propose the Multi-scale Polar Feature Fusion (MPFF) module. The module aligns multi-level features from the 3D sparse convolution branch and the 2D polar-coordinate U-Net branch within a unified polar BEV space, thereby establishing the correspondence between sparse voxel representations and BEV representations across different scales. Specifically, the 3D sparse convolution branch extracts multi-scale geometric representations $\{F_{3D}^{(s)}\}$, which are projected into the BEV space through polar coordinate transformation to obtain aligned features $\{F_{3D}^{proj(s)}\}$. Meanwhile, the 2D polar-coordinate U-Net branch generates semantic representations $\{F_{2D}^{(s)}\}$ at corresponding scales. For each BEV location (u, v), an adaptive fusion weight $\gamma^{(s)}(u, v)$ is calculated based on the two types of features and their polar coordinates (r, θ):

$$\gamma^{(s)}(u, v) = \sigma(\mathrm{MLP}([F_{2D}^{(s)}(u, v), F_{3D}^{proj(s)}(u, v), r, \theta])). \tag{3}$$

Subsequently, the two feature representations are adaptively aggregated according to the learned weights to obtain the fused representation $F_{fused}^{(s)}$:

$$F_{fused}^{(s)} = \gamma^{(s)} \odot F_{3D}^{proj(s)} + (1 - \gamma^{(s)}) \odot F_{2D}^{(s)}. \tag{4}$$

The above fusion process is performed independently at different feature scales, enabling collaborative modeling of local geometric details and global semantic contexts in a multi-scale manner. By jointly considering geometric features, semantic representations, and polar positional information, the proposed adaptive fusion mechanism dynamically adjusts the contribution of different feature modalities according to spatial distribution variations in point clouds. Consequently, it improves semantic representation capability in distant sparse regions and enhances boundary localization accuracy for nearby objects.

3.4 Back-Projection-based Instance Modeling and Mutual Feedback Strategy

For panoptic prediction, we first estimate an offset vector Δp for each voxel $p \in V^{th}$ belonging to the Thing classes and shift it toward its corresponding instance center:

$$V^{th'} = \{p + \Delta p \mid p \in V^{th}\}. \tag{5}$$

As illustrated in Fig. 3(a), the predicted offsets guide Thing voxels toward their corresponding instance centers, thereby reducing the spatial dispersion of points belonging to the same instance. The shifted voxels are then projected into the BEV space to construct a density heatmap $H(x, y)$. As shown in Fig. 3(c), the back-projection operation aggregates spatially related voxels and produces concentrated density regions, which facilitate reliable instance-center localization. A window-based maximum pooling operation is subsequently applied to identify local density peaks as candidate instance centers. The distribution of offset magnitudes is presented in Fig. 3(b), providing an overview of the spatial displacement required for instance aggregation. Each Thing voxel is subsequently assigned to its nearest center to obtain instance groups. The overall instance prediction procedure is summarized in Algorithm 1.

Algorithm 1 Back-Projection-based Instance Modeling

Require: Thing voxels $\{p_t\}_{t=1}^{M}$, predicted offsets $\{\Delta p_t\}_{t=1}^{M}$, BEV grid, window size w
Ensure: Instance assignments $\{I_t\}_{t=1}^{M}$

1: Shift Thing voxels:

$$\tilde{p}_t \leftarrow p_t + \Delta p_t, \quad t = 1, \ldots, M$$

2: Project $\{\tilde{p}_t\}$ into BEV space
3: Construct density heatmap $H(x, y)$ by voxel counting
4: Apply window-based maximum pooling to H
5: Extract local density peaks as candidate centers $\{c_j\}_{j=1}^{D}$
6: **for** $t \leftarrow 1$ to M **do**
7: $\quad I_t \leftarrow \arg\min_j \|\tilde{p}_t - c_j\|_2$
8: **end for**
9: **return** Instance assignments $\{I_t\}_{t=1}^{M}$

