GeoOpen3D: Geometry-guided training-free open-vocabulary 3D segmentation via visual foundation models

Shuai Zhang[1], Zhuoxiao Li[1], Jing Ou[1], Tengxi Wang[1], Zhecheng Shi[1,2], Wufan Zhao[1*]

[1]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511442, China
[2]School of Computer and Communication Engineering, Northeastern University, Qinhuangdao 066004, China
{szhang240, jou719, twang744}@connect.hkust-gz.edu.cn
{zhuoxiaoli, zhechengs, wufanzhao}@hkust-gz.edu.cn

Keywords: 3D Point Cloud Segmentation; Open-Vocabulary Segmentation; Training-Free Framework; Vision–Language Models; Urban Scene Analysis

Abstract

Open-vocabulary 3D segmentation offers an attractive alternative to closed-set scene parsing, yet directly transferring 2D vision-language models to outdoor point clouds remains difficult because projection disrupts geometric continuity and sparse sampling weakens mask quality. This paper presents GeoOpen3D, a geometry-guided and training-free framework for open-vocabulary 3D point cloud segmentation. GeoOpen3D constructs a geometry-preserving RGB-D representation through projection, super-sampling, and depth enhancement to improve alignment between 3D structure and 2D foundation models. It then combines GroundingDINO for language-driven proposal generation with SAM for mask extraction, while introducing depth-aware regularization to favor structurally coherent regions and clearer boundaries. The selected masks are back-projected to the original point cloud through pixel-to-point correspondence, yielding point-wise semantic labels without any 3D model training. Experiments on the SensatUrban dataset show that GeoOpen3D achieves 42.1% mIoU, including 98.5% IoU for buildings and 97.3% IoU for vegetation, outperforming existing training-free open-vocabulary baselines. Additional experiments on a custom island dataset further demonstrate promising transferability to unseen categories. These results indicate that geometry-guided 2D-to-3D transfer provides an effective and scalable path towards open-vocabulary understanding of large-scale outdoor scenes.

1. Introduction

Large-scale outdoor point clouds play a fundamental role in urban mapping (Wang et al., 2018), autonomous systems (Ma et al., 2025), and digital twin construction (Zhang et al., 2024). Despite their importance, semantic segmentation in such environments remains a challenging problem. Outdoor point clouds are characterized by irregular sampling, significant scale variation, and heterogeneous structural patterns, which complicate reliable point-wise classification (Weinmann et al., 2015; Hu et al., 2020a). Although supervised 3D segmentation methods have achieved substantial progress, they rely heavily on dense point-wise annotations, predefined semantic taxonomies, and dataset-specific training pipelines. These assumptions limit their applicability in real-world geospatial scenarios, where semantic categories evolve dynamically and large-scale annotation is both costly and impractical.

To overcome these limitations, open-vocabulary learning has emerged as a promising direction by replacing fixed label spaces with language-guided recognition. With the advent of vision-language pre-training, models such as CLIP (Radford et al., 2021) have demonstrated the ability to align visual features with rich semantic concepts described by natural language. In parallel, open-vocabulary 2D segmentation methods have shown that text supervision can support dense prediction beyond predefined taxonomies (Ghiasi et al., 2021; Xu et al., 2022). These advances have stimulated increasing interest in open-vocabulary 3D scene understanding, where semantic labels are inferred directly from language prompts rather than restricted to a closed training set. However, most existing approaches still depend on 3D feature distillation (Wu et al., 2025), additional adaptation modules (Thomas et al., 2019), or large-scale annotated training data (Wang et al., 2020). Such dependencies hinder scalability and reduce flexibility, particularly in large-scale outdoor environments.

Recent developments in visual foundation models provide a new opportunity for training-free semantic segmentation. GroundingDINO (Liu et al., 2023) enables text-driven object localization, while SAM (Kirillov et al., 2023) produces high-quality segmentation masks with strong zero-shot generalization capability. A straightforward solution is to project 3D point clouds into 2D views, perform segmentation using these models, and then map the results back to 3D space (Yang et al., 2023). However, this projection-based paradigm introduces a critical limitation. In outdoor scenes, sparse sampling leads to holes and fragmented observations, while depth discontinuities and occlusions are often blurred in the image plane. As a result, appearance-driven 2D masks frequently fail to capture the underlying geometric structure, leading to incorrect merging of adjacent objects and degraded boundary quality after back-projection.

These observations indicate that the core challenge lies not in the availability of strong 2D semantic priors, but in preserving geometric consistency during the 2D-to-3D transfer process. Once the projection step disrupts spatial structure, even powerful foundation models cannot produce stable and coherent 3D predictions. This issue is particularly pronounced in large-scale outdoor environments, where uneven point density, long-range depth variation, and occlusion effects jointly degrade the reliability of 2D predictions and their subsequent lifting. Therefore, a practical training-free framework must explicitly integrate ge-

* Corresponding author

ometric constraints into the segmentation pipeline, rather than treating projection as a purely visual pre-processing operation.

To address this problem, we propose **GeoOpen3D**, a geometry-guided, training-free framework for open-vocabulary semantic segmentation of large-scale outdoor point clouds. The framework constructs a geometry-preserving RGB–D representation through projection refinement, super-sampling, and depth enhancement, and integrates language-driven region proposal with geometry-aware mask selection. By enforcing depth-aware consistency and establishing explicit pixel-to-point correspondence, GeoOpen3D enables coherent 2D-to-3D semantic transfer without any 3D training.

The main contributions of this paper are summarized as follows:

1. A geometry-preserving RGB–D representation that integrates projection, super-sampling, and depth enhancement to improve cross-modal alignment between 2D observations and 3D structures.

2. A training-free open-vocabulary segmentation pipeline that combines language-driven proposal generation using GroundingDINO and mask prediction using SAM, with explicit lifting to 3D space via pixel-to-point correspondence.

