LiDAR Point Cloud Oversegmentation via SAM-based Knowledge Distillation

Dening Lu[1], Michael A. Chapman[2], Jonathan Li[1,3]

[1] Dept. of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada - d62lu@uwaterloo.ca
[2] Dept. of Civil Engineering, Toronto Metropolitan University, Toronto, Ontario, M5B 2K3, Canada - mchapman@torontomu.ca
[3] Dept. of Geography and Environmental Management, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada - junli@uwaterloo.ca

Keywords: LiDAR point cloud, Oversegmentation, Foundation model, Knowledge distillation.

Abstract

Large-scale LiDAR point clouds provide rich geometric information, yet learning effective structural representations remains challenging due to the misalignment between semantic categories and geometric structures. To address this issue, we propose a SAM-guided framework for point cloud oversegmentation. We transfer grouping knowledge from 2D vision by constructing a large-scale oversegmentation dataset using the Segment Anything Model (SAM) on bird's-eye-view projections. Based on these grouping priors, a structure-aware point cloud encoder is learned via a distillation objective that enforces intra-region compactness and inter-region separation in the embedding space. The proposed approach does not rely on semantic supervision and directly learns generalizable structural representations. Experiments on various benchmark datasets (STPLS3D, Toronto-3D, DALES, and S3DIS) demonstrate that the proposed method achieves competitive performance. In particular, it significantly improves boundary recall (e.g., 92.21% on STPLS3D and 93.47% on Toronto-3D) while maintaining high oracle accuracy (up to 97.62%). Moreover, the model generalizes well to unseen datasets without retraining, showing strong cross-dataset inference capability.

1. Introduction

With the fast development of LiDAR sensing technologies, large-scale point cloud data have become increasingly accessible across a wide range of applications. Modern acquisition platforms, such as UAVs and mobile laser scanners, are capable of capturing scenes containing tens of millions of points. While these data provide rich geometric information, the lack of scalable annotation strategies remains a fundamental limitation for learning-based methods (Meng et al., 2021; Xiao et al., 2024). A core difficulty arises from the fact that semantic categories do not always align well with geometric structures. Points belonging to different classes may exhibit similar spatial patterns, while instances within the same class can vary significantly in shape and distribution. As a result, directly operating on point-wise representations often leads to unstable grouping and ambiguous semantic boundaries, particularly in large and complex outdoor environments.

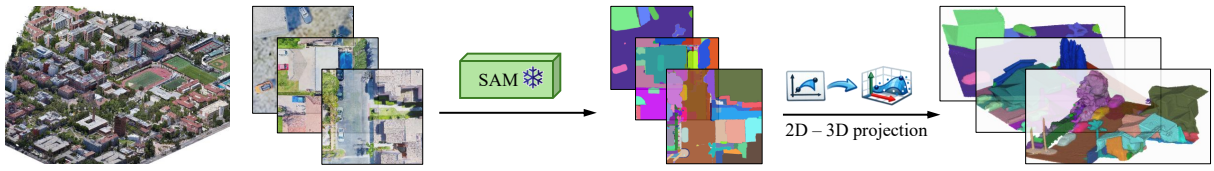
Such observation motivates a shift from point-level processing to structure-aware representations. Instead of assigning labels to individual points, it is more natural to first partition the scene into locally coherent regions that respect geometric continuity. Such an oversegmentation, commonly referred to as superpoints, provides a compact and robust description of the underlying scene structure. However, most existing approaches (Papon et al., 2013; Landrieu and Simonovsky, 2018; Landrieu and Boussaha, 2019; Hui et al., 2023; Robert et al., 2023; Lu et al., 2025) rely on handcrafted geometric heuristics or are trained with full semantic supervision, which limits their adaptability to new datasets and acquisition conditions.