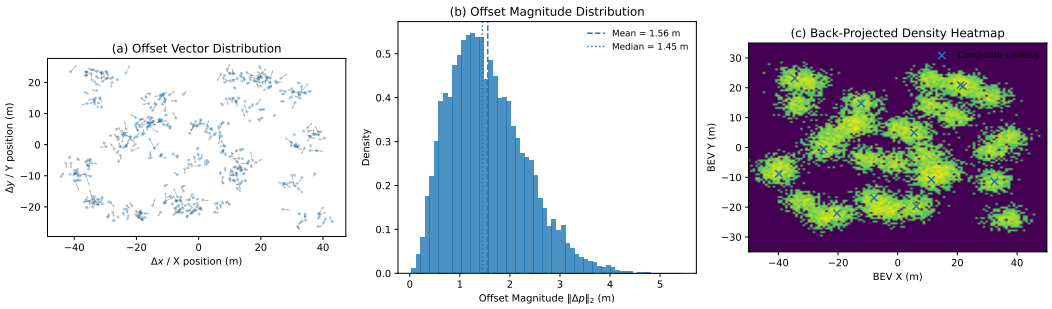


Figure 3: Visualization of the back-projection-based instance modeling. (a) Predicted offset vectors of Thing voxels, where the offsets guide voxels toward their corresponding instance centers. (b) Distribution of offset magnitudes. (c) BEV density heatmap after back-projection, where concentrated density regions facilitate the localization of candidate instance centers.

To jointly optimize semantic and instance representations, we employ a multi-task objective combining semantic segmentation, Lovász-Softmax, and offset regression losses:

$$L_{\text{total}} = \lambda_1 L_{\text{sem}} + \lambda_2 L_{\text{Lovász}} + \lambda_3 L_{\text{offset}}. \tag{6}$$

This joint optimization enables the semantic and instance branches to provide complementary supervision and improves the overall panoptic segmentation performance.

4 Experiments

Implementation Details All experiments are conducted on a workstation equipped with two Intel Xeon Silver 4110 CPUs, one NVIDIA Quadro RTX 6000 GPU, and 64 GB RAM. SpaGeo3D is built upon Panoptic-PolarNet (Zhou et al., 2021) and implemented in PyTorch. The model is optimized using Adam with an initial learning rate of 0.01, a batch size of 2, and 50 training epochs. Input point clouds are cropped to $[-51.2, 51.2]$ m horizontally and $[-5, 3]$ m vertically, and represented using a BEV grid of $480 \times 360 \times 64$. The training objective combines semantic segmentation, Lovász-Softmax, offset regression, and occupancy reconstruction losses. For instance prediction, the predicted offsets are projected into BEV space to construct center heatmaps, followed by local maximum suppression with a kernel size of 5 and at most 20 candidate centers.

Evaluation Metrics Following the official evaluation protocols of SemanticKITTI and nuScenes Behley et al. (2019); Caesar et al. (2020), we adopt Panoptic Quality (PQ) as the primary metric, together with Segmentation Quality (SQ) and Recognition Quality (RQ). We further report the corresponding metrics for *Thing* and *Stuff* classes, and use mean Intersection over Union (mIoU) to evaluate semantic segmentation.

Comparative Study Table 1 compares SpaGeo3D with recent self-supervised methods under three downstream evaluation protocols. Under linear evaluation, SpaGeo3D achieves 58.3% PQ on

Table 1: Comparison of self-supervised learning methods on outdoor scene datasets under linear evaluation, decoder evaluation, and full fine-tuning. Panoptic-PolarNet is included as the supervised baseline for reference.

Method	nuScenes val				SemanticKITTI val			
	mIoU	PQ	SQ	RQ	mIoU	PQ	SQ	RQ
Linear Evaluation								
SegContrast Nunes et al. (2022)	37.8	25.4	69.8	33.4	–	–	–	–
NOMAE Abdelsamad et al. (2025a)	64.7	45.5	77.0	56.4	29.8	17.1	54.7	24.9
Sonata Wu et al. (2025)	59.2	50.7	79.8	61.6	46.5	34.5	65.5	44.2
DOS Abdelsamad et al. (2025b)	74.1	57.4	82.8	68.5	67.5	49.6	71.8	60.9
SpaGeo3D	73.5	**58.3**	81.5	**71.5**	65.5	46.3	72.5	58.0
Decoder Evaluation								
SegContrast	73.1	60.7	83.9	71.7	–	–	–	–
NOMAE	80.1	69.0	85.6	79.3	64.3	52.4	73.6	62.3
Sonata	76.8	66.0	85.2	76.8	64.5	55.1	77.5	64.8
DOS	79.2	69.1	85.3	79.0	68.6	56.7	76.3	65.2
SpaGeo3D	79.5	68.8	**86.5**	**79.5**	68.5	57.7	77.5	**67.5**
Full Fine-tuning								
SegContrast	78.0	69.3	85.9	80.3	–	–	–	–
NOMAE	81.8	71.0	86.5	80.8	71.6	59.5	76.4	68.9
Sonata	79.6	70.0	86.4	80.5	69.6	58.2	78.5	67.7
DOS	81.5	70.5	86.0	81.1	73.1	59.2	76.3	68.6
SpaGeo3D	81.5	**72.2**	**88.0**	**82.0**	71.0	**61.5**	**79.1**	**71.7**