3. A depth-aware constraint mechanism that introduces implicit geometric regularization, improving boundary coherence and structural consistency in large-scale outdoor scenes.

4. A unified geometry-guided framework that reformulates 2D–3D semantic transfer as a consistency-driven process, enabling robust and scalable segmentation without any 3D training.

2. Related Work

2.1 Supervised 3D Segmentation

Supervised 3D semantic segmentation has progressed rapidly from early point-based networks to highly optimized sparse convolutional and transformer-based architectures. PointNet (Qi et al., 2017a) first demonstrated that unordered point sets can be processed directly through shared multilayer perceptrons and symmetric aggregation, while PointNet++ (Qi et al., 2017b) further introduced hierarchical neighborhood abstraction for modeling local geometry. Graph-based and hybrid representations then improved local relational modeling and multi-scale feature learning, as illustrated by Dynamic Graph CNN and Point-Voxel Transformer (Wang et al., 2020; Zhang et al., 2021). In parallel, voxel-based and sparse-convolution methods improved scalability by exploiting regular computation on sparse 3D structures, and KPConv (Thomas et al., 2019) achieved strong performance through continuous kernel point convolution in Euclidean space. More recent models such as Point Transformer, PointNeXt, and Point Transformer V3 further strengthened representation learning through attention-based local aggregation and improved scaling behavior (Zhao et al., 2021; Qian et al., 2022; Wu et al., 2024).

Although these methods report strong closed-set accuracy, their dependence on fully annotated datasets remains a major obstacle for large outdoor scenes. Point-wise labelling of urban point clouds is costly and slow, particularly when scenes contain billions of points and long-tail object categories. Recent self-supervised point representation learning has reduced this dependence to some extent (Wu et al., 2025), but the final prediction space in most pipelines is still fixed at training time. In addition, supervised models are typically tied to a fixed taxonomy defined at training time, making it difficult to accommodate unseen or user-specified concepts after deployment. These limitations motivate the search for more flexible open-vocabulary solutions.

2.2 Vision-Language Models for Open-Vocabulary Understanding

Vision-language pre-training has fundamentally changed the way open-vocabulary recognition is formulated. CLIP (Radford et al., 2021) showed that large-scale image-text supervision can align visual and linguistic semantics in a shared embedding space, enabling zero-shot recognition far beyond a predefined category list. Strong self-supervised image representations, such as those learned by DINO-style transformers, have further reinforced the transferability of visual encoders for open-world perception (Caron et al., 2021). Building on this idea, subsequent work explored dense open-vocabulary prediction using image-level or text supervision, showing that segmentation can emerge without conventional category-specific mask training (Ghiasi et al., 2021; Xu et al., 2022). Recent foundation models have further decomposed the problem into localization and segmentation. GroundingDINO (Liu et al., 2023) performs text-conditioned detection, whereas SAM (Kirillov et al., 2023) provides promptable mask generation with strong cross-domain generalization.

These models are especially attractive for geospatial applications because they can be used off the shelf, avoiding task-specific retraining. Nevertheless, they are fundamentally designed for 2D imagery. When directly applied to projected point clouds, their predictions can become unstable because image evidence alone does not fully encode depth continuity, point density variation, or the topology of 3D structures. As a result, the challenge in outdoor 3D segmentation is not simply whether 2D foundation models can recognize a category, but whether their masks remain geometrically meaningful after projection and back-projection.

2.3 Open-Vocabulary 3D Segmentation

Open-vocabulary 3D segmentation seeks to transfer the semantic richness of vision-language models into point cloud understanding. A representative line of work is based on cross-modal distillation. OpenScene (Peng et al., 2023), for example, projects multi-view 2D features into 3D and learns point representations aligned with CLIP semantics. Such methods establish strong open-vocabulary baselines, but they still require dedicated 3D training and often depend on heavy feature extraction and fusion pipelines. This makes them less attractive for lightweight deployment in large-scale outdoor mapping scenarios, where both scene size and vocabulary can change frequently after data acquisition.

Another line of research focuses on proposal-based reasoning. Methods such as OpenMask3D (Takmaz et al., 2023) and OpenIns3D (Huang et al., 2024) generate region proposals, associate them with language features, and then infer 3D masks from these candidate regions. While proposal-based methods can improve object-level reasoning, they typically

rely on trained 3D backbones, differentiable rendering, or instance-level supervision. Their performance may also degrade in outdoor scenes where point density, scale variation, and occlusion differ substantially from indoor benchmarks.

More recently, training-free pipelines have attempted to bypass 3D learning entirely by projecting point clouds into 2D, invoking foundation models directly, and mapping the predictions back to 3D (Yang et al., 2023). This strategy is attractive because it avoids the annotation bottleneck and simplifies deployment. However, without explicit geometric modeling, these methods often suffer from fragmented masks, blurred boundaries, and inconsistent region assignment after back-projection. GeoOpen3D follows the same training-free philosophy but strengthens the transfer process through geometry-preserving representation construction and depth-aware mask selection.

2.4 Geometry-Aware 3D Learning

Geometry is a fundamental cue in 3D scene understanding, as it provides information that cannot be recovered from appearance alone. In point cloud segmentation, depth variation, surface continuity, and spatial adjacency are essential for distinguishing neighboring structures with similar color or texture. This is particularly important in outdoor environments, where large depth ranges, irregular sampling, and complex object boundaries often make purely appearance-based reasoning unreliable.

To exploit such information, existing geometry-aware methods typically incorporate geometric priors into learned 3D representations. For example, RandLA-Net (Hu et al., 2020b) improves large-scale point cloud processing through random sampling and local feature aggregation while preserving neighborhood structure. Similarly, transformer-based point encoders explicitly model local geometric relationships to strengthen feature learning and contextual reasoning (Zhao et al., 2021; Wu et al., 2024). Other studies have explored boundary-aware filtering, multi-view consistency, or explicit geometric constraints to improve structural coherence in segmentation results.