In parallel, recent progress in 2D vision (Kirillov et al., 2023; Oquab et al., 2023) has shown that large-scale foundation models can learn transferable grouping cues without explicit semantic annotations. In particular, the Segment Anything Model (SAM) (Kirillov et al., 2023) demonstrates strong capability in producing high-quality region proposals across diverse visual domains. Despite this success, the idea of learning generalizable structural decomposition has not been sufficiently explored for point cloud data. Current point cloud foundation models mainly target semantic prediction tasks (Thengane et al., 2025), leaving structure-oriented modeling largely unaddressed.

To bridge this gap, we propose a SAM-guided framework for point cloud oversegmentation. The key idea is to transfer grouping knowledge from the image domain to 3D point clouds. We first project LiDAR data into bird's-eye-view (BEV) representations, where SAM is applied to generate dense segmentation masks. These 2D regions are subsequently lifted back to 3D space, forming a large collection of structure-consistent point groups that serve as supervision signals. Based on this dataset, we learn an oversegmentation model through a distillation objective that enforces consistency within each region while maintaining separation across different regions. Unlike conventional approaches that depend on semantic labels, the proposed method focuses purely on structural grouping, allowing the model to capture intrinsic geometric organization in the data. The learned model is able to produce reliable and consistent superpoints directly from raw point clouds, and generalizes well across datasets with different sensing modalities. By decoupling structure learning from semantic supervision, our approach offers a scalable solution for point cloud oversegmentation and provides a solid foundation for large-scale 3D scene understanding.

2. Related Work

Oversegmentation partitions point clouds into coherent regions that respect structural boundaries while remaining semantically agnostic. These regions serve as fundamental building blocks for various downstream tasks. Traditional methods (e.g., VCCS (Papon et al., 2013), SPG (Landrieu and Simonovsky, 2018)) rely only on geometric heuristics, requiring careful parameter tuning and often producing irregular segmentation on complex



(a) UAV LiDAR Scene (b) SAM-based Image Segmentation (c) Point Cloud Over-Segmentation

Figure 1. The pipeline of oversegmentation dataset construction.

structures. Learning-based methods have demonstrated improved consistency. SSP (Landrieu and Boussaha, 2019) proposed joint optimization of superpoint boundaries and semantic predictions. SPNet (Hui et al., 2021) integrated superpoint generation into end-to-end instance detection. More recently, SPCNet (Lu et al., 2025) introduced serialization-based oversegmentation with sequential region growing and learned affinity.

Despite these advances, existing oversegmentation methods still face several limitations. Most approaches are dataset-specific and require retraining or careful parameter tuning when applied to new domains. Learning-based methods are typically trained with semantic supervision, which couples structure learning with semantic labels and limits their use as general-purpose primitives. Moreover, no prior work has developed a model that produces consistent oversegmentation across different sensing platforms without adaptation.

Our work addresses these limitations by learning oversegmentation through structural and semantic knowledge distillation from SAM. As a result, it generalizes well across datasets and platforms. The resulting structure-consistent primitives further enable efficient annotation with substantially reduced human effort while maintaining high semantic quality.

3. Method

We focus on learning a structure-preserving oversegmentation model for large-scale LiDAR point clouds. The proposed framework consists of two key components: (1) construction of an oversegmentation dataset using SAM-derived grouping priors, and (2) a distillation-based model that learns structure-aware embeddings from these regions.

3.1 Oversegmentation Dataset Construction

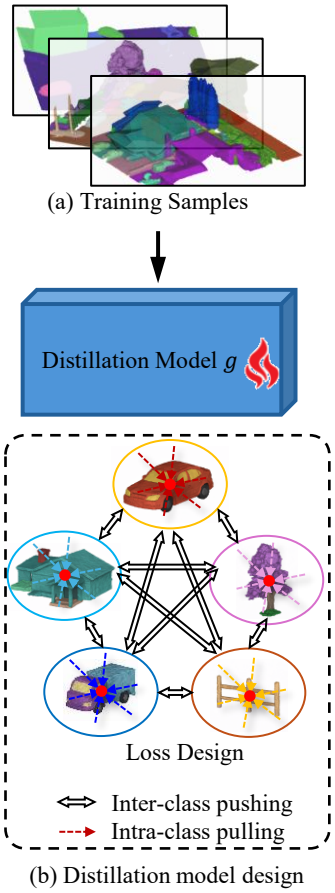
To obtain large-scale supervision for oversegmentation, we transfer grouping cues from 2D vision to 3D point clouds. Figure 1 shows the overall pipeline of oversegmentation dataset construction. Given a point cloud $\mathcal{P} = \{(\mathbf{x}_i, \mathbf{a}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^3$ denotes coordinates and $\mathbf{a}_i$ denotes point attributes, we first convert the 3D scene into a 2D representation.