Table 2: Comparison with state-of-the-art methods on the SemanticKITTI test set.

Method	PQ	RQ	SQ	PQ^{Th}	RQ^{Th}	SQ^{Th}	PQ^{St}	RQ^{St}	SQ^{St}
RangeNet++ Milioto et al. (2019)/PointPillars	37.1	45.9	47.0	20.2	75.2	25.2	49.3	62.8	76.5
KPConv Thomas et al. (2019)/PointPillars	44.5	54.4	80.0	32.7	38.7	81.5	53.1	65.9	79.0
Panoster Gasperini et al. (2021)	52.7	64.1	80.7	49.4	58.5	83.3	55.1	68.2	78.8
Panoptic-PolarNet Zhou et al. (2021)	54.3	65.0	81.4	53.3	60.6	87.2	54.8	68.1	77.2
DS-Net Hong et al. (2021)	55.9	66.7	82.3	55.1	62.8	87.2	56.5	69.5	78.7
EfficientLPS Sirohi et al. (2021)	57.4	68.7	83.0	53.1	60.5	87.8	60.5	74.6	79.5
GP-S3Net Razani et al. (2021)	60.0	72.1	82.0	65.0	74.5	86.6	56.4	70.4	78.7
SpaGeo3D	**61.5**	71.7	79.1	**66.9**	74.4	87.2	57.5	69.7	73.1

nuScenes, outperforming DOS by 0.9 points. Under decoder evaluation, it obtains 68.8% PQ on nuScenes and 57.7% on SemanticKITTI. After full fine-tuning, SpaGeo3D reaches 72.2% PQ on nuScenes and 61.5% on SemanticKITTI, demonstrating that the learned geometric representations transfer effectively across datasets and downstream evaluation settings.

Overall Performance on SemanticKITTI As shown in Table 2, SpaGeo3D achieves 61.5% PQ on the SemanticKITTI test set, improving over Panoptic-PolarNet by 7.2 percentage points. It further obtains the highest PQ^{Th} of 66.9%, indicating improved instance-level modeling for Thing classes. These results demonstrate the effectiveness of combining self-supervised geometric representation learning with multi-scale polar feature interaction.

Accuracy and Efficiency Table 3 compares semantic segmentation accuracy and inference speed. SpaGeo3D achieves 71.0% mIoU at 10.8 FPS. Compared with Panoptic-PolarNet, it improves mIoU by 12.2 percentage points while retaining comparable inference efficiency, indicating that the proposed geometric representation and feature interaction substantially improve semantic understanding without introducing excessive computational overhead.

Per-class Analysis Table 4 reports class-wise performance on SemanticKITTI. SpaGeo3D performs strongly on both Thing and Stuff categories, achieving high PQ values for road (92.1%), building (87.8%), vegetation (88.9%), car (91.5%), and bicyclist (90.7%). Performance remains lower for sparsely observed rare classes such as motorcyclist, highlighting the difficulty of limited geometric evidence and class imbalance.

Table 3: Semantic segmentation accuracy and inference speed comparison on SemanticKITTI.

Method	mIoU(%)	FPS
RangeNet++/PointPillars	52.4	2.4
LPSAD	51.0	11.8
KPConv/PointPillars	58.8	1.9
Panoster	60.4	–
Panoptic-PolarNet	58.8	11.6
DS-Net	61.6	2.1
EfficientLPS	61.4	–
GP-S3Net	70.8	–
SpaGeo3D	**71.0**	**10.8**

Table 4: Per-class panoptic segmentation results on SemanticKITTI.