However, most of these methods rely on dedicated 3D network training to encode geometric reasoning, and are therefore tied to fixed label spaces and dataset-specific supervision. Such designs are less suitable for training-free open-vocabulary segmentation, where the challenge is not only to exploit geometry, but to preserve geometric consistency while transferring semantic predictions from 2D foundation models back to 3D space. This limitation highlights the need for a geometry-aware framework that can support open-vocabulary 3D understanding without introducing an additional trained 3D backbone.

3. Methodology

3.1 Overview

Given a large-scale outdoor point cloud P with N points, where each point is represented as $(x_i, y_i, z_i, R_i, G_i, B_i)$, the goal of open-vocabulary semantic segmentation is to assign each point $p_i \in P$ a semantic label $L(p_i)$ inferred from a set of language queries $T = \{t_1, t_2, \ldots, t_M\}$, without relying on any predefined category set or 3D training. A straightforward solution is to project the 3D point cloud into 2D views, perform language-guided segmentation using vision-language models, and map the predictions back to 3D space. However, this paradigm often breaks geometric consistency during projection. In large-scale outdoor scenes, sparse sampling, occlusion, and depth discontinuities lead to incomplete observations and ambiguous boundaries in the image domain, making appearance-driven masks difficult to transfer reliably to 3D.

To address this issue, we propose GeoOpen3D, a geometry-guided framework that reformulates open-vocabulary segmentation as a geometry-constrained 2D–3D transfer problem (Figure 1). The framework consists of **four stages**: (1) geometry-preserving RGB–D representation construction, (2) language-driven proposal generation with GroundingDINO, (3) depth-aware mask selection based on SAM candidates, and (4) explicit back-projection to point-wise semantic labels. The key idea is to maintain geometric consistency throughout the transfer process, ensuring that semantically meaningful regions in 2D remain structurally coherent in 3D.

3.2 Geometry-Preserving RGB-D Representation

Given a 3D point cloud P with N points, we construct a geometry-preserving RGB-D representation through projection and enhancement. The goal of this stage is not merely to obtain a visual rendering, but to retain geometric structure during the 2D transformation so that subsequent segmentation remains consistent with the underlying 3D scene.

Direct projection of sparse outdoor point clouds often leads to incomplete and irregular observations in the image plane. Due to uneven sampling density, occlusions, and depth discontinuities, the projected RGB images typically contain holes, fragmented regions, and blurred object boundaries. Such artifacts degrade the reliability of appearance-based segmentation models and introduce ambiguity when lifting predictions back to 3D. Therefore, it is necessary to explicitly enhance both the spatial completeness and the geometric fidelity of the projected representation.

3.2.1 Projection and Super-Sampling The 3D coordinates and color attributes are projected into the image plane using the camera intrinsic matrix K and extrinsic matrix T:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = KT \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} \tag{1}$$

where (u_i, v_i) are the image coordinates of point i. Because outdoor point clouds often produce incomplete and uneven image coverage after projection, we apply a local super-sampling strategy to densify both depth and color observations:

$$D(u,v) = \frac{1}{|N(u,v)|} \sum_{(u',v') \in N(u,v)} z(u',v') \tag{2}$$

$$C(u,v) = \frac{1}{|N(u,v)|} \sum_{(u',v') \in N(u,v)} c(u',v') \tag{3}$$

where $N(u,v)$ denotes the local neighborhood around (u,v) and $|N(u,v)|$ is the number of valid projected points inside that neighborhood. This aggregation step alleviates sparsity-induced holes and produces a more continuous representation, which stabilizes subsequent proposal generation and mask extraction.

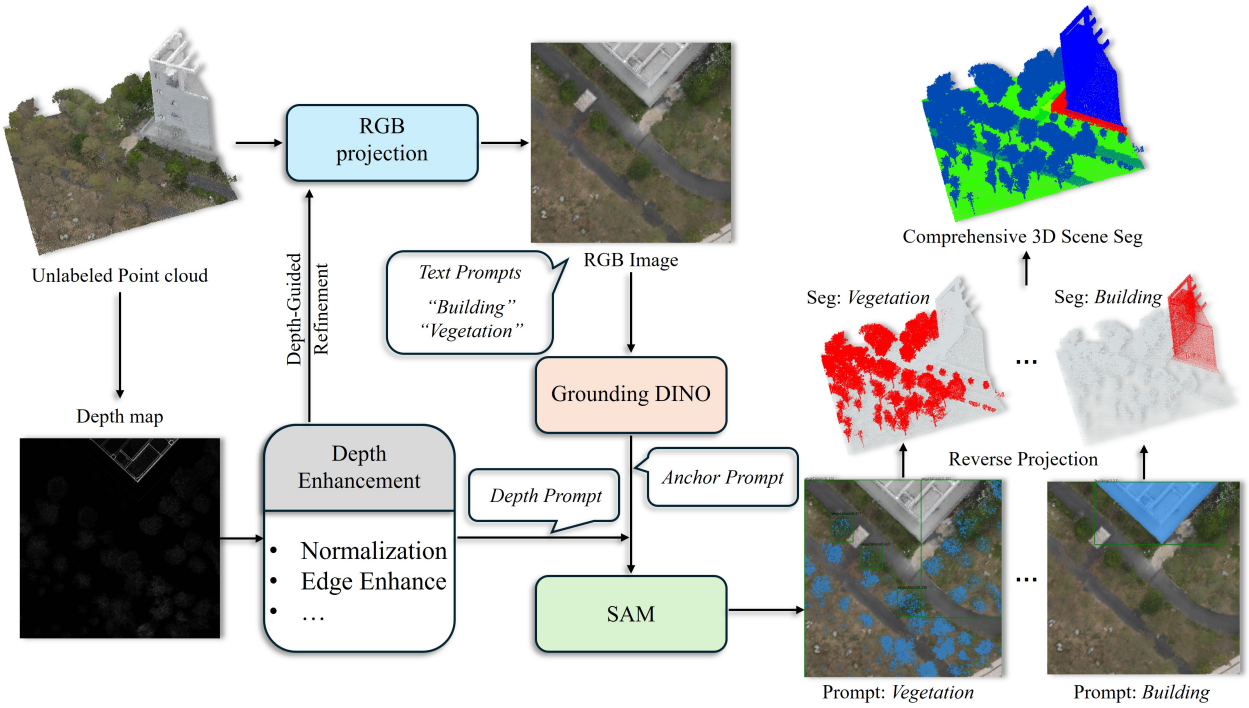

Figure 1. The workflow of the proposed geometry-guided training-free 3D segmentation framework.