Specifically, we partition the scene into local spatial blocks and project each block onto the bird’s-eye-view (BEV) plane using

$$\pi(x, y, z) = (x, y). \tag{1}$$

This projection preserves large-scale spatial layouts while enabling efficient processing in the image domain.

On the rendered BEV images, we apply the Segment Anything Model (SAM) to obtain dense region masks. These masks capture generic grouping structures but may not perfectly align with the underlying 3D geometry due to projection ambiguity.



(a) Training Samples

(b) Distillation model design

Figure 2. Distillation model and loss design.

To recover 3D consistency, each 2D region M_j is lifted back to the point cloud, forming a corresponding subset

$$\mathcal{S}_j = \{p_i \in \mathcal{P} \mid \pi(\mathbf{x}_i) \in M_j\}. \tag{2}$$

Since single-view projection may introduce mixed regions, we further refine these subsets in 3D space. When a subset contains multiple semantic categories, it is decomposed into smaller components according to ground-truth labels to remove inconsistencies. In addition, excessively small regions are merged with nearby subsets sharing the same semantic label to avoid fragmentation.

Through this process, we obtain a collection of fine-grained and structurally coherent regions $\{\mathcal{S}_j\}$. Although semantic labels are used during this refinement stage, they are only employed to improve region quality and are not used as supervision for model training. Consequently, the resulting dataset encodes structural grouping information rather than semantic categories, which is crucial for learning generalizable representations.

Table 1. Implementation details and hyperparameter settings.

Parameter	Description	Value
Optimizer	Optimizer for model training	Adam
Initial LR	Initial learning rate	0.001
LR Scheduler	Learning rate schedule	Cosine annealing
Training epochs	Number of training epochs	200
Embedding dimension d	Output feature dimension of f_θ	384
Margin m	Inter-segment separation margin	0.5
Loss weight λ	Weight for inter-segment loss	0.3

3.2 Structure-aware Distillation Model

Given the constructed oversegmentation dataset, our goal is to learn an encoder that produces structure-consistent partitions directly from raw point clouds. Figure 2 shows the pipeline of distillation model and loss design. Let $f_\theta(\cdot)$ denote a point cloud encoder (PointTransformerV3 (Wu et al., 2024) was used in our experiments) that maps each point p_i to a d-dimensional embedding

$$\mathbf{z}_i = f_\theta(p_i) \in \mathbb{R}^d. \tag{3}$$

Instead of treating $\{\mathcal{S}_j\}$ as semantic labels, we interpret them as grouping constraints in the embedding space. Points belonging to the same region are expected to be close to each other, while points from different regions should remain separable.

To enforce this property, we compute the centroid of each region

$$\boldsymbol{\mu}_j = \frac{1}{|\mathcal{S}_j|} \sum_{p_i \in \mathcal{S}_j} \mathbf{z}_i, \tag{4}$$

and define an intra-region objective that encourages compactness:

$$\mathcal{L}_{\text{intra}} = \sum_j \frac{1}{|\mathcal{S}_j|} \sum_{p_i \in \mathcal{S}_j} \left\| \mathbf{z}_i - \boldsymbol{\mu}_j \right\|_2^2. \tag{5}$$

To avoid degenerate solutions and promote discriminative partitions, we further introduce a separation term between different regions:

$$\mathcal{L}_{\text{inter}} = \frac{1}{J(J-1)} \sum_{j \neq k} \max\left(0,\, m - \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|_2\right), \tag{6}$$

where m is a margin controlling the minimum distance between region centers.