Class	Car	Bicycle	Motorcycle	Truck	Bus	Person	Bicyclist	Motorcyclist	Road	Building	Vegetation	Pole
PQ	91.5	58.0	78.4	61.1	53.0	85.8	90.7	41.0	92.1	87.8	88.9	61.3
SQ	93.0	78.3	93.5	93.0	88.4	91.6	93.7	84.4	93.4	91.0	88.9	75.2
RQ	98.4	74.0	83.9	65.7	60.0	94.1	96.8	48.5	98.6	96.3	100.0	81.5

Visualization Results Figures 4–6 show qualitative results on SemanticKITTI. SpaGeo3D preserves clear semantic boundaries, separates adjacent instances of the same category, and produces spatially coherent panoptic predictions across diverse driving scenes. Remaining errors mainly occur for extremely sparse or rare objects.

Effectiveness of Individual Components Table 5 evaluates the contribution of the proposed components. Back-projection-based instance modeling improves PQ from 54.3% to 57.0%, while MPFF increases PQ to 58.4% and mIoU from 58.8% to 64.2%. Combining self-supervised pretraining, MPFF, and back projection yields 61.5% PQ and 71.0% mIoU, confirming the complementary contributions of geometric pretraining, multi-scale feature fusion, and instance refinement.

Label Efficiency Table 6 evaluates SpaGeo3D under limited annotation settings on nuScenes. With only 0.1% labeled data, SpaGeo3D achieves 34.5% PQ, exceeding DOS by 0.9 points. At 1% labels, its PQ increases to 42.5%, 1.3 points higher than DOS. These results demonstrate that self-supervised geometric pretraining effectively improves label efficiency and reduces reliance on dense annotations.

Range-aware Analysis Table 7 investigates the complementary behavior of 3D voxel and polar BEV representations across different ranges. The 3D voxel branch performs better in nearby regions, whereas the polar BEV branch becomes more effective as the observation range increases. MPFF consistently improves performance across all intervals, achieving 70.2%, 64.3%, and 55.6% PQ at 0–20 m, 20–40 m, and 40–60 m, respectively, with an overall PQ of 61.5%. These results verify that adaptive multi-scale fusion effectively combines local geometric details with broader polar-domain contextual information.

5 Conclusion

This paper presents **SpaGeo3D**, a geometry-aware self-supervised framework for LiDAR panoptic segmentation in large-scale outdoor scenes. The proposed framework integrates a multi-scale sparse 3D CNN branch with a lightweight polar-coordinate 2D U-Net branch, enabling effective interaction between 3D geometric features and polar-domain contextual information through the proposed **Multi-scale Polar Feature Fusion (MPFF)** module. Furthermore, the self-supervised geometric pretraining strategy enhances the representation capability of sparse point clouds by learning geometric priors from unlabeled LiDAR data, while the offset–density refinement mechanism improves instance-center localization and object-level grouping.

Extensive experiments on the SemanticKITTI and nuScenes datasets demonstrate the effectiveness of **SpaGeo3D**. On the SemanticKITTI benchmark, the proposed method achieves a panoptic quality

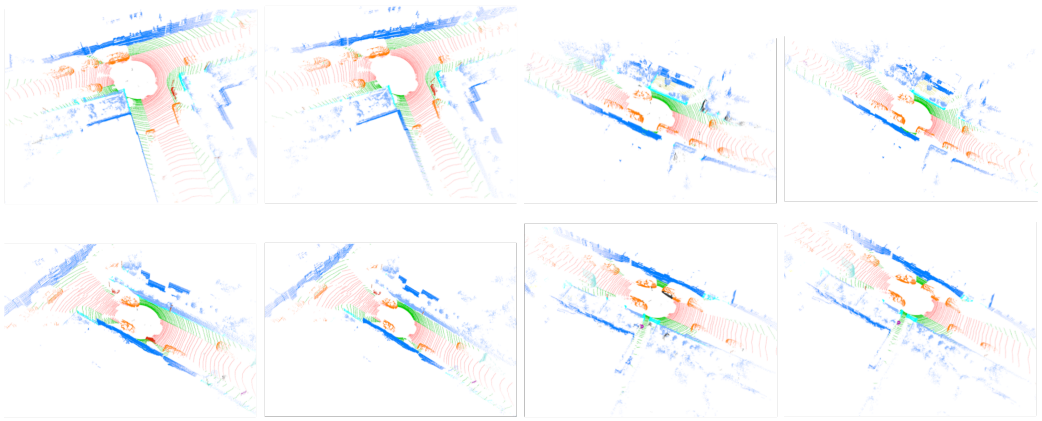


Figure 4: Semantic segmentation visualization results on the SemanticKITTI dataset.