3.2.2 Depth Enhancement While super-sampling improves completeness, it may smooth out local geometric variations. To preserve structural details, especially at object boundaries, we further refine the depth map using an edge-aware filtering strategy. This step aims to maintain depth discontinuities that correspond to real geometric separations in the scene, which are critical for distinguishing adjacent objects.

$$D_{\text{enhanced}}(u,v) = \sum_{(u',v')\in W} w(u',v') \cdot D(u',v') \tag{4}$$

where W is the neighborhood window, σ_c and σ_s control the color and spatial similarity, and $w(u',v')$ is the bilateral filter weight defined as:

$$w(u',v') = \exp\Big(- \frac{\|C(u,v) - C(u',v')\|^2}{2\sigma_c^2} - \frac{\|(u,v) - (u',v')\|^2}{2\sigma_s^2} \Big) \tag{5}$$

By jointly considering spatial proximity and appearance similarity, this filtering process preserves sharp depth transitions while suppressing noise in homogeneous regions. As a result, the enhanced depth map provides a more faithful representation of scene geometry.

The final RGB-D representation $I_{\text{RGB-D}} = [C, D_{\text{enhanced}}]$ integrates both appearance and geometric cues, serving as a reliable input for foundation models. Unlike plain RGB projections, this representation explicitly encodes structural information, enabling more stable and geometry-consistent semantic inference in subsequent stages.

3.3 Language-Driven Proposal Generation

Given a set of text prompts $T = \{t_1, t_2, \ldots, t_M\}$ describing candidate semantic categories, the objective of this stage is to localize semantically relevant regions in the projected image that correspond to the queried concepts. Rather than directly predicting pixel-wise labels, we first identify coarse region proposals that serve as anchors for subsequent mask refinement. This design decouples semantic grounding from fine-grained segmentation and improves robustness under noisy or incomplete observations.

In large-scale outdoor scenes, directly performing dense segmentation based on language can be unreliable due to cluttered backgrounds and ambiguous visual cues. Therefore, we formulate proposal generation as a language-conditioned region retrieval problem, where each query is used to retrieve a set of candidate regions that are semantically aligned but not yet geometrically refined.

To this end, we employ GroundingDINO (Liu et al., 2023), which aligns visual features with language embeddings and enables text-driven object localization. For each prompt t_j, GroundingDINO produces a set of bounding boxes $B_j = \{b_{j,k}\}$ with associated confidence scores $S_j = \{s_{j,k}\}$. These proposals provide a coarse but semantically meaningful partition of the scene, bridging the gap between textual descriptions and visual regions.

To suppress low-quality detections and reduce ambiguity, we retain only proposals whose confidence exceeds a threshold τ:

$$B_{\text{filtered}} = \{b_{j,k} \mid s_{j,k} > \tau, \forall j,k\} \tag{6}$$

The filtered proposals define a set of candidate regions that are likely to contain objects consistent with the input language

queries. These regions are then passed to the subsequent stage for geometry-aware mask prediction, where finer structural constraints are introduced to resolve ambiguities and improve spatial coherence.

3.4 Depth-Aware Mask Prediction

For each filtered bounding box $b \in B_{\text{filtered}}$, SAM (Kirillov et al., 2023) generates a set of candidate masks. While SAM provides strong appearance-based segmentation, its predictions are not explicitly constrained by scene geometry, which may lead to inconsistent regions in the presence of depth variation or object adjacency. Therefore, instead of directly adopting the default mask, we introduce a depth-aware selection mechanism that evaluates candidates based on geometric consistency.

Given a candidate mask M, we first measure depth consistency inside the mask region:

$$\mathcal{L}_{\text{depth}} = \frac{1}{|M|} \sum_{(u,v)\in M} \left| D(u,v) - \bar{D}_M \right| \tag{7}$$

where $\bar{D}_M$ is the mean depth within mask M and $|M|$ is the number of pixels in the mask. Lower values indicate that the region is geometrically more homogeneous.

We additionally penalize masks whose boundaries cross strong depth discontinuities:

$$\mathcal{L}_{\text{boundary}} = \sum_{(u,v)\in \partial M} \|\nabla D(u,v)\| \tag{8}$$

where ∂M denotes the boundary of mask M and $\nabla D(u,v)$ is the depth gradient. This term encourages mask boundaries to align with true geometric edges in the scene.

The final mask is selected by minimising the combined energy:

$$M^* = \arg\min_M \left(\mathcal{E}_{\text{sam}} + \lambda_1 \mathcal{L}_{\text{depth}} + \lambda_2 \mathcal{L}_{\text{boundary}} \right) \tag{9}$$

where $\mathcal{E}_{\text{sam}}$ denotes the cost derived from SAM's mask confidence, and λ_1, λ_2 are weighting factors. This formulation integrates semantic confidence with geometric regularization, favoring masks that are both visually plausible and structurally consistent.

3.5 Back-Projection to 3D Space

After selecting geometrically consistent masks, we map the 2D predictions back to the original 3D point cloud. This step restores semantic information in the spatial domain and completes the 2D–3D transfer process.