The final objective is given by

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{intra}} + \lambda \mathcal{L}_{\text{inter}}, \tag{7}$$

where λ balances compactness and separation. By optimizing this objective on large-scale LiDAR scenes, the model learns an embedding space that reflects structural grouping patterns. Importantly, the learned representation is independent of the 2D projection process and can be directly applied to raw point clouds at inference time. This enables consistent and generalizable oversegmentation across different datasets and sensing modalities.

4. Experiments

We evaluate the proposed method from the perspective of point cloud oversegmentation, focusing on the quality and generalization ability of the learned structural partitions.

4.1 Datasets

Experiments are conducted on various large-scale point cloud benchmarks with different acquisition settings, including STPLS3D (Chen et al., 2022), SensatUrban (Hu et al., 2022), Toronto-3D (Tan et al., 2020), DALES (Varney et al., 2020), and S3DIS (Armeni et al., 2016).

STPLS3D and SensatUrban are UAV-based datasets capturing large outdoor scenes from a top-down perspective. These datasets are used for training the oversegmentation model, from which more than 100K oversegmentation samples are constructed. Toronto-3D and DALES are collected using mobile and aerial LiDAR systems, respectively, representing street-level and large-area outdoor environments. S3DIS is an indoor RGB-D dataset, which is used to evaluate cross-domain generalization. We note the model is trained only on UAV LiDAR scenes (STPLS3D and SensatUrban) and directly evaluated on STPLS3D and other unseen test datasets without fine-tuning.

4.2 Evaluation Metrics

Following standard practice in point cloud oversegmentation (Lu et al., 2025), we adopt Oracle Overall Accuracy (OOA), Boundary Recall (BR), and Boundary Precision (BP) to assess segmentation quality.

OOA measures the upper-bound accuracy by assigning each superpoint the majority semantic label, reflecting region purity. BR evaluates the coverage of true semantic boundaries, while BP measures the precision of predicted boundaries. Together, these metrics provide a comprehensive evaluation of structural consistency and boundary alignment.

4.3 Implementation Details

The model is implemented in PyTorch and trained using the Adam optimizer with an initial learning rate of 0.001 and a cosine annealing schedule. Training is performed for 200 epochs on UAV LiDAR scenes. All hyperparameters shown in Table 1 are fixed across datasets without additional tuning.

4.4 Oversegmentation Results

We compare the proposed approach with representative oversegmentation methods, including both unsupervised (VCCS, SPG) and supervised (SSP, SPNet, SPCNet) approaches. For fair comparison, all methods are configured to produce approximately 1050 superpoints per scene.

As summarized in Table 2, the proposed method achieves competitive performance across all datasets. On STPLS3D, our model attains comparable OOA while significantly improving boundary recall, indicating that more true structural boundaries are successfully captured. This improvement can be attributed

Table 2. Comparison of point cloud oversegmentation performance (%) on different datasets.

Method	STPLS3D			Toronto-3D			DALES			S3DIS Area5		
	OOA	BR	BP	OOA	BR	BP	OOA	BR	BP	OOA	BR	BP
Unsupervised												
VCCS (Papon et al., 2013)	90.53	44.31	14.84	91.35	22.82	10.64	92.27	33.02	10.36	95.22	62.06	10.86
SPG (Landrieu and Simonovsky, 2018)	91.19	39.33	14.12	92.57	33.16	13.36	92.58	38.84	13.85	95.84	52.06	13.11
Supervised												
SSP (Landrieu and Boussaha, 2019)	94.07	77.49	12.37	92.37	70.19	10.25	94.38	75.12	12.66	**97.04**	80.73	13.02
SPNet (Hui et al., 2021)	94.36	79.44	12.53	93.45	76.68	13.17	93.45	77.64	12.91	96.50	84.75	13.14
SPCNet (Lu et al., 2025)	**96.71**	86.64	**22.61**	96.53	83.29	**24.10**	96.83	86.52	**28.74**	96.81	**89.50**	**26.78**
Ours	96.52	**92.21**	13.11	**97.62**	**93.47**	12.94	**97.10**	**89.64**	11.13	96.13	75.44	13.36