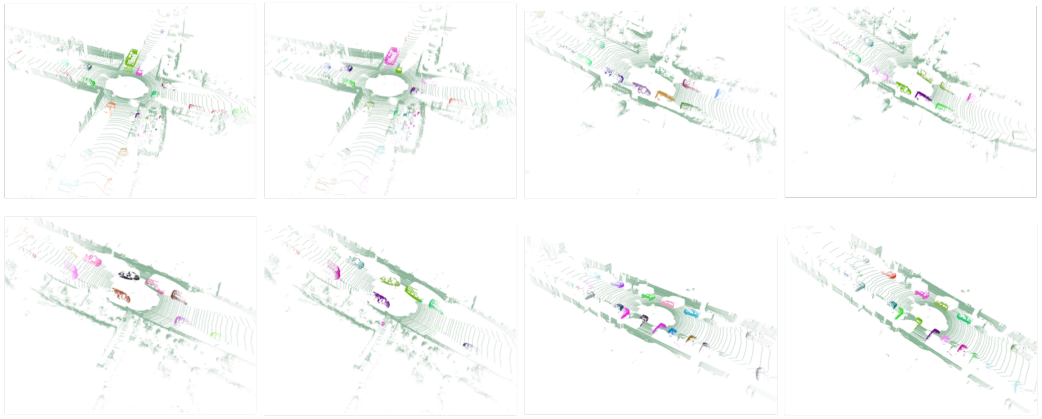


Figure 5: Instance segmentation visualization results on the SemanticKITTI dataset.

(PQ) of 61.5% and an mIoU of 71.0%, while achieving a PQ of 72.2% on the nuScenes validation set. These results demonstrate that self-supervised geometric pretraining can effectively exploit unlabeled LiDAR data to enhance geometric representation learning, while multi-scale polar interaction facilitates the integration of local geometric details and broader contextual information. The proposed framework achieves competitive panoptic segmentation performance while maintaining favorable computational efficiency. Building upon these results, future work will explore more robust geometric representation learning under varying levels of point sparsity and investigate adaptive representations for long-range and open-world LiDAR scene understanding.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript for auxiliary tasks, including language polishing, translation, structural revision, programming assistance, and the improvement of figure and table presentation. They were also used to help organize implementation notes, inspect experimental logs, and improve the clarity of technical descriptions. All experimental results, numerical values, and scientific conclusions reported in this paper were obtained from the actual implementation and experiments conducted by the authors. AI-generated suggestions were manually examined and, where applicable, cross-checked against the source code, experimental records, and generated outputs. The authors are fully responsible for the accuracy, validity, and integrity of the final manuscript.

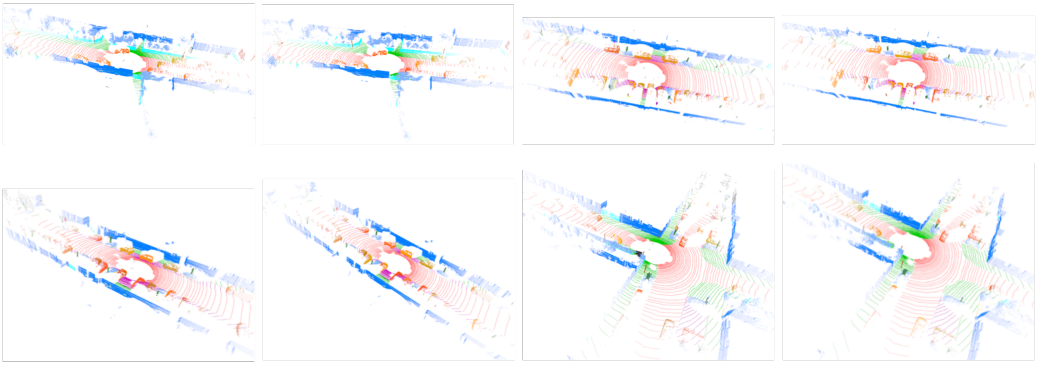


Figure 6: Panoptic segmentation visualization results on the SemanticKITTI dataset.

Table 5: Ablation study of different components on SemanticKITTI.