The back-projection is achieved through the explicit pixel-to-point correspondence established during projection. For each point p_i with image coordinates (u_i, v_i), we assign the semantic label of the mask covering that position:

$$L(p_i) = \text{label}(M^*(u_i, v_i)) \tag{10}$$

where $L(p_i)$ is the semantic label of point p_i. Because this correspondence is explicitly defined, the recovered 3D labels remain consistent with both the image-space predictions and the underlying geometry.

By combining geometry-aware mask selection with direct pixel-to-point mapping, the proposed framework ensures that semantic regions inferred in 2D are faithfully transferred to 3D space without introducing additional inconsistencies.

4. Experiments

4.1 Datasets

4.1.1 SensatUrban Dataset SensatUrban (Hu et al., 2020a) is a large-scale urban dataset consisting of photogrammetric point clouds acquired via unmanned aerial vehicles across multiple UK cities, covering approximately 7.6 km^2. It contains nearly three billion densely annotated points with RGB information and includes thirteen semantic classes, spanning both man-made structures (e.g., buildings, bridges, roads) and natural elements (e.g., vegetation and water). The dataset is characterized by large spatial extent, high density variation, and complex urban layouts, making it particularly suitable for evaluating open-vocabulary segmentation in realistic outdoor environments.

4.1.2 Island Dataset To further assess generalization ability, we evaluate GeoOpen3D on a custom island dataset containing diverse scene elements such as beaches, cliffs, squares, vegetation, and buildings. Compared to SensatUrban, this dataset introduces unseen categories and different spatial configurations, providing a more challenging test for open-vocabulary transfer without retraining. This setting allows us to evaluate the flexibility of the proposed framework beyond fixed benchmark taxonomies.

4.2 Evaluation Protocol

4.2.1 Evaluation Metrics We adopt standard metrics for 3D semantic segmentation, including Intersection over Union (IoU), Mean IoU (mIoU), Overall Accuracy (OA), and Boundary F1-score. IoU measures the overlap between predicted and ground-truth regions for each class, while mIoU provides an overall assessment across all categories. OA reflects the proportion of correctly classified points over the entire scene.

In addition to region-based metrics, Boundary F1-score is used to explicitly evaluate the quality of object boundaries. This metric is particularly important in our setting, as the proposed depth-aware constraints are designed to improve boundary consistency during mask selection. Therefore, Boundary F1-score complements mIoU by capturing improvements that may not be reflected in region-level overlap alone.

4.2.2 Baselines We compare GeoOpen3D with three categories of methods. First, fully-supervised approaches, including PointNet++, RandLA-Net, and KPConv, are used as upper-bound references for segmentation performance. Second, training-required open-vocabulary methods, such as OpenScene, PLA, RegionPLC, OpenMask3D, OpenIns3D, OpenYOLO3D, and OpenUrban3D, are included to evaluate the effectiveness of open-vocabulary learning strategies with supervision or distillation.

Finally, we construct a training-free baseline by removing depth enhancement and depth-aware constraints from our framework. This variant isolates the contribution of geometric regularization and allows us to directly quantify the impact of the proposed geometry-aware design.

4.3 Implementation Details

GeoOpen3D is implemented in Python using PyTorch, and directly leverages publicly available pre-trained models, including GroundingDINO and SAM, without any task-specific fine-tuning. For each scene, projection parameters are adaptively determined based on the spatial extent of the point cloud to ensure stable and consistent RGB-D rendering across varying scene scales. Unless otherwise specified, the proposal confidence threshold is set to $\tau = 0.3$, while the geometry-aware terms are weighted by $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$. All parameters are kept fixed across experiments to maintain a consistent evaluation setting and to highlight the training-free nature of the proposed framework.

4.4 Results on SensatUrban Dataset

Table 1 reports the quantitative comparison on the SensatUrban validation split. GeoOpen3D achieves 42.1% mIoU in a fully training-free setting. These results, obtained without any task-specific optimization, demonstrate that our method significantly outperforms existing training-free open-vocabulary baselines, as well as training-required approaches such as OpenUrban3D (39.6% mIoU).

Specifically, our method demonstrates exceptional performance on large-scale structural classes. It achieves 98.5% IoU for Building and 97.3% IoU for Vegetation, highlighting the effectiveness of the geometry-preserving RGB-D representation in recognizing large, homogeneous regions commonly found in urban environments. Moreover, our approach shows substantial improvements in identifying fine-grained and geometrically complex categories. For example, in classes such as Rail (15.5%), Street Furniture (14.9%), and Bridge (58.2%), where previous zero-shot methods fail to achieve meaningful results (0.0–2.0% IoU), our geometry-guided mask selection strategy effectively preserves critical semantic features and geometric coherence, enabling accurate segmentation and recognition.

The category-wise results reveal two main strengths. First, dominant urban structures benefit from stable alignment between language prompts, clear visual evidence, and large geometrically coherent regions. Second, the proposed geometry-guided strategy improves recognition of difficult classes that are often suppressed by naive projection, including thin structures such as rail and street furniture as well as elevated objects like bridge. Although the absolute IoU of these fine-grained categories remains modest compared to fully-supervised methods, the substantial gains over other training-free baselines indicate that explicit geometric regularization helps preserve structural details and boundary integrity.

Overall, the results support the main claim of this paper: competitive open-vocabulary outdoor segmentation can be achieved without training a dedicated 3D network, provided that geometry is explicitly modeled during projection and mask selection.