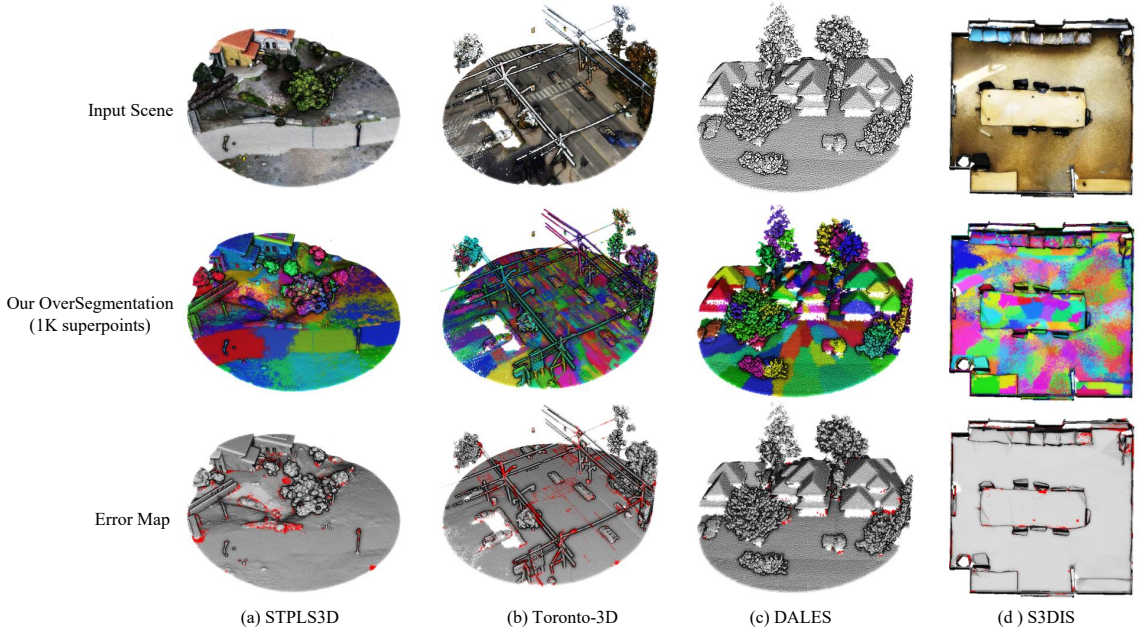


Figure 3. Oversegmentation results and corresponding error maps on different datasets (STPLS3D, Toronto-3D, DALES, and S3DIS).

to the learned structure-aware representations obtained through large-scale distillation.

More importantly, the proposed model generalizes well to unseen datasets. On Toronto-3D and DALES, which are not used during training, our method consistently achieves strong performance, demonstrating robustness across different sensing platforms and scene distributions. Even on S3DIS, which exhibits substantial domain differences from outdoor LiDAR data, the model still produces reasonable oversegmentation results. The visualization results are shown in Figure 3.

Compared with existing learning-based approaches that rely on semantic supervision from target datasets, our model is trained only once and directly applied to new domains without retraining. This highlights the advantage of decoupling structural learning from semantic labels. It is worth noting that the proposed method yields relatively lower boundary precision compared to some geometry-driven methods. This is because our model groups points based on learned feature embeddings rather than strictly enforcing local geometric continuity. As a result, spatially non-adjacent but structurally similar points may be assigned to the same superpoint, leading to less regular boundaries and reduced BP. Despite this, the higher BR and competitive OOA indicate that the overall structural consistency remains strong.