Method	MAE	MPFF	Back Projection	PQ(%)	mIoU(%)
Baseline	–	–	–	54.3	58.8
Baseline+BP	–	–	✓	57.0(+2.7)	58.7(-0.1)
Baseline+MPFF	–	✓	–	58.4(+4.1)	64.2(+5.4)
SpaGeo3D	✓	✓	✓	**61.5**(+7.2)	**71.0**(+12.2)

Table 6: Label-efficient training results on the nuScenes validation set.

Method	0.1% Labels			1% Labels		
	PQ	SQ	RQ	PQ	SQ	RQ
NOMAE	30.4	74.9	40.9	37.4	78.5	45.2
Sonata	27.1	71.4	34.0	34.7	72.5	42.3
DOS	33.6	74.3	41.2	41.2	75.7	49.2
SpaGeo3D	**34.5**	**75.4**	**45.8**	**42.5**	**82.2**	**51.7**

Table 7: Range-aware analysis of different feature representations on the SemanticKITTI validation set.

Method	0–20m PQ(%)	20–40m PQ(%)	40–60m PQ(%)	Overall PQ(%)
3D Voxel Branch only	67.8	58.9	45.6	56.2
Polar BEV Branch only	63.5	59.8	49.7	57.1
Simple Fusion (Concat)	68.4	61.7	52.1	59.4
MPFF (Ours)	**70.2**	**64.3**	**55.6**(+3.5)	**61.5**(+2.1)

Ethics Statement

This study focuses on self-supervised representation learning and panoptic segmentation for LiDAR point clouds in autonomous-driving environments. The experiments are conducted exclusively on publicly available benchmark datasets, including SemanticKITTI and nuScenes, and do not involve direct interaction with human participants or the collection of new personal data. No new dataset containing personally identifiable information was created as part of this work. As with other perception systems developed for autonomous platforms, practical deployment may require additional evaluation under diverse environmental conditions and appropriate consideration of safety, robustness, and potential failure cases.

Reproducibility Statement

We provide the main architectural design, training objectives, and experimental protocols required to reproduce SpaGeo3D in the main paper. The proposed self-supervised geometric pretraining procedure is described in Section 3, together with the visible-region reconstruction strategy, Multi-scale Polar Feature Fusion (MPFF), and the instance modeling procedure. The experimental section reports the datasets, optimization settings, input ranges, evaluation metrics, and major implementation

parameters used in our experiments. In addition, the paper includes algorithmic descriptions, ablation studies, label-efficiency experiments, and range-aware analyses to facilitate verification of the reported results. We use standard benchmark evaluation protocols for SemanticKITTI and nuScenes and report the corresponding settings consistently across all comparisons.

REFERENCES

Mohamed Abdelsamad, Michael Ulrich, Claudius Gläser, and Abhinav Valada. Multi-scale neighborhood occupancy masked autoencoder for self-supervised learning in lidar point clouds, 2025a. URL https://arxiv.org/abs/2502.20316.

Mohamed Abdelsamad, Michael Ulrich, Bin Yang, Miao Zhang, Yakov Miron, and Abhinav Valada. Dos: Distilling observable softmaps of zipfian prototypes for self-supervised point representation, 2025b. URL https://arxiv.org/abs/2512.11465.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences, 2019. URL https://arxiv.org/abs/1904.01416.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. URL https://arxiv.org/abs/1903.11027.

Stefano Gasperini, Mohammad-Ali Nikouei Mahani, Alvaro Marcos-Ramiro, Nassir Navab, and Federico Tombari. Panoster: End-to-end panoptic segmentation of lidar point clouds. *IEEE Robotics and Automation Letters*, 6(2):3216–3223, April 2021. ISSN 2377-3774. doi: 10.1109/lra.2021.3060405. URL http://dx.doi.org/10.1109/LRA.2021.3060405.

Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Gim Hee Liu. DS-Net: Dynamic Spatial Network for 3D LiDAR Panoptic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3844–3853, 2021. doi: 10.1109/CVPR46437.2021.00384.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation, 2020. URL https://arxiv.org/abs/2004.01658.

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation, 2023. URL https://arxiv.org/abs/2311.14405.

Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation, 2023. URL https://arxiv.org/abs/2303.05367.

Georg Krispel, David Schinagl, Christian Fruhwirth-Reisinger, Horst Possegger, and Horst Bischof. Maeli: Masked autoencoder for large-scale lidar point clouds, 2023. URL https://arxiv.org/abs/2212.07207.