The qualitative results in Figures 2 and 3 are consistent with the quantitative findings. The simple-scene example highlights clean boundaries over large homogeneous regions, while the

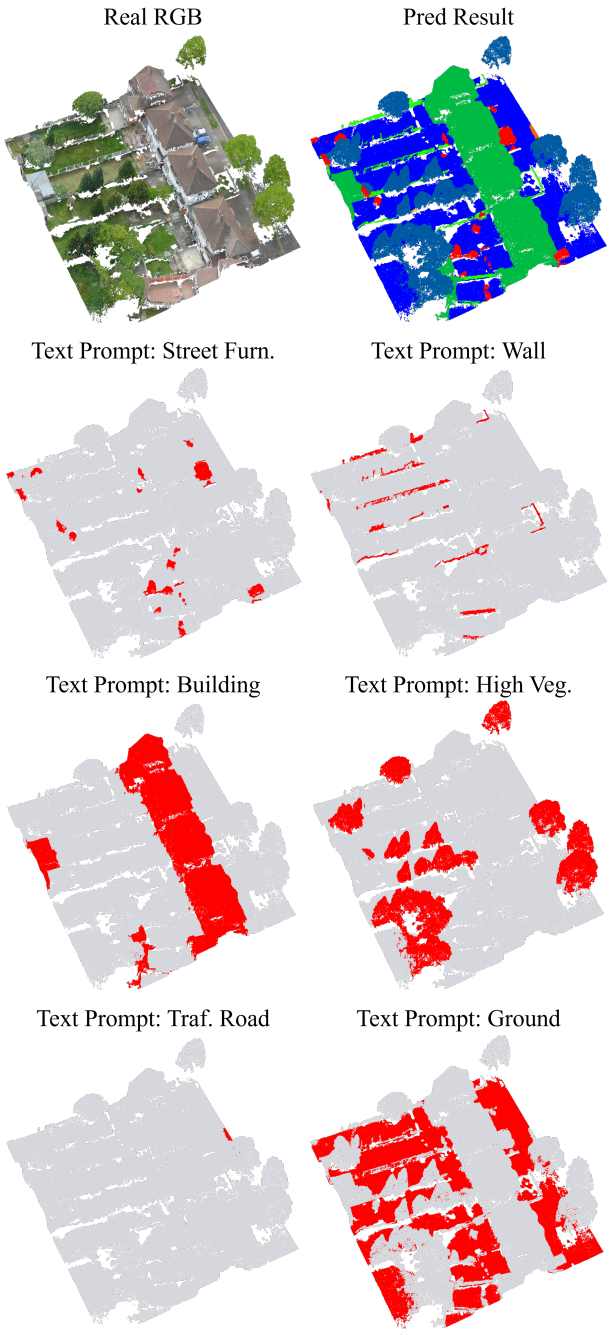

Figure 2. Simple scene result on SensatUrban. GeoOpen3D preserves clear boundaries over large homogeneous regions.

complex-scene example shows that the method remains stable in cluttered areas with multiple object categories. These observations further support the effectiveness of the geometry-preserving RGB-D construction and depth-aware mask selection.

4.5 Open-Vocabulary Evaluation

To further evaluate open-vocabulary capability, we test GeoOpen3D on the custom island dataset. We use prompts for visually familiar categories (for example, building and tree) and unseen categories such as beach, square, and cliff. As shown in Table 2, the method maintains strong performance on seen categories. It also achieves encouraging precision and recall on unseen ones. This result suggests that the proposed

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-1335-2026

Table 1. Semantic segmentation results on the **SensatUrban** validation dataset. Note: OpenUrban3D and OpenYOLO3D rely on trained 3D backbones (Distillation or Mask3D), whereas our method is strictly training-free. Numbers for some baselines are taken from their original papers.

Method	Type	mIoU (%)	OA (%)	mAcc (%)	Gro. (%)	Veg. (%)	Bui. (%)	Wal. (%)	Bri. (%)	Par. (%)	Rai. (%)	Tra. (%)	Str. (%)	Car (%)	Foo. (%)	Bik. (%)	Wat. (%)
Fully-supervised Methods																	
PointNet	Supervised	23.7	80.7	-	67.9	89.5	80.0	0.0	0.0	3.9	0.0	31.5	0.0	35.1	0.0	0.0	0.0
PointNet++	Supervised	32.9	84.3	-	72.4	94.2	84.7	2.7	2.0	25.7	0.0	31.5	11.4	38.8	7.1	0.0	56.9
SPGraph	Supervised	37.2	76.9	-	69.9	94.5	88.8	32.8	12.5	15.7	15.4	30.6	22.9	56.4	0.5	0.0	44.2
SparseConv	Supervised	42.6	85.2	-	74.1	97.9	94.2	63.3	7.5	24.2	0.0	30.1	34.0	74.4	0.0	0.0	54.8
MinkUNet	Supervised	46.2	90.8	**53.3**	81.5	97.7	94.8	59.6	0.1	30.7	0.0	55.0	33.1	77.0	22.9	0.0	49.1
RandLA-Net	Supervised	52.7	89.8	-	80.1	98.1	91.6	48.9	**40.8**	**51.6**	0.0	**56.7**	33.3	80.1	32.6	0.0	71.3
KPConv	Supervised	**57.6**	**93.2**	-	**87.1**	**98.3**	**95.3**	**74.4**	28.7	41.4	0.0	56.0	**54.4**	**85.7**	**40.4**	0.0	**86.3**
Training-Required Open-Vocabulary Methods																	
OpenScene	Distillation	12.6	51.6	21.4	41.8	55.6	39.1	0.0	0.0	0.0	0.0	18.3	0.0	1.9	0.8	0.0	6.0
PLA	Distillation	1.7	17.6	6.1	21.6	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RegionPLC	Distillation	2.5	25.2	8.6	28.1	0.0	2.0	1.2	0.0	0.0	0.0	0.2	0.0	1.3	0.1	0.0	0.0
OpenMask3D	Proposal-based	11.2	44.6	25.7	18.2	86.6	14.4	**2.2**	0.8	4.8	1.9	1.7	0.4	3.2	3.5	**0.4**	7.0
OpenIns3D(yoloworld)	Proposal-based	13.5	34.2	25.5	43.7	23.8	45.7	0.1	0.0	5.0	0.5	21.3	0.1	17.5	0.0	0.0	17.8
OpenIns3D(odise)	Proposal-based	24.5	57.9	57.3	45.3	70.7	47.4	0.0	0.0	0.0	0.0	23.8	0.2	33.4	0.0	0.0	0.0
OpenYOLO3D	Proposal-based	12.3	43.3	25.1	45.5	40.2	42.1	0.0	0.0	0.0	0.0	15.5	0.0	10.5	0.1	0.0	6.0
OpenUrban3D	Distillation	39.6	**84.7**	46.2	**70.8**	**90.8**	85.8	0.2	**85.2**	0.0	0.0	46.4	0.0	**72.6**	8.7	0.0	**53.7**
Training-Free Open-Vocabulary Methods																	
GeoOpen3D (Ours)	**Projection**	**42.1**	82.4	**50.2**	60.4	97.3	98.5	0.0	58.2	23.9	15.5	53.9	14.9	70.8	24.3	0.0	29.2