4.5 Ablation Study

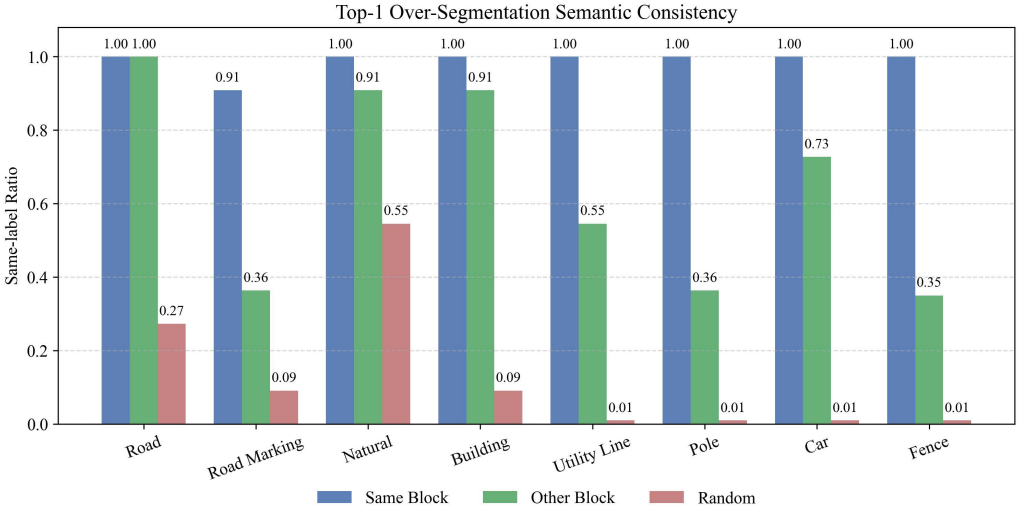
We further analyze key design components of the proposed oversegmentation model on the Toronto-3D dataset.

Effect of distillation objective. As shown in Table 3(a), removing the inter-region separation term leads to noticeable degradation in OOA and boundary metrics, indicating that enforcing separation between region embeddings is critical for preventing feature collapse and improving discriminability.

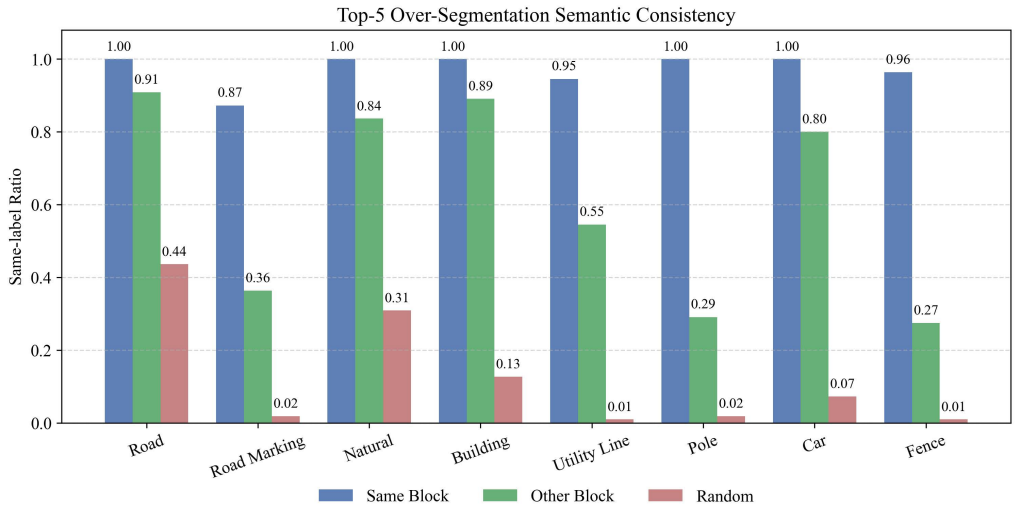
Backbone analysis. We evaluate different backbone networks, including SparseConv, KPConv, and PTV3, as shown in Table 3(b), the results show that the performance remains stable across architectures, suggesting that the learned structure-aware representation is largely independent of the backbone choice.