Xiaoyan Li, Gang Zhang, Boyue Wang, Yongli Hu, and Baocai Yin. Center focusing network for real-time lidar panoptic segmentation, 2023. URL https://arxiv.org/abs/2311.09499.

Zhiwei Lin, Yongtao Wang, Shengxiang Qi, Nan Dong, and Ming-Hsuan Yang. Bev-mae: Bird's eye view masked autoencoders for point cloud pre-training in autonomous driving scenarios, 2024. URL https://arxiv.org/abs/2212.05758.

Jiacheng Mei, Ying Yang, Yuxin Wang, et al. PANet: A Proposal Aggregation Network for 3D LiDAR Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4213–4220, 2019.

Rohit Mohan, Florian Drews, Yakov Miron, Daniele Cattaneo, and Abhinav Valada. Up-fuse: Uncertainty-guided lidar-camera fusion for 3d panoptic segmentation, 2026. URL https://arxiv.org/abs/2602.19349.

Lucas Nunes, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. *IEEE Robotics and Automation Letters*, 7(2):2116–2123, 2022. doi: 10.1109/LRA.2022.3142440.

Yining Pan, Qiongjie Cui, Xulei Yang, and Na Zhao. How do images align and complement lidar? towards a harmonized multi-modal 3d panoptic segmentation, 2025. URL https://arxiv.org/abs/2505.18956.

Yining Pan, Shijie Li, Yuchen Wu, Xulei Yang, and Na Zhao. Panda: Unsupervised domain adaptation for multimodal 3d panoptic segmentation in autonomous driving, 2026. URL https://arxiv.org/abs/2604.19379.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017. URL https://arxiv.org/abs/1612.00593.

Ryan Razani, Ran Cheng, Enxu Li, Ehsan Taghavi, Yuan Ren, and Liu Bingbing. Gp-s3net: Graph-based panoptic sparse semantic segmentation network, 2021. URL https://arxiv.org/abs/2108.08401.

Remco Royen, Leon Denis, and Adrian Munteanu. Protoseg: A prototype-based point cloud instance segmentation method, 2024. URL https://arxiv.org/abs/2410.02352.

Kshitij Sirohi, Rohit Mohan, Daniel Büscher, Wolfram Burgard, and Abhinav Valada. Efficientlps: Efficient lidar panoptic segmentation, 2021. URL https://arxiv.org/abs/2102.08009.

Kushal Sirohi, Ehsan Marvasti, Philippe Giguere, and Gregory Dudek. EvLPSNet: Evidential Learning for LiDAR Point Cloud Segmentation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1156–1163, 2022.

Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds, 2019. URL https://arxiv.org/abs/1904.08889.

Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on point clouds, 2022. URL https://arxiv.org/abs/2203.01509.

Thang Vu, Kookhoi Kim, Tung M. Luu, Thanh Nguyen, Junyeong Kim, and Chang D. Yoo. Scalable softgroup for 3d instance segmentation on point clouds, 2023. URL https://arxiv.org/abs/2209.08263.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations, 2025. URL https://arxiv.org/abs/2503.16429.

Tiange Xiang, Chaoxu Zhang, Yang Song, Jianbing Yu, and Weidong Cai. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12658–12667, 2020. doi: 10.1109/CVPR42600.2020.01267.

Ziyang Xiao et al. P3Former: Parallel 3D Point Cloud Transformer for Semantic and Instance Segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14419–14428, 2023.

Siming Yan, Yuqi Yang, Yuxiao Guo, Hao Pan, Peng shuai Wang, Xin Tong, Yang Liu, and Qixing Huang. 3d feature prediction for masked-autoencoder-based point cloud pretraining, 2024. URL https://arxiv.org/abs/2304.06911.

Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Unipad: A universal pre-training paradigm for autonomous driving, 2024. URL https://arxiv.org/abs/2310.08370.

Feihu Zhang, Chenye Guan, Jin Fang, Song Bai, Ruigang Yang, Philip H.S. Torr, and Victor Prisacariu. Instance segmentation of lidar point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9448–9455, 2020. doi: 10.1109/ICRA40945.2020.9196622.

Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation, 2021. URL https://arxiv.org/abs/2103.14962.

Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception, 2021. URL https://arxiv.org/abs/2109.05441.