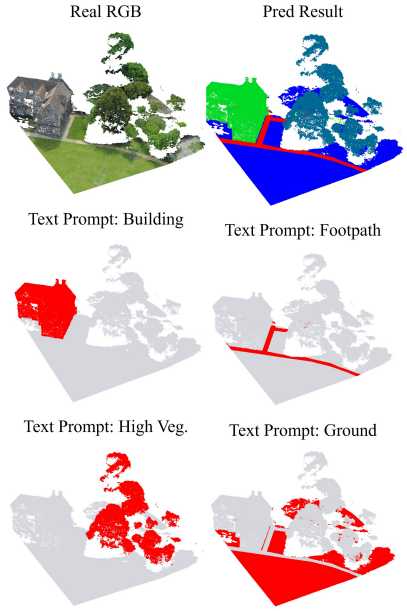

Figure 3. Complex scene result on SensatUrban. GeoOpen3D remains stable in cluttered areas with multiple object categories.

framework does not merely memorize the dominant classes of the benchmark dataset. It can also transfer language-guided recognition to new semantic concepts and scene layouts.

The quantitative results demonstrate that GeoOpen3D can recover meaningful regions for diverse natural and man-made structures, even when their geometry and appearance differ substantially from the urban scenes used in the main evaluation.

4.6 Ablation Study

We conduct ablation experiments to evaluate the contribution of each component in the proposed framework. The results are summarized in Table 3. Overall, all components consistently improve performance, validating the necessity of both geometry enhancement and geometry-aware mask selection for stable 2D–3D semantic transfer.

Removing super-sampling leads to one of the largest drops in mIoU, confirming that projection sparsity is a primary bottleneck in outdoor point cloud rendering. Without sufficient spatial densification, the projected representation becomes incomplete, which directly degrades downstream segmentation. Depth enhancement further improves both mIoU and boundary F1-score by refining local geometric structure, indicating that preserving depth continuity and discontinuity is critical for reliable region formation.

Removing depth-aware constraints results in a moderate decrease in mIoU but a more noticeable degradation in boundary F1-score, suggesting that geometric regularization mainly improves boundary precision rather than coarse region coverage. Finally, mask merging enhances completeness by reducing fragmented predictions from overlapping proposals, highlighting the importance of resolving redundancy in proposal-based segmentation.

4.7 Efficiency Analysis

We evaluate the computational efficiency of GeoOpen3D on a 50 m $\times$ 50 m urban block from the SensatUrban dataset. The full pipeline requires approximately 200 seconds with a peak GPU memory usage of around 16 GB. The runtime is dominated by foundation model inference, particularly the proposal generation and mask prediction stages.

Specifically, GroundingDINO and SAM together account for approximately 135 seconds, reflecting the computational cost of large-scale vision-language and segmentation models. In comparison, geometry processing stages, including projection and depth enhancement, introduce moderate overhead (approximately 45 seconds) while significantly improving segmentation quality. The remaining time (approximately 20 seconds) is spent on back-projection and post-processing.

The ablation results in Table 3 further indicate that super-sampling introduces the largest additional cost among geometric modules due to neighborhood aggregation, while depth enhancement incurs moderate overhead from edge-aware filtering. In contrast, depth-aware constraints are computationally lightweight, demonstrating that geometry-based regularization can be integrated efficiently. Overall, the results suggest that future acceleration should primarily focus on foundation model inference, while the proposed geometry-aware components achieve a favorable trade-off between efficiency and performance.

5. Discussion

5.1 Impact of Geometric Constraints in 2D–3D Transfer

The effectiveness of GeoOpen3D is largely attributed to the explicit incorporation of geometric constraints during the 2D–3D transfer process. Unlike purely appearance-driven approaches,

Category Type	Precision (%)	Recall (%)	F1-Score (%)
Seen Categories	89.2	87.5	88.3
Unseen Categories	76.8	72.3	74.5
Overall	85.4	82.9	84.1

Table 2. Open-vocabulary evaluation on custom island dataset.

Table 3. Ablation study on the SensatUrban dataset. The full model achieves the best performance across all metrics.

Variant	mIoU (%)	Boundary F1-score (%)	Time (s)
Full framework	**42.1**	**82.3**	200
w/o depth enhancement	37.7	78.5	178
w/o depth-aware constraints	39.8	79.8	195
w/o super-sampling	37.2	80.1	162
w/o mask merging	39.0	81.2	188

the framework evaluates segmentation candidates based on depth consistency and boundary alignment, which significantly improves structural coherence. This design reduces common errors such as fragmented regions and incorrect merging of adjacent objects, particularly in outdoor scenes with large depth variations. By enforcing geometric consistency during mask selection, the method ensures that semantically plausible regions are also spatially valid, which is essential for reliable 3D semantic understanding.