Feature consistency analysis. We evaluate the semantic consistency of learned embeddings by measuring k-nearest-neighbor (k equals 1, 5, 10, respectively) agreement in feature space. We consider three sampling settings: (1) neighbors selected within the same spatial block, (2) neighbors from different blocks, and (3) randomly sampled points from the entire scene. This design allows us to evaluate both local and global consistency of the learned representations.

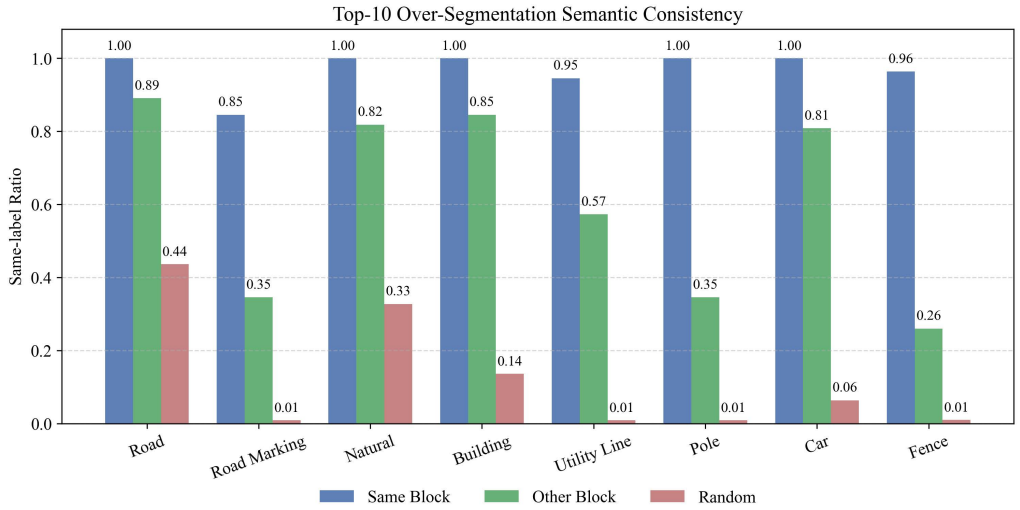
As shown in Figure 4, the results show that the proposed model



(a) Top-1 Semantic Consistency of Learned Over-Segmentation Features



(b) Top-5 Semantic Consistency of Learned Over-Segmentation Features



(c) Top-10 Semantic Consistency of Learned Over-Segmentation Features

Figure 4. Ablation study on semantic consistency of learned oversegmentation features. We evaluate the Top-1, Top-5, and Top-10 neighborhood consistency under different sampling settings, including same block, cross block, and random sampling. The results show that the learned features exhibit strong semantic consistency, which remains stable across spatial blocks.

Table 3. Ablation study (%) on the Toronto-3D test set.

Setting	OOA	BR	BP
(a) Distillation loss components			
w/o $\mathcal{L}_{\text{inter}}$	96.51	88.39	10.06
Full loss (ours)	**97.62**	**93.47**	**12.94**
(b) Backbone of oversegmentation model			
SparseConv (Graham et al., 2018)	96.47	92.06	**13.17**
KPConv (Thomas et al., 2019)	96.84	91.59	12.71
PTV3 (Wu et al., 2024)	**97.62**	**93.47**	12.94

achieves very high consistency in local neighborhoods. Under the same-block setting, the Top-1 and Top-5 consistency for most categories approach near-perfect values, indicating that the learned embeddings form highly compact clusters. More importantly, strong consistency is also observed across different spatial blocks, especially for large and structurally regular categories such as buildings and vegetation. This suggests that the learned representation is not limited to local geometric cues, but captures more global structural patterns that generalize across spatial regions. In contrast, when neighbors are randomly sampled from the entire scene, the consistency drops significantly. This confirms that the observed high consistency is not a trivial consequence of class imbalance, but reflects meaningful organization in the embedding space.