This observation is supported by the ablation results in Table 3. Removing depth-aware constraints leads to a clear decrease in boundary F1-score (from 82.3% to 79.8%), while the impact on mIoU is relatively moderate, indicating that geometric constraints primarily enhance boundary precision. Furthermore, removing depth enhancement causes a larger drop in both mIoU and boundary accuracy, confirming that preserving depth structure is critical for stable region formation. These results demonstrate that geometric reasoning plays a fundamental role in improving both the accuracy and robustness of 2D–3D semantic transfer.

5.2 Benefits of Separating Semantic and Geometric Processing

Another key aspect of the framework is the separation between semantic grounding and geometric refinement. Language-driven proposal generation provides coarse semantic localization, while depth-aware optimization refines spatial consistency in a subsequent stage. This decoupling avoids forcing a single model to simultaneously handle semantic reasoning and geometric precision, thereby reducing model complexity and improving robustness. It also enhances modularity, allowing different foundation models or geometric constraints to be integrated without retraining, which is particularly advantageous in open-vocabulary scenarios.

The effectiveness of this design is reflected in the ablation study. When mask merging is removed, the mIoU decreases from 42.1% to 39.0%, indicating that resolving redundant and overlapping proposals is necessary for maintaining global consistency. In addition, the larger performance drop observed when removing super-sampling compared to removing depth constraints suggests that coarse semantic localization alone is insufficient, and must be complemented by geometry-aware refinement. These findings confirm that separating semantic and geometric processing allows each stage to focus on its specific objective, leading to improved overall performance.

5.3 Limitations

Despite its advantages, the proposed method still depends on the quality of the projection. In extremely sparse regions or under severe occlusion, the projected RGB-D representation may lose geometric details, which can affect mask reliability, as reflected by the significant performance drop when super-sampling is removed. In addition, the reliance on large foundation models introduces computational overhead, with most runtime dominated by GroundingDINO and SAM inference, which may limit efficiency in large-scale or real-time applications.

6. Conclusion

This paper presents GeoOpen3D, a geometry-guided training-free framework for open-vocabulary 3D point cloud segmentation. The core contribution lies in explicitly preserving geometric structure throughout the 2D–3D transfer pipeline, including geometry-consistent RGB-D construction, language-driven proposal generation, depth-aware mask selection, and accurate back-projection to point-level labels. Experiments on SensatUrban demonstrate that GeoOpen3D achieves 42.1% mIoU, outperforming existing training-free open-vocabulary baselines, while maintaining strong performance on major urban categories such as buildings and vegetation. Additional results on a custom island dataset further confirm its generalization ability to unseen categories. These findings indicate that competitive outdoor 3D segmentation can be achieved without training dedicated 3D networks, by effectively coupling foundation models with explicit geometric regularization.

Future work will focus on extending the framework to dynamic and multi-temporal scenarios, as well as improving cross-view consistency for more robust large-scale 3D semantic understanding.

Acknowledgements

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 42401567), Guangdong Provincial Project (Grant No. 2024QN11G095), AI Research and Learning Base of Urban Culture, Guangdong Provincial Department of Education (Grant No. 2023WZJD008), and the Open Fund of the Technology Innovation Center for 3D Real Scene Construction and Urban Refined Governance, Ministry of Natural Resources (Grant No. 2024PF-1)

References

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Ghiasi, G., Gu, X., Cui, Y., Lin, T.-Y., 2021. Scaling open-vocabulary image segmentation with image-level labels. *Proceedings of the European Conference on Computer Vision*, 540–557.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2020a. Sensaturban: Learning to sense urban scenes from point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2316–2325.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020b. Randla-net: Efficient semantic segmentation of large-scale point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11108–11117.

Huang, Z., Wu, X., Chen, X., Zhao, H., Zhu, L., Lasenby, J., 2024. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. *European Conference on Computer Vision*, Springer, 169–185.

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. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J. et al., 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*.

Ma, H., Zhang, F., Chen, S., Yu, J., 2025. Individual Tree Segmentation Using Deep Learning and Climbing Algorithm: A Method for Achieving High-precision Single-tree Segmentation in High-density Forests under Complex Environments. *Photogrammetric Engineering & Remote Sensing*, 91(2), 101–110.

Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., 2023. Openscene: 3d scene understanding with open vocabularies. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 815–824.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B., 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35, 23192–23204.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.

Takmaz, A., Fedele, E., Sumner, R. W., Pollefeys, M., Tombari, F., Engelmann, F., 2023. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE/CVF international conference on computer vision*, 6411–6420.

Wang, R., Peethambaran, J., Chen, D., 2018. Lidar point clouds to 3-D urban models : A review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(2), 606–627.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., Solomon, J. M., 2020. Dynamic Graph CNN for Learning on Point Clouds. *ACM Transactions on Graphics*, 38(5), 1–12.

Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 286–304.

Wu, X., DeTone, D., Frost, D., Shen, T., Xie, C., Yang, N., Engel, J., Newcombe, R., Zhao, H., Straub, J., 2025. Sonata: Self-supervised learning of reliable point representations. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22193–22204.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H., 2024. Point transformer v3: Simpler faster stronger. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4840–4851.

Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X., 2022. Groupvit: Semantic segmentation emerges from text supervision. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18134–18144.

Yang, Y., Wu, X., He, T., Zhao, H., Liu, X., 2023. SAM3D: Segment Anything in 3D Scenes. *arXiv preprint arXiv:2306.03908*.

Zhang, C., Wan, H., Shen, X., Wu, Z., 2021. Pvt: Point-voxel transformer for 3d deep learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16169–16178.

Zhang, S., Chen, Y., Wang, B., Pan, D., Zhang, W., Li, A., 2024. SPTNet: Sparse Convolution and Transformer Network for Woody and Foliage Components Separation from Point Clouds. *IEEE Transactions on Geoscience and Remote Sensing*.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., Koltun, V., 2021. Point transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.