5. Conclusion

In this work, we propose a SAM-guided framework for large-scale LiDAR point cloud oversegmentation. By transferring grouping priors from 2D vision to 3D point clouds, we construct a large-scale oversegmentation dataset and learn a structure-aware encoder via a distillation objective. The proposed method focuses on structural grouping rather than semantic supervision, enabling the model to capture intrinsic geometric organization in point clouds. Extensive experiments on multiple benchmark datasets demonstrate that the proposed approach achieves competitive oversegmentation performance while significantly improving boundary recall. More importantly, the model exhibits strong generalization ability across different datasets and sensing platforms, even without retraining, highlighting the effectiveness of the learned structure-aware representations.

Nevertheless, the current oversegmentation dataset is primarily constructed from UAV-based outdoor LiDAR scenes. As a result, the performance on indoor datasets such as S3DIS is relatively limited due to the domain gap in structure, density, and acquisition patterns. This indicates that the current model does not fully capture the diversity of indoor spatial structures. In future work, we will extend the dataset construction pipeline to indoor point clouds and develop more comprehensive oversegmentation supervision across diverse environments. By incorporating both outdoor and indoor data, we aim to build a unified and generalizable point cloud oversegmentation foundation model that can adapt to a wide range of real-world scenarios.

References

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 1534–1543.

Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., McCullough, K., Ren, F., Soibelman, L., 2022. STPLS3D: A Large-Scale Synthetic and Real Aerial Photogrammetry 3D Point Cloud Dataset. *arXiv preprint arXiv:2203.09065*.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 9224–9232.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2022. SensatUrban: Learning Semantics from Urban-Scale Photogrammetric Point Clouds. *Int. J. Comput. Vis.*, 130(2), 316–343.

Hui, L., Tang, L., Dai, Y., Xie, J., Yang, J., 2023. Efficient LiDAR Point Cloud Oversegmentation Network. *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 17957–17966.

Hui, L., Yuan, J., Cheng, M., Xie, J., Zhang, X., Yang, J., 2021. Superpoint Network for Point Cloud Oversegmentation. *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 5490–5499.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment Anything. *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 4015–4026.

Landrieu, L., Boussaha, M., 2019. Point Cloud Oversegmentation With Graph-Structured Deep Metric Learning. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 7432–7441.

Landrieu, L., Simonovsky, M., 2018. Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 4558–4567.

Lu, C., Kwan, J., Li, D., Chen, Z., Guan, H., 2025. Serialization based Point Cloud Oversegmentation. *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 25831–25840.

Meng, Q., Wang, W., Zhou, T., Shen, J., Jia, Y., Van Gool, L., 2021. Towards a Weakly Supervised Framework for 3D Point Cloud Object Detection and Annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8), 4454–4468.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2023. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*.

Papon, J., Abramov, A., Schoeler, M., Worgotter, F., 2013. Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2027–2034.

Robert, D., Raguet, H., Landrieu, L., 2023. Efficient 3D Semantic Segmentation with Superpoint Transformer. *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 17149–17158.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K., Li, J., 2020. Toronto-3D: A Large-scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 797–806.

Thengane, V., Zhu, X., Bouzerdoum, S., Phung, S. L., Li, Y., 2025. Foundational Models for 3D Point Clouds: A Survey and Outlook. *arXiv preprint arXiv:2501.18594*.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. KPConv: Flexible and Deformable Convolution for Point Clouds. *Proc. IEEE Int. Conf. Comput. Vis.*, 6410–6419.

Varney, N., Asari, V. K., Graehling, Q., 2020. DALES: A Large-scale Aerial LiDAR Data Set for Semantic Segmentation. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 717–726.

Wu, X., Jiang, L., Wang, P., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H., 2024. Point Transformer V3: Simpler, Faster, Stronger. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4840–4851.

Xiao, A., Zhang, X., Shao, L., Lu, S., 2024. A Survey of Label-Efficient Deep Learning for 3D Point Clouds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12), 9139-9160.